тольяттинский
государственный
университет

О.М. Гущина, О.В. Аникина, Е.В. Желнина

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Министерство науки и высшего образования Российской Федерации
Тольяттинский государственный университет

О.М. Гущина, О.В. Аникина, Е.В. Желнина

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Учебно-методическое пособие

Тольятти
Издательство ТГУ
2025

УДК 004.6.04(075.8)
ББК 16.22я73
Г981

Рецензенты:
канд. пед. наук, доцент, доцент кафедры транспорта и технологий нефтегазового комплекса филиала Тюменского индустриального университета в городе Ноябрьске *С.В. Лаптева*;
д-р техн. наук, доцент, профессор кафедры «Прикладная математика и информатика» Тольяттинского государственного университета *С.В. Мкртычев*.

Г981 Гущина, О.М. Анализ и визуализация данных : учебно-методическое пособие / О.М. Гущина, О.В. Аникина, Е.В. Желнина. – Тольятти : Издательство ТГУ, 2025. – 204, [2] с. : обл.
ISBN 978-5-8259-1712-2.

Учебно-методическое пособие содержит теоретические представления о технологии анализа и визуализации данных. В нем приводится пример проектного решения, позволяющего сформировать практические навыки применения инструментов сбора, обработки, анализа и визуализации данных для решения задач профессиональной деятельности.

Предназначено для студентов, обучающихся по направлениям подготовки 01.04.02 «Прикладная математика и информатика», 09.04.03 «Прикладная информатика» очной и заочной форм обучения (в том числе с использованием дистанционной образовательной технологии).

УДК 004.6.04(075.8)
ББК 16.22я73

Рекомендовано к изданию научно-методическим советом Тольяттинского государственного университета.

ISBN 978-5-8259-1712-2

ВВЕДЕНИЕ

Учебно-методическое пособие «Анализ и визуализация данных» предназначено для студентов, обучающихся по направлениям подготовки 01.04.02 «Прикладная математика и информатика», 09.04.03 «Прикладная информатика» очной и заочной форм обучения (в том числе с использованием дистанционной образовательной технологии), и всех тех, кто хочет получить базу теоретических знаний в области анализа данных и практических навыков визуализации данных.

В пособии представлено описание основ анализа данных, которое включает описание основных техник и методик анализа данных в соответствии с рассматриваемыми видами анализа данных: корреляционный анализ, регрессионный анализ, факторный анализ и кластерный анализ. Дается общее представление о визуализации данных и техниках ее реализации в зависимости от выбранного инструментария и вида представления. Каждая тема пособия завершается контрольными вопросами и практическими заданиями, которые дают представление о том, как выполнить анализ и визуализацию данных.

Цель пособия – дать теоретические знания об основных видах анализа данных, техниках приведения данных к визуализации для получения большей аналитической информации инструментальными средствами. В рамках данного пособия дано описание проекта анализа данных и их визуализации.

Задачи учебно-методического пособия:

1. Дать общее представление об области анализа данных, способах реализации и сферах применения. Показать роль описательной статистики в анализе данных.

2. Рассмотреть процедуру проведения разных видов анализа данных, выделив концептуальные основы в подготовке данных для их последующей визуализации для принятия аналитического решения.

3. Проверить знание теоретического материала с помощью вычленения ответов на контрольные вопросы в конце глав и выполнения практических заданий.

4. Реализовать пилотный проект для проведения анализа данных с последующей визуализацией с использованием разных технологий.

Первая глава пособия направлена на знакомство с основными понятиями в области анализа данных и задачами получения генеральной и выборочной совокупности исследуемых данных. Вторая глава описывает понятия, сферы применения и алгоритмы основных видов анализа данных. Третья глава дает общее представление о визуализации данных, о наиболее распространенных ее типах и применяемых инструментах. Четвертая глава показывает, как можно выполнить анализ и визуализацию на основе применения современных подходов.

В результате изучения учебно-методического пособия обучающийся должен:

✓ *знать:*

– основные понятия анализа данных, виды, способы реализации и сферы применения;
– виды анализа данных, их структурные элементы, особенности и преимущества;
– алгоритмы проведения разных видов анализа данных;
– понятие визуализации данных и ее характеристики;
– сущность визуализации данных в понимании и передаче аналитических данных;
– распространенные типы визуализации данных и инструменты для их отображения;
– способы и техники визуализации данных;

✓ *уметь:*

– выбирать совокупность данных для анализа;
– проводить регрессионный анализ;
– проводить корреляционный анализ;
– проводить факторный анализ;
– проводить кластерный анализ;
– выбирать инструменты визуализации;
– выбирать тип визуализации;

✓ *владеть навыками:*

– проведения обработки данных для их последующего анализа;
– проведения анализа данных на основе алгоритмов, в зависимости от выбранного типа;
– работы с инструментальными средствами визуализации данных;
– создания разных типов диаграмм как форм представления визуализации данных;
– проведения визуализации данных в R;
– проведения визуализации данных в Python.

Глава 1. ОСНОВЫ АНАЛИЗА ДАННЫХ

Тема 1. Анализ данных: понятие, виды, способы реализации и сферы применения

Систематическое применение статистических и логических методов для описания объема данных, модуляризации структуры данных, сжатия представления данных, иллюстрации с помощью изображений, таблиц и графиков, а также оценки статистических наклонностей, вероятностных данных и получения значимых выводов, известных как анализ данных, позволяет сделать основной вывод из данных, устраняя из них ненужный шум и ошибки. Генерация данных является непрерывным процессом; это делает анализ данных непрерывным, итеративным процессом, в котором сбор и выполнение анализа происходят одновременно.

Существует несколько методов анализа данных, основными из которых считаются два: качественный и количественный анализ.

1. *Качественный анализ* в основном отвечает на вопросы «почему?», «что?» или «как?». Каждый из этих вопросов решается с помощью количественных методов, таких как вопросники, масштабирование отношения, стандартные результаты и многое другое. Такой анализ обычно проводится на основе текстов, которые могут также включать аудио- и видеоматериалы.

2. *Количественный анализ*, как правило, измеряется с точки зрения чисел. Данные здесь представлены в виде шкал измерения и расширяются для получения необходимой статистики.

Существуют и другие, менее распространенные, методы анализа данных.

3. *Анализ текста* – это метод анализа, применяемый для извлечения машиночитаемых фактов. Он направлен на создание структурированных данных из свободного и неструктурированного контента. Процесс состоит из нарезки неструктурированных, гетерогенных данных в легко читаемые, управляемые и интерпретируемые их фрагменты. Он также известен как интеллектуальный анализ текста, анализ текста и извлечение информации. Неоднозначность естественного языка является самой большой проблемой анализа текста.

4. *Статистический анализ.* Статистика включает в себя сбор, интерпретацию и проверку данных. Статистический анализ – это метод выполнения нескольких статистических операций для количественной оценки данных и применения статистических инструментов. Количественные данные включают в себя описательные данные, такие как опросы и данные наблюдений. Статистический анализ еще называют описательным анализом. Он включает в себя различные инструменты, такие как SAS (Система статистического анализа), SPSS (Статистический пакет для социальных наук), Stat soft и многое другое.

5. *Диагностический анализ* является следующим шагом после статистического анализа, обеспечивающим более глубокий анализ для ответа на исследовательские вопросы. Он также называется анализом первопричин, поскольку включает в себя такие процессы, как обнаружение данных, интеллектуальный анализ, а также детализацию. Диагностический анализ является более современным методом.

Функциями диагностической аналитики являются:

– выявление аномалии. После проведения статистического анализа аналитики должны определить области, требующие дальнейшего изучения, поскольку такие данные вызывают вопросы, на которые невозможно ответить, глядя на данные;

– углубление в аналитику (обнаружение). Идентификация источников данных помогает объяснить аномалии. Этот шаг часто требует поиска закономерностей за пределами существующих наборов данных. Это требует привлечения данных из внешних источников, тем самым выявляя корреляции и определяя, являются ли они причинно-следственными по своей природе;

– определение причинно-следственных связей. Скрытые связи раскрываются путем рассмотрения событий, которые могли привести к выявленным аномалиям. Теория вероятностей, регрессионный анализ, фильтрация и анализ данных временных рядов могут быть полезны для выявления скрытых историй в данных.

6. *Прогностический анализ* использует исторические данные и вводит их в модель машинного обучения для поиска критических закономерностей и тенденций. Модель применяется к текущим

данным, чтобы предсказать, что произойдет в будущем. Многие предпочитают этот вид анализа из-за его преимуществ, таких как объем и тип данных, более быстрые и дешевые компьютеры, простое в использовании программное обеспечение, более жесткие экономические условия и потребность в конкурентной дифференциации. Наиболее общими способами использования прогностического анализа являются:

– обнаружение мошенничества: некоторые методы аналитики улучшают обнаружение шаблонов и предотвращают преступное поведение;

– оптимизация маркетинговых кампаний: прогностические модели помогают компаниям привлекать, удерживать и развивать своих наиболее прибыльных клиентов;

– совершенствование операций: использование прогностических моделей также включает в себя прогнозирование запасов и управление ресурсами, например, авиакомпании используют прогностические модели для установки цен на билеты;

– снижение риска: кредитный рейтинг, используемый для оценки вероятности дефолта покупателя по покупкам, генерируется прогностической моделью, которая включает в себя все данные, относящиеся к кредитоспособности человека; другие виды использования, связанные с риском, включают страховые претензии и сборы.

7. *Предписывающий анализ* предлагает различные направления действий и описывает потенциальные результаты, которые могут быть достигнуты после прогностического анализа. Предписывающий анализ, генерирующий автоматизированные решения или рекомендации, требует конкретных и уникальных алгоритмических и четких указаний от тех, кто использует аналитические методы.

При сборе данных для анализа необходимо понимание, что нужно собрать большое количество информации. Из большого количества данных для обработки нужно выделить соответствующие данные для анализа, чтобы получить точный вывод и принять обоснованные решения.

Определение и сортировка данных для анализа включает следующие этапы:

1. Спецификация требований к данным: необходимо определить область исследования/изучения:

– определить короткие и простые вопросы, ответы на которые понадобятся для изучения какого-либо феномена или процесса или принятия решения;
– определить параметры измерения;
– определить, какой параметр принимается во внимание;
– определить единицу измерения, например время, валюта, зарплата и многое другое.

2. Сбор данных:

– собрать данные на основе параметров измерения;
– собрать данные из баз данных, с веб-сайтов и многих других источников; эти данные могут быть неструктурированными или однородными, что делает обязательным следующий шаг.

3. Обработка данных:

– организовать данные и обязательно добавить примечания, если таковые имеются;
– осуществить перекрестную проверку данных с надежными источниками;
– преобразовать данные в соответствии со шкалой измерения, определенной заранее;
– исключить нерелевантные данные.

4. Анализ данных:

– после сбора данных следует выполнить сортировку, построение графиков и определение корреляций;
– при манипулировании данными и их организации может потребоваться повторить шаги с самого начала; возможно, потребуется изменить вопрос, переопределить параметры и реорганизовать данные;
– следует использовать различные инструменты, доступные для анализа данных.

5. Вывод и интерпретация результатов:

– проверить, дают ли полученные результаты ответы на поставленные вопросы;

– проверить наличие всех параметров для принятия решения;
– рассмотреть вопрос о наличии каких-либо факторов, препятствующих выполнению решения;
– выбрать оптимальные методы визуализации данных (диаграммы, графики, цветовое кодирование и т. д.).

Существуют различные методы анализа данных в зависимости от рассматриваемого вопроса, типа и объема собранных данных. Каждый из них фокусируется на использовании новых данных, анализе идей и детализации информации для преобразования фактов и цифр в параметры принятия решений.

Соответственно, различные методы анализа данных можно классифицировать следующим образом:

1. Методы, основанные на математике и статистике.

Описательный анализ – расматривает исторические данные, ключевые показатели эффективности и производительность. Он учитывает прошлые тенденции и то, как они могут повлиять на будущую производительность.

Дисперсионный анализ – изучение дисперсии в области, на которую распространяется набор данных. Этот метод позволяет аналитикам данных определять изменчивость исследуемых факторов.

Регрессионный анализ. Этот метод работает путем моделирования отношений между зависимой переменной и одной или несколькими независимыми переменными. Регрессионная модель может быть линейной, множественной, логистической, хребтовой, нелинейной, жизненными данными и многим другим.

Факторный анализ. Этот метод помогает определить, существует ли какая-либо связь между набором переменных. Этот процесс выявляет другие факторы или переменные, которые описывают закономерности во взаимосвязях между исходными переменными. Факторный анализ позволяет перейти к полезным процедурам кластеризации и классификации.

Дискриминантный анализ является методом классификации в интеллектуальном анализе данных. Он идентифицирует различные точки в разных группах на основе переменных измерений, определяет, что отличает две группы друг от друга. Это помогает выявлять новые элементы.

Анализ временных рядов. В этом виде анализа измерения распределяются по времени, что дает коллекцию организованных данных, известных как временные ряды.

2. Методы, основанные на искусственном интеллекте и машинном обучении.

Искусственные нейронные сети. Нейронная сеть – это парадигма программирования, вдохновленная биологическими факторами, которая представляет собой метафору мозга для обработки информации. Нейронные сети устойчивы к шумным (ошибочным, замусоренным) данным и отличаются высокой точностью. Их можно считать очень надежными в приложениях бизнес-классификации и прогнозирования.

Дерево принятия решений. Как следует из названия, это древовидная структура, представляющая собой классификационную или регрессионную модель. Она делит набор данных на более мелкие подмножества, одновременно развиваясь в связанное дерево решений.

Эволюционное программирование. Этот метод сочетает в себе различные типы анализа данных с использованием эволюционных алгоритмов. Это независимый от предметной области метод, который может исследовать достаточное пространство поиска и очень эффективно управлять взаимодействием атрибутов.

Нечеткая логика – это метод анализа данных, основанный на вероятности, который помогает справиться с неопределенностями в методах интеллектуального анализа данных.

3. Методы, основанные на визуализации и графиках.

Гистограмма, линейчатая диаграмма. Обе указанные диаграммы используются для представления числовых различий между категориями. Гистограмма использует высоту столбцов, чтобы отразить различия. В случае линейчатой диаграммы происходит замена осей.

Линейная диаграмма. Эта диаграмма представляет изменение данных в течение непрерывного интервала времени.

Диаграмма площади. Эта концепция основана на линейном графике. Он также заполняет область между полилинией и осью цветом, представляя лучшую информацию о тренде.

Круговая диаграмма. Используется для представления доли различных классификаций. Она подходит только для одной серии дан-

ных. Однако ее можно сделать многоуровневой, чтобы представить долю данных в разных категориях.

Воро́нковая диаграмма. Эта диаграмма представляет пропорцию каждого этапа и отражает размер каждого модуля. Это помогает в сравнении рейтингов.

Облако тегов. Это визуальное представление текстовых данных. Оно требует большого объема данных. Степень дифференциации должна быть высокой, чтобы пользователи могли воспринимать наиболее заметную разницу. Это не очень точный аналитический метод.

Диаграмма Ганта показывает фактические сроки и ход выполнения деятельности по сравнению с требованиями.

Радиолокационная диаграмма используется для сравнения нескольких квантованных диаграмм. Этот графический метод показывает, какие переменные в данных имеют более высокие значения, а какие – более низкие. Радиолокационная диаграмма используется для сравнения и классификации наряду с пропорциональным представлением.

Точечная диаграмма показывает распределение переменных в точках по прямоугольной системе координат. Распределение в точках данных может выявить корреляцию между переменными.

Пузырьковая диаграмма является вариацией точечной диаграммы. Здесь, в дополнение к координатам x и y, область пузырька представляет собой 3-е значение.

Рамочная диаграмма – это визуальное представление иерархии в перевернутой древовидной структуре.

Древовидная диаграмма. Этот метод используется для представления иерархических отношений, но на одном уровне. Он эффективно использует пространство и изображает иерархические данные в виде площадей прямоугольников.

Карта:

– региональная карта использует цвет для представления распределения значений по разделу карты;
– карта точек представляет собой географическое распределение данных в точках на географическом фоне;

– карта потока представляет собой связь между областью притока и областью оттока; это линия, соединяющая геометрические центры тяжести пространственных элементов; использование динамических линий потока помогает уменьшить визуальный беспорядок;
– тепловая карта показывает вес каждой точки в географической области; цвет здесь обозначает плотность.

Перечислим и проанализируем некоторые инструменты, используемые при анализе данных в исследованиях. Сегодня доступно несколько инструментов анализа данных, каждый из которых имеет свой набор функций. Выбор инструментов всегда должен основываться на типе выполняемого анализа и типе обрабатываемых данных.

1. Excel. Данный инструмент имеет различные привлекательные функции и с установленными дополнительными плагинами может обрабатывать огромное количество данных. Таким образом, при наличии данных, которые не приближаются к значительному запасу данных (небольшие данные), Excel может быть универсальным инструментом для анализа данных.

2. Tableau. Данный инструмент подпадает под категорию BI Tool и имеет единственную цель – анализ данных. Tableau представляет собой сводную таблицу и сводную диаграмму и работает над представлением данных наиболее удобным для пользователя способом. Он также имеет функцию очистки данных наряду с блестящими аналитическими функциями.

3. Power BI. Первоначально данный инструмент задумывался как плагин для Excel, но позже дифференцировался от него, чтобы развиваться в качестве одного из самых эффективных инструментов анализа данных. Он поставляется в трех версиях: Free, Pro и Premium. Его языки Power Pivot и DAX могут реализовывать сложную расширенную аналитику, аналогичную написанию формул Excel.

4. Fine Report поставляется с простой операцией перетаскивания, которая помогает проектировать различные отчеты и создавать систему анализа решений данных. Он может напрямую подключаться ко всем видам баз данных, и его формат аналогичен формату Excel. Кроме того, данный инструмент также предо-

ставляет различные шаблоны панелей мониторинга и несколько самостоятельно разработанных визуальных библиотек подключаемых модулей.

5. R & Python. Это мощные и гибкие языки программирования, которые также применимы для анализа данных. Лучше всего они подходят для статистического анализа, такого как нормальное распределение, алгоритмы кластерной классификации и регрессионный анализ. Также с их помощью выполняется индивидуальный прогностический анализ, такой как поведение клиентов, расходы, предметы, предпочитаемые ими на основе истории просмотров, и многое другое. Они также включают в себя концепции машинного обучения и искусственного интеллекта.

6. SAS – это язык программирования для анализа данных и манипулирования данными, который может легко получить доступ к данным из любого источника. Компания SAS представила широкий набор продуктов для профилирования клиентов для веб-аналитики, социальных сетей и маркетинговой аналитики. Данный инструмент может прогнозировать их поведение, управлять им и оптимизировать коммуникации.

Анализ данных является ключом к любому бизнесу, будь то запуск нового предприятия, принятие маркетинговых решений, продолжение определенного курса действий или полное закрытие. Выводы и статистические вероятности, рассчитанные на основе анализа данных, помогают обосновывать наиболее важные решения, исключая все человеческие предубеждения. Различные аналитические инструменты имеют перекрывающиеся функции и различные ограничения, но они также являются взаимодополняющими инструментами. Прежде чем выбрать инструмент для анализа данных, важно рассмотреть объем работ, инфраструктурные ограничения, экономическую целесообразность и подготовить окончательный отчет.

Тема 2. Генеральная и выборочная совокупности, их значение в анализе данных

Генеральной совокупностью называется вся группа, которую необходимо исследовать и о которой необходимо сделать выводы. *Выборка* (или *выборочная совокупность*) представляет собой конкретную группу, из которой собираются эмпирические данные.

В реальных исследованиях генеральная совокупность используется в тех случаях, когда того требует исследовательский вопрос (всеобщая перепись населения) или когда есть доступ к данным от каждого члена совокупности (сплошное анкетирование всех работников предприятия или организации). Как правило, достаточно просто собрать данные по всей генеральной совокупности, когда она обладает тремя характеристиками: небольшая, доступная и компактная.

Для крупных и рассредоточенных генеральных совокупностей часто бывает трудно или невозможно собрать данные о каждом человеке. Например, каждые 10 лет правительства многих стран стремятся учитывать каждого человека, проживающего в стране, с помощью переписи населения. Эти данные используются для распределения финансирования по стране. Однако исторически сложилось так, что с маргинализованными группами и группами с низким доходом достаточно трудно установить контакт, найти их и побудить к участию в общегосударственной переписи. Из-за отсутствия ответов данных социальных групп населения страны подсчет оказывается неполным и смещенным в сторону других социальных групп, что приводит к непропорциональному финансированию по всей стране. В подобных случаях можно использовать выборочную совокупность (выборку), чтобы сделать более точные выводы о населении (генеральной совокупности).

В случае если генеральная совокупность велика по размеру, географически рассредоточена или с ней трудно установить контакт, необходимо использовать выборку. Статистический анализ позволяет использовать выборочные данные для оценки или проверки гипотез по всей генеральной совокупности.

Перечислим основные причины использования выборочной совокупности:

1. *Необходимость*. Иногда просто невозможно изучить всю генеральную совокупность из-за ее размера или труднодоступности.
2. *Практичность*. Проще и эффективнее собирать данные из выборки.
3. *Экономическая эффективност*ь. Меньше затрат на участников, лаборатории, оборудование и исследователей.
4. *Управляемость*. Хранение и выполнение статистического анализа небольших наборов данных намного проще и надежнее.

При исследовании большой группы людей достаточно редко существует возможность собрать данные о каждом человеке в этой группе. Вместо этого можно выбрать образец – уменьшенную модель этой группы. Выборка – это группа лиц, которые фактически участвуют в исследовании.

Чтобы получить значимые (валидные) выводы из эмпирических результатов, необходимо тщательно продумать, как будет собираться выборка, которая должна стать репрезентативной для группы в целом. Существует два типа методов формирования выборки:

– *вероятностная выборка*, которая предполагает случайный отбор, что позволяет делать убедительные (значимые, валидные) статистические выводы обо всей группе;
– *невероятностная выборка*, которая предполагает неслучайный отбор на основе удобства или других критериев, что позволяет легко собирать данные.

При этом необходимо четкое и понятное обоснование выборочной совокупности. Именно достаточность и убедительность данного обоснования позволят распространить статистические и аналитические выводы, сделанные по выборке, на всю генеральную совокупность.

Основой выборки служит фактический список лиц, из которых будет составлена выборка. В идеале он должен включать всю целевую генеральную совокупность (и никого, кто не является ее частью). Например, проводится исследование условий труда в компании X. Генеральную совокупность составляют все 1000 сотрудников компании. Таким образом, основой выборки станет база данных

отдела кадров компании, в которой перечислены имена и контактные данные каждого сотрудника. Количество лиц (единиц наблюдения), которых следует включить в выборку, зависит от различных факторов, в том числе от размера, дифференциации и изменчивости генеральной совокупности, а также от дизайна исследования (методов сбора данных, стратегии анализа и т. д.). Существуют различные калькуляторы размера выборки и формулы, которыми можно воспользоваться в зависимости от того, чего необходимо достичь с помощью статистического анализа.

Вероятностная выборка означает, что каждый член генеральной совокупности имеет шанс быть выбранным. Вероятностная выборка в основном используется в количественных исследованиях. Если необходимо получить результаты, репрезентативные для всей генеральной совокупности, то методы вероятностной выборки являются наиболее правильным выбором.

Существует четыре основных типа вероятностной выборки: простая случайная, систематическая, стратифицированная и кластерная. Каждый из этих типов разберем подробнее.

I. Простая случайная выборка

В простой случайной выборке каждый член генеральной совокупности имеет равные шансы быть отобранным. Основа выборки должна включать всех представителей генеральной совокупности. Для проведения этого типа выборки можно использовать такие инструменты, как генераторы случайных чисел или другие методы, полностью основанные на случайности.

Например, необходимо собрать простую случайную выборку из 100 сотрудников компании X. Следует назначить номер каждому сотруднику в базе данных компании от 1 до 1000 и использовать генератор случайных чисел, чтобы выбрать 100 номеров.

Простая случайная выборка используется для получения валидных статистических выводов о генеральной совокупности. Кроме того, при достаточно большом размере простая случайная выборка обладает высокой внешней валидностью: она может репрезентировать характеристики большей совокупности. Однако простая случайная выборка может быть сложной для практической реали-

зации. Чтобы использовать этот метод, необходимо соблюсти ряд условий:

– наличие полного списка членов генеральной совокупности;
– наличие связи или доступа к каждому члену генеральной совокупности;
– наличие времени и ресурсов для сбора данных из выборки необходимого размера.

Важно понимать, что простая случайная выборка работает лучше всего, если есть много времени и ресурсов для проведения исследования или если изучается ограниченная группа населения, которую можно легко отобрать. В некоторых других случаях может оказаться более целесообразным использовать другой тип вероятностной выборки.

Существует четыре ключевых шага для формирования простой случайной выборки:

1. *Определение генеральной совокупности.* Следует начать с определения генеральной совокупности, которую нужно изучить. Важно убедиться, что есть доступ к каждому отдельному члену генеральной совокупности, чтобы можно было собирать данные о всех тех, кто отобран для выборки.

2. *Определение размера выборки.* На данном этапе следует решить, насколько велик будет размер выборки. Хотя большие выборки обеспечивают бóльшую статистическую достоверность, они также стоят дороже и требуют гораздо больше работы. Существует несколько возможных способов определения размера выборки, но один из самых простых включает использование формулы с желаемым доверительным интервалом и уровнем достоверности, предполагаемым размером совокупности, с которой идет работа, и стандартным отклонением того, что необходимо измерить в генеральной совокупности.

Чаще всего используются доверительный интервал и уровень достоверности 0,05 и 0,95 соответственно. Поскольку не всегда известно стандартное отклонение изучаемой совокупности, то следует выбрать число достаточно высокое, чтобы учесть множество возможностей (например, 0,5). Далее можно использовать калькулятор размера выборки, чтобы оценить ее необходимый размер.

3. *Случайный выбор единиц наблюдения.* Это можно сделать одним из двух способов: лотереей или методом случайных чисел.

В лотерейном методе выбирается образец случайным образом путем «вытягивания из шляпы» или использования компьютерной программы, которая будет имитировать то же самое действие.

В методе случайных чисел каждому человеку из генеральной совокупности присваивается номер. Далее с помощью генератора случайных чисел или таблиц случайных чисел случайным образом выбирается необходимое подмножество. Можно использовать функцию случайных чисел (RAND) в Microsoft Excel для генерации случайных чисел.

4. *Сбор данных.* Чтобы убедиться в достоверности выводов, необходимо убедиться, что каждый выбранный человек действительно участвует в исследовании. Если кто-то выбывает или не участвует по причинам, связанным с изучаемым вопросом, это может существенно повлиять на выводы.

II. Систематическая выборка

Систематическая выборка формируется на основе регулярного интервала, а не полностью случайного выбора. Его также можно использовать, когда нет полного списка всей генеральной совокупности. Систематическая выборка похожа на простую случайную выборку, но обычно ее немного легче проводить. Каждому члену генеральной совокупности присваивается номер, но вместо того, чтобы генерировать случайным образом числа, люди выбираются через равные интервалы.

Существует три ключевых шага для формирования систематической выборки:

1. *Определение генеральной совокупности.* Как и при других методах выборки, необходимо определиться с изучаемой совокупностью. При систематической выборке существует два варианта сбора данных:

– заранее выбрать образец из списка, а затем обратиться к выбранным субъектам для сбора данных;
– обратиться к каждому *k*-му члену целевой группы, чтобы попросить их принять участие в исследовании.

Следует убедиться, что список содержит всех представителей генеральной совокупности и не находится в периодическом или циклическом порядке. В идеале он должен располагаться в случайном или похожем на случайный порядке. Это позволит имитировать преимущества рандомизации простой случайной выборки.

Если невозможно получить доступ к списку заранее, то можно физически наблюдать за представителями генеральной совокупности. Таким образом, будет использован метод систематической выборки для формирования необходимого количества единиц наблюдения непосредственно в момент сбора данных. В данном случае необходимо убедиться, что время и место проведения процедуры отбора охватывают всю генеральную совокупность. Это позволит избежать систематической ошибки в полученных результатах.

2. *Определение размера выборки и интервала отбора*. Прежде чем выбрать интервал, необходимо сначала определиться с размером выборки. Существует несколько различных способов выбора размера выборки. Один из наиболее распространенных – это использование калькулятора размера выборки. После того как выбрана желаемая погрешность и уровень достоверности, предполагаемый общий размер выборочной совокупности и стандартное отклонение переменных, которые подлежат измерению, калькулятор предоставит нужный размер выборки.

Когда известен целевой размер выборки, можно рассчитать интервал k, разделив общий предполагаемый размер генеральной совокупности на размер выборки. Это, безусловно, достаточно грубая, приблизительная оценка, а не точный расчет. Хотя заранее сложно точно сказать, сколько людей посетит магазин, но можно оценить общую численность посетителей, используя средний показатель посещаемости за предыдущие несколько недель. Предположим, что каждую неделю магазин посещают около 7500 человек, и на основе этой оценки рассчитывается идеальный размер выборки, равный 366. Таким образом, интервал выборки k равен 7500 / 366 = 20,49, который следует округлить, например, до 20.

3. Ф*ормирование выборки и сбор данных*. Если есть список генеральной совокупности, то случайным образом следует выбрать на-

чальную точку в списке и оттуда выбирать каждого k-го члена генеральной совокупности для включения в выборку.

Если списка нет, то необходимо выбирать каждого k-го члена генеральной совокупности для выборки одновременно со сбором данных.

III. Стратифицированная выборка

Стратифицированная выборка уместна, когда необходимо обеспечить пропорциональное представительство определенных характеристик в выборочной совокупности. Предполагается разделение совокупности на подгруппы, которые могут существенно различаться. Это позволяет сделать более точные выводы при условии, что каждая подгруппа правильно представлена в выборке.

Для формирования стратифицированной выборки генеральная совокупность на основе соответствующих характеристик делится на страты (подгруппы, слои и т. п.), а затем случайным образом единицы наблюдения выбираются из каждой из этих подгрупп (страт). Основываясь на общих пропорциях генеральной совокупности, следует рассчитать, сколько людей должно быть выбрано из каждой подгруппы. Затем используется случайная или систематическая выборка, чтобы выбрать конкретную единицу наблюдения из каждой сформированной подгруппы.

Чтобы использовать стратифицированную выборку, нужно иметь возможность разделения генеральной совокупности на взаимоисключающие и исчерпывающие подгруппы. Это означает, что каждого члена генеральной совокупности можно четко отнести ровно к одной подгруппе.

Стратифицированная выборка является лучшим выбором среди методов вероятностной выборки. Перечислим некоторые из преимуществ стратифицированной выборки:

– обеспечение разнообразия выборки;
– гарантированное включение субъектов из каждой подгруппы, что позволяет отразить разнообразие генеральной совокупности;
– обеспечение аналогичной дисперсии; если необходимо, чтобы данные, собранные из каждой подгруппы, имели одинаковый уровень дисперсии, нужно обеспечить одинаковый размер выборки для каждой подгруппы;

– при использовании других методов выборки можно получить небольшой размер выборки для определенных подгрупп, потому что они менее распространены в генеральной совокупности;
– снижение общей дисперсии генеральной совокупности; хотя генеральная совокупность может быть весьма неоднородной, но она может быть более однородной внутри определенных подгрупп;
– возможность использования различных методов сбора данных.

Существует четыре ключевых шага для формирования стратифицированной выборки:

1. *Определение генеральной совокупности и подгрупп*, которые в нее входят. Как и при использовании других методов вероятностной выборки, необходимо начать с четкого определения генеральной совокупности, из которой будет формироваться выборка.

Кроме того, следует выбрать одну или несколько характеристик, по которым выявленная генеральная совокупность будет разделяться на подгруппы.

2. *Деление генеральной совокупности на подгруппы* (слои, страты). Важно убедиться, что каждая страта является взаимоисключающей (между ними нет пересечения), но вместе они определяют всю генеральную совокупность целиком.

3. *Определение размера выборки для каждой страты*. В первую очередь необходимо принять решение о выборе требуемого типа выборки. В данном случае выбирать следует из двух вариантов:

– пропорциональная выборка; размер выборки этого типа каждой страты равен доле подгруппы в генеральной совокупности в целом; в результате формирования пропорциональной выборки подгруппы, менее представленные в бóльшей части генеральной совокупности, также будут менее представлены в сформированной выборке;

– непропорциональная выборка; в данном случае размеры выборки каждой страты непропорциональны их представленности в генеральной совокупности в целом; как правило, данный тип выборки используется в случае, если необходимо изучить особенно недопредставленную подгруппу, размер выборки которой в противном случае был бы слишком мал, чтобы позволить делать какие-либо статистические выводы.

После того как выбран предпочтительный тип выборки, необходимо определить общий размер выборочной совокупности, который должен быть достаточно большим, чтобы можно было сделать статистические выводы о каждой подгруппе.

Если известна желаемая погрешность и уровень достоверности, а также предполагаемый размер и стандартное отклонение имеющейся генеральной совокупности, то можно использовать калькулятор размера выборки для вычисления ее объема.

4. *Случайный отбор единиц наблюдения из каждой страты.* Необходимо использовать один из методов вероятностной выборки (простая случайная или систематическая выборка) для отбора единиц наблюдения из каждой страты. При правильном исполнении методов вероятностной выборки рандомизация, присущая этим методам, позволит получить выборку, репрезентативную для этой конкретной подгруппы.

IV. Кластерная выборка

Кластерная выборка уместна, когда нет возможности сформировать выборку из всей совокупности. В этом случае генеральная совокупность представляется в виде кластеров, которые ее характеризуют (отражают), а затем из случайного набора этих кластеров формируется выборка.

Кластерная выборка также предполагает разделение совокупности на подгруппы, но в данном случае каждая подгруппа должна иметь сходные характеристики со всей генеральной совокупностью. Вместо выборки отдельных лиц из каждой подгруппы необходимо случайным образом выбирать целые подгруппы.

Если это практически возможно, то можно включить в исследование каждого человека из каждого отобранного кластера. Если кластеры сами по себе большие, то также можно выбрать людей из каждого кластера, используя один из описанных выше методов вероятностной выборки. Эта процедура называется многоступенчатой выборкой. Данный метод хорош для работы с большими и рассредоточенными совокупностями, но риск ошибки в выборке повышается, поскольку между кластерами могут быть существенные различия. Достаточно сложно гарантировать, что выбранные кластеры действительно репрезентативны для всей генеральной совокупности.

Помимо вероятностных методов отбора следует рассмотреть и невероятностные методы формирования выборки.

В невероятностной выборке единицы наблюдения отбираются на основе неслучайных критериев, и не каждый человек имеет шанс быть включенным в исследование. Этот тип выборки легче и дешевле сформировать, но он имеет более высокий риск систематической ошибки выборки. Это означает, что выводы, которые можно сделать по выборочной совокупности, слабее, менее валидны и более ограниченны. Если в исследовании используется невероятностная выборка, то все равно необходимо стремиться сделать ее как можно более репрезентативной относительно генеральной совокупности.

Невероятностные методы формирования выборки часто используются в поисковых и качественных исследованиях. В этих типах исследований цель состоит не в том, чтобы проверить гипотезу о широкой генеральной совокупности, а в том, чтобы развить первоначальное понимание небольшой или недостаточно изученной популяции. Выделяют следующие типы выборок, формирующихся невероятностными методами:

1. *Удобная выборк*а. Данный тип выборки просто включает в исследование людей (единицы наблюдения), которые оказались наиболее доступными для исследователя. Это простой и недорогой способ сбора исходных данных. Но существует большой недостаток: нет возможности определить, является ли выборка репрезентативной для генеральной совокупности. Именно этот недостаток не позволяет получить обобщаемые для всей генеральной совокупности результаты и статистические выводы.

2. *Выборка добровольных ответов*. Подобно удобной выборке данный тип основан на простоте доступа. Выборки добровольных ответов, как правило, в бóльшей степени получаются предвзятыми, поскольку некоторые люди по своей природе более склонны к добровольному участию, чем другие.

3. *Целевая выборка*. Этот тип выборки предполагает, что исследователь использует свой опыт для формирования выборки, которая наиболее полезна для целей исследования. Он часто используется в качественных исследованиях, когда исследователь хочет получить

подробные сведения о конкретном явлении, а не делать статистические выводы, или когда генеральная совокупность очень мала и специфична. Эффективная целевая выборка должна иметь четкие критерии и обоснование включения.

4. *Выборка методом снежного кома.* Если доступ к генеральной совокупности затруднен, то можно использовать выборку методом снежного кома для набора участников через других участников.

В анализе данных результаты и выводы, полученные из выборок, используются для понимания больших генеральных совокупностей. Стандартная ошибка имеет значение, поскольку она помогает оценить, насколько хорошо данные выборки представляют всю генеральную совокупность.

С помощью вероятностной выборки, где элементы выборки выбираются случайным образом, можно собрать данные, которые с бóльшей вероятностью будут репрезентативными для генеральной совокупности. Однако даже при вероятностных выборках остается некоторая вероятность появления погрешности выборки. Это происходит потому, что выборка никогда не будет идеально соответствовать генеральной совокупности, из которой она формируется.

Вычисляя стандартную погрешность, можно оценить, насколько репрезентативна полученная выборка для генеральной совокупности. Именно на этой основе можно делать обоснованные аналитические выводы.

Тема 3. Описательная статистика и показатели изменчивости вариации

Описательная статистика поможет получить представление о «середине» и «распространении» данных с помощью мер центральной тенденции и изменчивости (рис. 1).

Меры центральной тенденции представлены модой, медианой и средним, а меры изменчивости – диапазоном (размахом), стандартным отклонением и дисперсией.

При измерении центральной тенденции или изменчивости набора данных именно уровень измерения определяет, какие методы можно использовать. Методы, которые можно применять,

являются кумулятивными: на более высоких уровнях можно применять все математические операции и меры, используемые на более низких уровнях.

Рис. 1. Группы и виды показателей вариации

Всего выделяют четыре уровня данных: номинальный, порядковый, интервальный и метрический (табл. 1).

Среднее арифметическое является наиболее часто используемым типом среднего. Геометрическое среднее – это метод, используемый для усреднения значений из шкал с широко варьирующимися диапазонами для отдельных субъектов. После этого появляется возможность сравнить средства предметного уровня друг с другом. В то время как среднее арифметическое основано на сложении значений, геометрическое среднее умножает значения.

Относительное стандартное отклонение рассчитывается как стандартное отклонение, деленное на среднее. Важно отметить, что если относительное стандартное отклонение используется для измерения температуры в градусах Цельсия, Фаренгейта и Кельвина, то будут получены три совершенно разных ответа. Но единственным значимым ответом является тот, который основан на шкале с истинным нулем, шкале Кельвина.

Меры центральной тенденции помогут найти середину, или среднее значение набора данных. Тремя наиболее распространенными мерами центральной тенденции являются мода, медиана и среднее. Подробнее рассмотрим каждую из них.

Таблица 1

Зависимость методов измерения изменчивости от типа данных

Тип данных	Математические операции	Меры центральной тенденции	Показатели изменчивости
Номинальный	Равенство (=, ≠)	Мода	Никакой
Порядковый	Равенство (=, ≠) Сравнение (>, <)	Мода Медиана	Диапазон Межквартильный диапазон
Интервальный	Равенство (=, ≠) Сравнение (>, <) Сложение, вычитание (+, –)	Мода Медиана Среднее арифметическое	Диапазон Межквартильный диапазон Стандартное отклонение Дисперсия
Метрический	Равенство (=, ≠) Сравнение (>, <) Сложение, вычитание (+, –) Умножение, деление (×, ÷)	Мода Медиана Среднее арифметическое Среднее геометрическое	Диапазон Межквартильный диапазон Стандартное отклонение Дисперсия Относительное стандартное отклонение

Мода является наиболее часто встречающимся значением в наборе данных. Возможно отсутствие моды, одна мода или более одной моды.

Чтобы найти моду, нужно отсортировать набор данных по номерам или категориям и выбрать наиболее часто встречающийся ответ. Если данные принимают форму числовых значений – упорядочить их от низких до высоких. Если данные представлены в форме категорий или групп, следует отсортировать значения по группам в любом порядке.

Моду легко обнаружить в гистограмме, потому что это значение с самой высокой полосой (рис. 2).

Набор данных часто может не иметь моды вообще, иметь одну моду (унимодальный набор данных) или более одной моды (бимодальный – с двумя модами, тримодальный – с тремя модами или

мультимодальный – с четырьмя модами и более). Это зависит от того, сколько различных значений повторяется чаще всего.

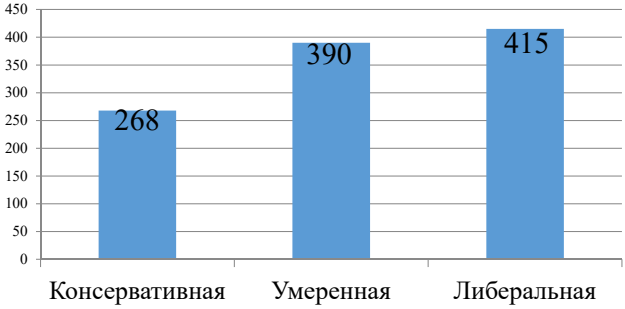


Рис. 2. Распределение респондентов по политическим идеологиям

Уровень измерения переменных определяет использование моды. Мода лучше всего работает с категориальными (номинальными) данными. Это единственный показатель центральной тенденции для номинальных переменных, где он может отражать наиболее часто встречающуюся характеристику (например, демографическую информацию). Мода также полезна при оценке порядковых переменных, например для отражения наиболее популярного ответа по ранжированной шкале.

Для количественных данных, таких как время реакции или высота, мода может не быть полезной мерой центральной тенденции. Это связано с тем, что часто существует гораздо больше возможных значений для количественных данных, чем для категориальных данных, поэтому маловероятно, что значения в наборе данных будут повторяться.

Рассмотрим пример количественных данных, в котором нет моды. Идет сбор данных о времени реакции в компьютерной задаче, и набор данных содержит значения, которые отличаются друг от друга (табл. 2).

В наборе данных, представленном в табл. 2, нет моды, поскольку каждое значение встречается только один раз. В случае графического представления этого набора данных (рис. 3) также станет понятно, что мы имеем дело с безмодальным массивом.

Таблица 2

Набор данных без моды

Респондент	Данные								
	1	2	3	4	5	6	7	8	9
Время реакции (миллисекунды)	267	345	421	324	401	312	382	298	303

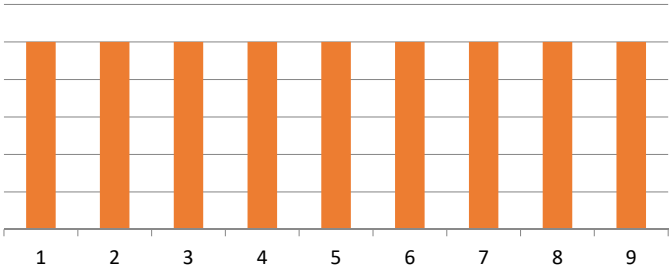


Рис. 3. Графическое представление набора данных без моды

Медиана набора данных показывает значение, которое находится точно посередине, в случае, когда все данные массива упорядочены от низкого к высокому.

Пример поиска медианы. Происходит замер времени реакции семи участников на компьютерную задачу и классификация их на три группы: медленная реакция, средняя или быстрая (табл. 3).

Таблица 3

Замер времени реакции семи участников на компьютерную задачу

Участник	1	2	3	4	5	6	7
Скорость	Средняя	Медленная	Быстрая	Быстрая	Средняя	Быстрая	Медленная

Чтобы найти медиану, сначала следует упорядочить все значения от наименьшего до наибольшего (табл. 4).

Таблица 4

Таблица поиска медианы – упорядоченный набор данных

Упорядоченный набор данных	Медленная	Медленная	Средняя	Средняя	Быстрая	Быстрая	Быстрая

Исходя из анализа данных табл. 4, можно сделать вывод, что модой в данной выборке является вариант «Средняя». Затем необходимо найти значение в середине упорядоченного набора данных, в данном случае значение в 4-й позиции. Таким образом, медианой в представленной выборке будет значение «Средняя».

В больших наборах данных намного проще использовать простые формулы для определения положения медианы в распределении. Важно использовать различные методы для нахождения медианы набора данных в зависимости от того, является ли общее количество значений четным или нечетным.

Для набора данных с нечетным номером следует найти значение, лежащее в позиции $(n + 1) / 2$, где n – число значений в наборе данных. Например, нужно измерить в миллисекундах время реакции пяти участников. Сначала следует упорядочить набор полученных данных по возрастанию значений: время реакции (миллисекунды): 287, 298, 345, 365, 380. Средняя позиция вычисляется с использованием $(n + 1) / 2$, где $n = 5$: $(5 + 1) / 2 = 3$. Это означает, что медиана является третьим значением в упорядоченном наборе данных, соответственно, медианное значение составляет 345 миллисекунд.

Для четного набора данных нужно найти два значения в середине набора данных: значения в позициях $n / 2$ и $(n / 2) + 1$. Затем следует найти их среднее арифметическое. Например, нужно измерить время реакции шести участников. Упорядочиваем полученный набор данных: время реакции (миллисекунды): 287, 298, 345, 357, 365, 380. Средние позиции вычисляются с использованием $n / 2$ и $(n / 2) + 1$, где $n = 6$: $6 / 2 = 3$ и $(6 / 2) + 1 = 4$. Это означает, что средними значениями являются третье по порядку значение (составляет 345) и четвертое значение (составляет 357). Чтобы получить медиану, нужно вычислить среднее арифметическое из двух полученных средних значений, сложив их вместе и разделив на два: $(345 + 357) / 2 = 351$. Таким образом, медиана в данной выборке составляет 351 миллисекунду.

Медиана является наиболее информативным показателем центральной тенденции для искаженных распределений или распределений с выбросами. В искаженных распределениях больше значений приходится на одну сторону от центра, чем на другую,

а среднее, медиана и мода отличаются друг от друга. В положительно искаженном распределении есть группа более низких оценок и растянутый хвост справа. В отрицательно искаженном распределении есть группа более высоких баллов и растянутый хвост слева.

Поскольку медиана использует только одно или два значения из середины набора данных, на нее не влияют экстремальные выбросы или несимметричные распределения баллов. Напротив, положения среднего и моды могут варьироваться в искаженных распределениях. По этой причине медиана часто сообщается как мера центральной тенденции для таких переменных, как доход, поскольку эти распределения обычно положительно искажены.

Уровень измерения переменной также определяет, можно ли использовать медиану. Медиана может быть использована только для данных, которые могут быть упорядочены, то есть из порядкового, интервального и метрического уровней измерения.

Выделим характеристики медианы:

– медианой является значение, которое находится точно в середине набора данных при его упорядочении;
– медиана отделяет самые низкие 50 % от самых высоких 50 % значений;
– алгоритм поиска медианы различается в зависимости от того, нечетное или четное количество точек данных;
– если в середине набора данных есть два числа, их среднее значение равно медиане;
– медиана обычно используется с количественными данными (где значения являются числовыми), но иногда можно найти медиану для набора порядковых данных (где значения ранжируются категориями).

Еще одной мерой центральной тенденции является среднее. Среднее арифметическое набора данных (важно, что оно существенно отличается от геометрического среднего) представляет собой сумму всех значений, деленную на общее число значений. Это наиболее часто используемая мера центральной тенденции, потому что в расчете используются все значения.

Для примера поиска среднего значения воспользуемся уже известной исследовательской ситуацией изучения скорости реакции респондентов (участников) в компьютерной игре (табл. 5).

Таблица 5

Набор имеющихся данных для расчета среднего

Участник	1	2	3	4	5
Время реакции (миллисекунды)	287	345	365	298	380

Расчет среднего арифметического производится по формуле

$$\bar{X} = \sum x/n, \quad (1)$$

где $\bar{X}$ – среднее арифметическое; x – значения ряда данных; n – размер выборки (общее количество единиц наблюдения).

Используя формулу (1), произведем следующие расчеты:

среднее $(\bar{x})$ = (287 + 345 + 365 + 298 + 380) / 5 = 1675 / 5 = 335.

Таким образом, среднее арифметическое представленного ряда данных составляет 335 миллисекунд.

Очень важно понимать и учитывать влияние выбросов на среднее значение. Выбросы могут значительно увеличивать или уменьшать среднее значение, когда они включены в расчет. Поскольку все значения используются для расчета среднего значения, на него могут влиять экстремальные выбросы. Выбросом называется значение, которое значительно отличается от других значений в наборе данных. При расчете среднего в имеющемся наборе данных, в котором одно значение было заменено значением с крайним выбросом, получается, что среднее становится намного выше, хотя все остальные числа в наборе данных остаются прежними. Например:

среднее $(\bar{x})$ = (832 + 345 + 365 + 298 + 380) / 5 = 2220 / 5 = 444.

Показательно, что добавление всего одного выброса в набор данных существенно увеличило среднее значение. В этом случае более подходящим был бы другой показатель центральной тенденции, например, такой как медиана.

Набор данных содержит значения из выборки (или генеральной совокупности). Генеральная совокупность – это вся группа, которая изучается, в то время как выборка – это только подмножество этой генеральной совокупности. Безусловно, данные из выборки могут помочь сделать выводы о генеральной совокупности,

но только полные данные о генеральной совокупности могут дать истинно полную картину.

В анализе данных обозначение среднего выборочной совокупности и среднего генеральной совокупности и формулы их расчета различны только в математической нотации. Но важно отметить, что процедуры расчета одинаковы. В обозначении генеральной совокупности используются заглавные буквы, в то время как в обозначении выборки – строчные.

Формула вычисления среднего по генеральной совокупности выглядит следующим образом:

$$\bar{X} = \frac{\sum X}{N}, \tag{2}$$

где $\bar{X}$ – среднее значение генеральной совокупности; $\sum X$ – сумма каждого значения в генеральной совокупности; N – число значений в генеральной совокупности.

Среднее значение генеральной совокупности в некоторых источниках также может быть обозначено как μ.

Формула вычисления среднего по выборке выглядит следующим образом:

$$\bar{x} = \frac{\sum n}{n}, \tag{3}$$

где $\bar{x}$ – среднее значение выборки; Σn – сумма каждого значения в выборке; n – число значений в выборке.

Среднее значение выборки в некоторых источниках также может быть обозначено как M.

Три основных показателя центральной тенденции лучше всего использовать в сочетании друг с другом, потому что они имеют взаимодополняющие сильные стороны и ограничения. Но иногда только один или два из них применимы к имеющемуся набору данных, в зависимости от уровня измерения переменной.

Моду можно использовать для любого уровня измерения, но она наиболее значима для номинальных и порядковых уровней.

Медиана может быть использована только для данных, которые могут быть упорядочены, то есть из порядкового и интервального уровней и уровня соотношений.

Среднее значение может использоваться только на интервальных и метрических уровнях измерения, поскольку оно требует равного расстояния между соседними значениями или баллами в шкале.

Чтобы решить, какие меры центральной тенденции использовать, следует также рассмотреть распределение набора данных. Для нормально распределенных данных все три меры центральной тенденции дадут один и тот же ответ. В искаженных распределениях медиана является лучшей мерой, поскольку на нее не влияют экстремальные выбросы или несимметричные распределения баллов. Среднее значение и мода могут варьироваться в искаженных распределениях.

Переходим к рассмотрению показателей изменчивости: дисперсии, стандартного отклонения и диапазона. Но перед этим важно отметить, что изменчивость описывает, как далеко лежат точки данных друг от друга и от центра распределения. Наряду с мерами центральной тенденции меры изменчивости дают описательную статистику, которая обобщает данные.

В то время как центральная тенденция говорит, где находится большинство точек, изменчивость показывает, насколько они далеки друг от друга. Это важно, потому что величина изменчивости определяет, насколько качественно можно обобщить результаты выборки для генеральной совокупности.

Небольшое значение изменчивости идеально, потому что это означает, что велика вероятность качественного диагноза и прогноза состояния генеральной совокупности, составленного на основе выборочных данных. Высокое значение изменчивости означает, что данные менее последовательны, поэтому на их основе сложнее делать выводы и прогнозы относительно генеральной совокупности.

Наборы данных могут иметь одну и ту же центральную тенденцию, но разные уровни изменчивости, или наоборот. Если известен только один из этих показателей (центральная тенденция или изменчивость), то совершенно невозможно что-либо сказать о другом показателе. Оба они вместе дают полную картину данных.

Чтобы получить четкое представление об изменчивости данных, диапазон лучше всего использовать в сочетании с другими мерами изменчивости, такими как межквартильный диапазон и стан-

дартное отклонение. В то время как диапазон дает разброс всего набора данных, межквартильный диапазон представляет диапазон средней половины набора данных. Для любого распределения, упорядоченного от меньшего к бóльшему, межквартильный диапазон содержит половину значений. В то время как первый квартиль (Q1) содержит первые 25 % значений, четвертый квартиль (Q4) содержит последние 25 % значений (рис. 4).

Квартили сегментируют любое распределение, упорядоченное от низкого к высокому, на четыре равные части. Межквартильный диапазон (IQR) содержит второй и третий квартили, или среднюю половину набора данных.

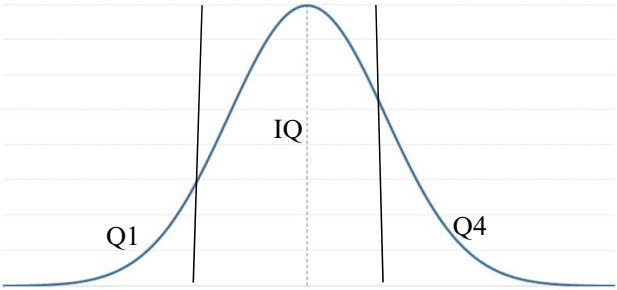


Рис. 4. Нормальное распределение данных

Межквартильный диапазон определяется путем вычитания значения Q1 из значения Q3:

$$IQR = Q3 - Q1, \quad (4)$$

где IQR – межквартильный диапазон; Q3 – 3-й квартиль или 75-й процентиль; Q1 – 1-й квартиль или 25-й процентиль.

Квартиль Q1 – это значение, ниже которого лежит 25 % распределения выборочной совокупности, в то время как квартиль Q3 – это значение, ниже которого лежит 75 % распределения. Можно предположить, что квартиль Q1 является медианой первой половины, а квартиль Q3 – медианой второй половины распределения.

Несмотря на то что существует только одна формула вычисления межквартильного размаха, применяются различные методы идентификации квартилей. Можно получить разное значение

межквартильного диапазона в зависимости от используемого метода. Рассмотрим два наиболее часто используемых метода, которые различаются в зависимости от того, как они используют медиану, – исключающий и включающий методы.

Исключающий метод исключает медиану при определении Q1 и Q3, в то время как инклюзивный метод включает медиану при определении квартилей. Процедура нахождения медианы отличается в зависимости от того, является ли набор данных нечетным или четным.

Когда имеется нечетное количество точек данных, медиана – это значение в середине этого набора данных. В этом случае можно выбрать между инклюзивным и эксклюзивным методом. При четном количестве точек данных в середине есть два значения, поэтому медиана – это их среднее значение. В этом случае чаще всего используется эксклюзивный метод.

Несмотря на то что нет единого мнения о наилучшем методе определения межквартильного диапазона, значение исключительного межквартильного диапазона всегда больше, чем значение инклюзивного.

Исключительный межквартильный диапазон может быть более подходящим для больших выборок, в то время как для небольших выборок инклюзивный межквартильный диапазон может быть более репрезентативным, поскольку это более узкий диапазон.

Рассмотрим алгоритм применения исключительного метода для расчета межквартильного диапазона. Чтобы проанализировать, как работает исключительный метод вручную, мы будем использовать два примера: один с четным числом точек данных и один с нечетным числом.

Четный набор данных. Необходимо пройти четыре шага алгоритма, используя образец набора данных с 10 значениями.

Шаг 1: упорядочение значений от наименьшего к наибольшему.

Шаг 2: поиск медианы и отделение значений под ней от значений над ней – деление набора данных на первую и вторую половины (рис. 5).

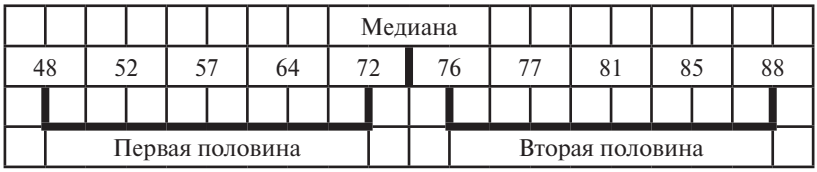


Рис. 5. Результат выполнения шага 2 алгоритма нахождения межквартильного диапазона в четном наборе данных

При четном наборе данных медиана является средним из двух значений, поэтому следует просто поделить набор данных на две половины.

Шаг 3: поиск Q1 и Q3. Q1 – медиана первой половины, Q3 – медиана второй половины (рис. 6). Поскольку каждая из этих половин имеет нечетное число значений, в середине каждой половины есть только одно значение.

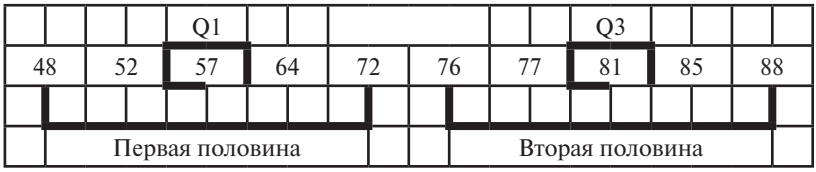


Рис. 6. Результат выполнения шага 3 алгоритма нахождения межквартильного диапазона в четном наборе данных

Шаг 4: произвести расчет межквартильного диапазона по формуле (4):

$$IQR = 81 - 57 = 24.$$

В результате последовательного выполнения алгоритма был вычислен межквартильный диапазон представленного четного набора данных исключительным методом.

Что касается набора данных с нечетными номерами, то расчет межквартильного диапазона будет происходить по следующему алгоритму.

Шаг 1: упорядочение значений от наименьшего к наибольшему.

Шаг 2: поиск медианы и отделение значений под ней от значений над ней (первая и вторая половины набора данных). В нечетном наборе данных медиана – это число в середине списка. Сама медиана исключается из обеих половин: одна половина содержит

все значения ниже медианы, а другая содержит все значения выше нее (рис. 7).

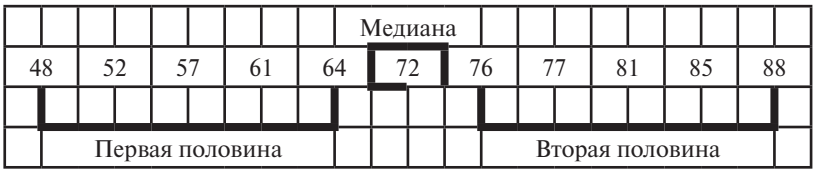


Рис. 7. Результат выполнения шага 2 алгоритма нахождения межквартильного диапазона в нечетном наборе данных

Шаг 3: поиск Q1 и Q3. Q1 – медиана первой половины, Q3 – медиана второй половины (рис. 8). Поскольку каждая из этих половинок имеет нечетный размер, в середине каждой половины есть только одно значение.

Рис. 8. Результат выполнения шага 3 алгоритма нахождения межквартильного диапазона в нечетном наборе данных

Шаг 4: произвести расчет межквартильного диапазона по формуле (4):

$$IQR = 81 - 57 = 24.$$

В результате последовательного выполнения алгоритма был вычислен межквартильный диапазон представленного нечетного набора данных исключительным методом.

Теперь рассмотрим алгоритм расчета межквартильного диапазона для реализации инклюзивного метода. Практически все этапы для инклюзивного и эксклюзивного методов идентичны. Разница заключается в том, как набор данных разделен на две половины. Инклюзивный метод иногда предпочтительнее для нечетных наборов данных, поскольку он не игнорирует медиану – реальное значение в этом типе набора данных.

Шаг 1: упорядочение значений от наименьшего к наибольшему.

Шаг 2: поиск медианы. Следует разделить список на две половины и включить медиану в обе половины (рис. 9). Медиана включается как самое высокое значение в первой половине и самое низкое значение во второй половине.

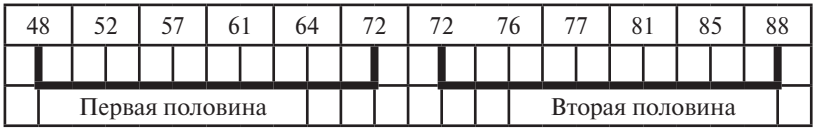


Рис. 9. Результат выполнения шага 2 алгоритма нахождения межквартильного диапазона в четном наборе данных

Шаг 3: поиск Q1 и Q3. Q1 – медиана первой половины, Q3 – медиана второй половины. Поскольку каждая из двух половин содержит четное число значений, Q1 и Q3 вычисляются как средние значения:

$$Q1 = \frac{57 + 61}{2} = 59;$$

$$Q3 = \frac{77 + 81}{2} = 79.$$

Шаг 4: произвести расчет межквартильного диапазона по формуле (4):

$$IQR = 79 - 59 = 20.$$

Из этих примеров мы видим, что использование инклюзивного метода дает ме́ньший IQR. При том же наборе данных эксклюзивный IQR равен 24, а инклюзивный IQR – 20.

В результате последовательного выполнения алгоритма был вычислен межквартильный диапазон представленного четного набора данных инклюзивным методом.

Межквартильный диапазон (IQR) является особенно полезной мерой изменчивости для искаженных распределений. Для этих распределений медиана является лучшей мерой центральной тенденции, потому что это значение точно посередине, когда все значения упорядочены от наименьшего к наибольшему.

Наряду с медианой IQR поможет увидеть, где находится большинство характеристик выборки и насколько они сгруппированы.

IQR также полезен для наборов данных с выбросами, поскольку он основан на средней половине распределения и на него меньше влияют экстремальные значения.

Еще одной мерой изменчивости является стандартное отклонение – средняя величина изменчивости в наборе данных. Эта мера показывает в среднем, как далеко каждое значение лежит от среднего. Чем больше стандартное отклонение, тем более изменчивым является набор данных.

Существует шесть шагов для определения стандартного отклонения вручную:

1. Перечислите каждый балл (значение шкалы) и найдите их среднее значение.
2. Вычтите среднее значение из каждого балла, чтобы получить отклонение от среднего.
3. Вычислите квадрат каждого из этих отклонений.
4. Сложите все квадратные отклонения.
5. Разделите сумму квадратов отклонений на $n - 1$ (для небольшой выборки, до 50) или N (для выборки, которая больше 50).
6. Найдите квадратный корень из найденного числа.

Пример вычисления стандартного отклонения представлен в табл. 6.

Таблица 6

Пример вычисления стандартного отклонения

Шаг 1 Данные (минуты)	Шаг 2 Отклонение от среднего	Шаги 3 + 4 Квадратное отклонение	Шаги 5 + 6
72	72 – 207,5 = –135,5	18360,25	63904 / (8 – 1) = = 9129 √9129 ≈ 96 Стандартное отклонение ≈ 96
110	110 – 207,5 = –97,5	9506,25	
134	134 – 207,5 = –73,5	5402,25	
190	190 – 207,5 = –17,5	306,25	
238	238 – 207,5 = 30,5	930,25	
287	287 – 207,5 = 79,5	6320,25	
305	305 – 207,5 = 97,5	9506,25	
324	324 – 207,5 = 116,5	13572,25	
Среднее = 207,5	Сумма = 0	Сумма квадратов = 63 904	

Рассчитав, что стандартное отклонение (SD) = 96, можно сказать, что каждый балл (ответ, оценка) отклоняется от среднего на 96 баллов в среднем.

Для вычисления стандартных отклонений используются различные формулы в зависимости от того, имеются данные из всей генеральной совокупности или из выборки.

Когда имеются данные от каждого члена генеральной совокупности, можно получить точное значение стандартного отклонения этой генеральной совокупности. Формула стандартного отклонения генеральной совокупности выглядит следующим образом:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}, \tag{5}$$

где σ – стандартное отклонение генеральной совокупности; X – каждое значение; μ – среднее значение по генеральной совокупности; N – число значений в генеральной совокупности.

При сборе данных из выборки стандартное отклонение выборки используется для оценки или выводов о стандартном отклонении популяции. Формула стандартного отклонения для расчета на выборке выглядит следующим образом:

$$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}, \tag{6}$$

где s – стандартное отклонение выборки; X – каждое значение; $\bar{x}$ – среднее значение выборки; n – число значений в выборке.

В случае с выборкой необходимо использовать $n - 1$ в формуле, потому что использование n даст предвзятую оценку, которая будет вести к последовательной недооценке изменчивости. Стандартное отклонение выборки, как правило, ниже, чем реальное стандартное отклонение генеральной совокупности. Уменьшение выборки n до $n - 1$ делает стандартное отклонение искусственно большим, предоставляя консервативную оценку изменчивости. Хотя это и не объективная оценка стандартного отклонения, но менее предвзятая: лучше переоценить изменчивость выборок, чем недооценить.

Высокое значение стандартного отклонения означает, что данные, как правило, далеки от среднего, в то время как низкое значе-

ние стандартного отклонения указывает на то, что значения сгруппированы близко к среднему.

Более высокое стандартное отклонение говорит, что распределение не только более рассредоточено, но и более неравномерно. Однако не стоит забывать, что стандартное отклонение чувствительно к выбросам.

Дисперсия представляет собой среднее значение квадратов отклонений от среднего значения. Отклонение от среднего значения – это то, как далеко оценка лежит от среднего.

Дисперсия – это квадрат стандартного отклонения. Это означает, что дисперсионные единицы намного больше, чем единицы типичного значения набора данных. Дисперсионное число достаточно трудно интуитивно интерпретировать, но важно рассчитать дисперсию для сравнения различных наборов данных в статистических тестах и аналитических процедурах.

Дисперсия отражает степень распространения в наборе данных. Чем больше распространение данных, тем больше дисперсия по отношению к среднему значению. Чтобы получить дисперсию, следует возвести стандартное отклонение в квадрат. Например, стандартное отклонение (с) равно 95,5. Дисперсия (*s*) будет составлять 95,52 = 9129,14. Таким образом, отклонение представленных данных составляет 9129,14 (см. табл. 6).

Для того чтобы найти дисперсию, следует выполнить все шаги для стандартного отклонения, кроме самого последнего шага. Формула расчета дисперсии для генеральной совокупности выглядит следующим образом:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}, \tag{7}$$

где σ^2 – дисперсия генеральной совокупности; X – каждое значение; μ – среднее значение генеральной совокупности; N – число значений в генеральной совокупности.

Формула расчета дисперсии для выборки выглядит как

$$s^2 = \frac{\sum(X - \bar{x})^2}{n}, \tag{8}$$

где s^2 – дисперсия выборочной совокупности; X – каждое значение; $\bar{x}$ – среднее значение выборочной совокупности; n – число значений в выборочной совокупности.

В связи с тем что стандартное отклонение выборки происходит от нахождения квадратного корня дисперсии выборки, то именно поэтому стандартное отклонение выборки нельзя считать непредвзятой оценкой. Поскольку квадратный корень не является линейной операцией, такой как сложение или вычитание, непредвзятость формулы дисперсии выборки не переносится через формулу стандартного отклонения выборки.

Перечислим причины высокой значимости дисперсии в анализе данных:

– параметрические статистические тесты чувствительны к дисперсии;
– сравнение дисперсии выборок помогает оценить групповые различия;
– однородность дисперсии четко прослеживается в статистических тестах.

Дисперсию важно учитывать перед выполнением параметрических тестов. Эти тесты требуют равных или аналогичных дисперсий, также называемых однородностью дисперсии, или гомоскедастичностью, при сравнении различных образцов. Неравномерные дисперсии между образцами приводят к смещению и искажению результатов. При неравномерных отклонениях между выборками более уместны непараметрические тесты.

Стандартное отклонение выводится из дисперсии и показывает в среднем, насколько далеко каждое значение лежит от среднего. Это квадратный корень дисперсии. Оба показателя отражают изменчивость распределения, но их единицы различаются:

– стандартное отклонение выражается в тех же единицах, что и исходные значения (например, метры);
– дисперсия выражается в гораздо бóльших единицах (например, квадратных метрах).

Поскольку единицы дисперсии намного больше, чем единицы типичного значения набора данных, трудно интуитивно интерпретировать число дисперсии. Вот почему стандартное отклонение часто предпочтительнее в качестве основной меры изменчивости. Тем не менее дисперсия более информативна относительно измен-

чивости, чем стандартное отклонение, и она используется для статистических выводов.

Наилучшая мера изменчивости определяется в зависимости от уровня измерения и типа распределения.

Уровень измерения. Для данных, измеренных на порядковом уровне, диапазон и межквартильный диапазон являются единственными подходящими мерами изменчивости. Для более сложных уровней интервалов и соотношений также применимы стандартное отклонение и дисперсия.

Тип распределения. Для нормального распределения могут использоваться все меры. Стандартное отклонение и дисперсия предпочтительны, потому что они учитывают весь набор данных, но это также означает, что на них легко влияют выбросы. Для искаженных распределений или наборов данных с выбросами межквартильный диапазон является лучшим показателем. На него меньше всего влияют экстремальные значения, поскольку он фокусируется на показателях, находящихся в середине набора данных.

Контрольные вопросы

1. Какие два основных метода существуют в анализе данных? Определите и охарактеризуйте каждый из них.
2. Функции диагностической аналитики делятся на три категории. Назовите и дайте характеристику каждой из трех.
3. Каковы наиболее общие способы использования прогностического анализа?
4. Из каких этапов состоит процесс анализа данных? Каково наполнение данных этапов, каковы их цели и контрольные точки?
5. Каким образом можно классифицировать разнообразные методы анализа данных? Представьте развернутую классификацию.
6. Охарактеризуйте дисперсионный анализ как метод определения изменчивости исследуемых факторов.
7. Как работает регрессионный анализ данных? Какие виды регрессионных моделей бывают?
8. Опишите факторный анализ: его понятие, исходные данные и результаты.

9. Перечислите достоинства дискриминантного анализа как метода классификации в интеллектуальном анализе данных.
10. Что такое генеральная совокупность? Какое значение она имеет для анализа данных?
11. Что такое выборочная совокупность? Какими характеристиками она обладает? Какая роль отдается выборочной совокупности в аналитике данных?
12. Перечислите и кратко охарактеризуйте четыре основных типа вероятностной выборки.
13. Каковы основные преимущества простой случайной выборки? Охарактеризуйте алгоритм ее формирования.
14. Каковы недостатки кластерной выборки? Чем кластерная выборка отличается от стратифицированной? Охарактеризуйте алгоритм формирования кластерной выборки.
15. Какие группы и виды показателей вариации применяются в статистике и анализе данных?
16. Определите понятие диапазона. Как он рассчитывается? Перечислите его достоинства и недостатки. Приведите примеры использования диапазона в статистике и анализе данных.
17. Определите понятие стандартного отклонения. Как оно рассчитывается? Перечислите его достоинства и недостатки. Приведите примеры использования стандартного отклонения в статистике и анализе данных.
18. Определите понятие дисперсии. Как она рассчитывается? Перечислите ее достоинства и недостатки. Приведите примеры использования дисперсии в статистике и анализе данных.

Тесты для самоконтроля

1. Как называется метод выполнения нескольких статистических операций для количественной оценки данных и применения статистического анализа? (*один вариант ответа*)

1) статистический анализ
2) анализ текста
3) диагностический анализ
4) инвент-анализ

2. Какие действия включает этап «Анализ данных» в процедуре анализа данных? (*несколько вариантов ответа*)

1) сортировка, построение графиков и определение корреляций
2) возможно, изменение вопроса, переопределение параметров и реорганизация данных
3) сбор данных на основе параметров измерения
4) проверка того, дают ли полученные результаты ответы на поставленные вопросы
5) выбор оптимальных методов визуализации данных (диаграммы, графики, цветовое кодирование и т. д.)

3. Как называется парадигма программирования, вдохновленная биологическими факторами, которая представляет собой метафору мозга для обработки информации, – система, которая изменяет свою структуру на основе информации, которая проходит через сеть? (*один вариант ответа*)

1) нейронная сеть
2) нечеткая логика
3) дерево принятия решений
4) эволюционное программирование

4. Какие действия включает этап «Сбор данных» в процедуре анализа данных? (*несколько вариантов ответа*)

1) сбор данных на основе параметров измерения
2) сбор данных из баз данных, веб-сайтов и многих других источников
3) проверка того, дают ли полученные результаты ответы на поставленные вопросы
4) выбор оптимальных методов визуализации данных (диаграммы, графики, цветовое кодирование и т. д.)
5) определение параметров измерения

5. Какой анализ учитывает исторические данные, ключевые показатели эффективности и производительность, учитывает прошлые тенденции и то, как они могут повлиять на будущую производительность? (*один вариант ответа*)

1) описательный анализ
2) анализ временных рядов

3) дисперсионный анализ
4) факторный анализ
5) дискриминантный анализ

6. Какой метод работает путем моделирования отношений между зависимой переменной и одной или несколькими независимыми переменными? (*один вариант ответа*)
1) регрессионный анализ
2) описательный анализ
3) анализ временных рядов
4) дисперсионный анализ
5) дискриминантный анализ

7. Какой метод анализа данных (основанный на вероятности) помогает справиться с неопределенностями в методах интеллектуального анализа данных? (*один вариант ответа*)
1) нечеткая логика
2) нейронные сети
3) деревья принятия решений
4) эволюционное программирование

8. Какая диаграмма представляет изменение данных в течение непрерывного интервала времени? (*один вариант ответа*)
1) линейная диаграмма
2) гистограмма
3) круговая диаграмма
4) воро́нковая диаграмма

9. Какая диаграмма основана на линейном графике, но еще заполняет область между полилинией и осью цветом, представляя лучшую информацию о тренде? (*один вариант ответа*)
1) диаграмма площади
2) линейная диаграмма
3) облако тегов
4) диаграмма Ганта

10. Какая диаграмма показывает распределение переменных в точках по прямоугольной системе координат, что позволяет выявить корреляцию между переменными? (*один вариант ответа*)

1) точечная диаграмма
2) гистограмма
3) линейная диаграмма
4) диаграмма площади

11. Как называется анализ данных, предполагающий создание древовидной модели, представляющей классификационную или регрессионную модель, которая делит набор данных на более мелкие подмножества? (*один вариант ответа*)

1) деревья принятия решений
2) нейронные сети
3) нечеткая логика
4) эволюционное программирование

12. Какие четыре уровня изменения существуют в анализе данных? (*несколько вариантов ответа*)

1) номинальный уровень
2) порядковый уровень
3) уровень интервалов
4) уровень соотношения
5) нулевой уровень
6) сравнительный уровень

13. Как называется вся группа, которую необходимо исследовать и о которой необходимо сделать статистические и аналитические выводы? (*один вариант ответа*)

1) генеральная совокупность
2) выборка
3) выборочная совокупность
4) аналитическая совокупность

14. Что такое генеральная совокупность? (*один вариант ответа*)

1) вся группа, которую необходимо исследовать и о которой необходимо сделать выводы
2) конкретная группа, из которой собираются эмпирические данные

3) массив собранных эмпирических данных
4) аналитические и статистические выводы, сформированные о выборочной совокупности

15. Для того чтобы по выборке можно было делать выводы о свойствах генеральной совокупности, выборка должна быть случайной и... (*один вариант ответа*)
1) репрезентативной
2) маленькой
3) большой
4) качественной

16. Как называется конкретная группа, из которой собираются эмпирические данные? (*один вариант ответа*)
1) выборка
2) генеральная совокупность
3) массив данных
4) аналитическая совокупность

17. Какая диаграмма используется для представления числовых различий между категориями; принимает высоту столбцов, чтобы отразить различия? (*один вариант ответа*)
1) гистограмма
2) круговая диаграмма
3) облако тегов
4) диаграмма Ганта

18. Как в анализе данных называется понятие, которое обозначает все, что может принимать различные значения в наборе данных (например, высоту или результаты тестов)? (*один вариант ответа*)
1) переменная
2) совокупность
3) массив
4) коэффициент

19. Какие четыре уровня изменений существуют в анализе данных? (*несколько вариантов ответа*)
1) номинальный уровень
2) порядковый уровень

3) уровень интервалов
4) уровень соотношения
5) нулевой уровень
6) сравнительный уровень

20. Среди функций диагностической аналитики есть та, которая осуществляет поиск закономерностей за пределами существующих наборов данных. Как называется эта функция? (*один вариант ответа*)
1) углубление в аналитику (обнаружение)
2) выявление аномалии
3) определение причинно-следственных связей
4) сбор выборочной совокупности

21. Какой вид анализа данных еще называется анализом первопричин, поскольку он включает в себя такие процессы, как обнаружение данных, интеллектуальный анализ, а также детализация? (*один вариант ответа*)
1) диагностический анализ
2) анализ текста
3) статистический анализ
4) инвент-анализ

22. Какие функции диагностической аналитики раскрывают скрытые связи путем рассмотрения событий, которые могли привести к выявленным аномалиям? (*один вариант ответа*)
1) определение причинно-следственных связей
2) выявление аномалии
3) углубление в аналитику (обнаружение)
4) сбор выборочной совокупности

23. Меры центральной тенденции представлены (*несколько вариантов ответа*)
1) модой
2) медианой
3) средним
4) квартилем

24. Меры изменчивости представлены (*несколько вариантов ответа*)
1) диапазоном (размахом)
2) стандартным отклонением
3) дисперсией
4) модой

25. Какая мера центральной тенденции показывает значение, которое находится точно посередине, в случае, когда все данные массива упорядочены от низкого к высокому? (*один вариант ответа*)
1) медиана
2) среднее
3) дисперсия
4) стандартное отклонение

26. Что именно определяет использование моды при анализе данных? (*один вариант ответа*)
1) уровень измерения переменных
2) качество измерения переменных
3) величина измерения переменных
4) способ измерения переменных

27. Какая мера центральной тенденции может не быть полезной для количественных данных, таких как время реакции или высота? (*один вариант ответа*)
1) мода
2) среднее
3) дисперсия
4) стандартное отклонение

28. Если в середине набора данных есть два числа, их ... равно медиане. (*один вариант ответа*)
1) среднее значение
2) максимальное значение
3) минимальное значение
4) экстремальное значение

29. Укажите, что такое среднее арифметическое набора данных? (*один вариант ответа*)

1) значение, которое находится посредством деления на общее число значений суммы всех значений
2) значение, которое находится точно в середине набора данных при его упорядочении
3) значение, которое находится точно посередине, в случае, когда все данные массива упорядочены по алфавиту
4) значение, которое находится посредством деления на сумму общего числа значений

30. Что нужно сделать, чтобы найти моду? (*один вариант ответа*)

1) отсортировать набор данных по номерам или категориям и выбрать наиболее часто встречающийся ответ
2) отсортировать набор данных по времени и выбрать наиболее редко встречающийся ответ
3) отсортировать набор данных по номерам или категориям и выбрать наиболее редко встречающийся ответ
4) отсортировать набор данных по респондентам и выбрать наиболее редко встречающийся ответ

Практическое задание

Проведите анализ данных из отчетных документов средствами MS Excel, Python или R.

Методические указания

1. Прочитайте методические рекомендации к заданию.

2. Решите задачу. В ходе исследования результатов олимпиады из отчетных документов были выбраны данные о числе победителей в различных номинациях среди случайных 100 участников, победивших в одной номинации или более. Среди отобранных не вышли в победители 50 человек, а по остальным число дипломов по номинациям оказалось таким: 1, 1, 1, 2, 3, 1, 1, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 3, 4, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, то есть 35 человек победили в одной номинации, 10 человек – в двух, 4 человека – в трех и 1 человек стал победителем в четырех

номинациях. Чтобы составить представление о закономерности варьирования чисел в «неизвестной» генеральной совокупности, результаты выборочных наблюдений группируют. Для этого:

– создайте лист Excel с названием **Данные**.

	A	B	C	D	E	F	G	H
1								
2		**Количество номинаций (xi)**	**0**	**1**	**2**	**3**	**4**	**Итого**
3		**Количество человек (mi)**	50	35	10	4	1	100
4		**Опытная вероятность (pi~)**	0.5	0.35	0.1	0.04	0.01	1
5		**(число людей в %)**	50%	35%	10%	4%	1%	100%
6		**Вероятность Пуассона (pi)**	0.49	0.35	0.12	0.03	0.01	
7								

Данные | Выборка | Статистика | Гистограмма

Диапазон C2:G3 заполняется начальными данными.

В ячейку H3 введите формулу =СУММ(C3:G3).

В ячейку H4 введите формулу =СУММ(C4:G4).

В ячейку C4 введите формулу =C3/H3 (скопировать формулу в ячейки диапазона D4:G4).

В ячейку C5 введите формулу =C4 (скопировать формулу в ячейки диапазона D5:H5 и установить формат ячеек «Процентный»).

В ячейку C6 введите формулу

=0,71^C2/ФАКТР(C2)*EXP(-0,71)

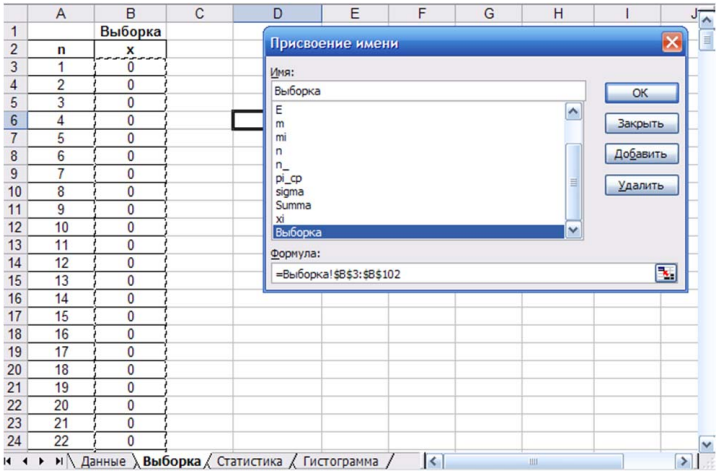
(скопировать формулу в ячейки диапазона D6:G6);

– создайте лист *Выборка*. Расположите на листе *Выборка* исходную выборку данных (в диапазоне B3:B102) следующим образом:

A	B	C
	Выборка	
n	**x**	
1	0	
…	0	
50	0	
51	1	
…	1	
85	1	
86	2	
…	2	
95	2	
96	3	
…	3	
99	3	
100	4	

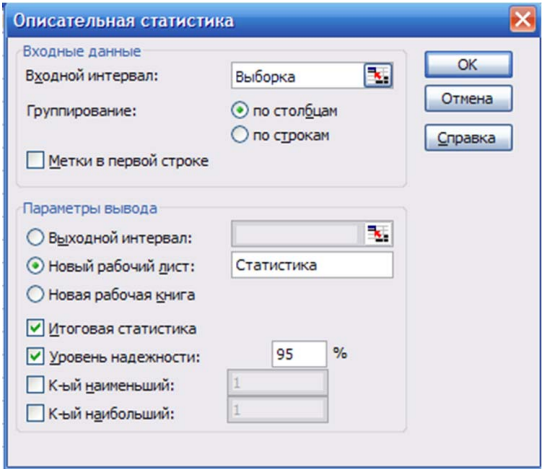
Данные | **Выборка**

– присвойте диапазону B3:B102 исходной выборки данных на листе Выборка имя *Выборка* (Меню ***Вставка / Имя / Присвоить***):



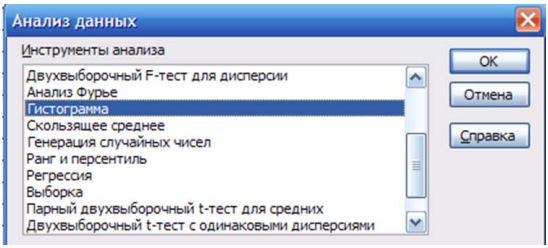
– выполните команду меню ***Сервис / Анализ данных…***, в диалоговом окне ***Анализ данных*** выберите инструмент *Описательная статистика* и нажмите кнопку ***OK***;

– в диалоговом окне ***Описательная статистика*** задайте входной интервал *Выборка*, имя нового рабочего листа *Статистика*, установите флажки *Уровень надежности, Итоговая статистика*.

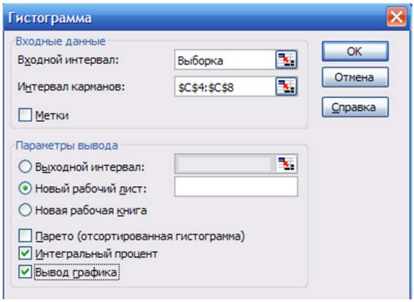


После нажатия кнопки ***OK*** автоматически создается новый лист с результатом описательной статистики;

– постройте гистограмму, для этого выберите меню ***Сервис / Анализ данных / Гистограмма***.



В диалоговом окне задайте входной интервал *Выборка*, *Интервал карманов*, поставьте флажки *Интегральный процент* и *Вывод графика*.



Нажмите кнопку ***OK***.

3. Сохраните файл. Представьте отчет, который будет содержать результаты всех выполненных пунктов в задании. К отчету должны прилагаться созданные файлы. Результат выполненного задания предоставляется архивным файлом.

Методические рекомендации

В Excel инструмент получения описательной статистики вызывается командой меню ***Сервис / Анализ данных / Описательная статистика***.

Пакет анализа – это надстройка Excel, которая предоставляет дополнительные возможности статистического анализа. Она позволяет выполнить более обширный анализ, чем с помощью обыч-

ных средств Excel. Для установки средства ***Пакет анализа*** необходимо выполнить следующие действия:

1. Выбрать команду ***Сервис / Надстройки***.

2. В диалоговом окне ***Надстройки*** установить флажок ***Пакет анализа.***

3. Нажать кнопку ***OK***.

Чтобы создать диаграмму с помощью ***Мастера диаграмм***:

– выделите диапазон ячеек таблицы, содержащих данные для построения диаграммы, в том числе заголовки строк и (или) столбцов, которые вы хотите использовать в диаграмме. Не следует выделять пустые ячейки вне строк и столбцов;

– щелкните на кнопке ***Мастер диаграмм*** стандартной панели инструментов и следуйте дальше инструкциям ***Мастера диаграмм***.

Глава 2. ВИДЫ АНАЛИЗА ДАННЫХ

Тема 4. Понятие и процедура корреляционного анализа

Корреляционные исследования предусматривают использование количественных методов для изучения отношений и связей между переменными. Корреляционный анализ применяется для проверки прочности связи между переменными. Переменные в рамках корреляционного анализа наблюдаются без каких-либо манипуляций или вмешательства со стороны исследователей. Стоит отметить, что корреляционный анализ предполагает ограниченный контроль, поэтому необходимо учитывать, что на рассматриваемые переменные могут влиять другие переменные. Корреляционный анализ обеспечивает высокую внешнюю валидность. Другими словами, можно с большой долей уверенности отнести сделанные по выборочной совокупности выводы ко всей генеральной совокупности или подобным совокупностям.

Корреляционные исследования идеально подходят для быстрого сбора данных и обобщения выводов о реальных жизненных ситуациях внешне обоснованным образом. Есть несколько ситуаций, когда корреляционное исследование является подходящим выбором: исследование непричинных связей, изучение причинно-следственных связей между переменными и тестирование новых средств измерения.

Исследование непричинных связей происходит в случае, когда необходимо выяснить, есть ли связь между двумя переменными, но не ожидается нахождение причинно-следственной связи между ними. Корреляционные исследования могут дать представление о сложных реальных отношениях, тем самым помогая исследователям разрабатывать теории и делать прогнозы.

Изучение причинно-следственных связей между переменными реализуется в случае, если исследователь предполагает, что существует причинно-следственная связь между двумя переменными, но непрактично, неэтично или слишком дорого проводить исследования, которые манипулируют одной из переменных. Таким образом, корреляционные исследования могут обеспечить пер-

воначальные показания или дополнительную поддержку теорий о причинно-следственных связях.

Есть много различных методов, которые можно использовать в корреляционных исследованиях. Важно внимательно выбирать методы, чтобы обеспечить надежность и обоснованность результатов, и тщательно выбирать репрезентативную выборку, чтобы полученные данные точно отражали интересующую генеральную совокупность.

Для проведения корреляционного анализа можно использовать данные, которые уже были собраны для других целей, таких как официальные отчеты, опросы или предыдущие исследования, – так называемые вторичные данные. Использование вторичных данных обеспечивает дешевизну и быстроту исследования, поскольку нет необходимости тратить ресурсы на сбор данных. Однако важно понимать, что вторичные данные могут быть ненадежными, неполными или не совсем актуальными, а также отсутствует возможность контроля надежности или достоверности процедур сбора данных. Именно поэтому при использовании вторичных данных необходимо доверять их источнику и обосновать выбор именно этого источника вторичных данных.

После сбора данных нужно статистически проанализировать взаимосвязь между переменными с помощью корреляционного анализа, а также визуализировать отношения между переменными с помощью точечной диаграммы. Различные типы коэффициентов корреляционного анализа подходят для данных в зависимости от их уровней измерения и распределения.

Используя корреляционный анализ, можно обобщить взаимосвязь между переменными в коэффициенте корреляции – одном числе, описывающем силу и направление взаимосвязи между переменными. С помощью этого числа можно будет количественно оценивать степень взаимосвязи между переменными. В процессе корреляционного анализа можно использовать самые разные коэффициенты корреляции. Коэффициент корреляции произведения Пирсона, также известный как *r* Пирсона, обычно используется для оценки линейной зависимости между двумя количественными переменными.

Важно помнить, что корреляция не подразумевает причинно-следственную связь. То, что найдена корреляция между двумя переменными, не означает, что можно сделать вывод о том, что одна из них вызывает другую по каким-либо причинам. Именно в этом заключается проблема направленности. Если две переменные коррелируют, это может быть связано с тем, что одна из них является причиной, а другая – следствием. Но корреляционный анализ не позволяет сделать вывод, что есть что. Поэтому исследователи не делают вывод о причинно-следственной связи, исходя из результатов корреляционного анализа.

В корреляционном анализе существует задача третьей переменной. Таковой называют смешивающую переменную, которая влияет на две другие переменные (коррелирующие между собой), заставляя их казаться причинно связанными, даже если это не так. Вместо этого существуют отдельные причинно-следственные связи между смешивающей переменной и каждой исследуемой переменной в отдельности.

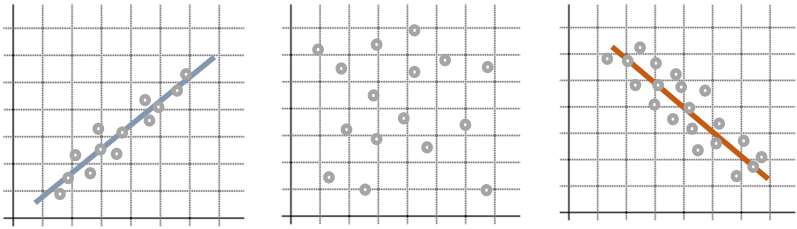
В корреляционных исследованиях существует ограниченный или вообще отсутствует контроль исследователя над переменными. Даже если статистически контролируются некоторые потенциальные связи, все равно могут быть другие скрытые переменные, которые маскируют связь между переменными. Хотя корреляционное исследование не может продемонстрировать причинно-следственную связь само по себе, оно может помочь разработать причинно-следственную гипотезу, которая будет проверяться в последующих исследованиях.

Важным понятием в корреляционном анализе является коэффициент корреляции – число между –1 и 1, которое сообщает силу и направление отношения между переменными. Другими словами, он отражает, насколько похожи измерения двух или более переменных в изучаемом наборе данных (табл. 7). Показательно графическое представление различных значений коэффициента корреляции (рис. 10).

Таблица 7

Коэффициенты корреляции, их тип и значение

Значение коэффициента корреляции	Тип корреляции	Значение
1	Идеальная положительная корреляция	Одна переменная изменяется, другая переменная изменяется в том же направлении (рис. 10, *а*)
0	Нулевая корреляция	Между переменными нет связи (рис. 10, *б*)
–1	Идеальная отрицательная корреляция	Одна переменная изменяется, другая переменная изменяется в противоположном направлении (рис. 10, *в*)



а) *б*) *в*)

Рис. 10. Графическое представление различных значений корреляции: *а* – идеальная положительная корреляция; *б* – нулевая корреляция; *в* – идеальная отрицательная корреляция

Коэффициенты корреляции обобщают данные и помогают сравнивать результаты между переменными или исследованиями. Коэффициент корреляции является описательной статистикой. Это означает, что он суммирует выборочные данные, не позволяя сделать какие-либо выводы о генеральной совокупности. Коэффициент корреляции может представлять:

– двухмерную статистику в случае, если анализируется связь между двумя переменными;
– многомерную статистику, когда анализируется более двух переменных.

Если полученный коэффициент корреляции основан на выборочных данных, то, чтобы распространить результаты на всю гене-

ральную совокупность, понадобится осуществить выводную статистику. В данном случае можно использовать *F*-тест или *t*-тест для расчета тестовой статистики, которая позволяет сделать выводы о статистической значимости полученного результата.

Коэффициенты корреляции являются безудельными, что позволяет напрямую сравнивать коэффициенты разных переменных и разных исследований.

После реализации процедуры сбора данных можно их визуализировать с помощью точечной диаграммы, построив одну переменную на оси *X*, а другую на оси *Y*. Не имеет значения, какая переменная размещается на какой оси. Визуальный осмотр графика позволит выявить наличие тренда (шаблона) и сделать вывод о том, какая связь существует между переменными – линейная или нелинейная. Линейная связь означает, что можно поместить прямую линию наилучшего соответствия между точками данных, в то время как нелинейный или криволинейный узор может принимать всевозможные формы, такие как *U*-образная форма или линия с кривой.

Значение коэффициента корреляции всегда находится в диапазоне от 1 до –1 и всегда рассматривается как общий показатель силы взаимосвязи между переменными.

Знак коэффициента корреляции отражает, изменяются ли переменные в одном или в противоположных направлениях:

– положительное значение означает, что переменные изменяются вместе в одном направлении;
– отрицательное значение означает, что переменные изменяются одновременно в противоположных направлениях.

Абсолютное значение числа равно числу без его знака. Абсолютное значение коэффициента корреляции показывает величину корреляции: чем больше абсолютное значение, тем сильнее корреляция. Приведенную ниже табл. 8 можно использовать в качестве общего ориентира для интерпретации силы корреляции по значению коэффициента корреляции.

Существует много различных рекомендаций по интерпретации коэффициента корреляции. Сложность заключается в том, что коэффициенты могут сильно различаться в разных областях исследо-

вания. Например, если большинство исследований в изучаемой области имеют коэффициенты корреляции, приближающиеся к 0,9, то коэффициент корреляции 0,58 может быть низким в этом контексте.

Таблица 8

Интерпретация значений коэффициента корреляции

Значение коэффициента корреляции	Сила связи (прочность корреляции)	Тип корреляции
От –0,7 до –1	Очень сильная	Отрицательная
От –0,5 до –0,7	Сильная	Отрицательная
От –0,3 до –0,5	Умеренная	Отрицательная
От 0 до –0,3	Слабая	Отрицательная
0	Никакой	Ноль
От 0 до 0,3	Слабая	Положительная
От 0,3 до 0,5	Умеренная	Положительная
От 0,5 до 0,7	Сильная	Положительная
От 0,7 до 1	Очень сильная	Положительная

Коэффициент корреляции показывает, насколько близко анализируемые данные помещаются на линии. Если выявлена линейная зависимость, то можно нарисовать прямую линию наилучшего соответствия (рис. 11), которая учитывает все точки данных на точечной диаграмме. Чем ближе точки к этой линии, тем выше абсолютное значение коэффициента корреляции и тем сильнее ваша линейная корреляция (сравните графики «а» и «б», «г» и «д» на рис. 11). Если все точки находятся на построенной линии, то обнаружена идеальная корреляция (рис. 11, *в* и *е*).

Если все точки находятся близко к этой линии, абсолютное значение коэффициента корреляции высокое (рис. 11, *а* и *г*). Если эти точки разбросаны далеко от этой линии, абсолютное значение коэффициента корреляции низкое (рис. 11, *б* и *д*).

Важно отметить, что крутизна или наклон линии не связаны со значением коэффициента корреляции. Коэффициент корреляции не помогает предсказать, насколько одна переменная будет изменяться на основе данного изменения в другой, поэтому два набора

данных с одинаковым значением коэффициента корреляции могут иметь линии с очень разными наклонами (рис. 12).

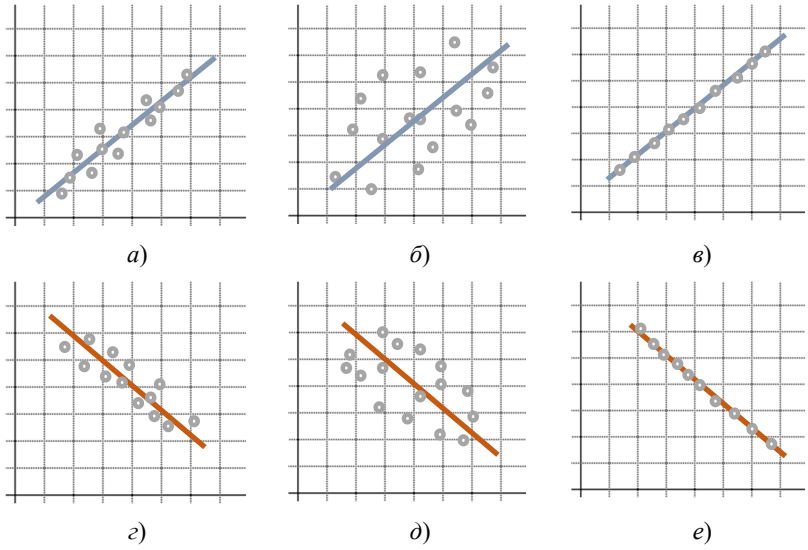


Рис. 11. Графическое представление различных значений корреляции: *а* – сильная положительная корреляция; *б* – слабая положительная корреляция; *в* – идеальная положительная корреляция; *г* – сильная отрицательная корреляция; *д* – слабая отрицательная корреляция; *е* – идеальная отрицательная корреляция

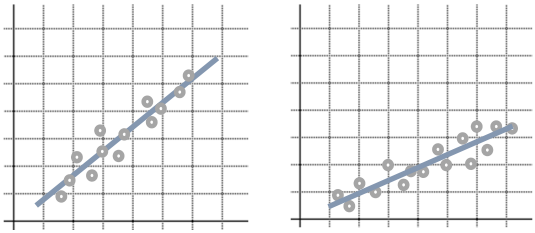


Рис. 12. Графическое представление различных значений корреляции

В зависимости от линейности отношения, уровня измерения переменных и распределения данных можно выбрать из множества различных наиболее приемлемый коэффициент корреляции (табл. 9). Для бóльшей статистической достоверности и точности

лучше всего использовать коэффициент корреляции, который наиболее подходит для исследуемых данных.

Наиболее часто используемым коэффициентом корреляции является *r* Пирсона, потому что он позволяет делать стабильные достоверные выводы. Он параметричен и измеряет линейные отношения. Но если исследуемые данные не соответствуют всем предположениям для этого теста, нужно будет использовать непараметрический тест.

Таблица 9

Подборка часто используемых коэффициентов корреляции

Коэффициент корреляции	Тип отношений	Уровни измерения	Распределение данных
r Пирсона	Линейный	Две количественные (интервальная или относительная) переменные	Нормальное распределение
ρ Спирмена	Нелинейный	Две порядковые, интервальные переменные или переменные соотношения	Любая дистрибуция
Точечно-бисериальный	Линейный	Одна дихотомическая (двоичная) переменная и одна количественная (интервальная или относительная) переменная	Нормальное распределение
φ Крамера	Нелинейный	Две номинальные переменные	Любая дистрибуция
τ Кендалла	Нелинейный	Две порядковые, интервальные переменные или переменные соотношения	Любая дистрибуция

Непараметрические тесты коэффициентов ранговой корреляции обобщают нелинейные соотношения между переменными. ρ Спирмена (ро Спирмена) и τ Кендалла (тау Кендалла) имеют одинаковые условия использования, но τ Кендалла, как правило, предпочтительнее для небольших образцов, тогда как ρ Спирмена используется более широко.

Коэффициент корреляции Пирсона (также известный как *r* Пирсона) описывает линейную связь между двумя количественны-

ми переменными. Условия, которым должны соответствовать данные, чтобы можно было применить к ним *r* Пирсона:

– обе переменные находятся на интервальном или коэффициентном уровне измерения;
– данные из обеих переменных следуют нормальным распределениям;
– данные не имеют выбросов;
– данные взяты из случайной или репрезентативной выборки;
– ожидается линейная зависимость между двумя переменными.

Коэффициент *r* Пирсона является параметрическим тестом, поэтому он обладает высокой значимостью. Но это неподходящая мера корреляции, если переменные имеют нелинейную зависимость или если данные имеют выбросы, искаженные распределения или исходят из категориальных (номинальных) переменных. Если какое-либо из этих предположений нарушено, необходимо использовать меру корреляции ранга.

Формула для *r* Пирсона сложна, но большинство компьютерных программ могут быстро рассчитать коэффициент корреляции из имеющихся данных. В более простой форме формула делит ковариацию между переменными на произведение их стандартных отклонений:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}, \quad (9)$$

где r_{xy} – сила корреляции между переменными *x* и *y*; *n* – размер выборки; *X* – каждое значение *x*-переменной; *Y* – каждое значение *y*-переменной; *XY* – произведение каждой оценки *x*-переменной и соответствующей оценки *y*-переменной.

При использовании формулы коэффициента корреляции Пирсона важно учитывать, откуда взяты данные для анализа: из выборки или из генеральной совокупности. Формулы выборки и генеральной совокупности различаются по символам и входным данным. Коэффициент корреляции выборки называется *r*, в то время как коэффициент корреляции генеральной совокупности обозначается греческой буквой ρ (ро).

Коэффициент корреляции выборки использует ковариацию выборки между переменными и стандартные отклонения выборки:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{S_x S_y}, \quad (10)$$

где r_{xy} – сила корреляции между переменными x и y; cov(x, y) – ковариация x и y; S_x – стандартное отклонение выборки x; S_y – стандартное отклонение выборки y.

Коэффициент корреляции генеральной совокупности использует популяционную ковариацию между переменными и их стандартные отклонения:

$$\rho_{XY} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}, \quad (11)$$

где ρ_{XY} – сила корреляции между переменными X и Y; cov(X, Y) – ковариация X и Y; σ_X – стандартное отклонение выборки X; σ_Y – стандартное отклонение выборки Y.

ρ Спирмена, или коэффициент ранговой корреляции Спирмена, является наиболее распространенной альтернативой r Пирсона. Это коэффициент ранговой корреляции, потому что он использует ранжирование данных из каждой переменной (например, от самой низкой к самой высокой), а не сами необработанные данные.

Коэффициент ранговой корреляции Спирмена можно использовать, когда данные не соответствуют условиям r Пирсона: когда по крайней мере одна из переменных находится на порядковом уровне измерения или когда данные из одной или обеих переменных не следуют нормальным распределениям.

В то время как коэффициент корреляции Пирсона измеряет линейность отношений, коэффициент корреляции Спирмена измеряет монотонность отношений.

В линейной зависимости каждая переменная изменяется в одном направлении с одинаковой скоростью во всем диапазоне данных. В монотонных отношениях каждая переменная также всегда изменяется только в одном направлении, но не обязательно с одинаковой скоростью:

– положительная монотонность: одна переменная увеличивается, другая также увеличивается;

– отрицательная монотонность: одна переменная увеличивается, другая уменьшается.

Формула коэффициента ранговой корреляции Спирмена выглядит следующим образом:

$$r_S = 1 - \frac{6 \sum d_i^2}{(n^3 - n)}, \quad (12)$$

где r_s – сила корреляции между переменными; d_i – разница между рангом x-переменной и рангом y-переменной для каждой пары данных; $\sum d_i^2$ – сумма квадратов разностей между рангами x- и y-переменных; n – размер выборки.

Чтобы использовать эту формулу, необходимо вначале проранжировать данные из каждой переменной отдельно от наименьшего до наибольшего значения: каждая точка данных получает ранг: первый, второй, третий и т. д. Затем находятся разности (d_i) между рангами переменных для каждой пары данных, и эти разности являются основными входными данными для формулы.

Если коэффициент корреляции равен единице (1), то это значит, что все рейтинги для каждой переменной совпадают для каждой пары данных. Если коэффициент корреляции равен минус единице (1), то это значит, что рейтинг для одной переменной является полной противоположностью ранжированию другой переменной. Коэффициент корреляции, близкий к нулю, означает, что между переменными рейтингами нет монотонной связи.

Тема 5. Понятие и структурные элементы регрессионного анализа

Регрессионные модели описывают взаимосвязь между переменными путем подгонки линии тренда к наблюдаемым данным. В регрессионном анализе важным понятием является линейная регрессия – это регрессионная модель, которая использует прямую линию для описания отношений между переменными. Находится линия наилучшего соответствия данных путем поиска значения коэффициента регрессии, который минимизирует общую погрешность модели. Модели линейной регрессии используют прямую линию, в то время как логистические и нелинейные регрессионные

модели используют изогнутую линию. Регрессия позволяет оценить, как изменяется зависимая переменная по мере изменения независимой переменной.

Существует два основных типа линейной регрессии:

– простая линейная регрессия, которая использует только одну независимую переменную;
– множественная линейная регрессия, использующая две независимых переменных или более.

Простая линейная регрессия применяется для оценки взаимосвязи между двумя количественными переменными. Она может быть использована, когда требуется узнать:

– насколько сильна связь между двумя переменными;
– значение зависимой переменной при определенном значении независимой переменной.

Простая линейная регрессия является параметрическим тестом, что означает, что она накладывает определенные требования к данным. Эти требования заключаются в следующем:

– однородность дисперсии (гомоскедастичность): размер погрешности в прогнозе существенно не изменяется по значениям независимой переменной;
– независимость наблюдений: наблюдения в наборе данных собираются с использованием статистически достоверных методов выборки, и между наблюдениями нет скрытых связей;
– нормальность: данные следуют нормальному распределению.

Линейная регрессия также позволяет сделать одно дополнительное предположение: связь между независимой и зависимой переменными является линейной: линия наилучшего соответствия через точки данных представляет собой прямую линию (а не кривую или какой-то фактор группировки).

Если имеющиеся данные не соответствуют предположениям гомоскедастичности или нормальности, то можно использовать вместо этого непараметрический тест, такой как тест ранга Спирмена. Кроме того, если данные нарушают предположение о независимости наблюдений, то можно выполнить линейную модель смешанных эффектов, которая учитывает дополнительную структуру в данных.

Говоря о регрессионном анализе, необходимо рассмотреть понятие регрессионных тестов. Регрессионные тесты показывают, вызывают ли изменения в предикторных переменных изменения в переменной результата. На основе анализа количества и типа переменных, которые используются в качестве предикторов и результатов, можно решить, какой регрессионный тест использовать (табл. 10).

Таблица 10

Зависимость типа регрессионных тестов от видов переменных

Тип регрессионного теста	Предикторная переменная	Переменная результата
Простая линейная регрессия	1 переменная интервала/соотношения	1 переменная интервала/соотношения
Множественная линейная регрессия	2+ переменных интервала/соотношения	1 переменная интервала/соотношения
Логистическая регрессия	1+ любая переменная (переменные)	1 двоичная переменная
Номинальная регрессия	1+ любая переменная (переменные)	1 номинальная переменная
Порядковая регрессия	1+ любая переменная (переменные)	1 порядковая переменная

Большинство часто используемых регрессионных тестов являются параметрическими. Но в случае, если данные не распределены (что проявляется достаточно часто), можно выполнить преобразование данных. Преобразования данных помогают сделать набор данных нормально распределенным с помощью математических операций, таких как вычисление квадратного корня из каждого значения.

Для того чтобы вычислить простую линейную регрессию, можно воспользоваться формулой

$$y = \beta_0 + \beta_1 X + \varepsilon, \quad (13)$$

где y – прогнозируемое значение зависимой переменной (y) для любого заданного значения независимой переменной (x); β_0 – перехват, предсказанное значение y, когда x равно 0; β_1 – коэффициент регрессии – насколько ожидаемо, что y изменится по мере увеличения x; X – независимая переменная (ожидаемая перемен-

ная влияет на y); ε – погрешность оценки, или величина вариаций в оценке коэффициента регрессии.

Линейная регрессия позволяет найти линию наилучшего соответствия через данные путем поиска коэффициента регрессии (β_1), которая сводит к минимуму общую погрешность (ε) модели.

При возведении в квадрат коэффициента корреляции получается коэффициент детерминации (r^2). Коэффициент детерминации показывает долю общей дисперсии между переменными. Коэффициент детерминации всегда находится в интервале между 0 и 1. Как правило, коэффициент детерминации выражается в процентах. Формула расчета коэффициента детерминации очень проста – это коэффициент корреляции, умноженный на себя: r^2.

Коэффициент детерминации используется в регрессионных моделях для измерения того, какая часть дисперсии одной переменной объясняется дисперсией другой переменной.

Регрессионный анализ помогает найти уравнение для линии наилучшего соответствия, которое можно использовать для прогнозирования значения одной переменной с учетом значения для другой переменной.

Высокий коэффициент детерминации означает, что бóльшая величина изменчивости в одной переменной определяется ее отношением к другой переменной. Низкий коэффициент детерминации означает, что лишь небольшая часть изменчивости одной переменной объясняется ее отношением к другой переменной; отношения с другими переменными с бóльшей вероятностью объясняют дисперсию переменной.

Безусловно, линейную регрессию можно выполнить вручную, хотя это утомительный и затратный по времени процесс. Поэтому большинство исследователей используют статистические программы, которые могут помочь быстро проанализировать данные.

Множественная линейная регрессия используется для оценки взаимосвязи между двумя или более независимыми переменными и одной зависимой переменной. Можно использовать множественную линейную регрессию, когда необходимо узнать:

– насколько сильна связь между двумя или более независимыми переменными и одной зависимой переменной;

– значение зависимой переменной при определенном значении независимых переменных.

Множественная линейная регрессия накладывает те же требования к данным, что и простая линейная регрессия. Эти требования заключаются в следующем:

– однородность дисперсии (гомоскедастичность): размер погрешности в прогнозе существенно не изменяется по значениям независимой переменной;
– независимость наблюдений: наблюдения в наборе данных собираются с использованием статистически достоверных методов, и между переменными нет скрытых связей. При множественной линейной регрессии возможно, что некоторые из независимых переменных фактически коррелируют друг с другом, поэтому важно проверить их перед разработкой регрессионной модели. Если две независимые переменные слишком сильно коррелируют (r > ~0,6), то в регрессионной модели следует использовать только одну из них;
– нормальность: данные следуют нормальному распределению;
– линейность: линия наилучшего соответствия через точки данных является прямой линией, а не кривой или каким-либо фактором группировки.

Для того чтобы вычислить множественную линейную регрессию, можно воспользоваться формулой

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon, \qquad (14)$$

где y – прогнозируемое значение зависимой переменной; β_0 – перехват, предсказанное значение y, когда x равно 0; $\beta_1 X_1$ – коэффициент регрессии (β_1) первой независимой переменной (X_1), также известный как влияние, которое увеличение значения независимой переменной оказывает на прогнозируемое значение (y); ... – то же самое для любого количества тестируемых независимых переменных; $\beta_n X_n$ – коэффициент регрессии последней независимой переменной; ε – ошибка модели (также известная какая вариация в оценке y).

Чтобы найти наиболее подходящую линию для каждой независимой переменной, множественная линейная регрессия вычисляет три вещи:

– коэффициенты регрессии, которые приводят к наименьшей общей погрешности модели;
– *t*-статистику общей модели;
– связанное *p*-значение (насколько вероятно, что t-статистика возникла бы случайно, если бы была верна нулевая гипотеза об отсутствии связи между независимыми и зависимыми переменными).

Затем вычисляется *t*-статистика и *p*-значение для каждого коэффициента регрессии в модели.

При отчете о результатах следует указать предполагаемый эффект (то есть коэффициент регрессии), стандартную погрешность оценки и *p*-значение. Также необходимо интерпретировать полученные цифры, чтобы было понятно, что означает полученный коэффициент регрессии.

Тема 6. Особенности и преимущества факторного анализа

Факторный анализ является мощным методом сокращения данных, позволяющим исследовать переменные, которые не могут быть легко измерены напрямую, – латентные переменные. Сводя большое количество переменных к нескольким понятным базовым факторам, факторный анализ приводит к простым для понимания, действенным данным. Применяя этот метод, можно быстрее и качественнее выявлять тенденции и обнаруживать латентные переменные во всех имеющихся наборах данных, что позволяет узнать, что общего у различных единиц исследования.

В отличие от статистических методов, таких как регрессионный анализ, факторный анализ не требует определенных переменных. Факторный анализ чаще всего используется для определения взаимосвязи между всеми переменными, включенными в изучаемый набор данных. В метафорическом ключе факторный анализ можно представить как термоусадочную пленку: при применении к большому объему данных он сжимает этот набор в ме́ньший набор, который гораздо более управляем и прост для понимания.

Общую цель факторного анализа можно разбить на четыре более мелкие:

1. Понимание того, какое количество факторов необходимо для объяснения латентных переменных среди имеющегося набора данных.
2. Определение степени, в которой каждая переменная в наборе данных связана с латентной переменной или фактором.
3. Обеспечение интерпретации общих факторов в наборе данных.
4. Определение степени, в которой каждая наблюдаемая точка данных представляет каждую латентную переменную или фактор.

Достаточно сложно определить, какие конкретные статистические методы нужно использовать, чтобы получить максимальную информацию из имеющихся данных. Рассматривая факторный анализ, важно постоянно обращаться к цели исследования. Существует три основные формы факторного анализа:

– исследовательский факторный анализ, который следует использовать, когда необходимо разработать гипотезу о взаимосвязи между переменными;
– подтверждающий факторный анализ, который используется для проверки гипотезы о взаимосвязи между переменными;
– оценивающий факторный анализ, использующийся для проверки степени, в которой опрос фактически измеряет то, для измерения чего он предназначен.

Если цель исследования соответствует любой из этих форм, то следует выбрать факторный анализ в качестве статистического метода выбора.

Таким образом, факторный анализ – это метод, который используется для уменьшения большого количества переменных в меньшее количество факторов. Этот метод извлекает максимальную общую дисперсию из всех переменных и помещает их в общую оценку. Эта оценка используется для дальнейшего анализа в качестве индекса всех переменных.

Факторный анализ предполагает несколько требований к набору данных:

– наличие в наборе данных соответствующих переменных, пригодных для анализа; используемые переменные должны быть только

метрическими; иные типы переменных также могут быть рассмотрены, но только в особых случаях;
– размер выборки более 200; в некоторых случаях размер выборки может быть рассмотрен для 5 наблюдений на переменную;
– однородность выборки; нарушение этого требования увеличивает размер выборки по мере увеличения числа переменных; для проверки однородности между переменными проводится анализ надежности;
– истинная корреляция между переменными и факторами: между исследовательскими переменными требуется не менее 0,30 корреляции;
– отсутствие выбросов в данных;
– наличие линейной зависимости.

Существуют различные методы, используемые для извлечения факторов из набора данных. Рассмотрим подробнее некоторые из них.

1. Анализ главных компонентов. Это наиболее распространенный метод, используемый исследователями. Анализ главных компонентов извлекает максимальные дисперсии и помещает их в первый фактор. После этого он удаляет дисперсию, объясняемую первыми факторами, а затем начинает извлекать максимальную дисперсию для второго фактора. Этот процесс идет последовательно от первого к последнему фактору.

2. Общий факторный анализ. Это второй наиболее предпочтительный метод исследователей, он извлекает общие дисперсии и помещает их в факторы. Этот метод не включает уникальную дисперсию всех переменных.

3. Факторинг изображений. Этот метод основан на корреляционной матрице. Для прогнозирования фактора в факторинге изображения используется метод регрессии.

4. Метод максимальной вероятности. Этот метод также работает на метрике корреляции, но использует максимальную вероятность для факторизации.

5. Другие методы факторного анализа: альфа-факторинг – перевешивает наименьшие квадраты; весовой квадрат – основан на регрессии, используемой для факторинга.

Чтобы применять факторный анализ для изучения набора данных, необходимо убедиться, что метод исследования, посредством которого собран массив данных, оптимизирован для факторного анализа.

Безусловно, такие характеристики, как размер генеральной совокупности и тема исследования, будут влиять на необходимое количество респондентов, но все-таки в факторном анализе лучше придерживаться мнения «Чем больше респондентов, тем лучше».

Исследования обычно призваны выяснить, какое влияние оказывает одна переменная на другую. Для примера возьмем исследовательскую ситуацию выявления влияния добавления соли на рост растений. Исследователь манипулирует независимой переменной (той, которая, по его мнению, может быть причиной – количество соли), а затем измеряет зависимую переменную (ту, которая, по его мнению, может быть эффектом – рост растений), чтобы выяснить, каким может быть этот эффект. Кроме того, у исследователя, вероятно, также будут переменные, которые он держит постоянными (контрольные переменные, управляющие переменные), чтобы сосредоточиться на цели исследования.

Таблица 11

Сравнение независимых, зависимых и управляющих переменных

Тип переменной	Определение	Пример (эксперимент по толерантности к соли)
Независимые переменные (они же переменные лечения)	Переменные, которыми манипулируют, чтобы повлиять на результат эксперимента	Количество соли, добавляемой в воду каждого растения
Зависимые переменные (они же переменные ответа, переменные результата)	Переменные, представляющие результат эксперимента	Любое измерение здоровья и роста растений: в данном случае высота растений и увядание
Управляющие переменные	Переменные, которые остаются постоянными на протяжении всего эксперимента	Температура и свет в помещении, в котором содержатся растения, и объем воды, отдаваемый каждому растению

Конечная цель факторного анализа состоит в том, чтобы взять обширную концепцию и упростить ее, рассмотрев более детальную, контекстуальную информацию. Таким образом, этот подход предоставляет требуемые и необходимые результаты (табл. 11).

Если планируется использовать факторный анализ, следует избегать открытых вопросов в анкете опроса. Для факторного анализа наиболее приемлемым является предоставление респондентам вариантов ответов в виде шкал (будь то шкалы Лайкерта, числовые шкалы или даже шкалы «да/нет»). Важно убедиться, что достаточно часто используются одни и те же варианты масштабированных ответов.

Факторный анализ является полезным инструментом для исследования отношений переменных в сложных понятиях, таких как социально-экономический статус, диетические модели или психологические масштабы. Это позволяет исследователям изучать концепции, которые невозможно измерить напрямую. Факторный анализ реализует это, используя большое количество переменных для определения нескольких интерпретируемых базовых факторов.

Ключевая концепция факторного анализа заключается в том, что множественные наблюдаемые переменные имеют сходные паттерны ответов, потому что все они связаны со скрытой переменной (то есть не измеряются напрямую). В каждом факторном анализе факторов должно быть на один фактор меньше, чем переменных. Каждый фактор фиксирует определенную величину общей дисперсии в наблюдаемых переменных. Факторы всегда перечисляются в порядке того, насколько они объясняют вариации.

Собственное значение является мерой того, какую часть общей дисперсии наблюдаемых переменных объясняет фактор. Любой фактор с собственным значением ≥1 объясняет бóльшую дисперсию, чем одна наблюдаемая переменная. Факторы, которые объясняют наименьшее количество дисперсии, как правило, отбрасываются.

В рамках факторного анализа применяется понятие «собственные значения» (в некоторых источниках – характерные корни). Собственные значения показывают дисперсию, объясняемую этим конкретным фактором из общей дисперсии. Из собственных значений можно узнать, насколько дисперсия объясняется первым фактором из общей дисперсии.

Оценка фактора также называется оценкой компонента. Эта оценка относится ко всем строкам и столбцам, которые могут быть использованы в качестве индекса всех переменных, а также для дальнейшего анализа. Можно стандартизировать эту оценку, умножив общий термин. Эта факторная оценка предполагает, что все переменные будут вести себя как факторные баллы и будут двигаться (изменяться) в предполагаемых направлениях.

Несколько подробнее рассмотрим процесс выполнения подтверждающего факторного анализа (CFA). Он представляет собой многомерную статистическую процедуру, которая используется для проверки того, насколько хорошо измеренные переменные представляют латентную переменную. Подтверждающий факторный анализ и исследовательский факторный анализ (EFA) являются аналогичными методами, но в исследовательском факторном анализе данные просто исследуются и предоставляют информацию о количестве факторов, необходимых для представления данных. В исследовательском факторном анализе все измеренные переменные связаны с каждой латентной переменной. Но в подтверждающем факторном анализе исследователи могут указать количество факторов, требуемых в данных, и какая измеряемая переменная связана с какой латентной переменной. Подтверждающий факторный анализ – это инструмент, который используется для подтверждения или отклонения теории измерения.

В самом начале необходимо определить индивидуальную конструкцию. Это включает в себя предварительный тест для оценки элементов конструкции и подтверждающий тест модели измерения, который проводится с использованием подтверждающего факторного анализа.

Далее в подтверждающем факторном анализе необходимо рассмотреть концепцию одномерности и оценить разницу между дисперсией конструкции и дисперсией ошибки конструкции. В исследовании должно присутствовать не менее четырех конструкций и трех факторов на конструкцию.

После выполнения вышеизложенного осуществляется разработка исследования для получения эмпирических результатов.

Важно, чтобы была указана модель измерения. Чаще всего значение одной оценки нагрузки должно быть единицей на конструкцию. Для идентификации доступны два метода:

– условие ранга;

– условие порядка.

Следующим этапом осуществления подтверждающего факторного анализа является оценка достоверности модели измерения. Она происходит, когда теоретическая модель измерения сравнивается с моделью реальности, чтобы увидеть, насколько хорошо подходят данные. Проверить валидность модели измерения помогает значение коэффициента. Например, коэффициент загрузки латентной переменной должен быть больше 0,7. Тест хи-квадрата и другие статистические данные о соответствии являются некоторыми ключевыми показателями, которые помогают в измерении достоверности модели.

Допущения CFA включают многомерную нормальность, достаточный размер выборки ($n > 200$), правильную априорную спецификацию модели и то, что данные должны поступать из случайной выборки.

Рассмотрим подробнее исследовательский факторный анализ. Это статистический метод, который используется для сведения данных к меньшему набору суммарных переменных и изучения базовой теоретической структуры явлений. Он также применяется для выявления структуры взаимосвязи между переменной и респондентом. Исследовательский факторный анализ может быть двух видов:

– *R*-типа: факторы вычисляются из корреляционной матрицы;

– *Q*-типа: факторы рассчитываются от отдельного респондента.

Существует два метода для определения движущего (главного) фактора:

1. Метод анализа главных компонентных факторов. Этот метод используется, когда необходимо управлять минимальным количеством факторов и объяснить максимальную часть дисперсии в исходной переменной.

2. Общий факторный анализ. Этот метод используется, когда исследователи не знают природу фактора, подлежащего извлечению, и общую дисперсию ошибок.

Определим критерии практической и статистической значимости факторных нагрузок. Факторные нагрузки можно классифицировать по их значению:

– +0,30 – минимальный уровень;
– +0,40 – более важный уровень;
– +0,50 – практически значимый уровень.

Исследователь может определить статистическую мощность и уровень значимости. Например, для достижения коэффициентной нагрузки 0,55 при мощности 0,80 необходима выборка объемом 100.

Подведем некоторые итоги. Подобно кластерному анализу, включающему группировку похожих случаев, факторный анализ включает в себя группировку похожих переменных по измерениям. Этот процесс используется для идентификации латентных переменных (или конструкций). Целью факторного анализа является сведение многих отдельных элементов к меньшему числу измерений. Факторный анализ может быть использован для упрощения данных и уменьшения числа переменных в регрессионных моделях.

Чаще всего факторы чередуются после экстракции. Факторный анализ имеет несколько различных методов вращения, и некоторые из них гарантируют, что факторы являются ортогональными (то есть некоррелированными), что устраняет проблемы многоколлинеарности в регрессионном анализе.

Факторный анализ также используется для проверки построения масштаба. В таких приложениях элементы, составляющие каждое измерение, указываются заранее. Эта форма факторного анализа чаще всего используется в контексте моделирования структурных уравнений и называется подтверждающим факторным анализом. Например, подтверждающий факторный анализ может быть выполнен, если исследователь хочет проверить факторную структуру черт личности «Большой пятерки» с помощью соответствующей анкеты.

Факторный анализ также может быть использован для построения индексов. Наиболее распространенным способом построения индекса является простое суммирование всех элементов в индексе. Однако некоторые переменные, составляющие индекс, могут иметь бóльшую объяснительную силу, чем другие. Факторный анализ может быть использован для обоснования отказа от вопросов с целью сокращения вопросников.

Факторный анализ можно выполнить в специализированной программе обработки статистических данных SPSS, нажав на *Анализ* в меню, а затем выбрав *Фактор* из опции сокращения данных.

Тема 7. Понятие, сфера применения и алгоритм проведения кластерного анализа

Кластерный анализ, или кластеризация – это задача группировки набора объектов таким образом, чтобы объекты в одной группе (называемой кластером) были более похожи друг на друга, чем на объекты в других группах (кластерах). Это основная задача исследовательского интеллектуального анализа данных и общий метод статистического анализа данных, используемый во многих областях, включая машинное обучение, распознавание образов, анализ изображений и сжатие данных.

Кластеризация может быть реализована различными алгоритмами, которые существенно различаются в понимании того, что представляют собой кластеры и как их эффективно найти. Популярные представления о кластерах включают группы с небольшими расстояниями между членами кластера, плотными областями пространства данных, интервалами или определенными распределениями.

Кластеризация автоматически разбивает набор данных на группы на основе их сходства (рис. 13).

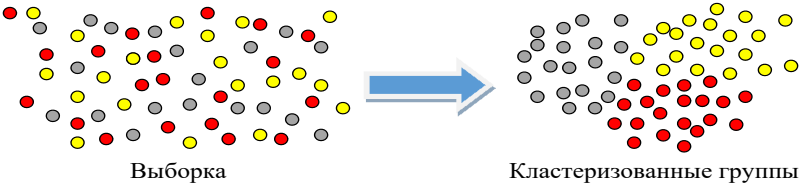


Рис. 13. Графическое представление кластеризованных данных

Перечислим некоторые ситуации приложения кластерного анализа:

– обнаружение аномалий – позволяет обнаружить необычные точки данных в наборе данных. Это полезно для поиска мошеннических транзакций;

– интеллектуальный анализ связей – определяет наборы элементов, которые часто встречаются вместе в наборе данных;
– латентные переменные модели – широко используются для предварительной обработки данных; например, уменьшение количества объектов в наборе данных или декомпозиция набора данных на несколько компонентов.

Важно понимать, что кластеры, найденные различными алгоритмами, значительно различаются по своим свойствам. Понимание этих «кластерных моделей» является ключом к пониманию различий между разными алгоритмами.

Существуют следующие типичные модели кластеров:
– модели подключения (иерархическая кластеризация): кластеризация строит модели на основе уровней иерархии;
– центроидные модели: алгоритм *k*-средних представляет каждый кластер одним средним вектором;
– модели распределения: кластеры моделируются с использованием статистических распределений, таких как многомерные нормальные распределения;
– модели плотности: DBSCAN и OPTICS определяют кластеры как связанные плотные области в пространстве данных;
– модели подпространств: кластеры моделируются как с членами кластера, так и с соответствующими атрибутами (свойствами);
– групповые модели: некоторые алгоритмы не предоставляют уточненную модель для своих результатов, а просто дают информацию о группировке;
– графовые модели: подмножество узлов в графе, такое, что каждые два узла в подмножестве соединены ребром, можно рассматривать как прототипную форму кластера;
– модели подписанных графов: каждый путь в знаковом графе имеет знак от произведения знаков по краям; более слабая «аксиома кластерности» дает результаты с более чем двумя кластерами или подграфами только с положительными ребрами;
– нейронные модели: наиболее известной неконтролируемой нейронной сетью является самоорганизующаяся карта, и эти модели обычно можно охарактеризовать как похожие на одну или несколько из вышеуказанных моделей, включая модели подпро-

странства, когда нейронные сети реализуют форму анализа главных компонентов или независимого компонентного анализа.

Кластеризация – это, по сути, набор таких кластеров, обычно содержащий все объекты в наборе данных. Кроме того, он может показывать отношение кластеров друг к другу, например, иерархию кластеров, встроенных друг в друга.

Кластеризацию можно условно разделить на две категории:

– жесткая кластеризация: каждый объект полностью принадлежит одному кластеру;
– мягкая кластеризация (нечеткая кластеризация): каждый объект принадлежит каждому кластеру в определенной степени (например, принадлежность к кластеру с некоторой вероятностью).

Возможны и более тонкие различия, например:

– строгое секционирование кластеризации: каждый объект принадлежит только одному кластеру;
– строгое секционирование кластеров с выбросами: объекты, которые не могут принадлежать ни к одному кластеру, считаются выбросами;
– перекрывающаяся кластеризация (альтернативная кластеризация, кластеризация с несколькими представлениями): объекты могут принадлежать более чем одному кластеру; обычно с участием жестких кластеров;
– иерархическая кластеризация: объекты, принадлежащие дочернему кластеру, также принадлежат родительскому кластеру;
– кластеризация подпространств: при перекрывающейся кластеризации в пределах однозначно определенного подпространства ожидается, что кластеры не должны перекрываться.

Рассмотрим подробнее некоторые виды кластеризации.

Кластеризация на основе подключения (иерархическая кластеризация, кластеризация на основе интерактивности, связности) основана на идее о том, что объекты больше связаны с соседними объектами, чем с объектами, находящимися дальше. Этот алгоритм соединяет объекты для формирования кластеров на основе их расстояния. Кластер может быть описан в основном максимальным расстоянием, необходимым для соединения частей кластера. На разных расстояниях будут образовываться разные кластеры, ко-

торые могут быть представлены с помощью дендрограммы, объясняющей, откуда взялось общее название «иерархическая кластеризация». Преимущество этого алгоритма заключается в том, что он обнаруживает обширную иерархию кластеров, которые сливаются друг с другом на определенных расстояниях. Недостаток – он не обеспечивает единого секционирования набора данных. Также этот алгоритм создает не уникальное секционирование набора данных, а иерархию, из которой пользователю по-прежнему необходимо выбрать соответствующие кластеры. Кластеры на основе подключения не очень устойчивы к выбросам, которые либо будут отображаться как дополнительные кластеры, либо даже приведут к слиянию других кластеров.

Иерархическая кластеризация опирается на поиск иерархии кластеров, которая напоминает древовидную структуру, называемую дендрограммой. Иерархическая кластеризация – это иерархическая декомпозиция данных на основе группового сходства (рис. 14).

Рис. 14. Иерархическая кластеризация
на основе континентов

Существует два метода определения верхнего уровня для поиска этих иерархических кластеров:

– агломеративная кластеризация использует подход «снизу вверх», при котором каждая точка данных начинается в своем собственном кластере; затем эти кластеры соединяются путем слияния наиболее похожих кластеров;
– разделяющая кластеризация использует нисходящий подход, при котором все точки данных начинаются в одном кластере. Затем можно использовать параметрический алгоритм кластеризации, такой как *k*-средние, чтобы разделить кластер на два; каждый кластер последовательно делится, пока не будет достигнуто нужное количество кластеров.

Оба этих подхода основаны на построении матрицы сходства между всеми точками данных, которая обычно вычисляется по косинусу или расстоянию Жаккара.

Следующий вид кластеризации, который будет рассмотрен, – это *центроидные модели* (*k*-средние). Кластеры на основе центроидов представлены центральным вектором, который не обязательно может быть членом набора данных. Когда число кластеров фиксируется равным *k*, кластеризация *k*-средних дает формальное определение в качестве задачи оптимизации: найти *k* кластерных центров и назначить объекты ближайшему центру кластера таким образом, чтобы квадратные расстояния от кластера были сведены к минимуму.

Сама задача оптимизации сложна, поэтому общий подход заключается в поиске только приближенных решений. Наиболее известным приближенным методом является алгоритм Ллойда, часто называемый просто «алгоритмом *k*-средних». Однако он находит только локальный оптимум и обычно запускается несколько раз с различными случайными инициализациями. Вариации *k*-средних часто включают в себя такие оптимизации, как:

– выбор лучшего из нескольких запусков, но также ограничение центроидов членами набора данных (*k*-медоиды);
– выбор медиан (кластеризация *k*-медиан);
– выбор начальных центров менее случайным образом (*k*-средние++);
– разрешение нечеткого кластерного назначения (нечеткие *c*-средние).

Большинство алгоритмов типа *k*-средних требуют, чтобы количество кластеров (*k*) указывалось заранее, что считается одним из самых больших недостатков этих алгоритмов. Кроме того, алгоритмы предпочитают кластеры примерно одинакового размера, так как они всегда будут присваивать объект ближайшему центроиду. Это часто приводит к неправильному нарезанию границ кластеров (что неудивительно, поскольку алгоритм оптимизирует кластерные центры, а не границы кластеров).

k-среднее обладает рядом интересных теоретических свойств:

– разбивает пространство данных на ячейки, образующие структуру, известную как диаграмма Вороного;
– концептуально близко к классификации ближайших соседей и как таковое популярно в машинном обучении;
– может рассматриваться как вариация кластеризации на основе модели, а алгоритм Ллойда – как вариация алгоритма максимизации ожиданий для этой модели.

Следующей рассмотрим *модель дистрибуции (распределения)*. Кластерная модель, наиболее тесно связанная со статистикой, основана на моделях распределения (дистрибуции). Кластеры можно легко определить как объекты, принадлежащие, скорее всего, одному и тому же распределению. Удобным свойством этого подхода является то, что он очень похож на способ генерации искусственных наборов данных путем выборки случайных объектов из распределения.

Несмотря на то что теоретическая основа этого метода понятна, он имеет одну ключевую проблему, известную как чрезмерная подгонка. Данную проблему можно избежать при назначении ограничений на сложность модели. Более сложная модель, как правило, сможет лучше объяснить данные.

Одним из известных методов являются модели смеси Гаусса (с использованием алгоритма максимизации ожиданий). Здесь набор данных обычно моделируется с фиксированным (чтобы избежать перенастройки) числом гауссовских распределений, которые инициализируются случайным образом и параметры которых итеративно оптимизированы для лучшего соответствия набору данных. Это соответствие будет сводиться к локальному оптимуму, поэтому несколько запусков могут привести к различным результатам. Что-

бы получить жесткую кластеризацию, объекты часто приближаются к гауссовскому распределению, к которому они, скорее всего, принадлежат; для мягких кластеров в этом нет необходимости.

Кластеризация на основе распределения создает сложные модели для кластеров, которые могут фиксировать корреляцию и зависимость между атрибутами. Однако эти алгоритмы создают дополнительную нагрузку на пользователя: для многих реальных наборов данных может не быть кратко определенной математической модели (например, требование о распределениях Гаусса является довольно жестким требованием к данным).

Еще одна модель кластеризации, которую мы рассмотрим, – это *модель плотности*. Кластеры на основе плотности определяются как области с более высокой плотностью, чем остальная часть набора данных. Объекты в этих разреженных областях, которые необходимы для разделения кластеров, обычно считаются шумовыми и пограничными точками.

Наиболее популярным методом кластеризации на основе плотности является DBSCAN. В отличие от многих новых методов он имеет четко определенную кластерную модель, называемую «плотность-достижимость». Подобно кластеризации на основе связей он основан на точках соединения в пределах определенных пороговых значений расстояния. Однако он соединяет только точки, которые удовлетворяют критерию плотности в исходном варианте, определяемом как минимальное количество других объектов в пределах этого радиуса. Кластер состоит из всех объектов, связанных плотностью (которые могут образовывать кластер произвольной формы, в отличие от многих других методов), плюс все объекты, которые находятся в пределах диапазона этих объектов. Еще одно интересное свойство DBSCAN заключается в том, что сложность его построения довольно низкая. Данный метод требует линейного количества запросов диапазона к базе данных. Он будет обнаруживать, по существу, одни и те же результаты (он детерминирован для основных и шумовых точек, но не для пограничных точек) в каждом запуске. Именно поэтому нет необходимости запускать его несколько раз.

OPTICS – это обобщение DBSCAN, которое устраняет необходимость выбора соответствующего значения для параметра range

и дает иерархический результат, связанный с кластеризацией связей. DeLi-Clu, Density-Link-Clustering сочетают в себе идеи кластеризации с одной связью и OPTICS, полностью исключая параметр range и предлагая повышение производительности по сравнению с OPTICS за счет использования индекса *R*-дерева.

Ключевым недостатком DBSCAN и OPTICS является то, что они ожидают какого-то падения плотности для обнаружения границ кластера. На наборах данных с, например, перекрывающимися распределениями Гаусса (достаточно распространенный вариант использования в искусственных данных) границы кластера, создаваемые этими алгоритмами, часто будут выглядеть произвольными, поскольку плотность кластера постоянно уменьшается. В наборе данных, состоящем из смесей Гаусса, эти алгоритмы почти всегда превосходят такие методы, как кластеризация ЭМ, которые способны точно моделировать такого рода данные.

Средний сдвиг – это кластерный подход, при котором каждый объект перемещается в самую плотную область в его окрестностях, основываясь на оценке плотности ядра. В конце концов, объекты сходятся к локальным максимумам плотности. Подобно кластеризации *k*-средних эти «аттракторы плотности» могут служить представителями для набора данных, но среднее смещение может обнаруживать кластеры произвольной формы, похожие на DBSCAN. Из-за дорогостоящей итеративной процедуры и оценки плотности средний сдвиг обычно медленнее, чем DBSCAN или *k*-средних. Кроме того, применимость алгоритма среднего сдвига к многомерным данным затруднена неравномерностью оценки плотности ядра, что приводит к чрезмерной фрагментации хвостов кластера.

Алгоритм DBSCAN требует наличия двух параметров:

– eps – определяет окрестности вокруг точки данных, то есть если расстояние между двумя точками меньше или равно eps, то они считаются соседями; если значение eps выбрано слишком малым, то большая часть данных будет рассматриваться как выбросы; если очень большим, то кластеры объединятся и большинство точек данных будут находиться в одних кластерах; один из способов найти значение eps основан на графике *k*-расстояния;

– minPts – минимальное количество соседей (точек данных) в радиусе eps; чем больше набор данных, тем больше значение minPts; как правило, минимальный minPts может быть получен из числа измерений D в наборе данных как minPts >= D + 1; минимальное значение minPts должно быть выбрано не менее 3.

В рассматриваемом алгоритме у нас есть три типа точек данных (рис. 15): основная точка, пограничная точка, шум (или выброс).

Кластеризация на основе сетки – еще один метод кластеризации, который предполагает создание сетчатой структуры. Далее на сетках (также известных как ячейки) выполняется сравнение. Метод на основе сетки является быстрым и имеет низкую вычислительную сложность. Существует два типа кластеризации на основе сетки: STING и CLIQUE.

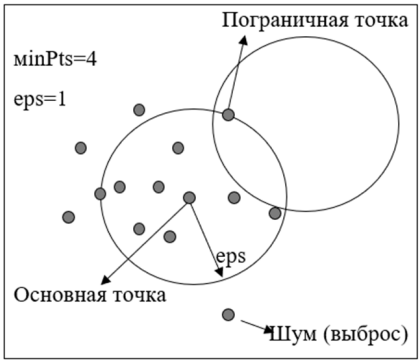


Рис. 15. Виды точек в DBSCAN

Кластерный подход на основе сетки отличается от обычных алгоритмов кластеризации тем, что он касается не точек данных, а пространства значений, которое окружает точки данных. В целом типичный алгоритм кластеризации на основе сетки состоит из следующих пяти основных шагов:

1. Создание структуры сетки, то есть разбиение пространства данных на конечное количество ячеек.
2. Расчет плотности каждой ячейки.
3. Сортировка клеток по их плотности.
4. Идентификация кластерных центров.
5. Обход соседних ячеек.

Рассмотрим различные приложения алгоритмов кластеризации.

Алгоритм кластеризации в медицине. Алгоритм кластеризации может быть использован при идентификации набора данных о каком-либо заболевании. Первоначально берутся и замеряются два набора данных с информацией о зараженных и здоровых образцах. Затем случайным образом смешиваются обе выборки и применяются различные алгоритмы кластеризации в смешанном наборе (фаза обучения алгоритма кластеризации). Далее, соответственно, проверяется результат на предмет соответствия исходным данным (поскольку это известные образцы, то результаты уже известны заранее). Следовательно, можно рассчитать процент полученных правильных результатов. Если этот результат устраивает исследователей, он применяется для кластеризации другого набора данных.

Алгоритм кластеризации в поисковых системах. Алгоритм кластеризации является основой поисковых систем. Поисковые системы пытаются сгруппировать похожие объекты в один кластер, а непохожие объекты – далеко друг от друга. Он предоставляет результат для искомых данных в соответствии с ближайшим аналогичным объектом, которые группируются вокруг данных для поиска. Чем качественнее используемый алгоритм кластеризации, тем лучше шансы получить требуемый результат. Следовательно, определение похожего объекта играет решающую роль в получении результатов поиска: чем качественнее определение аналогичного объекта, тем лучше результат.

Алгоритм кластеризации в образовании. Возможность контролировать успеваемость обучающихся всегда была важной задачей педагогического сообщества. Алгоритм кластеризации может быть использован для мониторинга успеваемости студентов. На основе оценок студентов они группируются в различные кластеры (с использованием k-средних, нечетких c-средних и т. д.), где каждый кластер обозначает разный уровень производительности. Зная количество студентов в каждом кластере, можно узнать среднюю производительность группы в целом.

Сегментация клиентов. Чаще всего кластеризация применяется для сегментации клиентов. Эта стратегия может быть применима

во всех сферах жизнедеятельности, включая телекоммуникации, электронную коммерцию, спорт, рекламу, продажи и т. д.

Кластерный анализ – это отличный способ узнать что-то новое из известных имеющихся данных, что может помочь разобраться в проблеме или исследовательском вопросе. Одна из самых полезных вещей, которую предоставляет кластеризация, заключается в том, что ее можно использовать в процессах контролируемого машинного обучения.

Тема 8. Методы проверки гипотез о взаимосвязи переменных

Гипотезой является утверждение, которое может быть проверено научными исследованиями. Если необходимо проверить связь между двумя или более вещами, нужно подготовить гипотезы перед началом сбора данных. Гипотеза может выглядеть следующим образом: ежедневное потребление яблок приводит к меньшему количеству посещений врача.

Гипотеза содержит прогнозы о том, что выявят проведенные исследования. Это предварительный ответ на исследовательский вопрос, который еще не был подтвержден эмпирически. Для некоторых исследовательских проектов, возможно, придется написать несколько гипотез, которые касаются различных аспектов исследовательского вопроса. Гипотеза – это не просто догадка. Гипотеза должна основываться на существующих теориях и знаниях. Гипотеза также должна быть проверяемой, что означает, что можно поддержать или опровергнуть ее с помощью научных методов исследования (таких как эксперименты, наблюдения и статистический анализ данных).

Гипотеза предполагает наличие связи между переменными:

– независимая переменная – это то, что исследователь изменяет или контролирует;
– зависимая переменная – это то, что исследователь наблюдает и измеряет.

Разработка гипотезы может происходить по определенному алгоритму:

1. Постановка вопроса. Написание гипотезы начинается с исследовательского вопроса, на который необходимо ответить. Вопрос должен быть сфокусированным, конкретным и исследуемым в рамках проекта.

2. Проведение предварительного исследования. Первоначальный ответ на вопрос должен основываться на том, что уже известно о теме. Необходимо искать теории или предыдущие исследования, которые помогут сформировать обоснованные предположения о том, что обнаружит проводимое исследование. На этом этапе можно построить концептуальную структуру, чтобы определить, какие переменные будут изучаться и каковы отношения между ними.

3. Формулировка гипотезы. После выполнения предыдущих этапов у исследователя должно сформироваться некоторое представление о том, что он ожидает найти. Важно представить первоначальный ответ на вопрос в ясном, лаконичном предложении.

4. Уточнение гипотезы. Необходимо убедиться, что сформулированная гипотеза специфична и проверяема. Существуют различные способы формулировки гипотезы, но все термины, которые используются, должны иметь четкие определения. Гипотеза должна содержать:

– соответствующие переменные;
– конкретную изучаемую группу;
– прогнозируемый результат эксперимента или анализа.

5. Окончательная формулировка гипотезы. Это можно сделать тремя способами:

1) чтобы идентифицировать переменные, можно написать простой прогноз в форме «если... то затем...». В первой части предложения указывается независимая переменная, а во второй части – зависимая переменная. Например: если студент первого курса начнет посещать больше лекций, то его экзаменационные баллы улучшатся;

2) гипотеза может формулироваться в терминах корреляций или эффектов, где напрямую заявляется предсказанная связь между переменными. Например: количество лекций, которые посещают студенты первого курса, положительно влияет на их экзаменационные баллы;

3) если сравниваются две группы, гипотеза может указать, какая разница ожидается между ними. Например: студенты первого курса, которые посетили большинство лекций, будут иметь лучшие экзаменационные баллы, чем те, кто посетил несколько лекций.

6. Написание нулевой гипотезы. Если исследование включает статистическую проверку гипотез, также придется написать нулевую гипотезу. Нулевая гипотеза – это позиция по умолчанию, согласно которой между переменными нет никакой связи (табл. 12). Нулевая гипотеза записывается как H_0, в то время как альтернативной гипотезой является H_1 или H_a.

Примеры сформулированных гипотез:

– H_0: количество лекций, которые посещают студенты первого курса, не влияет на их итоговые экзаменационные баллы;
– H_1: количество лекций, которые посещают студенты первого курса, положительно влияет на их итоговые экзаменационные баллы.

Таблица 12

Примеры гипотез на основе исследовательского вопроса

Исследовательский вопрос	Гипотеза	Нулевая гипотеза
Какова польза для здоровья от употребления яблока в день?	Увеличение потребления яблок в возрасте старше 60 лет приведет к снижению частоты посещений врача	Увеличение потребления яблок в возрасте старше 60 лет не повлияет на частоту посещений врача
У каких авиакомпаний больше всего задержек?	Бюджетные авиакомпании с большей вероятностью будут иметь задержки, чем премиальные авиакомпании	Бюджетные и премиальные авиакомпании с одинаковой вероятностью будут иметь задержки
Может ли гибкий график работы повысить удовлетворенность работой?	Сотрудники, которые имеют гибкий рабочий график, будут сообщать о большей удовлетворенности работой, чем сотрудники, которые работают по фиксированному графику	Нет никакой связи между гибкостью рабочего времени и удовлетворенностью работой

Окончание табл. 12

Исследовательский вопрос	Гипотеза	Нулевая гипотеза
Насколько эффективно половое воспитание в средней школе для сокращения подростковых беременностей?	Подростки, которые получили уроки полового воспитания в средней школе, будут иметь более низкие показатели незапланированной беременности, чем подростки, которые не получали никакого полового воспитания	Половое воспитание в средней школе не влияет на показатели подростковой беременности
Какое влияние оказывает ежедневное использование социальных сетей на концентрацию внимания детей в возрасте до 16 лет?	Существует отрицательная корреляция между временем, проведенным в социальных сетях, и концентрацией внимания в возрасте до 16 лет	Нет никакой связи между использованием социальных сетей и концентрацией внимания у детей в возрасте до 16 лет

Проверка гипотез – это формальная процедура исследования представлений о мире с помощью статистики. Она используется учеными для проверки конкретных предсказаний, называемых гипотезами, путем вычисления того, насколько вероятно, что паттерн или связь между переменными могли возникнуть случайно.

Для сравнения средних значений двух групп используется *t*-тест. Этот статистический тест часто применяется для проверки гипотез, чтобы определить, действительно ли переменные связаны. *t*-тест предполагает наличие двух видов гипотез:

– нулевая гипотеза (H_0) заключается в том, что истинная разница между этими групповыми средними равна нулю;
– альтернативная гипотеза (H_a) заключается в том, что истинная разница отличается от нуля.

t-тест может использоваться только при сравнении средних значений двух групп (также известно как попарное сравнение). Если необходимо сравнить более двух групп или если нужно выполнить несколько парных сравнений, используется тест ANOVA или пост-специальный тест.

t-тест – это параметрический тест различий, означающий, что он делает те же предположения об имеющихся данных, что и другие параметрические тесты. *t*-тест накладывает определенные требования к набору данных:

– данные должны быть независимыми;
– данные должны быть нормально распределенными (хотя бы приблизительно);
– данные должны иметь одинаковые дисперсии в каждой сравниваемой группе (так называемая однородность дисперсии).

Если данные не соответствуют этим требованиям, можно попробовать непараметрическую альтернативу *t*-тесту, например тест Wilcoxon Signed-Rank для данных с неравными дисперсиями.

При выборе *t*-теста нужно будет учитывать две вещи:

– происходят ли сравниваемые группы из одной популяции или двух разных популяций;
– необходимо ли проверить разницу в определенном направлении.

При различных исследовательских условиях применяются различные виды *t*-тестов.

Если группы взяты из одной популяции (например, измерение до и после экспериментального лечения), то следует выполнить парный *t*-тест.

Если группы происходят из двух разных популяций (например, двух разных видов или людей из двух разных городов), то выполняется *t*-тест с двумя выборками (он же независимый *t*-тест).

Если сравнивается одна группа со стандартным значением (например, сравнивается кислотность жидкости с нейтральным значением pH 7), проводится *t*-тест с одной выборкой.

Если в исследовании выясняется только то, отличаются ли две популяции друг от друга, – выполняется двухвостый *t*-тест.

Если необходимо узнать, является ли одно среднее значение популяции больше или меньше другого, – выполняется однохвостый *t*-тест.

Рассмотрим пример определения зависимости длины лепестка от вида цветка.

Наблюдения получены из двух отдельных популяций (отдельных видов), поэтому выполняется *t*-тест с двумя выборками.

Исследователя не волнует направление разницы, только то, есть ли разница, поэтому используется двухвостый *t*-тест.

Рассмотрим процедуру выполнения *t*-теста. *t*-тест оценивает истинную разницу между данными двух групп, используя отношение разности в групповых средних к объединенной стандартной погрешности обеих групп. Можно рассчитать эту разницу вручную с помощью формулы или использовать программное обеспечение для статистического анализа.

Формула для *t*-теста с двумя выборками (также известного как *t*-тест Стьюдента) приведена ниже:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}}, \quad (15)$$

где t – *t*-значение; x_1 и x_2 – средства сравнения двух групп; s – стандартная ошибка объединения двух групп; n_1 и n_2 – количество наблюдений в каждой из групп.

Бóльшее *t*-значение показывает, что разница между групповыми средними больше, чем стандартная ошибка набора данных, что указывает на более существенную разницу между группами.

Можно сравнить вычисленное *t*-значение со значениями в таблице критических значений, чтобы определить, является ли полученное *t*-значение больше ожидаемого. Если это так, то можно отвергнуть нулевую гипотезу и заключить, что эти две группы на самом деле разные.

Большинство статистических программ (R, SPSS и т. д.) включают функцию *t*-теста. Эта встроенная функция использует необработанные данные и рассчитывает *t*-значение. Затем сравнивает его с критическим значением и рассчитывает *p*-значение. Таким образом, можно быстро увидеть, отличаются ли изучаемые группы статистически.

При представлении результатов *t*-теста наиболее важными значениями, которые необходимо включить, являются *t*-значение, *p*-значение и степени свободы для теста. Они сообщат, является ли разница между двумя группами статистически значимой (например, что это вряд ли произошло случайно).

t-значение может быть и отрицательным. В большинстве случаев исследователя интересует только абсолютное значение разницы, или расстояние от 0, не важно, в каком направлении.

Степень свободы связана с размером выборки и показывает, сколько точек данных доступно в тесте для проведения сравнений. Чем больше степень свободы, тем лучше будет работать статистический тест.

p-значение описывает вероятность того, что и при повторном анализе получится такое же *t*-значение, как и это.

Доверительный интервал — это диапазон чисел, в пределах которого истинная разница в средних будет составлять указанный процент. Как правило, используется 95%-ный интервал.

В нашем примере с цветочными лепестками сообщить о результатах можно следующим образом: разница в длине лепестков между видом 1 (среднее = 1,46; $SD = 0{,}206$) и видом 2 (среднее = 5,54; $SD = 0{,}569$) оказалась значительной ($t(30) = -33{,}7190$; $p < 2{,}2e - 16$).

Контрольные вопросы

1. Для чего используется корреляционный анализ? Что включено в процедуру проведения корреляционного анализа?
2. Для каких ситуаций подходят корреляционные исследования? Приведите примеры для аргументации своего ответа.
3. Как осуществляется в корреляционном анализе контроль исследователя над переменными?
4. Что такое коэффициент корреляции? Каков диапазон коэффициента корреляции? Что он отражает?
5. Что такое идеальная отрицательная корреляция? Какое значение коэффициента корреляции должно быть в случае идеальной отрицательной корреляции?
6. Что описывают регрессионные модели? Какие типы регрессионных моделей существуют? Перечислите и охарактеризуйте каждую из них.
7. Какими характеристиками должен обладать набор данных, чтобы была возможность применить к нему метод простой линейной регрессии?

8. Что такое регрессионные тесты? Каково значение регрессионных тестов для анализа данных?
9. Когда в аналитике данных используется множественная линейная регрессия? Что позволяет узнать о наборе данных множественная линейная регрессия?
10. В чем суть проведения факторного анализа? Какие три формы факторного анализа существуют? Перечислите и охарактеризуйте каждую из них.
11. Что представляет собой кластерный анализ, или кластеризация? В чем суть данной аналитической процедуры? Как можно охарактеризовать понятие кластера? В чем его исследовательская значимость?
12. Какие типичные модели кластеров существуют? Перечислите и охарактеризуйте некоторые из них.
13. Что представляет собой кластеризация на основе подключения (иерархическая кластеризация, кластеризация на основе интерактивности)? Каков алгоритм реализации данной модели кластеризации? Каковы ее преимущества и недостатки?
14. Что представляют собой центроидные модели (*k*-средние) кластеризации? Каков алгоритм реализации данной модели кластеризации? Каковы ее преимущества и недостатки?
15. Что такое гипотеза? Для чего она необходима? По какому алгоритму происходит разработка гипотезы? Для чего необходимо проводить предварительное исследование при работе с гипотезой?

Тесты для самоконтроля

1. В корреляционных исследованиях используются ... методы для изучения отношений и связей между переменными. (*один вариант ответа*)

1) количественные
2) номинальные
3) текстовые
4) графические

2. Корреляционный анализ используется для проверки ... между переменными. (*один вариант ответа*)
1) прочности связи
2) причинно-следственных отношений
3) дисперсии
4) среднеквадратического интервала

3. Как называются данные, которые уже были собраны для других целей, таких как официальные отчеты, опросы или предыдущие исследования, и которые можно использовать в исследовании? (*один вариант ответа*)
1) вторичные данные
2) государственная статистика
3) аналитические данные
4) проверенные данные

4. Есть несколько ситуаций, когда корреляционное исследование является подходящим выбором: исследование непричинных связей, изучение причинно-следственных связей между переменными и... (*один вариант ответа*)
1) тестирование новых средств измерений
2) определение воздействующих факторов
3) расчет критериев воздействия главных факторов
4) обоснование выборочной совокупности

5. Какие исследования могут быть использованы для оценки того, насколько последовательно или точно новый инструмент отражает концепцию, которую он стремится измерить? (*один вариант ответа*)
1) корреляционные исследования
2) аналитические исследования
3) графологические исследования
4) графические исследования

6. В каких исследованиях используются анкеты для измерения изучаемых переменных? (*один вариант ответа*)
1) в опросных
2) в контент-анализе

3) в интеллектуальном анализе данных
4) в корреляционных

7. Что используется в опросных исследованиях для измерения изучаемых переменных? (*один вариант ответа*)
1) анкеты
2) гайды
3) единицы счета
4) коэффициенты

8. Что такое вторичные данные? (*один вариант ответа*)
1) данные, которые уже были собраны для других целей (официальные отчеты, опросы или предыдущие исследования)
2) данные, которые были проверены и отобраны для проведения аналитического исследования
3) данные официальной государственной статистики
4) статистические данные, которые собираются организациями, предприятиями и учреждениями

9. Существует два основных типа линейной регрессии. Укажите их. (*несколько вариантов ответа*)
1) простая линейная регрессия
2) множественная линейная регрессия
3) сложная линейная регрессия
4) вторичная линейная регрессия

10. Что использует множественная линейная регрессия? (*один вариант ответа*)
1) две и более независимых переменных
2) только одну независимую переменную
3) не менее трех независимых переменных
4) более одной независимой переменной

11. Какой коэффициент необходимо возвести в квадрат, чтобы получить коэффициент детерминации? (*один вариант ответа*)
1) коэффициент корреляции
2) коэффициент регрессии
3) коэффициент осцилляции
4) коэффициент модуляции

12. Каким видом статистического теста является простая линейная регрессия, если она накладывает определенные требования к данным? (*один вариант ответа*)

1) параметрическим тестом
2) непараметрическим тестом
3) аналитическим тестом
4) графическим тестом

13. Какой вид анализа помогает найти уравнение для линии наилучшего соответствия, которое можно использовать для прогнозирования значения одной переменной с учетом значения для другой переменной? (*один вариант ответа*)

1) регрессионный анализ
2) статистический анализ
3) корреляционный анализ
4) графический анализ

14. Как называется требование простой линейной регрессии к данным, заключающееся в том, что данные должны следовать нормальному распределению? (*один вариант ответа*)

1) нормальность распределения данных
2) однородность дисперсии
3) независимость наблюдений
4) репрезентативность выборки

15. Что показывают регрессионные модели? (*один вариант ответа*)

1) взаимосвязь между переменными путем подгонки линии тренда к наблюдаемым данным
2) взаимосвязь между переменными путем расчета групповой дисперсии
3) взаимосвязь между переменными путем расчета специальных статистических коэффициентов
4) взаимосвязь между переменными путем распределения данных по выборке на данные по генеральной совокупности

16. Существует три основных формы факторного анализа. Укажите их. (*несколько вариантов ответа*)
1) исследовательский факторный анализ
2) подтверждающий факторный анализ
3) оценивающий факторный анализ
4) статистический факторный анализ

17. Факторный анализ предполагает ряд основных требований к набору данных. Укажите два из них. (*несколько вариантов ответа*)
1) выборка должна быть однородной
2) существует истинная корреляция между переменными и факторами
3) набор данных для анализа включает переменные разного типа
4) размер выборки должен быть менее 200

18. В чем заключается требование к данным для применения к ним факторного анализа, называемое «набор данных для анализа включает соответствующие переменные»? (*один вариант ответа*)
1) набор данных содержит релевантные для проведения анализа переменные. В анализ включены соответствующие переменные из набора данных
2) выборка должна быть однородной
3) используемые переменные должны быть только метрическими
4) набор данных для анализа включает номинальные переменные

19. Что влечет за собой нарушение требования об однородности выборки для применения факторного анализа? (*один вариант ответа*)
1) увеличение размера выборки по мере увеличения числа переменных
2) увеличение размера выборки по мере уменьшения числа респондентов
3) уменьшение размера выборки по мере увеличения числа переменных
4) уменьшение размера выборки по мере уменьшения числа респондентов

20. В чем заключается требование к данным для применения к ним факторного анализа, называемое «существует истинная корреляция между переменными и факторами»? (*один вариант ответа*)
1) переменные и факторы действительно связаны
2) используемые переменные должны быть только метрическими
3) набор данных для анализа включает соответствующие переменные
4) используемые переменные должны быть только метрическими

21. Существуют различные типы методов, используемые для извлечения факторов из набора данных. Укажите два из них. (*несколько вариантов ответа*)
1) анализ главных компонентов
2) общий факторный анализ
3) среднеквадратическая нормализация
4) линейная регрессия

22. Факторный анализ предполагает ряд основных требований к набору данных. Укажите два из них. (*несколько вариантов ответа*)
1) отсутствие выбросов в данных
2) наличие линейной зависимости
3) набор данных для анализа включает переменные разного типа
4) размер выборки должен быть менее 200

23. Существуют различные типы методов, используемых для извлечения факторов из набора данных. Укажите два из них. (*несколько вариантов ответа*)
1) анализ главных компонентов
2) метод максимальной вероятности
3) среднеквадратическая нормализация
4) линейная регрессия

24. Задачей какого анализа является группировка набора объектов таким образом, чтобы объекты в одной группе были более похожи друг на друга, чем на объекты в других группах? (*один вариант ответа*)
1) кластерный анализ
2) регрессионный анализ

3) факторный анализ
4) статистический анализ

25. Какой процесс автоматически разбивает набор данных на группы на основе их сходства? (*один вариант ответа*)
1) кластеризация
2) факторинг
3) группировка
4) корреляция

26. Иерархическая кластеризация опирается на поиск иерархии кластеров, которая напоминает древовидную структуру, называемую... (*один вариант ответа*)
1) дендрограммой
2) деревом принятия решений
3) гистограммой
4) графом

27. Как называется утверждение, которое может быть проверено научными исследованиями? (*один вариант ответа*)
1) гипотеза
2) вопрос
3) цель
4) задача

28. Гипотеза предполагает связь между двумя или более переменными. Как называются эти переменные? (*несколько вариантов ответа*)
1) независимая переменная
2) зависимая переменная
3) ключевая переменная
4) основная переменная

29. На каком этапе происходит построение концептуальной структуры исследования и определяется, какие переменные будут изучаться и каковы отношения между ними? (*один вариант ответа*)
1) проведение предварительного исследования
2) постановка вопроса

3) формулировка гипотезы
4) уточнение гипотезы

30. Что предполагает этап проведения предварительного исследования? (*один вариант ответа*)

1) построение концептуальной структуры, чтобы определить, какие переменные будут изучаться и каковы отношения между ними
2) формальную процедуру исследования представлений о мире с помощью статистики
3) поиск теорий или исследований, которые помогут сформировать обоснованные предположения о том, что обнаружит проводимое исследование
4) формулирование исследовательского вопроса, на который необходимо ответить

Практическое задание

Проведите статистический анализ данных средствами MS Excel, Python или *R*.

Методические указания

1. Подготовьте данные, на основе которых будет проводиться статистический анализ. Представьте данные в виде таблицы.

2. Найдите числовые характеристики выборки (выборочное среднее, медиана, мода, наибольшее, наименьшее, дисперсия, среднее квадратичное отклонение и т. д.).

3. Выполните описательную статистику выборки.

4. Проведите дисперсионный анализ данных.

5. Проведите корреляционный и регрессионный анализ данных.

6. Предоставьте отчет, который содержит этапы выполнения задания с выводами.

В качестве примера рассмотрим статистическую модель. Существует двухмерная выборка, состоящая из выборочных значений x_{ij}; индекс i соответствует уровню β_i фактора β, индекс j соответствует уровню γ_j фактора γ. Пусть фактор β имеет r уровней, а фактор $\gamma - t$ уровней; выборка имеет размерность $r \times t$. Таким образом, каждое выборочное значение x_{ij} можно представить в виде

$$x_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij},$$

где μ – константа (общее среднее); ε_{ij} – случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и одинаковыми дисперсиями. Все величины ε_{ij} независимы.

Гипотезы

А. Равенство значений уровней фактора β

H_0: $\beta_1 = \beta_2 = ... = \beta_r$;

H1: не все значения уровней равны.

Б. Равенство значений уровней фактора γ

H_0: $\gamma_1 = \gamma_2 = ... = \gamma_t$;

H1: не все значения уровней равны.

Задан уровень значимости α.

Если нулевая гипотеза отклоняется, значит не все значения уровней фактора одинаковы. Для того чтобы определить, какие значения уровней фактора отличаются от других, следует применить метод множественных сравнений Шеффе:

1. В таблицы MS Excel представим выборку, по значениям которой вычисляются средние по строкам и столбцам и общее среднее.

Двухфакторный дисперсионный анализ, рабочий лист:

	A	B	C	D	E	F	G	H	I
1		Выборка					r =	5	
2		Бета 1	Бета 2	Бета 3	Бета 4	Средние	t =	4	
3	Гамма 1	7,026	39,023	44,031	35,4835	31,390875	Уровень значимости		
4	Гамма 2	22,4409	6,3142	39,578	15,5309	20,966000		0,05	
5	Гамма 3	8,02589	24,941	20,6053	9,954184	15,881594	T_β =	0,6892	
6	Гамма 4	8,93419	14,128	4,12927	41,70589	17,224338	$T\gamma$ =	0,2899	
7	Гамма 5	39,6024	44,138	17,1554	9,302168	27,549492	Критические значения		
8	Средние	17,20588	25,70884	25,099794	22,3953284	22,602460	t_β =	3,25917	
9	Дисперсионная таблица						$t\gamma$ =	3,49029	
10	Σ квадратов		df	Дисперсия	Гипотеза β	принимается			
11	Фактор β	713,93	4	178,4825	Гипотеза γ	принимается			
12	Фактор γ	225,262	3	75,0873					
13	Остаток	3107,79	12	258,9825					
14	Полная	4046,982	19	212,9991					
15									

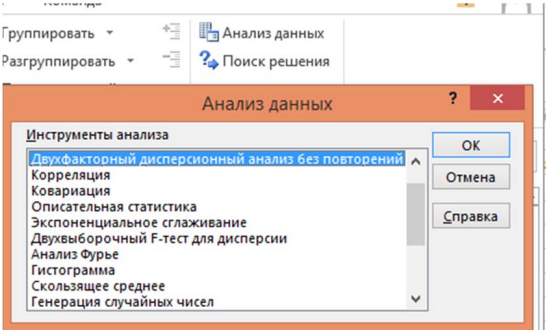
Ниже приведены формулы, необходимые для вычисления критерия.

F3 = СРЗНАЧ(B3:E3) (с помощью маркера автозаполнения нужно заполнить ячейки диапазона F4:F7);

H5 = D11/D13;

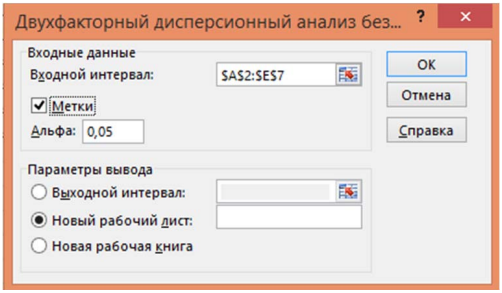
H6 = D12/D13;
H8 = FРАСПОБР(H4;C11;C13);
H9 = FРАСПОБР(H4;C12;C13);
C11 = H1-1;
C12 = H2-1;
C13 = C11*C12;
C14 = H1*H2-1;
D11 = B11/C11 (с помощью маркера автозаполнения нужно заполнить ячейки диапазона D12:D14);
F10 = ЕСЛИ(H5<H8;»принимается»;»отвергается»);
F11 = ЕСЛИ(H6<H9;»принимается»;»отвергается»).

2. В MS Excel данный критерий также можно реализовать с помощью средства **Двухфакторный дисперсионный анализ** без повторений из пакета анализа. Для этого на вкладке **Данные** выберите **Анализ данных**, в появившемся диалоговом окне выберите **Двухфакторный дисперсионный анализ без повторений**:



В поле **Входной интервал** укажите диапазон ячеек, содержащий входные данные. Если в этот диапазон включены заголовки строк и столбцов, нужно установить флажок опции **Метки**. Для рассматриваемого примера входным интервалом является диапазон ячеек A2:E7.

Исходные данные и **Двухфакторный дисперсионный анализ без повторений**:



Результат будем выводить на новый рабочий лист.

Результаты двухфакторного дисперсионного анализа без повторений:

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
2							
3	*ИТОГИ*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
4	Гамма 1	4	125,5635	31,390875	276,1396017		
5	Гамма 2	4	83,864	20,966	197,5991431		
6	Гамма 3	4	63,526374	15,881594	67,07750336		
7	Гамма 4	4	68,89735	17,224338	283,047019		
8	Гамма 5	4	110,197968	27,549492	287,1510569		
9							
10	Бета 1	5	86,02938	17,205876	196,3331428		
11	Бета 2	5	128,5442	25,70884	256,9395722		
12	Бета 3	5	125,49897	25,099794	272,7714397		
13	Бета 4	5	111,976642	22,395328	229,3828314		
14							
15							
16	Дисперсионный анализ						
17	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
18	Строки	713,9265	4	178,48161	0,68916665	0,613292561	3,259166727
19	Столбцы	225,2615	3	75,08716	0,2899322	0,83185666	3,490294819
20	Погрешность	3107,781	12	258,98179			
21							
22	Итого	4046,969	19				

Выходные результаты сгруппированы в две таблицы. Здесь принимаются обе нулевые гипотезы об отсутствии влияния факторов β и γ.

Глава 3. ОБЩЕЕ ПРЕДСТАВЛЕНИЕ О ВИЗУАЛИЗАЦИИ ДАННЫХ

Тема 9. Визуализация данных и ее характеристики

Как легко догадаться из названия, визуализация данных – это графическое представление каких-либо данных.

Визуализация данных позволяет лицам, принимающим решения, видеть аналитику, представленную визуально, чтобы они могли понять сложные концепции или выявить новые закономерности. В рамках интерактивной визуализации можно использовать технологию построения диаграмм и графиков для получения более подробной информации, интерактивного изменения данных.

С учетом того, что человеческий мозг от природы склонен обрабатывать визуальную информацию, использование диаграмм или графиков для визуализации больших объемов сложных данных более эффективно, чем составление электронных таблиц или отчетов. Визуализация данных – это быстрый и простой способ универсальной передачи концепций, что позволяет экспериментировать с различными сценариями, внося небольшие коррективы.

Визуализация данных также может:

– определить области, которые требуют внимания или улучшения;
– уточнить, какие факторы влияют на поведение клиентов;
– помочь понять, какие продукты нужны и где их разместить;
– спрогнозировать объемы продаж.

Изображения, используемые при визуализации данных, обладают интерактивными и динамическими возможностями, которые позволяют пользователям манипулировать ими, извлекать данные для запросов и глубокого анализа.

Визуализация данных не ограничивается стандартными диаграммами и графиками, созданными в электронной таблице Microsoft Excel. Существует множество других доступных способов отображения данных, таких как шкалы и датчики, географические карты, инфографика, тепловые карты, гистограммы и круговые диаграммы и т. д.

Для создания хорошей визуализации данных требуются тщательно отобранные, полные и чистые данные. Затем выбирается правильная диаграмма. После чего осуществляется проектирование и настройка визуализации в соответствии с простыми предпочтениями. После завершения визуализации ее публикуют для зрителей.

Визуализация данных является частью многих инструментов бизнес-аналитики и ключом к расширенной аналитике. Она помогает оценивать значение информации или данных, которые генерируются сегодня. При визуализации данных информация представляется в графической форме, в виде круговой диаграммы, графика или другого типа визуального представления.

Этот метод предназначен для более глубокого изучения предоставленной информации, особенно в интерактивном режиме.

Визуализация данных – это также процесс взаимодействия с клиентами для понимания закономерностей, тенденций и идей путем преобразования данных в визуальный контекст. Она сосредоточена на отчетных данных и означает взаимодействие с данными клиентов. Визуализация данных обычно решает аналитическую задачу. Интересующая клиентов информация должна быть представлена как решение, достаточное для выполнения этой задачи.

Таким образом, необработанная информация – тексты, числа или символы – визуализируется с четкой целью: показать логические корреляции между единицами и определить склонности, тенденции и закономерности.

Визуальные эффекты создаются с помощью соответствующего программного обеспечения, будь то PowerPoint или Photoshop. Но основное назначение визуальных эффектов – решение задач аналитики. По этой причине визуализация данных или дата-анализ стали стандартными способами представления информации пользователям через интерфейс бизнес-аналитики как инструмент представления данных.

Проще говоря, визуализация данных – это умный и простой способ передать концепции определенным образом, чтобы можно было экспериментировать с различными данными, делая небольшие манипуляции и корректировки и получая разные визуальные эффекты.

Есть несколько классов визуализации:

– обычное визуальное представление количественной информации в схематическом виде. В эту группу входят все известные круговые и линейные диаграммы, гистограммы и спектрограммы, таблицы и различные точечные диаграммы;

– преобразование данных в форму, которая улучшает восприятие и упрощает анализ этой информации. Это может быть карта и полярный график, временная шкала и график с параллельными осями, диаграмма Эйлера;

– концептуальная визуализация, которая позволяет разрабатывать сложные концепции, идеи и планы с помощью концептуальных карт, диаграмм Ганта, графиков с минимальными путями и других подобного типа диаграмм;

– стратегическая визуализация – представление различных данных по всем аспектам работы организации. Это всевозможные диаграммы производительности, жизненного цикла и графики структур организаций;

– метафорическая визуализация, которая поможет графически организовать структурную информацию с помощью пирамид, деревьев и карточек данных, ярким примером ее является карта метро;

– комбинированная визуализация, позволяющая объединить несколько сложных графиков в одну схему, как на карте с прогнозом погоды.

Один из способов классифицировать визуализацию данных – это подсчитать, сколько различных *измерений данных* она представляет. Под этим мы подразумеваем количество дискретных типов информации, которые визуально закодированы на диаграмме. Например, простой линейный график может отображать *цену* акций компании в разные *дни*: это два измерения данных.

По количеству измерений данных можно определить уровень *сложности* визуализации. По мере того как визуализации становятся более сложными, их становится все сложнее проектировать и на них становится труднее учиться. По этой причине наиболее распространены визуализации с не более чем тремя или четырьмя измерениями данных, хотя можно найти визуализации с шестью, семью или более измерениями.

При разработке более сложных визуализаций возникают две основные проблемы.

Во-первых, чем больше измерений нужно визуально кодировать, тем больше индивидуальных визуальных свойств нужно использовать. Выбрать свойства легко для первых нескольких измерений, когда большинство визуальных свойств не используется. Однако по мере добавления дополнительных измерений становится все труднее найти подходящие неиспользуемые визуальные свойства. Визуализация показывает не только типы информации, но и взаимосвязи между этими типами информации.

Путь к успеху перед лицом этой проблемы состоит в том, чтобы целенаправленно выбирать, какое свойство использовать для каждого измерения, и повторять или изменять кодировки по мере развития дизайна.

Вторая проблема при разработке более сложных визуализаций заключается в том, что хорошо известных соглашений, метафор, значений по умолчанию и передовых методов, на которые можно положиться, относительно мало.

Существует четыре типа визуальной коммуникации в анализе данных (рис. 16).

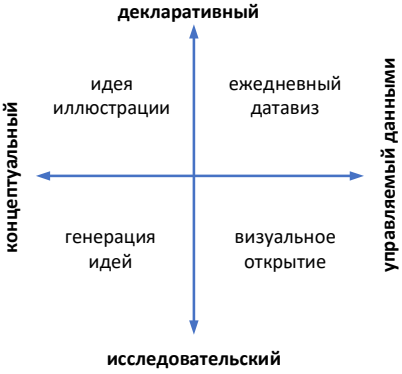


Рис. 16. Типы визуальной коммуникации

Идея иллюстрации. Можно назвать этот сектор «уголком консультантов». Консультанты не могут устоять перед диаграммами процессов, диаграммами циклов и т. п. В лучшем случае идеи

иллюстраций проясняют сложные идеи, опираясь на способность понимать метафоры и простые дизайнерские соглашения. Организационные диаграммы и деревья решений – классические примеры иллюстрации идей. Акцент в определении идеи должен быть сделан на четкой коммуникации, структуре и логике идей.

Генерация идей опирается на концептуальные метафоры, но происходит в более неформальной обстановке. Она используется для поиска новых способов увидеть, как работает бизнес, и для решения сложных управленческих задач: реструктуризация организации, разработка нового бизнес-процесса, кодификация системы принятия решений.

Визуальное открытие. Это наиболее сложный квадрант, потому что на самом деле он состоит из двух категорий: объем данных и тип диаграммы.

Объем данных, как правило, управляем, а типы диаграмм являются общими.

Визуальное исследование – это открытые визуализации на основе данных, которыми занимаются специалисты по данным и бизнес-аналитики.

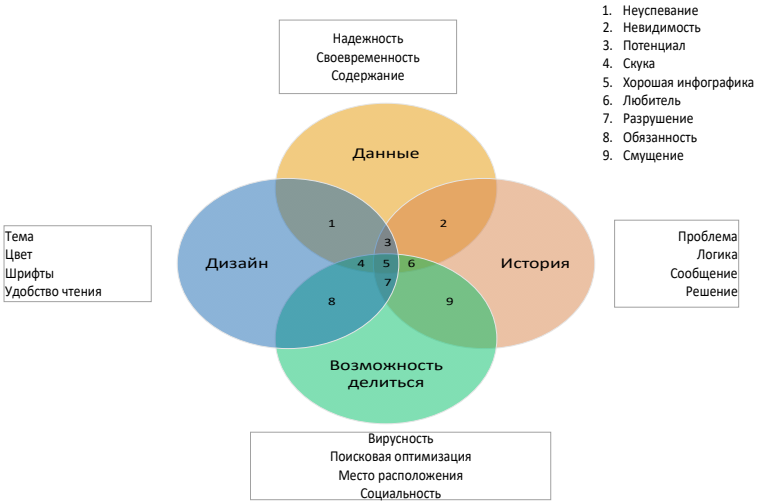


Рис. 17. Техника хорошей визуализации

Хорошая визуализация данных создается на стыке коммуникации, науки о данных и дизайна. Правильно выполненная визуализация данных позволяет легко понять, что представляют собой сложные наборы данных (рис. 17). Американский статистик и профессор Йельского университета Эдвард Тафте считает, что отличная визуализация данных состоит из «сложных идей, переданных с ясностью, точностью и эффективностью».

Чтобы создать хорошую визуализацию данных, нужно начать с чистых полных данных, которые имеют хороший источник. Когда данные готовы к визуализации, нужно выбрать правильную диаграмму. Это может быть сложно, но существует множество ресурсов, которые помогут выбрать правильный тип диаграммы для данных.

Рассмотрим, почему так важна визуализация данных.

Лучшее принятие решений. Сегодня различные организации активно используют визуализацию данных и инструменты обработки данных, чтобы задавать более точные вопросы и принимать более обоснованные решения. Новые компьютерные технологии и новые удобные для пользователя программы позволяют легко узнать больше о компании и принимать более обоснованные бизнес-решения на основе полученных данных.

Осмысленное повествование. Визуализация данных и информационная графика (инфографика) стали важным инструментом для современных СМИ.

Профессор Высшей школы бизнеса Стэнфордского университета Дженнифер Аакер сказал: «Когда данные и истории используются вместе, они находят отклик у аудитории как на интеллектуальном, так и на эмоциональном уровне».

Грамотность данных. Визуализация данных стала объективной необходимостью, которая вызвана наличием большого потока данных. Поскольку инструменты и ресурсы визуализации данных стали легкодоступными, ожидается, что все больше и больше нетехнических специалистов смогут собирать информацию из данных.

Генеральный директор Infogr.am Микко Джарвенпаа объясняет: «Мы считаем, что более информированные люди принимают более правильные решения, а люди, которые умеют как читать, так и создавать коммуникации на основе данных, играют ключевую роль в этом».

Тема 10. Визуализация данных и визуализация информации

Рассмотрим основные характеристики *визуализации данных* и *визуализации информации*:

– отрисовывается алгоритмически, с помощью компьютеризированных методов;
– легко регенерируется с использованием разных данных: одну и ту же форму можно изменить для представления разных наборов данных с аналогичными размерами или характеристиками;
– ориентирована на большой объем данных.

Визуализации данных изначально разрабатываются человеком, но затем создаются алгоритмически с помощью программного обеспечения для построения графиков или диаграмм.

Преимущество этого подхода в том, что относительно просто обновить или восстановить визуализацию с добавлением большего количества новых данных. Хотя информационные визуализации могут отображать большие объемы данных, они зачастую менее эстетичны, чем инфографика.

Существует две категории визуализации данных: исследование и объяснение. Они служат разным целям, поэтому использую разные инструменты и разные методы.

Для исследования не важно, что содержится в огромном объеме данных.

Исследование включает детализацию. В данных может быть много шума, но, если упростить или удалить слишком много информации, можно упустить что-то важное. Этот тип визуализации обычно является частью фазы анализа данных и используется для поиска истории, которую данные должны рассказать.

Объяснение применяется для рассказа истории. История, которая рассказывается, известна автору с самого начала, и поэтому можно сразу разработать дизайн, специально адаптированный под данную историю.

Исследование – это часть фазы анализа данных, а объяснение – часть фазы представления.

Для визуализации данных используются обе категории: сначала анализ (исследование), затем представление (объяснение).

Стоит отметить, что существует своего рода гибридная категория, которая включает тщательно подобранный набор данных, который предназначен и для анализа (исследования), и для объяснения полученных результатов. Эти визуализации обычно интерактивны через какой-то графический интерфейс, который позволяет выбирать и ограничивать определенные параметры, тем самым открывая для себя любые идеи, которые может предложить набор данных. Таким образом, в гибридных проектах присутствует определенный аспект свободы открытия в представленной информации.

Полезно представить себе эффективную визуализацию пояснительных данных (рис. 18), поддерживаемую схемой, состоящей из дизайнера, читателя и данных. Каждый элемент схемы имеет уникальное отношение к двум другим. Хотя необходимо учитывать потребности и перспективы всех трех в каждом проекте визуализации, доминирующая взаимосвязь в конечном итоге определит, какая категория визуализации необходима.

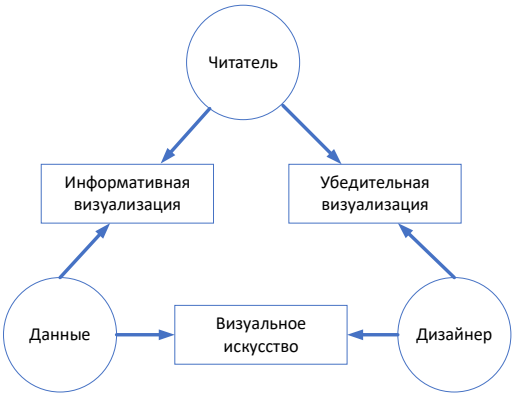


Рис. 18. Эффективная визуализация пояснительных данных

Информативная визуализация определяет отношения *между читателем и данными*. Она нацелена на нейтральное изложение фактов таким образом, чтобы обучить читателя. Информативные визуализации часто связаны с широкими наборами данных и стремятся преобразовать контент в удобную для использования форму.

В идеале они образуют основную часть визуализаций, с которыми обычный человек сталкивается ежедневно.

Убедительная визуализация определяет отношения *между дизайнером и читателем*. Это полезно, когда дизайнер хочет изменить мнение читателя о чем-то. *Убедительная визуализация* представляет собой очень конкретную точку зрения и призывает к изменению мнения или действия со стороны читателя. В этой категории визуализации представленные данные специально выбраны с целью поддержки точки зрения дизайнера и служат для убеждения читателя.

Третья категория, *визуальное искусство*, в первую очередь обслуживает отношения *между дизайнером и данными*. Визуальное искусство отличается от предыдущих двух категорий тем, что часто влечет за собой *однонаправленное* кодирование информации, а это означает, что читатель может не иметь возможности декодировать визуальное представление, чтобы понять основную информацию.

В то время как информативные и убедительные визуализации являются легко декодируемыми – *двунаправленными* в их кодировании, визуальное искусство просто переводит данные в визуальную форму.

Визуализация данных предназначена для представления данных таким образом, чтобы информация была легко усвоена и понятна с первого взгляда, поскольку данные могут быть представлены разными способами. Следует тщательно выбирать диаграмму, которая лучше всего подходит для визуализации.

Использование визуализации данных дает множество преимуществ:

– предоставляет возможность быстро усваивать информацию, улучшать понимание и принимать более быстрые и точные решения;
– позволяет отслеживать ключевые показатели эффективности компании;
– повышает понимание того, что показывают данные и какие будут следующие шаги;
– позволяет отслеживать и визуализировать статические и живые метрики;
– предоставляет возможность делиться информацией со всеми, поскольку легко распространяет информацию;

– устраняет необходимость для бизнеса полагаться на специалиста по обработке данных для понимания данных;
– позволяет компании добиться успеха быстрее и с меньшим количеством ошибок;
– определяет области бизнеса или бизнес-плана, которым нужно уделять больше внимания.

Визуальная информация лучше воспринимается и позволяет быстро и эффективно донести до зрителя собственные мысли и идеи. Физиологически восприятие визуальной информации имеет фундаментальное значение для людей. Особенно сегодня, когда мы имеем дело с огромными объемами данных и нам нужно как можно скорее извлечь самое важное.

Визуализация данных идеально подходит для разделения длинных блоков текста и эффективного выделения важной статистики. Можно использовать этот тип графики в следующих документах:
– бизнес-отчеты;
– информационные бюллетени;
– электронные письма.

При разработке визуализации данных нужно убедиться, что у нее есть следующие характеристики:
– легкость понимания с первого взгляда;
– краткость;
– аккуратность в представлении;
– сосредоточение на одной идее.

Использование визуализации данных – графического и числового изображения данных и агрегатов данных, таких как числа, таблицы и диаграммы, – является ключом к быстрой, эффективной и интуитивно понятной передаче информации. Другими словами, визуализация данных превращает обычные данные в информацию.

При этом сам процесс создания визуализации сопровождается решением некоторых проблем.

Проблема 1: использование неправильного формата визуализации. Очень легко заблудиться в типах графиков, диаграмм и карт, поэтому потребуется некоторое время, чтобы изучить необходимый минимум данных для бизнеса.

Проблема 2: использование неправильного типа данных. Очень похожая на первую проблема, но нужно понять, какой тип данных можно применить к вашему проверенному dataviz.

Проблема 3: инструменты dataviz не создают отчеты. Только несколько дорогих инструментов могут интерпретировать некоторую часть информации за человека.

Проблема 4: неправильный выбор инструмента.

Учитывая все возможности и проблемы, можно реализовать визуальную аналитику.

Тема 11. Визуализация данных в понимании и передаче аналитических данных

Визуализация данных очень часто является первым шагом в понимании и передаче аналитических данных, потому что люди гораздо лучше понимают данные, когда они представлены графически, а не численно. Когда данные визуализируются, легче увидеть возникающие тенденции.

Визуализация данных как довольно интуитивно понятный формат – это мощный способ сообщить о результатах, позволяющий упростить сотрудничество и ускорить инновации. С повсеместным распространением данных технология визуализации данных все шире используется и распространяется во многих дисциплинах.

Визуализация данных – это один из этапов процесса анализа данных, который гласит, что после того, как данные были собраны, обработаны и смоделированы, они должны быть визуализированы, чтобы можно было сделать выводы. Визуализация данных также является элементом более широкой дисциплины – архитектуры представления данных, которая направлена на выявление, определение местоположения, управление, форматирование и доставку данных наиболее эффективным способом.

Один из самых привычных способов использования визуализации данных – презентация информации в виде диаграмм или инфографики.

При анализе данных с помощью визуализации используют так называемое быстрое прототипирование, то есть создание большого

количества различных визуальных представлений одних и тех же данных. Делается это для нахождения скрытых, на первый взгляд, взаимосвязей и зависимостей, а также первичной оценки набора данных для возможности применения в дальнейшем более сложных инструментов анализа. Этот подход называется Exploratory data analysis (EDA), что в переводе означает «разведочный анализ данных». Основное отличие от презентации данных – визуализация здесь может быть черновой и некрасивой, но выполняется быстро и одним человеком или небольшой рабочей группой. Для этого чаще всего используют Excel, R или Matlab.

Визуализация активно используется в бизнесе. Принцип «говорите с данными» помогает компаниям зарабатывать больше, а клиентам – получать лучший сервис. Для разового анализа обычно используется Excel или R. Однако это неудобно, если необходимо следить за какими-то показателями (KPI) на постоянной основе. Для отслеживания рутинных KPI используют дашборды – дисплеи, на которых выведены все необходимые показатели в одном месте в виде графиков, диаграмм и таблиц.

Проектирование эффективных дашбордов – сложная и неординарная задача. Зачастую их перегружают ненужной информацией или стараются использовать все возможные типы шаблонных графиков. Часто для того, чтобы спроектировать хороший дашборд, необходимо создать новые типы визуализации информации.

Аудит корпоративных бизнес-процессов, мониторинг тенденций, отслеживание их регрессий и корреляций, поиск болевых точек, многоступенчатый анализ – группа сложнейших аналитических задач. Не упускать из виду массив данных и при этом иметь возможность быстро считывать информацию помогают специальные инструменты – программы для создания дашбордов, которые являются частью систем Business Intelligence (BI).

Эффективные дашборды для решения задач аналитики и мониторинга выглядят как обычная инфографика. В этом и кроется их преимущество: сложная задача – простой вид решения. Здесь могут быть каскадные, круговые и пузырьковые диаграммы, древовидные карты, гистограммы, графики и многое другое.

Хорошая визуализация данных важна для анализа данных и принятия решений на их основе. Это позволяет людям выявлять

закономерности и взаимосвязи, которые могут остаться незамеченными при использовании электронной таблицы с данными.

Хорошо продуманная графика может не только предоставить информацию, но и усилить воздействие этой информации с помощью яркого представления для привлечения внимания и поддержания интереса.

Представление данных для анализа в удобной и доступной форме становится необходимостью. Таким образом, технологии визуализации данных (BI) предоставляют следующие преимущества:

– быстрое и легкое восприятие данных;
– анализ большого набора данных со сложной структурой;
– эстетическая привлекательность;
– визуализация данных побуждает зрителя думать о сути, а не о методологии, помогает избежать искажения того, что должны сказать данные;
– возможность контроля и самоконтроля;
– отсутствие или уменьшение ошибок в визуализированной информации.

Превращение информации в изображение не является целью. На более высоком уровне данные легче понять, когда они представлены в виде наглядного изображения. Но на более низком уровне визуал – это инструмент для передачи связей между различными единицами.

Каждый тип визуала точно соответствует представлению о том, какие данные и какой тип связи он может интерпретировать: отношения, сравнение.

Визуализация данных – это первый шаг в понимании аналитических данных, потому что людям проще понять данные, представленные графически.

Являясь довольно интуитивно понятным форматом, визуализация данных – это мощный способ сообщить о результатах, упрощая совместную работу и ускоряя инновации.

Выделим три вида аналитики:

1. Визуальная аналитика. Управляемая искусственным интеллектом визуальная аналитика и информационные панели могут быть использованы для быстрого и удобного создания и представ-

ления ценной информации. Эти информационные панели позволяют всем сотрудникам организации использовать идеи для более эффективного принятия решений.

2. Встроенная аналитика. Способность предоставлять информацию в контексте, который бизнес-приложения, сотрудники и клиенты уже используют, имеет решающее значение. Для этого система должна иметь возможность беспрепятственно интегрировать интерактивные отчеты и визуализации в продукт или услугу.

3. Корпоративная отчетность. Система должна модернизировать корпоративную платформу отчетности и предоставлять экономически эффективные отчеты с точностью до пикселя. Создавая отчеты, которые связывают данные из нескольких источников, можно донести критически важные идеи до всех сотрудников организации.

Эффективная визуализация данных – это тонкий баланс между формой и функцией. Данные и визуальные эффекты должны работать вместе.

Чтобы понять, как происходит сбор, обработка и визуализация данных, обратимся к работе психологов А́моса Тве́рски и Дэ́ниела Кана, которые выделили два метода формирования мысли:

– **метод I:** быстрая и автоматическая обработка мыслей;

– **метод II:** медленные, логичные и нечастые вычисления.

Определив эти два метода, Кан смог объяснить, почему людям часто трудно мыслить статистически. Он считает, что мышление, лежащее в основе метода I, основано на эвристиках и предубеждениях.

В рамках мыслительного процесса применяется метод I, а затем включается метод II, который позволяет правильно анализировать данные.

Поскольку с помощью зрения мы обрабатываем больше информации, чем с помощью каких-либо других органов чувств, визуализация данных является идеальным способом передачи шаблонов и идей, которые можно получить с помощью больших объемов данных.

Для их визуализации используются тепловые карты, диаграммы Ганта, точечные диаграммы, прямоугольные диаграммы, пузырьковые облака, картограммы, временные шкалы, древовидные карты, облака слов и многое другое.

Выбор типа будет зависеть от того, какую тенденцию или паттерн необходимо продемонстрировать. Например, на линейной диаграмме отображаются восходящие и нисходящие тенденции, а на круговой диаграмме легко отображаются пропорции.

Независимо от того, какой тип визуализации данных будет выбран для демонстрации определенных закономерностей и тенденций, он должен состоять из трех основных элементов:

1) чистые, точные данные;
2) элементы дизайна, которые визуально демонстрируют данные;
3) возможность делиться с конкретной аудиторией.

Тема 12. Наиболее распространенные типы визуализации данных

Визуализация данных относится к методам, используемым для передачи данных или информации путем кодирования их в виде визуальных объектов: точек, линий или полос, содержащихся в графике.

Визуализация данных – это процесс отображения данных или информации в графических диаграммах, цифрах и столбцах.

Рассмотрим основные типы визуализации данных:

1. *Временной*. Визуализации данных относятся к временной категории, если они удовлетворяют двум условиям: они линейны и одномерны. Временные визуализации обычно представляют собой линии, которые либо стоят отдельно, либо перекрывают друг друга, с указанием времени начала и окончания.

Примеры визуализации временных данных:

– диаграммы разброса;
– диаграммы полярных областей;
– последовательности временных рядов;
– сроки;
– линейные графики.

2. *Иерархический*. Иерархические визуализации лучше всего подходят, если требуется отобразить кластеры информации.

Обратной стороной этих графиков является то, что они, как правило, более сложные и трудные для чтения, поэтому чаще всего используется древовидная диаграмма.

Примеры визуализации иерархических данных:
– древовидные диаграммы;
– кольцевые диаграммы;
– диаграммы солнечных лучей.

3. *Сетевой.* Наборы данных тесно связаны с другими наборами данных. Визуализации сетевых данных показывают, как они связаны друг с другом в сети. Другими словами, это демонстрация отношений между наборами данных без словесных объяснений.

Примеры визуализации сетевых данных:
– матричные диаграммы;
– диаграммы узловых соединений;
– облака слов;
– аллювиальные диаграммы.

4. *Многомерный*. Визуализации многомерных данных имеют несколько измерений. Это означает, что в миксе всегда есть две или более переменных для создания трехмерной визуализации данных. Из-за множества одновременных слоев и наборов данных эти типы визуализаций, как правило, являются наиболее яркими и привлекательными.

Примеры визуализации многомерных данных:
– диаграммы разброса;
– круговые диаграммы;
– диаграммы Венна;
– гистограммы.

5. *Геопространственный.* Визуализации геопространственных или пространственных данных предназначены для анализа реального физического местоположения путем наложения на знакомые карты различных точек данных. Этот тип визуализации данных обычно используется для отображения продаж или приобретений.

Примеры визуализации геопространственных данных:
– карта потока;
– карта плотности;
– картограмма;
– тепловая карта.

Для простого анализа и визуализации необходимо учесть следующие основные аспекты:

1. **Аудитория.** Важно адаптировать представление данных к конкретной целевой аудитории. Например, пользователи мобильных приложений для фитнеса, просматривающие свой прогресс, могут легко работать с несложными визуализациями. С другой стороны, если аналитика данных предназначена для исследователей и опытных лиц, принимающих решения, которые регулярно работают с данными, то часто приходится выходить за рамки простых диаграмм.

2. **Контент.** Тип данных, которые нужно визуализировать, будет определять тактику. Например, если это показатели временных рядов, во многих случаях следует использовать линейные диаграммы, чтобы показать динамику. Чтобы показать взаимосвязь между двумя элементами, часто используются диаграммы рассеяния. В свою очередь, гистограммы хорошо подходят для сравнительного анализа.

3. **Контекст.** Можно использовать разные подходы к визуализации данных и читать данные в зависимости от контекста. Чтобы подчеркнуть определенную цифру, например значительный рост прибыли, можно использовать оттенки одного цвета на графике и выделить самое высокое значение самым ярким. Напротив, чтобы различать элементы, можно использовать контрастные цвета.

4. **Динамика.** Существуют разные типы данных, и каждый тип имеет разную скорость изменения. Например, финансовые результаты можно измерять ежемесячно или ежегодно, в то время как временные ряды и данные отслеживания постоянно меняются. В зависимости от скорости изменения можно рассмотреть методы динамического представления или статической визуализации при интеллектуальном анализе данных.

5. **Цель.** Цель визуализации данных определяет способ ее реализации. Для проведения комплексного анализа визуализации объединяются в динамические и управляемые информационные панели, которые работают как методы и инструменты визуального анализа данных. Однако панели мониторинга не нужны для отображения единичных или случайных данных.

В зависимости от этих аспектов можно выбирать различные методы визуализации данных и настраивать их функции. Вот общие типы методов визуализации:

– **диаграммы** – самый простой способ показать развитие одного или нескольких наборов данных. Диаграммы варьируются

от гистограмм и линейных диаграмм, которые показывают взаимосвязь между элементами во времени, до круговых диаграмм, демонстрирующих компоненты или пропорции между элементами одного целого;

– **графики** – они позволяют распределить два или более наборов данных в двухмерном или даже трехмерном пространстве, чтобы показать взаимосвязь между этими наборами и параметрами на графике. Графики тоже разнятся. Точечные и пузырьковые диаграммы – одни из наиболее широко используемых визуализаций. Когда дело доходит до больших данных, аналитики часто используют более сложные ящичные диаграммы, которые помогают визуализировать взаимосвязь между большими объемами данных;

– **карты** – это популярный способ визуализации данных, используемый в различных отраслях. Они позволяют размещать элементы на соответствующих объектах и территориях – географических картах, планах зданий, макетах веб-сайтов и т. д. Среди наиболее популярных визуализаций карт – тепловые карты, карты распределения точек, картограммы;

– **диаграммы и матрицы.** Диаграммы обычно используются для демонстрации сложных взаимосвязей данных и включают различные типы данных в одну визуализацию. Они могут быть иерархическими, многомерными, древовидными. Матрица – это один из передовых методов визуализации данных, который помогает определить корреляцию между множеством постоянно обновляемых наборов данных.

Рассмотрим некоторые типы визуализации, общие для любого бизнеса.

1. Блок-схемы позволяют упорядочивать процесс шаг за шагом, от начала до конца, с целью его анализа, проектирования, документирования или управления. В то время как простая блок-схема применяется для документирования базовых процессов от A до B и C, диаграммы чаще используются для иллюстрации более сложных последовательностей с множеством решений или условий на этом пути. Каждый раз, когда условие выполняется, на диаграмме отображаются различные варианты, а затем путь продолжается после каждого выбора.

2. Контрольные диаграммы. Известная как диаграмма поведения процесса, контрольная диаграмма помогает определить, находится ли набор данных в пределах среднего или заранее определенного диапазона контроля. Типичная контрольная диаграмма, часто используемая в процессах контроля качества, состоит из точек, нанесенных на две оси, представляющих измерения образцов.

3. Графики акций помогают отслеживать рынки, чтобы определять прибыли и убытки, а также принимать решения о покупке и продаже. Несмотря на то что для представления рыночных изменений используются различные графики, наиболее распространенной является базовая гистограмма с повернутой линейной диаграммой.

4. Диаграммы Ганта – это особый тип гистограмм, который используется для построения диаграмм проектов и расписаний.

5. Каскадные диаграммы показывают, как на исходное значение положительно и отрицательно влияют различные факторы.

6. Диаграммы иерархии. Похожая по внешнему виду на блок-схему, иерархическая диаграмма, также известная как организационная диаграмма или органиграмма, иллюстрирует структуру организации, а также взаимосвязи внутри нее.

Визуализацию данных можно использовать как инструмент для понимания концептуальных процессов и процессов разработки идей. Важные преимущества визуализации данных для управления проектами:

1) улучшение общения;
2) укрепление сотрудничества;
3) улучшение восприятия.

Чтобы заставить команды сотрудничать и лучше общаться, менеджеры проектов могут использовать такие визуальные инструменты и методы, как видеостена проекта, стена совместной работы над проектом, социальные сети проекта, трехмерные среды проекта, геймификация проекта и т. д.

Рассмотрим наиболее распространенные типы визуализаций.

Гистограмма (рис. 19) – это один из основных способов сравнения единиц данных друг с другом. Из-за простой графической формы гистограмма часто используется в бизнес-аналитике в качестве интерактивного элемента страницы. Гистограммы достаточно

универсальны, чтобы их можно было изменять и отображать более сложные модели данных.

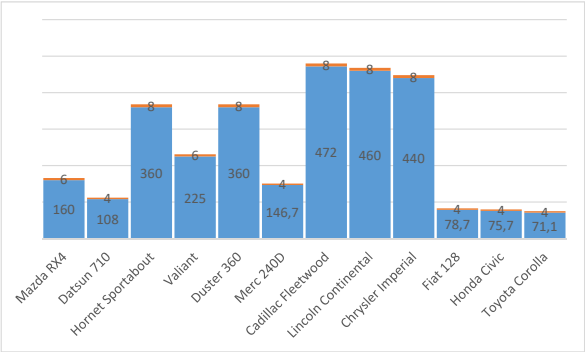


Рис. 19. Гистограмма

Когда использовать: сравнение объектов, числовая информация.

Круговая диаграмма (рис. 20). Этот тип диаграммы используется в любом отделе маркетинга или продаж, поскольку он позволяет легко продемонстрировать состав объектов или сравнение единиц.

Когда использовать: составление объекта, сравнение частей со всем объектом.

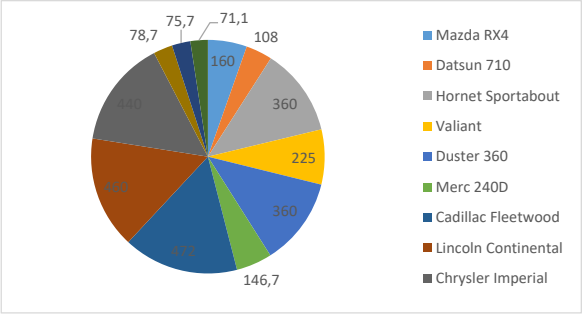


Рис. 20. Круговая диаграмма

Линейный график (рис. 21). Этот тип визуализации использует горизонтальную и вертикальную оси для отображения количественных показателей. Линейные графики также можно комбинировать с гистограммами для представления данных из нескольких измерений.

Когда использовать: отображение поведения объекта на временной шкале.

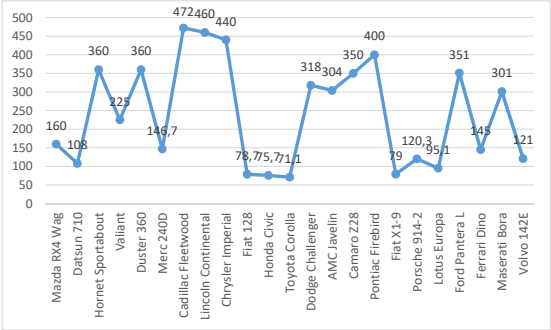


Рис. 21. Линейный график

Ящичковая диаграмма (рис. 22). Основные элементы здесь – это минимум, максимум и медиана, расположенные между первым и третьим квартилем, а также распределение объектов и их отклонение от медианы.

Когда использовать: распределение сложного объекта по отношению к медиане.

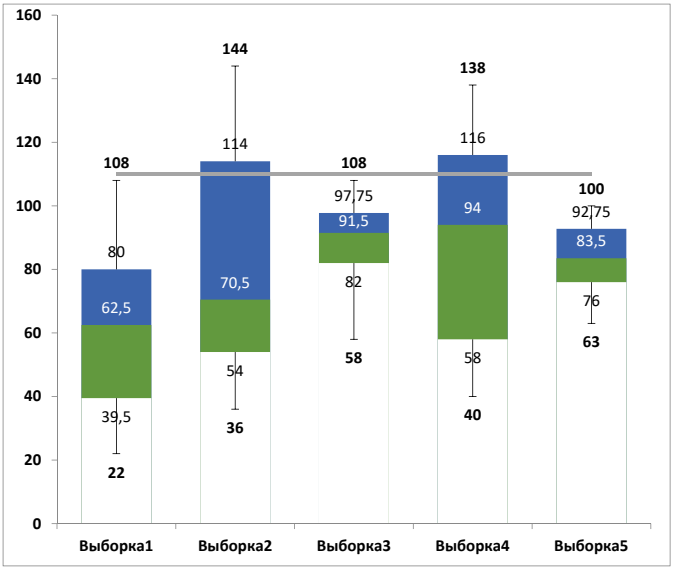


Рис. 22. Ящичковая диаграмма

Диаграмма разброса (рис. 23). Этот тип визуализации построен по осям X и Y. Между ними расположены точки, обозначающие объекты. Положение точки на графике обозначает, какими качествами обладает объект. Единственное ограничение этого типа визуализации – количество осей.

Когда использовать: отображение распределения объектов, определение качества каждого объекта на графике.

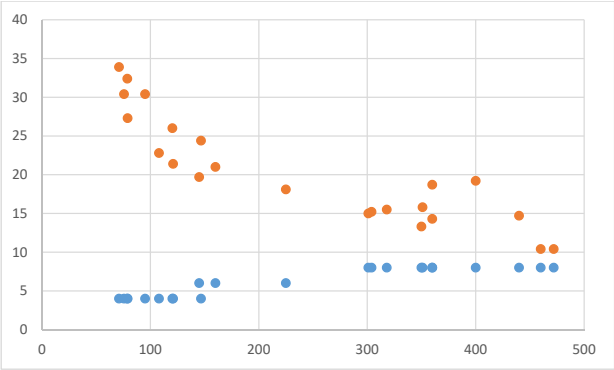


Рис. 23. Диаграмма разброса

Радар, или паук-карта (рис. 24). Этот тип диаграммы в основном представляет собой линейную диаграмму, нарисованную радиально. Он имеет форму паутины, которая создается несколькими осями и переменными.

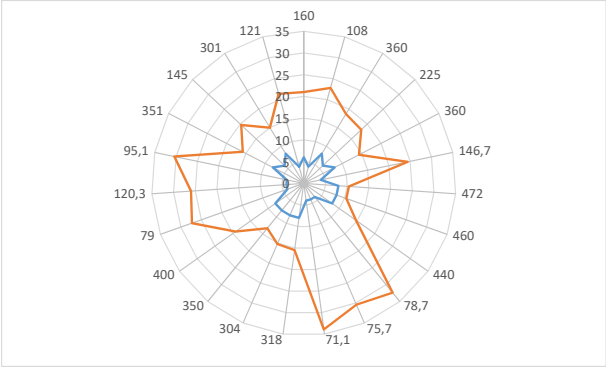


Рис. 24. График-радар

Его назначение такое же, как и у линейного графика. Но из-за количества осей можно сравнивать единицы под разными углами и отображать наклоны графически.

Когда использовать: описание качества данных, сравнение нескольких объектов друг с другом по разным измерениям.

Точечная карта, или карта плотности (рис. 25). Этот тип визуализации применяется для отображения географических данных. Карты плотности строятся с помощью точек, размещаемых на карте и показывающих местоположение каждого объекта. Точка может обозначать одну единицу или ряд объектов в определенной области. Такой формат позволяет легко отобразить плотность, но может дать нулевое значение, если требуются точные числа.

Когда использовать: изображение распределения или плотности объектов.

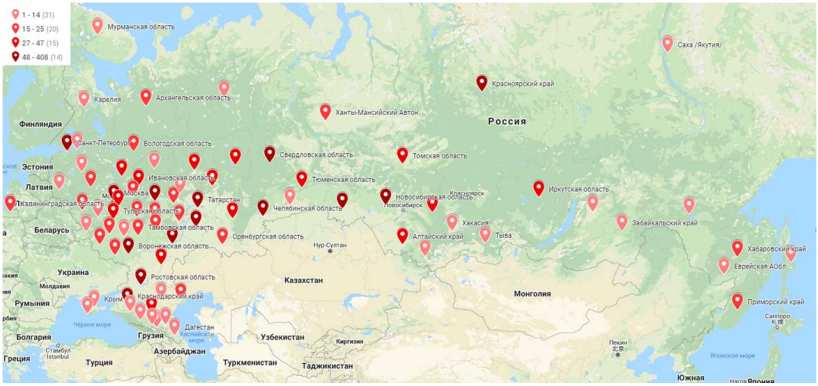


Рис. 25. Карта плотности

Диаграммы воронок (рис. 26). Они идеально подходят для демонстрации сужения корреляции между различными группами элементов. В большинстве случаев при построении воронок используют как геометрическую форму, так и цветовое кодирование для различения предметов. Этот тип диаграммы также удобен, когда процесс состоит из нескольких этапов.

Когда использовать: отображение этапов процесса с уменьшением процента стоимости / объектов.

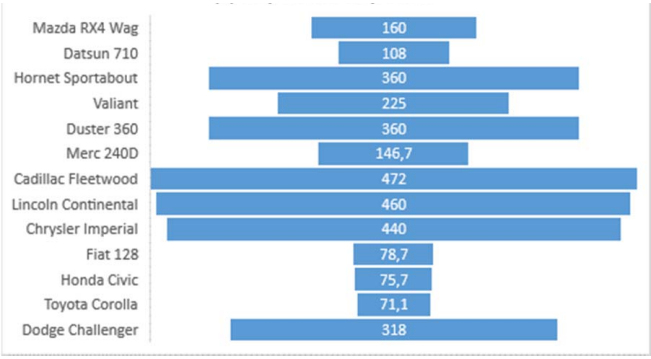


Рис. 26. Диаграмма воронок

При выборе типа визуализации нужно убедиться, что есть четкое понимание следующих моментов:

1) **специфика набора данных** – область знаний или подразделение в компании;

2) **аудитория** – люди, для которых предназначена информация;

3) **логика подключения** – сравнение объектов, распределение, взаимосвязь, описание процесса и т. д.;

4) **вывод** – цель представления информации.

Джин Желязны в книге «Говори на языке диаграмм» пишет, что почти каждая идея может быть выражена с помощью сравнения. Требуется лишь определить тип сравнения данных:

– покомпонентное: процент от целого;
– позиционное: соотношение объектов;
– временное: изменения во времени;
– частотное: число объектов в интервалах;
– корреляционное: зависимость между переменными.

При визуализации данных следует:

1) выбрать правильный график в зависимости от того, какая поставлена цель;
2) убедиться, что посыл графика подходит аудитории.
3) оформить график в правильном дизайне.

Тема 13. Инструменты визуализации данных

В начале развития визуализации наиболее распространенной техникой было использование электронной таблицы Microsoft Excel для преобразования информации в таблицу, гистограмму или круговую диаграмму.

Современные инструменты визуализации данных могут подключаться к источникам данных, например к реляционным базам данных. Эти данные, которые могут храниться локально или в облаке, извлекаются для анализа. Затем пользователи могут выбрать лучший способ представления данных из множества вариантов. Некоторые инструменты автоматически предлагают рекомендации по отображению в зависимости от типа представленных данных.

В большинстве случаев инструмент для визуализации данных представляет собой настольное приложение, представленное панелью управления командами. Интеграция с источниками данных осуществляется через API, поэтому наборы данных будут загружаться автоматически.

Каждый шаблон визуализации имеет свои настройки со свойствами данных и типами корреляции. Комбинируя разные типы визуальных элементов, можно построить отчет. В зависимости от функциональности отчеты могут быть в дальнейшем экспортированы в виде файлов CSV или предоставлены другим пользователям в системе.

Рассмотрим основные инструменты визуализации.

MS Excel – инструмент визуализации данных в системе бизнес-аналитики.

Looker – платформа для анализа данных, предлагающая подключение как к веб-источникам данных, так и к источникам данных SQL.

Возможности Looker:

– пользовательский интерфейс перетаскивания;
– настраиваемые информационные панели;
– экспортные отчеты;
– интеграция API со сторонними источниками;
– запрос данных из базы данных;
– кроссплатформенный доступ.

Tableau – этот инструмент предназначен не только для визуализации данных, но и для подготовки данных для преобразования, очистки и отображения информации без знания кода.

Особенности Tableau:

– пользовательский интерфейс перетаскивания;
– широкий список встроенных интеграций с источниками данных;
– общий доступ к аналитике;
– встраивание отчета;
– кроссплатформенный доступ;
– постоянно обновляемый поток данных;
– инструмент управления метаданными;
– встроенная панель инструментов для комментирования и выделения.

Tableau – это инструмент визуализации данных, который помогает упростить данные и преобразовать их в легко понятные форматы. Он позволяет создавать информационные панели и рабочие таблицы, а также быстро анализировать данные.

С Tableau можно:

– создавать графики и информационные панели, показывающие KPI;
– делиться отчетами и панелями мониторинга с членами команды, у которых есть доступ к инструменту;
– легко исследовать данные благодаря интерактивному пользовательскому интерфейсу.

Tableau Public – редактор, предоставляющий расширенную функциональность по визуализации данных.

QlikSense – инструмент визуализации, который позволяет нетехническим пользователям просматривать данные и задавать вопросы.

QlikSense предоставляет возможность:

– изучать визуальные представления информации с помощью простых взаимодействий;
– задавать любой вопрос по визуализации данных;
– создавать интеллектуальные, инновационные, полностью интерактивные и отзывчивые визуализации;

– осуществлять интеллектуальный поиск;
– сотрудничать и обмениваться информацией.

QlikView – поставщик бизнес-аналитики, который предлагает большую гибкость командам, желающим создавать индивидуальное программное обеспечение.

Возможности QlikView:
– настраиваемые отчеты;
– система доступа на основе ролей и разрешений;
– открытый API;
– открытый доступ к базе данных;
– кроссплатформенный доступ;
– общий доступ к аналитике.

Power BI – инструмент бизнес-аналитики, который дает возможность объединять данные из разных источников и создавать модель, позволяющую связать их для последующей визуализации.

Использование визуальных объектов в Power BI – один из самых простых способов предоставления данных в виде отчетов. Он имеет несколько стандартных визуальных элементов, которые включаются в отчеты без необходимости программирования: графика, таблицы, интерактивные карты и другое. А также дает возможность создавать собственные пользовательские визуальные объекты и упаковывать их в файл, чтобы их можно было импортировать в любой отчет.

Возможности Power BI:
– интерфейс перетаскивания;
– настольное приложение;
– широкий список встроенных интеграций с источниками данных;
– настраиваемые отчеты;
– дополнительные обновления данных;
– полная экосистема бизнес-аналитики.

Data Studio Google – это самый доступный вариант для пользователя, который хочет создавать визуализированные отчеты. Возможности Data Studio:
– веб-приложение;
– интерфейс перетаскивания;
– встроенная интеграция с Google Analytics и другими продуктами Google;

– настраиваемые отчеты;
– инструменты преобразования данных;
– общий доступ к аналитике.

Многие приложения, включая Excel, QlikView и Tableau, имеют прямые соединители для доступа к базе данных, которые позволяют создавать отчеты на основе содержащихся в них данных.

Проблема в том, что такой подход предлагает ограниченную видимость того, что на самом деле происходит в бизнесе. Чтобы получить необходимые отчеты, нужно будет сопоставить отчеты из нескольких систем, выявить расхождения между ними и найти способы определить механизм представления подтвержденных данных.

Более серьезная проблема, с которой сталкивается большинство крупных компаний, связана с очисткой и гармонизацией данных. Компании, которые расширились географически или за счет приобретений, обычно имеют несколько транзакционных систем, которые решают схожие задачи. Иногда транзакции начинаются в одной системе, а заканчиваются в другой, даже если они используют одного и того же поставщика программного обеспечения. И модули не могут быть взаимосвязаны таким образом, чтобы гарантировать, что данные, которые используются для вычислений, были стандартизированы и очищены.

Третья проблема носит более технический характер, связанный со скоростью обработки данных и быстрым устареванием информации в отчетах. Транзакционные системы отлично справляются с построчным вводом данных, но они не очень эффективны при создании агрегатов или аналитических операций, подобных тем, которые в настоящее время запрашиваются руководителем. Например, отчет, диаграмма или таблица, в которых суммируются все заказы за последние два месяца, требуют вычисления итогов для этих строк данных одновременно с добавлением и удалением новых строк.

Хранилища данных решают эти проблемы, отделяя транзакционные базы данных от аналитических и обеспечивая более быструю аналитику и более чистые, более функциональные данные, которые могут охватывать несколько транзакционных приложений.

Существуют также отдельные инструменты, которые можно использовать для создания визуальных эффектов определенного типа. Большинство из них требует знания языка программирования, а иногда и фреймворков.

D3.js – это бесплатная библиотека JavaScript для создания визуальных элементов путем соединения данных с объектной моделью документа через API, манипулирования документами как объектами.

Dygraphs – это бесплатная библиотека JavaScript с открытым исходным кодом. Она подходит для работы с огромными наборами данных для построения интерактивных диаграмм и графиков.

Chartist.js – еще один инструмент, основанный на JavaScript, который используется для построения графиков и диаграмм путем их стилизации с помощью CSS.

Gleam – это пакет Python для построения визуализации точечной диаграммы с помощью CSS и HTML.

Leather – это библиотека Python для создания диаграмм в их простейшей форме и сохранения их в виде файлов SVG для дальнейшего использования.

Matplotlib – еще одна библиотека Python с открытым исходным кодом, предназначенная для создания 2D-визуализаций.

Все упомянутые библиотеки и инструменты можно свободно использовать и интегрировать с существующим программным обеспечением для создания новых типов визуальных элементов данных в качестве шаблонов.

Контрольные вопросы

1. Что собой представляет визуализация данных? Для чего она предназначена?
2. Что требуется для создания хорошей визуализации данных? Какие виды визуализации применяются на практике?
3. Какие виды визуализации вы знаете? Дайте характеристику каждому известному виду. Что включают способы визуализации данных?
4. Какие проблемы возникают при разработке более сложных визуализаций?

5. Выделите типы визуальной коммуникации. Чем они схожи и в чем их различие?
6. Назовите основные положения, определяющие важность визуализации данных.
7. Перечислите основные условия для визуализации данных и визуализации информации.
8. Назовите основные категории визуализации данных. Определите сходства и различия между ними.
9. Какие преимущества дает использование визуализации данных? Приведите примеры использования визуализации данных, выделив преимущества.
10. Назовите основные типы визуализации данных. Приведите примеры.
11. Назовите основные методы визуализации. Дайте им характеристику.
12. Как определить тип визуализации для представления данных?
13. Перечислите основные инструменты визуализации. Дайте их описание, приведите примеры использования.
14. Какие проблемы связаны с выбором инструментов визуализации данных? Каковы способы их решения?
15. Перечислите основные факторы стандартизации инструментов визуализации.

Тесты для самоконтроля

1. Как называется метод представления данных в наглядном или графическом формате? (*один вариант ответа*)

1) инфографика
2) визуализация данных
3) визуализация информации
4) презентация

2. Какой метод предназначен для более глубокого изучения предоставленной информации, особенно в интерактивном режиме? (*один вариант ответа*)

1) инфографика
2) визуализация данных

3) визуализация информации
4) презентация

3. С чем связана визуализация данных? (*несколько вариантов ответа*)
1) с инфографикой
2) с визуализацией информации
3) с исследовательским анализом данных
4) с презентацией

4. Какие виды визуализации вы знаете? (*несколько вариантов ответа*)
1) обычное визуальное представление
2) концептуальная визуализация
3) стратегическая визуализация
4) комбинированная визуализация
5) интерактивная визуализация

5. Чем характеризуются данные визуализации? (*несколько вариантов ответа*)
1) содержание
2) логика
3) надежность
4) социальность

6. Назовите основные положения, определяющие важность визуализации данных. (*несколько вариантов ответа*)
1) лучшее принятие решений
2) осмысленное повествование
3) грамотность данных
4) наглядность представления

7. Какая категория визуализации данных применяется, когда нужно понять, что находится внутри набора данных? (*один вариант ответа*)
1) исследование
2) объяснение
3) разъяснение
4) развитие

8. Какая категория визуализации переводит данные в визуальную форму? (*один вариант ответа*)

1) информативная визуализация
2) убедительная визуализация
3) визуальное искусство
4) искусственная визуализация

9. Выберите признаки, характеризующие визуализацию данных. (*несколько вариантов ответа*)

1) предназначена для представления данных таким образом, чтобы информация была легко усвоена и понятна с первого взгляда
2) превращает большие и маленькие наборы данных в графическую форму
3) включает в себя некоторые специфические особенности
4) предоставляет закодированные данные, которые нужно программно обрабатывать

10. Каковы преимущества визуального представления данных? (*несколько вариантов ответа*)

1) данные можно понять быстрее и проще
2) заинтересованные стороны, не имеющие технического образования, могут извлекать из данных информацию
3) визуализация данных может выявить тенденции, ранжирование, отношения и выбросы
4) красивый дизайн дает эстетическое восприятие

11. Для чего нужны инструменты визуализации? (*несколько вариантов ответа*)

1) помогают читать информацию быстро
2) помогают быстро обработать информацию
3) помогают принимать обоснованные решения, подкрепленные данными
4) помогают отображать аналитические отчеты

12. Какие функции выполняет BI? (*несколько вариантов ответа*)

1) извлечение необработанных данных из источника
2) преобразование необработанных данных
3) загрузка необработанных данных в единую систему хранения
4) представление данных для восприятия

13. Какие задачи решает визуализация данных? (*несколько вариантов ответа*)

1) обеспечивает быструю и эффективную передачу информации
2) помогает предприятиям определить, какие факторы влияют на поведение клиентов
3) выявляет области, которые нужно улучшить или которые требуют большего внимания
4) наглядно представляет статистику

14. Какие преимущества предоставляют технологии визуализации данных BI? (*несколько вариантов ответа*)

1) быстрое и легкое восприятие данных
2) анализ большого набора данных со сложной структурой
3) отсутствие или уменьшение ошибок в визуализированной информации
4) глобальный анализ данных, который представляется на обозрение пользователям

15. Какие основные элементы включает визуализация данных? (*несколько вариантов ответа*)

1) чистые, точные и обработанные данные
2) элементы дизайна, которые визуально демонстрируют данные
3) возможность делиться с конкретной аудиторией
4) аудитория

16. Как называется средство наглядного графического представления количественных данных, помогающее их анализировать? (*один вариант ответа*)

1) чертеж
2) таблица
3) диаграмма
4) схема

17. Какие виды диаграмм позволяют отслеживать динамику изменения данных? (*один вариант ответа*)

1) гистограммы
2) круговые диаграммы
3) графики
4) столбиковые диаграммы

18. Какой тип визуализации позволяет распределить два или более наборов данных в двухмерном или даже трехмерном пространстве, чтобы показать взаимосвязь между этими наборами и параметрами на графике? (*один вариант ответа*)

1) диаграммы
2) графики
3) карты
4) матрицы

19. Какой метод визуализации помогает определить корреляцию между множеством постоянно обновляемых наборов данных? (*один вариант ответа*)

1) диаграммы
2) графики
3) карты
4) матрицы

20. Какой тип визуализации помогает определить, находится ли набор данных в пределах среднего или заранее определенного диапазона контроля? (*один вариант ответа*)

1) блок-схемы
2) контрольные диаграммы
3) графики акций
4) диаграммы Ганта
5) диаграммы иерархии

21. Какой тип визуализации помогает отслеживать рынки, чтобы определять прибыли и убытки, а также принимать решения о покупке и продаже? (*один вариант ответа*)

1) блок-схемы
2) контрольные диаграммы
3) графики акций
4) диаграммы Ганта
5) диаграммы иерархии

22. Какой тип визуализации использует горизонтальную и вертикальную оси для отображения значения единицы с течением времени? (*один вариант ответа*)

1) гистограмма
2) круговая диаграмма
3) линейный график
4) ящичковая диаграмма

23. Какой тип визуализации можно комбинировать с гистограммами для представления данных из нескольких измерений? (*один вариант ответа*)

1) гистограмма
2) круговая диаграмма
3) линейный график
4) ящичковая диаграмма

24. Какой тип визуализации используется для отображения распределения объектов, определения качества каждого объекта на графике? (*один вариант ответа*)

1) диаграмма разброса
2) радар, или паук-карта
3) точечная карта, или карта плотности
4) диаграммы воронок

25. Что определяет выбор типа визуализации? (*несколько вариантов ответа*)

1) специфика набора данных
2) аудитория
3) логика подключения
4) данные

26. Каковы основные правила визуализации данных? (*несколько вариантов ответа*)

1) нужно выбирать график в зависимости от того, какая у вас цель
2) необходимо убедиться, что выбранный тип графика подходит аудитории

3) следует оформить график в правильном дизайне
4) нужно проанализировать исходные данные

27. Какой метод визуализации показывает иерархические данные во вложенном формате? (*один вариант ответа*)
1) графики
2) диаграммы с областями
3) графики разброса
4) карты деревьев

28. Каковы ограничения инструментов визуализации данных, необходимых для анализа данных? (*несколько вариантов ответа*)
1) отсутствие объяснений
2) получение разных идей от разных пользователей
3) отсутствие руководства
4) безопасность данных

29. Какой из инструментов визуализации данных является еще и инструментом подготовки данных для преобразования, очистки и отображения информации без знания кода? (*один вариант ответа*)
1) MS Excel
2) Tableau
3) QlikSense
4) Looker

30. Назовите факторы стандартизации инструментов визуализации. (*несколько вариантов ответа*)
1) легкость использования
2) поддержка визуализаций
3) мобильная поддержка
4) интерактивность обработки

Практическое задание

Визуализируйте данные, представив разные типы диаграмм с помощью Matlab, Excel, Python или R.

Методические указания

1. Подготовьте данные, на основе которых будет проводиться исследование с последующей визуализацией полученных результатов анализа. Представьте данные в виде таблицы.

2. Постройте линейный график, показывающий общую прибыль за все месяцы.

Данные об общей прибыли должны быть предоставлены за каждый месяц. Созданный линейный график должен включать следующие свойства:

– X label name = номер месяца;
– Y label name = прибыль в рублях.

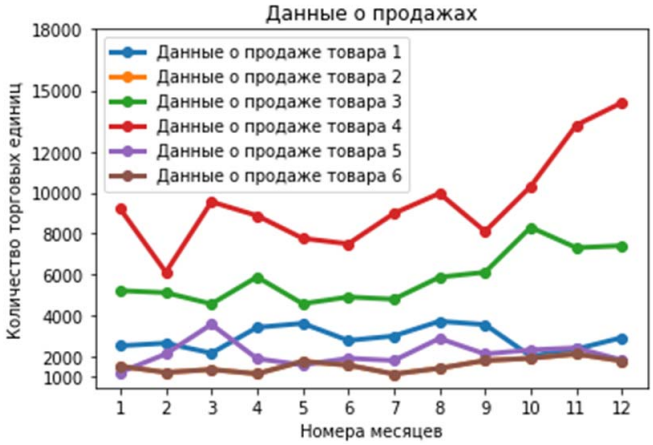
3. Измените линейный график, включив следующие свойства стиля:

– стиль линии – пунктирный, цвет линии – красный;
– легенда – в правом нижнем углу;
– X label name = номера месяцев;
– Y label name = прибыль в рублях;
– ширина линии – 3 пт.

График должен выглядеть следующим образом.

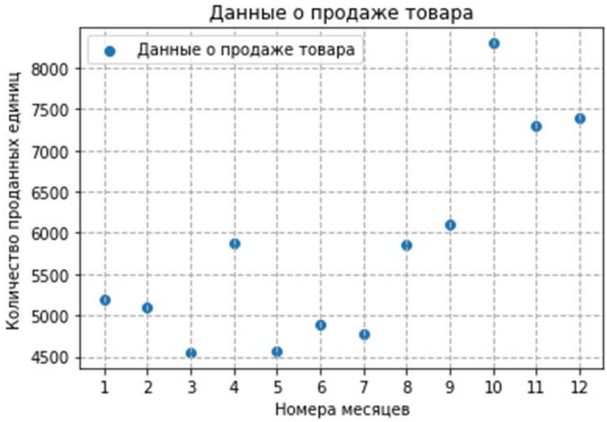
4. Представьте многострочные графики, на которых отображаются все данные о продажах продуктов. Отобразите количество единиц, проданных в месяц, для каждого продукта с помощью многострочных графиков (то есть отдельный график для каждого продукта).

График должен выглядеть следующим образом.



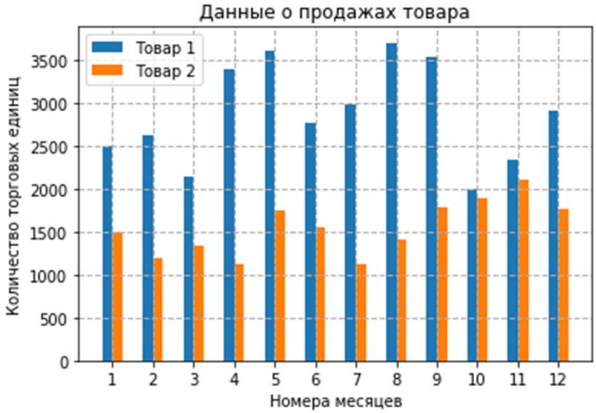
5. Представьте диаграмму рассеяния, в которой отобразите данные о продаже одного вида товаров за каждый месяц. Также добавьте сетку на сюжет. Стиль линии сетки должен быть «–».

График должен выглядеть следующим образом.



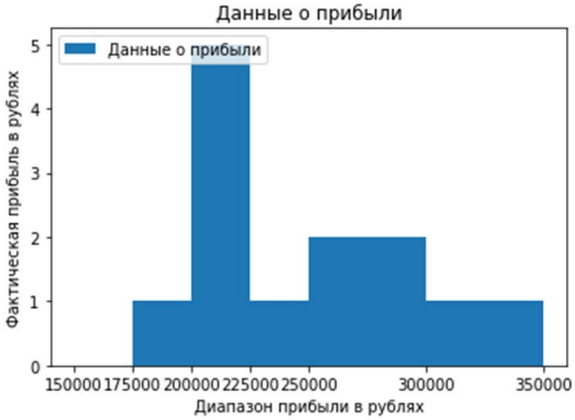
6. Создайте гистограмму по анализу данных для двух видов разных продуктов. Гистограмма должна отображать количество проданных единиц в месяц для каждого продукта. Добавьте отдельную полосу для каждого продукта в той же таблице.

График должен выглядеть следующим образом.



7. Разработайте гистограмму, показывающую общую прибыль каждого месяца, чтобы увидеть наиболее распространенные диапазоны прибыли.

График должен выглядеть следующим образом.

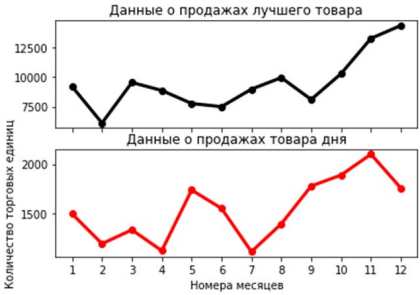


8. Покажите на круговой диаграмме общие данные о продажах за прошлый год для каждого продукта. Отобразите количество проданных единиц каждого продукта в год в процентах.

График должен выглядеть следующим образом.

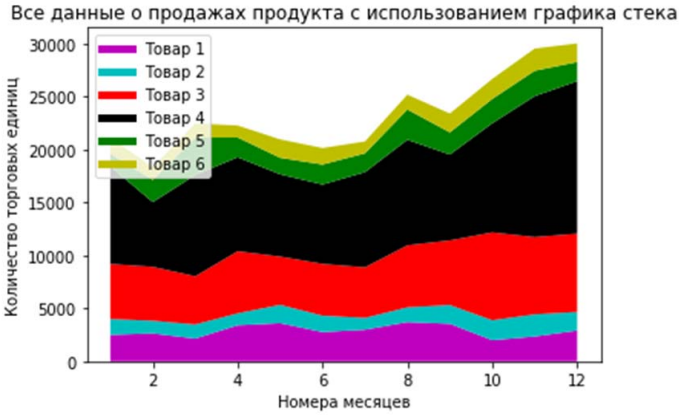
9. Добавьте вспомогательный график данных о продажах дня за все месяцы.

График должен выглядеть следующим образом.



10. Разработайте стек графиков для показа всех данных о продажах продуктов.

График должен выглядеть следующим образом.



11. Представьте отчет, который должен содержать текст программного кода выполненных заданий и скрин результатов работы.

Глава 4. ПРОЕКТ АНАЛИЗА И ВИЗУАЛИЗАЦИИ ДАННЫХ

Тема 14. Этапы анализа данных для их визуализации

Анализ данных можно описать как процесс, состоящий из нескольких шагов, по ходу которого сырые данные превращаются и обрабатываются с целью создать визуализации и сделать предсказания на основе математической модели.

Анализ данных – это всего лишь последовательность шагов, каждый из которых играет ключевую роль для последующих. Этот процесс похож на цепь последовательных, связанных между собой этапов:

– определение проблемы;
– извлечение данных;
– подготовка данных – очистка данных;
– подготовка данных – преобразование данных;
– исследование и визуализация данных;
– предсказательная модель;
– проверка модели, тестирование;
– развертывание – визуализация и интерпретация результатов;
– развертывание – развертывание решения.

Процесс анализа данных начинается задолго до сбора сырых данных. **В первую очередь определяется проблема**, которую необходимо решить. Определить ее можно, только сосредоточившись на изучаемой системе: механизме, приложении или процессе в целом. Исследование может быть предназначено для лучшего понимания функционирования системы, но его лучше спроектировать так, чтобы понять принципы поведения и впоследствии делать предсказания или осознанный выбор.

На самом деле всеобъемлющее и исчерпывающее исследование системы – это сложный процесс, и почти всегда нет достаточного количества информации, с которой можно начать. Поэтому определение проблемы и особенно планирование приводят к появлению руководящих принципов, которым необходимо следовать в течение всего проекта.

Когда проблема определена и задокументирована, можно двигаться к этапу планирования проекта по анализу данных. Планирование необходимо для понимания того, какие профессионалы и ресурсы понадобятся для максимально эффективного выполнения требований проекта. Таким образом, на данном этапе необходимо рассмотреть те вопросы, которые касаются решения этой проблемы. Необходимо найти специалистов с разными интересами и установить ПО, нужное для анализа данных.

Построение эффективной команды – одно из ключевых условий успешного анализа данных. Такие команды должны быть междисциплинарными, чтобы у них была возможность решать проблемы, рассматривая данные с разных точек зрения.

Первый шаг для проведения анализа – получение данных. Они должны быть выбраны с одной базовой целью – построение предсказательной модели. Поэтому выбор данных – также важный момент для успешного анализа.

Данные должны максимально отражать реальный мир. Использование больших наборов сырых данных, которые были собраны неграмотно, может привести либо к неудаче, либо к неопределенности. Поэтому недостаточное внимание, уделенное выбору данных, приведет к тому, что модели не будут соответствовать изучаемым системам.

Поиск и извлечение данных часто требуют интуиции, границы которой лежат за пределами технических исследований и извлечения данных. Этот процесс требует понимания природы и формы данных, предоставить которое может только опыт и знания практической области проблемы.

Следующий этап анализа – подготовка данных. Этот этап кажется наименее проблемным, но на самом деле требует наибольшего количества ресурсов и времени для завершения. Данные часто собираются из разных источников, каждый из которых может предлагать их в собственном виде или формате. Их нужно подготовить для процесса анализа.

Подготовка данных включает следующие процессы:

– получение,

– очистка,

– нормализация,
– превращение в оптимизированный набор данных.

Данные процессы можно реализовать в табличной форме.

Проблемы могут возникнуть при появлении недействительных, двусмысленных или недостающих значений, повторении полей или данных, не соответствующих допустимому интервалу.

Далее следует этап исследования данных – их анализ в графической или статистической репрезентации с целью поиска моделей или взаимосвязей. Визуализация – лучший инструмент для выделения подобных моделей.

За последние годы визуализация данных развилась так сильно, что стала независимой дисциплиной. Многочисленные технологии используются исключительно для отображения данных, а многие типы отображения работают так, чтобы получать только лучшую информацию из набора данных.

Исследование включает предварительное изучение, которое необходимо для понимания типа и значения собранной информации. Вместе с информацией, собранной при определении проблемы, такая категоризация определяет, какой метод анализа данных лучше всего подойдет для определения модели.

Исследование данных состоит из следующих шагов:
– обобщение данных;
– группировка данных;
– исследование отношений между разными атрибутами;
– определение моделей и тенденций;
– построение моделей регрессионного анализа;
– построение моделей классификации.

Как правило, анализ данных требует обобщения заявлений относительно изучаемых данных.

Обобщение – процесс, при котором количество данных для интерпретации уменьшается без потери важной информации.

Кластерный анализ – метод анализа данных, используемый для поиска групп, объединенных общими атрибутами (также называется группировкой).

Еще один важный этап анализа – идентификация отношений, тенденций и аномалий в данных. Для поиска такой информации

часто нужно использовать инструменты и проводить дополнительные этапы анализа, но уже на визуализациях.

Другие методы поиска данных, такие как деревья решений и ассоциативные правила, автоматически извлекают важные факты или правила из данных. Эти подходы используются параллельно с визуализацией для поиска взаимоотношений данных.

Предсказательная аналитика – это процесс в анализе данных, который нужен для создания или поиска подходящей статистической модели для предсказания вероятности результата.

После изучения данных следует этап создания математической модели, которая кодирует отношения между данными. Эти модели полезны для понимания изучаемой системы и используются в двух направлениях.

Первое – предсказания о значениях данных, которые создает система. В этом случае речь идет о регрессионных моделях.

Второе – классификация новых продуктов. Это уже модели классификации, или модели кластерного анализа. Их можно разделить на три группы в соответствии с результатами, к которым они приводят:

– модели классификации – если полученный результат – качественная переменная;
– регрессионные модели – если полученный результат числовой;
– кластерные модели – если полученный результат описательный.

Простые методы генерации этих моделей включают следующие техники:

– линейная регрессия;
– логистическая регрессия;
– классификация;
– дерево решений;
– метод *k*-ближайших соседей.

Но таких методов много, и у каждого есть свои характеристики, которые делают их подходящими для определенных типов данных и анализа. Каждый из них приводит к появлению определенной модели, а их выбор соответствует природе модели продукта.

Некоторые из методов будут предоставлять значения, относящиеся к реальной системе и структурам данных. Они смогут объяс-

нить некоторые характеристики изучаемой системы простым способом. Другие будут делать хорошие предсказания, но их структура будет оставаться «черным ящиком» с ограниченной способностью объяснить характеристики системы.

Проверка (валидация) модели, то есть фаза тестирования, – это следующий важный этап. Он позволяет проверить модель, построенную на основе начальных данных, узнать достоверность данных, созданных моделью, сравнив их с реальной системой.

Этот процесс позволяет не только в числовом виде оценивать эффективность модели, но также сравнивать ее с другими. Есть несколько подобных техник, самая известная – перекрестная проверка (кросс-валидация). Она основана на разделении учебного набора на разные части. Каждая из них, в свою очередь, будет использоваться в качестве валидационного набора. Все остальные – как тренировочные.

Развертывание (деплой) – это финальный шаг процесса анализа, задача которого – предоставить результаты, то есть выводы анализа. В процессе развертывания бизнес-среды анализ является выгодой, которую получит клиент, заказавший его.

Развертывание – это процесс использования на практике результатов анализа данных.

Есть несколько способов развертывания результатов анализа данных, или майнинга данных. Обычно развертывание состоит из написания отчета для руководства или клиента. Этот документ концептуально описывает полученные результаты. Он должен быть направлен руководству, которое будет принимать решения и затем использовать выводы на практике.

В документации от аналитика должны быть подробно рассмотрены следующие темы:

– результаты анализа;
– развертывание решения;
– анализ рисков;
– измерения влияния на бизнес.

Когда результаты проекта включают генерацию предсказательных моделей, они могут быть использованы в качестве отдельных приложений или встроены в ПО.

Прежде чем приступить к визуализации, рассмотрим все этапы анализа данных.

Формулирование цели. Каждое исследование должно отвечать на ряд поставленных вопросов – не нужно плодить исследования для исследований.

Сбор данных. На этом этапе аналитик или работает с уже собранными данными, или участвует в процессе постановки задания на сбор данных (фактически решает, какая информация ему необходима и в каком виде).

В первом случае особое внимание стоит уделить правильной интерпретации данных, которые записаны в базу, и зачастую смириться с существующим форматом данных, дизайном таблиц и т. д. Во втором случае аналитик сталкивается с проблемой построения грамотного сценария сбора данных – он может особенно перестараться в планировании A/B-тестов, логировании событий и т. п. Здесь важна коммуникация с программистами, которые могут помочь в понимании процессов и оценке масштабов планируемых записей.

Подготовка данных. «Мусор на входе – мусор на выходе» – правило, о котором всегда нужно помнить. Структурирование, устранение ошибок, изменение форматов содержимого, разбор аномальных результатов, очистка от выбросов, устранение дубликатов, интеграция данных из разных источников – одни из важнейших пунктов в анализе данных.

Иногда требуется расширение метрик, например добавление вычислительной информации (прирост, ранг, номер и т. п.). Иногда следует сократить количество признаков (переменных) или перейти к вспомогательным переменным, принимающим одно из двух значений: true (1) / false (0).

На этом этапе сырые данные превращаются в полезную входную информацию для моделирования и анализа.

Исследование данных. Для правильной интерпретации многомерных данных необходимо посмотреть на них в разрезе как конкретного признака, так и группы признаков. Также следует представить ключевые показатели в динамике с планами и фактическими результатами. Именно на этом этапе подбирается формат будущей визуализации.

Визуализация и построение выводов. Каждое исследование должно заканчиваться результатами и выводами. Даже если они негативные, их стоит проговорить и обсудить. При этом правильная постановка задачи, методика проведения сбора данных, правильная интерпретация результатов, выявленные ошибки и многое другое должны послужить базой для последующих исследований.

Время, затрачиваемое аналитиком на каждую фазу, зависит от многих переменных: начиная от опыта работы и уровня знания данных, заканчивая перечнем используемых инструментов и технических характеристик ПК.

Необходимо понимать, что процесс анализа данных имеет итерационный характер и может быть представлен циклом (рис. 27).

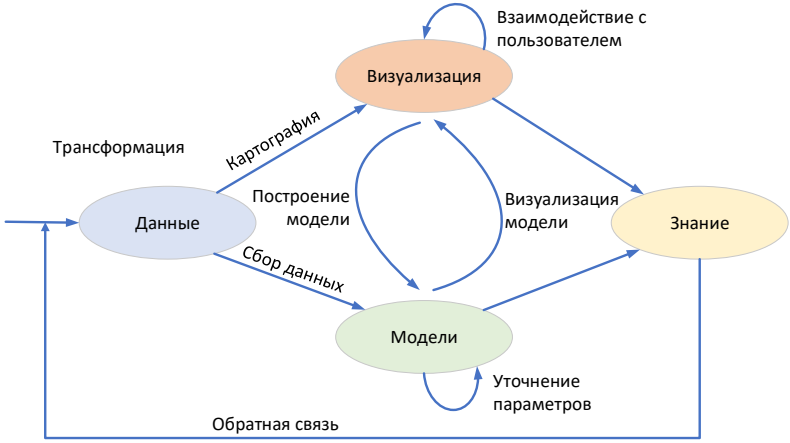


Рис. 27. Циклический процесс анализа данных

Чтобы сформулировать окончательные выводы, иногда необходимо пройтись по циклу несколько раз. Каждый раз уточняя данные, перестраивая модели обработки и представления данных, получая всё новые знания об анализируемой сущности. Чтобы уменьшить количество итераций этого цикла и время, затрачиваемое на анализ, аналитик должен быть не только специалистом в области Big Data, но и хорошо знать свою предметную область. При этом только опыт помогает понять, какие данные и методы анализа нужны в каждом отдельном случае и как интерпретировать результаты.

В каждом проекте визуализации данных необходимо учитывать множество факторов, чтобы минимизировать риски и обеспечить успешный результат. Чтобы обеспечить основу для организации работы и получить четкое представление о данных, полезно рассматривать ее как цикл с определенной последовательностью этапов.

Шаг 1: понимание бизнес-вопросов.

В начале проекта основное внимание уделяется получению четкого понимания общего объема работы, бизнес-целей, информации, которую ищут заинтересованные стороны, типа анализа, который они хотят, чтобы был использован, и основных результатов. Определение этих элементов до начала анализа очень важно, так как другой возможности задать вопросы до завершения проекта может не быть.

Шаг 2: понимание набора данных.

Этот этап начинается с первоначального сбора данных и продолжается такими действиями, как проверка качества данных, исследование данных для обнаружения первых идей в данных или обнаружение интересных подмножеств для формирования гипотез скрытой информации. Есть множество инструментов, которые можно использовать для анализа данных. В зависимости от размера набора данных можно использовать Excel для управляемых наборов данных или более жесткие инструменты, такие как R, Python, Alteryx, Tableau, для изучения и подготовки данных для дальнейшего анализа.

Важный момент, о котором следует помнить, – это определение ключевых переменных, представляющих интерес для изучения данных, поиска ошибок: пропущенных данных, данных, которые не имеют логического смысла, повторяющихся строк или даже орфографических ошибок, или любых отсутствующих переменных, которые необходимо исправить, чтобы можно было правильно очистить данные.

Здесь важно отметить, что при работе в корпоративной или бизнес-среде полезно привлекать специалистов с глубокими знаниями исходной системы, например администратора базы данных, который может помочь с пониманием и извлечением данных.

Шаг 3: подготовка данных.

После того как данные будут организованы и все ключевые переменные определены, можно приступить к очистке набора данных. Здесь нужно будет обрабатывать пропущенные значения: заменять наиболее логичными значениями, создавать новые переменные, чтобы помочь классифицировать данные, и удалять дубликаты. Задачи подготовки данных, вероятно, будут выполняться несколько раз. После этого шага окончательный набор данных готов для передачи в инструмент моделирования для дальнейшего анализа.

С точки зрения бизнеса на протяжении всего процесса подготовки данных необходимо развивать все более глубокое понимание структуры данных, их содержания, отношений и правил получения. Крайне важно убедиться, что данные существуют в пригодном для использования состоянии, и понять, что нужно, чтобы преобразовать их в полезный набор данных для отчетов и визуализации. В таком сценарии использование профилирования данных может помочь изучить фактический контент и взаимосвязи в исходных системах предприятия. Профилирование данных может быть простым, как написание некоторых операторов SQL, или сложным, как специализированный инструмент.

Шаг 4: моделирование.

На этом этапе используются различные методы моделирования для проверки данных и поиска ответов на поставленные вопросы. Как правило, существует несколько методов для решения однотипных проблем интеллектуального анализа данных с некоторыми конкретными требованиями к форме данных. Общие модели включают, помимо прочего, линейную регрессию, деревья решений и моделирование случайного леса.

Шаг 5: проверка.

Закончив построение модели и перейдя к окончательному развертыванию, крайне важно тщательно оценить модель и проанализировать шаги, выполненные для ее построения, чтобы убедиться, что она достигает бизнес-целей. Модель работает правильно? Требуется ли дополнительная очистка данных? Получен ожидаемый результат? Если нет, возможно, придется повторить предыдущие шаги еще раз.

На этом этапе важно определить причины отрицательного результата и задокументировать их для использования в будущем других методов. Такая документация полезна с точки зрения бизнеса для будущих пользователей. Ведение списка проблем и устранение новых проблем, возникающих во время проверки данных, может значительно повысить качество проекта, помочь расширить возможности для будущих улучшений и определить потребности бизнеса в инфраструктуре.

Шаг 6: визуализация.

Создание модели, как правило, не означает завершение проекта. Даже если задача модели – расширить знания о данных, полученная информация должна быть организована и представлена таким образом, чтобы это было полезно для клиента. В зависимости от требований этот шаг может быть таким же простым, как создание отчета, или сложным, как реализация повторяющейся оценки данных или процесса интеллектуального анализа данных.

Во многих случаях визуализация данных будет иметь решающее значение для передачи выводов клиенту. Не все клиенты разбираются в данных, и инструменты интерактивной визуализации, такие как Tableau, чрезвычайно полезны для иллюстрации выводов клиентам. Очень важно иметь возможность рассказать историю с полученными данными. Это поможет объяснить клиенту ценность сделанных открытий.

В любом проекте важно четко определить бизнес-цели. Разделение процесса на этапы – гарантия получения наилучших результатов для клиента.

Шаг 7: документация.

Важной составляющей на всех этапах реализации проекта визуализации данных является документация. Она должна содержать краткое описание проекта, источников данных, профиля и качества данных, ограничений данных, ключевых преобразований и представленных моделей, а также их влияния или полезности для повышения качества визуализации. Наконец, в этой документации также должны быть указаны проблемы, возникающие при работе с данными или создании определенной визуализации, которые могут быть решены в будущем.

Рассмотрим важные предпосылки успешного проекта визуализации данных.

Определение цели проекта. За каждым проектом стоит организационная потребность. Она может быть такой же простой, как еженедельная панель показателей продаж, или сложной, как механизм прогнозных рекомендаций. Удовлетворение этих потребностей с помощью определения конкретных измеримых целей обеспечивает правильную основу для предоставления нужной информации правильным образом. Передача ключевых показателей эффективности конечного продукта очень важна для потребителей. Для этого необходимо собрать требования, настроить процессы проектирования, запланировать регулярные обсуждения с пользователями и продолжать эти встречи до окончательного развертывания проекта. Примеры вопросов, которые помогут лучше понять проект:

– Какие потребности организации вы пытаетесь решить?
– К каким основным источникам данных вам нужно получить доступ?
– Являются ли данные актуальными или будут предоставляться обновленные данные через регулярные промежутки времени?
– Какой тип визуализации данных работает лучше всего?
– Есть ли измеримая цель, которую вы хотите достичь?
– Какие ключевые показатели эффективности следует сообщить пользователям, если таковые имеются?
– Какой тип визуализации KPI подходит для этой цели?

Понимание аудитории. Необходимо понимать, как клиенты будут воспринимать эту визуализацию. Визуализация для научных работников кардинально отличается от визуализации, предназначенной для законодателей или для широкой публики. Большинство пользователей хотят видеть ключевые показатели эффективности, которые являются основными драйверами визуализации. Рассмотрим различные виды визуализаций, которые можно использовать для отображения ключевых показателей эффективности:

– количество или мера: пример – количество лайков или комментариев;
– тенденции и изменения во времени: временные ряды; пример – изменение количества продаж с течением времени;

– доли и пропорции: отображение отношений между частями и целым; пример – разбивка портфеля акций по активам;
– ранжированный список: хотя и не является реальной визуализацией данных, но может достичь необходимой цели;
– географическое положение: дает пользователю представление о расположении в пространстве объекта визуализации.

Понимание данных, которые нужно визуализировать: их форма, размер, временной ряд, отношения между объектами, категориальные атрибуты – также является важным предварительным условием. Это данные из одного источника или из нескольких источников? Если данные поступают из нескольких источников, их необходимо объединить по общему критерию. При сборе данных следует помнить о конечной цели.

Выбор лучшего изображения. Одна из самых больших проблем для бизнес-пользователей – решить, какой визуальный элемент следует использовать для наилучшего представления информации: таблица, линейная диаграмма, ареальная диаграмма, гистограмма, точечная диаграмма, круговая диаграмма, древовидные карты, тепловые карты и т. д.

Выбор инструмента для проекта. Выбор инструмента зависит от человека, выполняющего или разрабатывающего визуализацию, и платформы, в которую он хочет интегрировать свою работу, а также от способностей пользователя и его потребностей.

Рассмотрим основные подходы к выбору инструментов визуализации по уровню программирования или кодирования, необходимого для работы с ними.

Без кодирования: при использовании таких инструментов визуализации требуется минимальное программирование либо оно может полностью отсутствовать. Они обычно имеют интуитивно понятный интерфейс, позволяющий пользователям создавать визуализации путем перетаскивания элементов или выбора из предустановленных шаблонов. Такие инструменты идеально подходят для людей без опыта программирования или для тех, кому нужно быстро создать простые визуализации. Для начала можно воспользоваться MS Excel, применяя функцию сводных таблиц для создания качественных диаграмм. Если же у вас уже имеются определенные данные

и вам требуется мощный инструмент для анализа этих данных визуальным способом, то следует обратить внимание на Tableau.

Немного кодирования: эти инструменты требуют некоторого уровня программирования или кодирования, но они все еще достаточно просты в использовании и предназначены для пользователей, имеющих базовые навыки программирования. Они часто предоставляют пользовательские интерфейсы для настройки визуализаций, а также возможность добавления дополнительного функционала через скрипты или API. Если кто-то желает испытать свои силы в кодировании для создания диаграмм, можно использовать язык программирования R. В R доступно множество библиотек для визуального анализа данных.

Больше кодирования: эти инструменты требуют значительного уровня программирования или кодирования. Они обычно используются профессионалами или экспертами в области визуализации данных, которые хотят иметь полный контроль над созданием визуализаций и способны написать сложный код для достижения своих целей. Можно применять Python как мощный инструмент, популярный для анализа данных.

Взаимодействие с пользователем. Если проект будет интерактивным, стоит рассмотреть возможность сделать визуализацию динамичной с помощью фильтров, различных возможностей детализации. Это сделает ее более интересной для пользователя и повысит его вовлеченность.

Придумывая визуализацию, нужно учитывать влияние цветов, используемых в диаграмме. Цвет может привлекать внимание людей, регулировать настроение и влиять на восприятие.

Перед началом любого проекта крайне важно привлечь нужных людей. Это могут быть владельцы бизнеса, которые заказали проект визуализации данных, или ключевые заинтересованные стороны, которые будут активно использовать визуализацию данных. Участие представителей бизнеса имеет большое значение для определения потребности в проекте и достижения общего понимания требований и критериев успеха. Такое сотрудничество значительно увеличивает шансы на то, что полученная визуализация будет отвечать потребностям бизнеса. Процесс создания визуализации должен быть итеративным и динамичным.

Реализация проекта визуализации данных может быть как простой, так и сложной. Как и в любом другом проекте, этап планирования важен для создания эффективной визуализации. Хорошо спланированный проект помогает снизить количество итераций или повторений в процессе создания визуализаций и попыток привести их в соответствие с историей, которую они рассказывают.

Вся визуализация данных строится на наборе данных. У каждого набора данных есть свои специфические потребности в представлении, и цель, для которой используется набор данных, влияет на эти потребности так же, как и сами данные. Понимание характеристик набора данных поможет определить, какую визуализацию нужно использовать.

При планировании проекта визуализации следует учитывать следующие важные моменты.

Избыток информации. Стремясь предоставить аудитории наиболее полную информацию, всегда нужно помнить, что ее не должно быть слишком много, иначе визуализация станет сложной для понимания и запутает читателей больше, чем поможет прояснить ситуацию.

Сбор данных. Следует убедиться, что данные получены из надежного источника и что их объем и глубина достаточны для того, чтобы поверить в то, что рассказывается.

При запуске проекта визуализации важно помнить, что эффективная визуализация информации начинается не с набора данных, а с вопросов. Зачем были собраны данные, что в них интересного и какие истории они могут рассказать? Нужно подумать о том, как будут использоваться данные.

Задавая конкретные и лаконичные вопросы, вы получите четкие результаты. Если же ваши вопросы будут более общими, как на этапе исследовательского анализа данных, то и ответы будут такими же и, скорее всего, они будут адресованы тем, кто хорошо разбирается в данной теме.

Правильная визуализация – это своего рода рассказ, который дает четкий ответ на вопрос, не вдаваясь в детали. Сосредотачиваясь на первоначальной цели вопроса, вы можете исключить ненужную информацию, поскольку вопрос определяет, что является важным, а что нет.

Визуализации позволяют обнаруживать паттерны и идеи, которые могут быть уже известными и очевидными или совершенно новыми. Результаты могут быть различными. Чтобы эффективно искать идеи на основе данных и визуализаций, рекомендуется повторять следующие шаги:

1. Визуализация. Она позволяет особым образом взглянуть на набор данных и может быть выполнена несколькими способами, такими как диаграммы, таблицы, карты и графики. Ценность хорошей визуализации данных заключается в том, что она не только передает информацию, но и помогает увидеть то, что другие люди, возможно, не замечают.

2. Анализ и интерпретация того, что видите. На этом этапе следует задать себе такие вопросы: «Что мы можем видеть на этом изображении?», «Это то, что мы ожидали?», «Есть какие-нибудь интересные паттерны?», «Что это означает в контексте данных?». Эти вопросы помогут не только понять смысл визуализации, но и выявить, что она не раскрывает важной информации о данных, несмотря на то, что выглядит привлекательно.

3. Документирование наблюдений и шагов. Документация – самый важный, но самый пропускаемый шаг. Документация предоставляет контекст, в котором были созданы диаграммы, тем самым устраняя любую путаницу, которая может возникнуть при просмотре нескольких наборов диаграмм. При документировании следует отметить следующие моменты: 1) почему мы создали эту диаграмму; 2) что мы сделали с данными, чтобы их создать; 3) что нам говорит эта диаграмма.

4. Преобразование наборов данных. Этот шаг позволяет глубже изучить взаимосвязи и результаты. Возможные дополнительные вопросы, связанные с данными или результатами, могут возникнуть на основании выводов, сделанных на предыдущих стадиях, и может понадобиться дополнительная проверка или анализ. Это можно осуществить путем преобразования, например масштабирования, фильтрации и удаления выбросов.

Рассмотрим некоторые сложности, с которыми можно столкнуться при работе над проектами по визуализации данных:

1. Игнорирование конечных пользователей. Конечные пользователи часто не участвуют напрямую в определении потребностей в проектах визуализации. Это основная причина, по которой визуальные информационные панели часто не используются после развертывания. Важно составить карту пользовательских историй и услышать, как пользователи подходят к бизнес-задачам. Это практический опыт пользователя, который трудно передать и который тесно связан с умением действовать. Желательно согласовать свои действия с конечными пользователями и подробно узнать их бизнес-представления, чтобы включить их в информационные панели. Формируйте мысли пользователей с помощью интервью, составляйте карту пути пользователя, осторожно исследуя их, и совместно набрасывайте бизнес-сценарии «как есть». Также полезно составить список вопросов, на которые будет отвечать визуализация, и уточнить те, на которые не будет ответа.

2. Попытка включить в приложение все функции: чем больше функций загружается в приложение, тем меньше оно будет использоваться. Когда речь заходит о расстановке приоритетов, даже самые осведомленные пользователи могут испытывать затруднения при принятии сложных решений. В таких случаях важно выступить в роли консультанта и помочь составить список функций, включающий только самое важное. Хотя технически пространство экрана не ограничено, полезно ввести ограничения на плотность данных. Для этого необходимо заинтересовать ключевых лиц, которые понимают приоритетность задач и могут принимать сложные решения, а также обладают достаточным авторитетом, чтобы убедить остальных пользователей.

3. Отсутствие необходимости в исследовании данных: модернизация данных является основной причиной того, что в итоге получаются бездействующие информационные панели или странно выглядящие диаграммы. Без исследовательского анализа карты могут быть искажены выбросами или, что еще хуже, остаться без шаблонов. Данные также влияют на выбор диаграмм. В процессе планирования проекта важно заранее учесть наличие данных. Получение заголовков строк – хорошее начало, однако полные данные требуются для принятия ключевых дизайнерских решений. Клиенты должны быть осведомлены о том, что данные играют критически

важную роль в процессе визуализации и что их понимание непосредственно влияет на конструктивные решения.

4. Стремление сделать все интерактивным. При создании навигации и интерактивности главная идея заключается в том, чтобы сделать все максимально удобным для пользователя. Функциональный интерфейс не подразумевает использования большого количества инструментов, он содержит лишь необходимые и интуитивно понятные элементы.

Тема 15. Визуализация данных в R

Различные инструменты визуализации данных используются для предоставления информации о данных с помощью визуальных подсказок, таких как графики, диаграммы, карты и многие другие. Это помогает интуитивно и легко понять большие объемы данных и принять более взвешенные решения относительно них.

Доступными и популярными инструментами визуализации данных являются Tableau, Plotly, R, Google Charts, Infogram и Kibana. Различные платформы визуализации данных имеют разные функциональные возможности и варианты использования. Они также требуют дополнительного набора навыков.

R – это язык, разработанный для статистических вычислений, графического анализа данных и научных исследований. Обычно его предпочитают для визуализации данных, поскольку он обеспечивает гибкость и минимум необходимого кодирования благодаря своим пакетам.

Язык программирования R обладает обширными возможностями для проведения всех этапов анализа, включая работу с данными. С помощью R можно легко повторно использовать скрипт анализа, а затем просто импортировать в него новые данные.

Анализ данных в R состоит из импорта данных в R, а затем выполнения функций для визуализации и моделирования данных. R имеет мощные функции для охвата всего процесса, начиная от сырых данных до передачи результатов. То есть пользователям не нужно переключаться между приложениями на различных этапах рабочего процесса анализа. Они просто вводят код, позволяют R оценить его и получить результат.

Полный анализ от исходных данных до отчета может включать множество небольших шагов – преобразование переменных в данных, построение графиков, вычисление сводок, моделирование и тестирование, – которые часто выполняются итеративно. Для создания «сценариев R» нужна интегрированная среда разработки (IDE), в которой есть текстовый редактор и консоль в одном месте – RStudio.

R предлагает несколько базовых пакетов, которые устанавливаются по умолчанию. Один из них – графический пакет, который содержит около 100 функций для создания традиционной графики, применяемой для визуализации данных. Простые функции позволяют быстро создавать изображения, такие как диаграммы рассеяния, ящичные диаграммы или гистограммы. Эти функции особенно полезны для быстрого изучения данных.

Например, если применить функцию plot() к набору данных радужной оболочки глаза, то можно получить матрицу графиков разброса, которая соответствует матрице корреляции всех столбцов. Это полезно для очень простого обзора взаимосвязей между переменными (рис. 28).

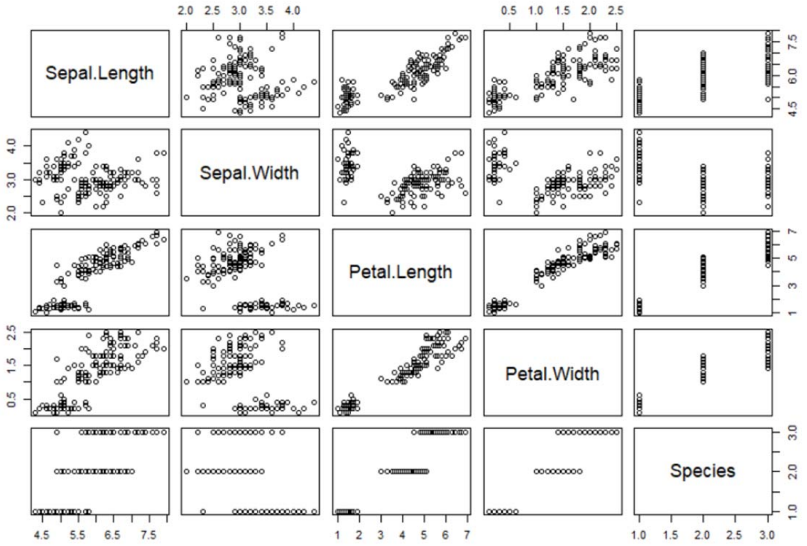


Рис. 28. Визуализация переменных

На рисунке можно видеть, например, что переменные Petal. Length и Petal.Width положительно коррелируют друг с другом.

В дополнение к универсальным функциям R также предлагает многочисленные библиотеки, такие как ggplot2, lattice или plotly, для создания различных типов графики: визуально привлекательных или даже интерактивных.

Библиотека ggplot2 предлагает элегантную и универсальную систему для создания графики. Ее отличает многоуровневый подход, который позволяет создавать изображения шаг за шагом: взять данные, добавить эстетику (например, оси, положение точек данных в графике), а затем любой другой элемент стиля, такой как линии, шкалы или доверительные интервалы.

Создание простой графики на основе набора данных Iris иллюстрирует эту философию. Сначала мы создаем базовую структуру нашей фигуры, которая содержит данные и оси. Кроме того, здесь указано, что данные представлены сгруппированными по видам. Следующим шагом является добавление точек данных. Затем мы добавляем заголовок рисунка, меняем маркировку оси и фон рисунка (рис. 29):

```
    ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, col =
Species)) +
    geom_point() +
    labs(title = “A Nice Iris Dataset Graphic”, x = “Sepal
Length”, y = “Sepal Width”) +
    theme_minimal()
```

Конечно, R также можно использовать для создания значительно более сложных изображений. Можно предложить множество вариантов визуализации данных с помощью R.

В R построение двумерного графика можно выполнить в декартовой или полярной (пакет расширения plotrix) системах координат.

Для того чтобы построить график $y(x)$ в *прямоугольной системе координат*, необходимо сформировать два массива (вектора) x и y одинаковой размерности, а затем обратиться к функции **plot(x,y)**. График формируется путем соединения соседних точек прямыми линиями. Чем больше будет интервал между соседними точками, тем больше будет заметно, что график представляет собой ломаную линию.

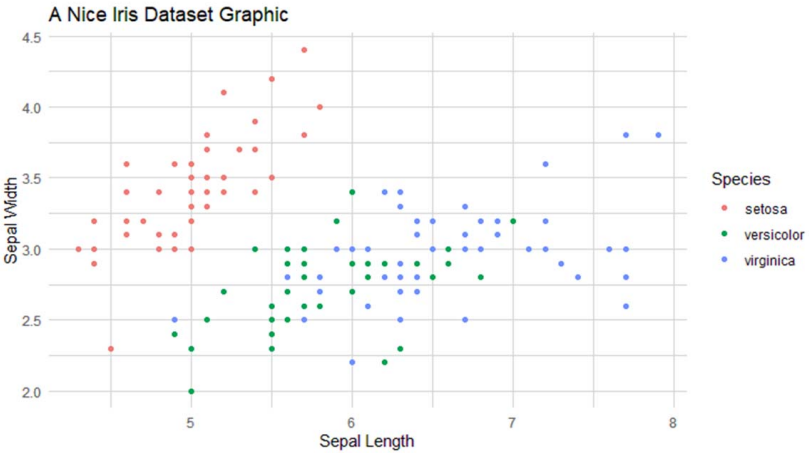


Рис. 29. Диаграмма рассеяния на R

Например, график функции (рис. 30) y = cos(x) при x = [2ϖ, 2ϖ] выглядит следующим образом:

```
h <- pi/10
x <- seq(-2*pi, 2*pi, h)
y <- cos(x)
plot(x,y, type="l",xlab="x", ylab="y = cos(x)")
abline(h=0, v=0)
```

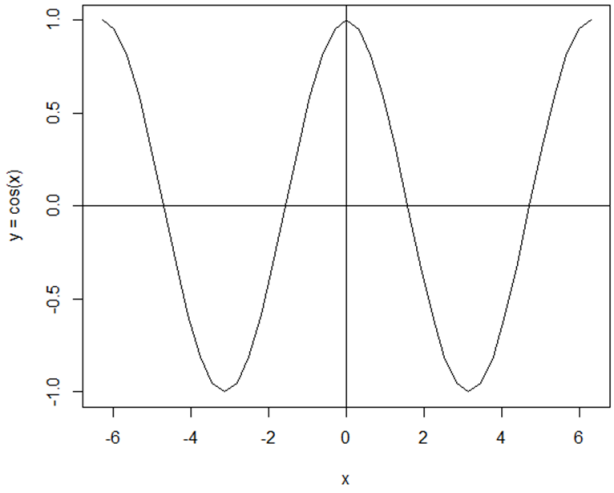


Рис. 30. График функции

С помощью графических опций можно настроить такие атрибуты графиков, как шрифты, цвета, оси, заголовки.

Параметр **pch** используется для указания символов, используемых при построении точек (рис. 31).

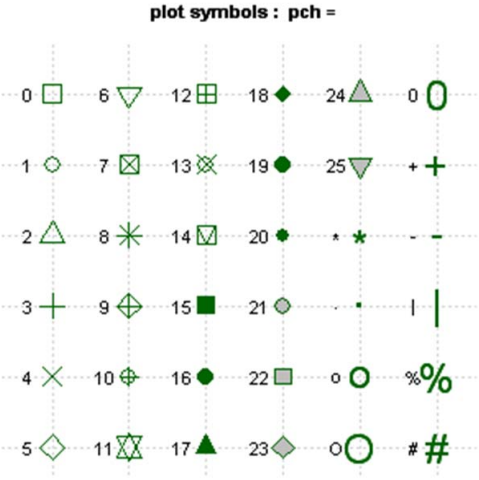


Рис. 31. Типы маркеров

Можно указать цвета в R по индексу, имени, шестнадцатеричному коду или компонентам RGB. Например, col = 1, col = «white» и col = «#FFFFFF» эквивалентны.

R предоставляет функцию задания цвета **rgb()**. Например, желтый цвет можно определить следующим образом:

```
> rgb(1,1,0)
[1] «#FFFF00»
```

Функция **colors()** выводит вектор всех известных цветов R. Нужные цвета могут быть извлечены, если известны их позиции в векторе, следующим образом:

```
> colors()[c(552,254,26)]
[1] “red” “green” “blue”
```

Вывести текущую палитру цветов можно с помощью функции **palette()**:

```
> palette()
[1] “black” “red” “green3” “blue” “cyan” “magenta” “yellow” “gray”
```

Получить информацию о других цветовых палитрах можно с помощью запроса **?rainbow**.

Функция **col2rgb()** может использоваться для извлечения RGB компонентов цвета, например:

```
> col2rgb("green")
[,1]
red                0
green              255
blue               0
```

Полную цветовую палитру цветов R можно найти на сайте: http://research.stowers.org/mcm/efg/R/Color/Chart/.

Использование цветовых параметров показано на примере построения графика (рис. 32):

```
h <- pi/10
A <- 2
x <- seq(0, 2*pi, h)
y <- A*sin(x)
plot(x,y, main="Синусоида", xlab="x", ylab="y = sin(x)", col.axis="red2", col.lab="blue", cex.main=1.5, font.lab=2, pch=22, bg="red", cex=0.7, mar=c(1,3,3,3))
abline(h=0, v=0)
text(pi,0.15, "Точка перегиба", pos=4)
```

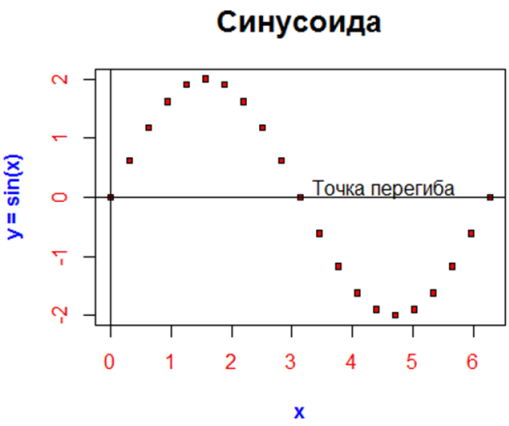


Рис. 32. График синусоиды

R предоставляет также возможность построить несколько областей в графическом окне и вывести на каждую из них свои гра-

фики. Для этого предназначена функция **par(mfrow,mfcol)**. Параметры **mfrow** и **mfcol** определяют количество графиков по вертикали и горизонтали.

На рис. 33 представлен пример графика функций $y = \sin(x)$, $y = x \cdot \sin(x)$, $y = \cos(x)$, $y = |x|$ на интервале $[-\pi, \pi]$.

```
n <- 200
a <- -pi
b <- pi
h <- (b-a)/n
par(mfrow=c(2, 2), cex=0.7)
x <- (0:n)*h+a
y <- sin(x)
plot(x,y, pch=16, cex=0.5)
points(-1.5,0)
y <- x*sin(x)
plot(x,y, pch=16, cex=0.5)
y <- cos(x)
plot(x,y, pch=16, cex=0.5)
y <- abs(x)
plot(x,y, pch=16, cex=0.5)
```

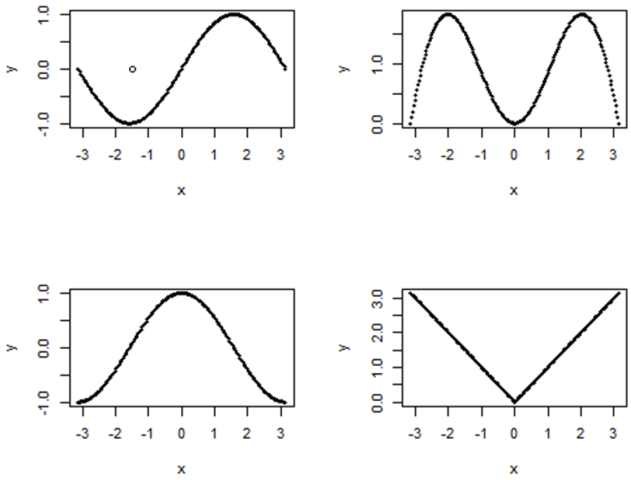


Рис. 33. График функций

Для возможности построения графиков в полярной системе координат в RStudio необходимо подключить пакет plotrix, обеспечивающий способ быстрого получения множества видов графиков без изучения специализированного синтаксиса.

На рис. 34 представлен пример графика функции $\rho = 3\varphi$ на интервале $[0, 6\pi]$.

```
library(plotrix)
n <- 200
a <- 0
b <- 180*6
h <- (b-a)/n
fi <- (0:n)*h+a
fi.rad <- fi*pi/18
r <- 3*fi.rad # спираль Архимеда
polar.plot(r,fi,rp.type=”p”, lwd=2)
```

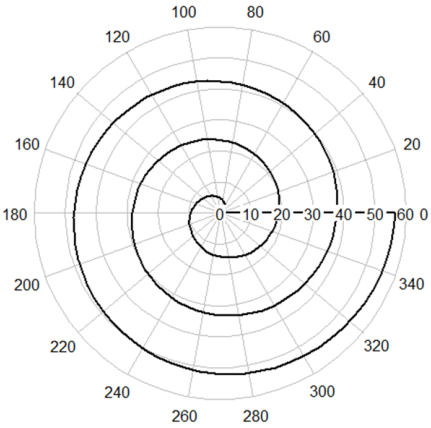


Рис. 34. График функции $\rho = 3\varphi$

Задание функции y(x) с помощью равенств $x = f(t)$ и $y = g(t)$ называют *параметрическим*, а вспомогательную величину t – *параметром*. Построение графика функции, заданной параметрически, можно осуществлять следующим образом:

1. Определить массив t.
2. Определить массивы $x = f(t)$ и $y = g(t)$.
3. Построить график функции $y(x)$ с помощью функции **plot(x,y)**.

Рассмотрим пример (рис. 35) параметрического графика функции $x(t) = t \cdot \sin(t)$; $y(t) = t \cdot \cos(t)$ при $t = [0, 5\pi]$ с шагом $h = 0.01 \cdot \pi$.

```
h <- 0.01*pi
t <- seq(from=0, to=5*pi, by = h)
x <- t*sin(t)
y <- t*cos(t)
plot(x,y,type=”l”,lwd=2,pch=16, cex=0.5)
```

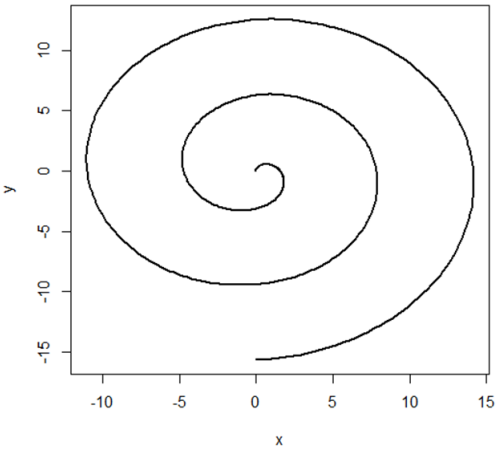


Рис. 35. Параметрический график функции
$x(t) = t \cdot \sin(t)$; $y(t) = t \cdot \cos(t)$

График поверхности (трехмерный, или 3D-график) – это график, положение точки в котором определяется значениями трех координат. Для построения графиков поверхностей в R используются функции **expand.grid()** и **wireframe()**.

Для построения графика двух переменных $z = f(x, y)$ необходимо выполнить следующие действия:

1. Сформировать ряд значений $a = [x.\min, x.\max]$ с шагом $a.h$.
2. Сформировать ряд значений $b = [y.\min, y.\max]$ с шагом $b.h$.
3. С помощью функции expand.grid() сформировать сетку значений x и y.
4. Вычислить значение $z(x, y)$ на этой сетке.
5. Используя функцию wireframe(), построить поверхность $z(x, y)$.

Рассмотрим пример (рис. 36) трехмерного графика функции $z(x, y) = y^2 - x^2$, где $x \in [-2, 2]$, $y \in [-3, 3]$, используя функцию expand.grid().

```
# --- 3D-поверхность z(x,y)=y^2-x^2 ---
library(lattice)
a <- seq(from=-2, to=2,by=0.1)
b <- seq(from=-3, to=3,by=0.1)
xy.grid <- expand.grid(x=a,y=b) # формирует решетку значений x и y
z <- list()
```

```
    z$z <- xy.grid$y^2-xy.grid$x^2 # рассчитывает значения z
```
в узлах решетки
```
    wireframe(z~xy.grid$x*xy.grid$y, z, shade = TRUE, aspect =
c(1, 1), light.source = c(10,-10,10), main = "z(x,y)=y^2-x^2",
    scales = list(z.ticks=10,arrows=FALSE, col="black",
font=15, tck=1), screen = list(z = 40, x = -75, y = 0))
```

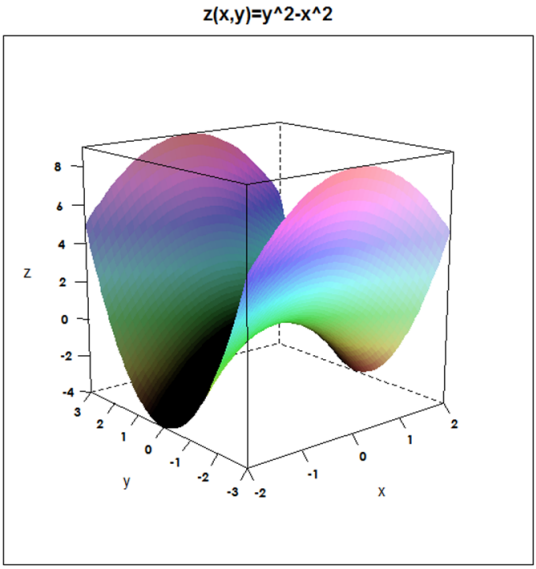


Рис. 36. Трехмерный график функции $z(x, y) = y^2 - x^2$

Рассмотрим пример (рис. 37) трехмерного графика функции $z(x, y) = y^2 + x^2$, где $x \in [-2, 2]$, $y \in [-3, 3]$, используя функцию expand. grid().

```
    # --- 3D-поверхность z(x,y)=y^2+x^2 ---
    library(lattice)
    a <- seq(from=-2, to=2,by=0.1)
    b <- seq(from=-3, to=3,by=0.1)
    xy.grid <- expand.grid(x=a,y=b)
    z <- list()
    z$z <- xy.grid$y^2+xy.grid$x^2 # рассчитывает значения z
в узлах решетки
    wireframe(z~xy.grid$x*xy.grid$y, z, shade = TRUE, aspect =
c(1, 1), light.source = c(10,-10,10), main = "z(x,y)=y^2+x^2",
    scales = list(z.ticks=10,arrows=FALSE, col="black",
font=15, tck=1), screen = list(z = 40, x = -75, y = 0))
```

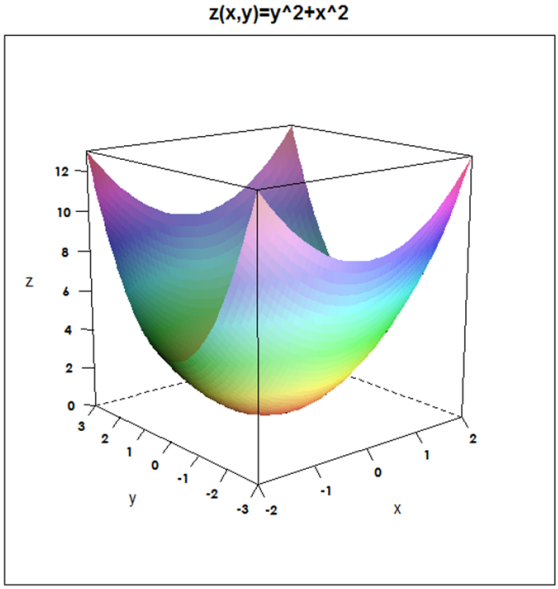


Рис. 37. Трехмерный график функции $z(x, y) = y^2 + x^2$

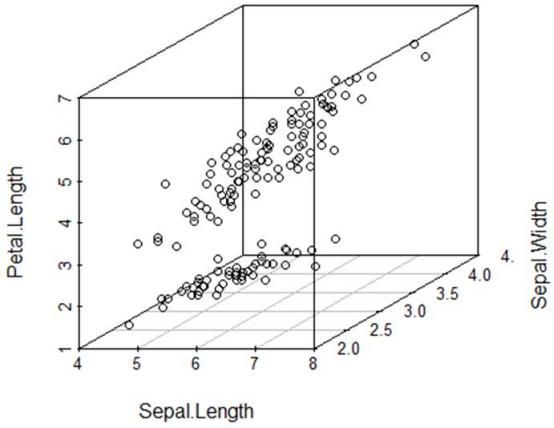


Рис. 38. Трехмерный график рассеяния выборки iris

Рассмотрим пример (рис. 38) трехмерного графика рассеяния встроенной в R выборки iris. Для трехмерной визуализации данных подключить пакет **scatterplot3d**:

```
library(scatterplot3d)
scatterplot3d(iris[,1:3])
```

R имеет следующие *преимущества* перед другими инструментами визуализации данных:

- предлагает широкий набор библиотек визуализации, а также подробное онлайн-руководство по их использованию;
- предлагает визуализацию данных в виде 3D-моделей и многопанельных диаграмм;
- с его помощью можно легко настроить визуализацию данных, изменив оси, шрифты, легенды, аннотации и метки.

R также имеет следующие *недостатки*:

- предпочтителен для визуализации данных только тогда, когда она выполняется на отдельном автономном сервере;
- визуализация данных с его использованием выполняется медленно для больших объемов данных по сравнению с другими аналогами.

Области применения:

- предоставление аналитических выводов данных неаналитическим отделам компании;
- использование визуализации данных устройствами для мониторинга здоровья для отслеживания любых аномалий артериального давления, холестерина и других;
- обнаружение повторяющихся закономерностей и тенденций в потребительских и маркетинговых данных;
- использование визуализации данных метеорологами для оценки преобладающих изменений погоды во всем мире;
- карты реального времени и системы геопозиционирования можно использовать для визуализации и мониторинга трафика и оценки времени в пути.

Тема 16. Визуализация данных в Python

Python – один из важнейших языков программирования в области науки о данных. Начать работу с этим языком программирования немного проще по сравнению с языком программирования R, например, потому что Python имеет простой для понимания синтаксис.

Важным шагом в исследовании данных, а затем и в публикации и передаче информации является визуализация данных, поскольку это позволяет находить и прояснять аномалии и взаимосвязи, которые остаются скрытыми, с помощью чисто описательных оценок.

Все решения по визуализации в Python можно ограничить тремя основными элементами:

– *кодировка* (или *эстетика*): какие данные или какие переменные должны быть перенесены в какие визуальные элементы (например, ось *X*, ось *Y*, цвета);
– *геометрия*: как эти визуальные элементы должны быть представлены (например, в виде точек, линий, полос) в двумерном пространстве, доступном с помощью графического представления;
– *весы*: в каких масштабах следует отображать соответствующие элементы.

Наиболее часто используемой библиотекой для визуализации данных в Python является Matplotlib.

Диаграмма рассеяния с графическим заголовком и маркировкой осей может быть создана при помощи Matplotlib с использованием функции **scatter()** (рис. 39).

```
import matplotlib.pyplot as plt
girls_grades = [89, 90, 70, 89, 100, 80, 90, 100, 80, 34]
boys_grades = [30, 29, 49, 48, 100, 48, 38, 45, 20, 30]
grades_range = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
fig=plt.figure()
ax=fig.add_axes([0,0,1,1])
ax.scatter(grades_range, girls_grades, color='r')
ax.scatter(grades_range, boys_grades, color='b')
ax.set_xlabel('Ось X')
ax.set_ylabel('Ось Y')
ax.set_title('Диаграмма рассеяния')
plt.show()
```

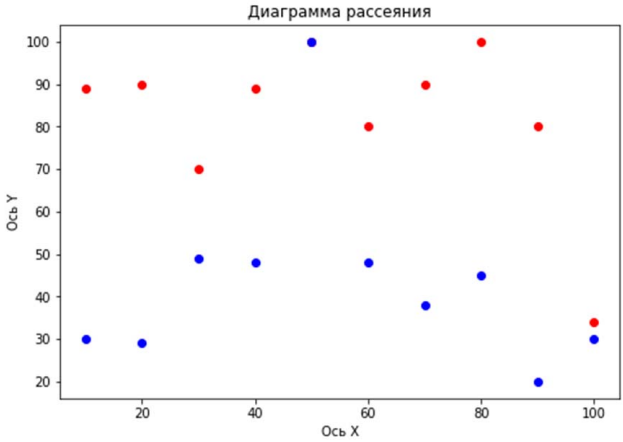


Рис. 39. Диаграмма рассеяния

Рассмотрим пример получения трехмерной диаграммы (рис. 40).

```
iimport matplotlib.pyplot as plt
import numpy as np
from mpl_toolkits.mplot3d import axes3d
ax = axes3d.Axes3D(plt.figure())
i = np.arange(-1, 1, 0.01)
X, Y = np.meshgrid(i, i)
Z = X**2 - Y**2
ax.plot_wireframe(X, Y, Z, rstride=10, cstride=10)
plt.show()
```

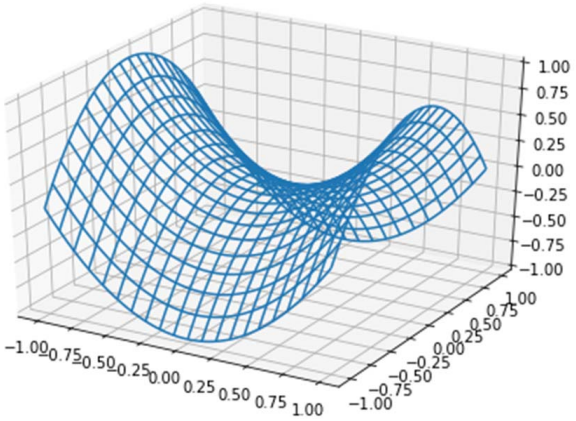


Рис. 40. Трехмерный график

Рассмотрим пример получения столбиковой диаграммы (рис. 41) с возможностью группирования данных и добавлением легенды.

```
    # a stacked bar plot with errorbars
    import numpy as np
    import matplotlib.pyplot as plt
    N = 5
    menMeans = (20, 35, 30, 35, 27)
    womenMeans = (25, 32, 34, 20, 25)
    menStd = (2, 3, 4, 1, 2)
    womenStd = (3, 5, 2, 3, 3)
    ind = np.arange(N) # the x locations for the groups
    width = 0.35 # the width of the bars: can also be len(x)
sequence
    p1 = plt.bar(ind, menMeans, width, color='#d62728',
yerr=menStd)
    p2 = plt.bar(ind, womenMeans, width,
                 bottom=menMeans, yerr=womenStd)
    plt.ylabel('скорость')
    plt.title('Скорость по группам›)
    plt.xticks(ind, ('G1', 'G2', 'G3', 'G4', 'G5'))
    plt.yticks(np.arange(0, 81, 10))
    plt.legend((p1[0], p2[0]), ('Мужчины', 'Женщины'))
    plt.show()
```

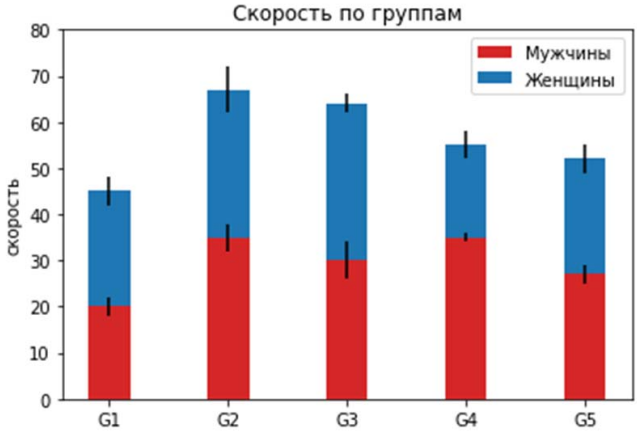


Рис. 41. Столбиковая диаграмма с группировкой

Рассмотрим пример получения ящичковой диаграммы (рис. 42).

```
# example of a box and whisker plot
from numpy.random import seed
from numpy.random import randn
from matplotlib import pyplot
# seed the random number generator
seed(1)
# random numbers drawn from a Gaussian distribution
x = [randn(1000), 5 * randn(1000), 10 * randn(1000)]
# create box and whisker plot
pyplot.boxplot(x)
# show line plot
pyplot.show()
```

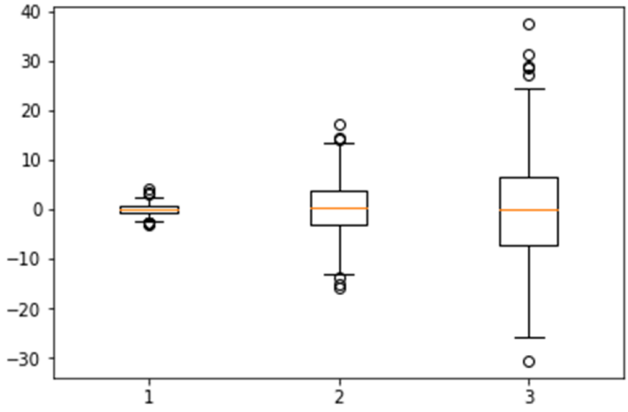


Рис. 42. Ящичковая диаграмма

Seaborn – библиотека на Python, которая часто используется в дополнение к Matplotlib. Она должна упростить создание изображений. Seaborn подходит для получения первоначального обзора данных в рамках их исследования. Например, можно использовать функцию **pairplot()** (рис. 43).

```
import seaborn as sns
df = sns.load_dataset(«penguins»)
sns.pairplot(df, hue=”species”)
```

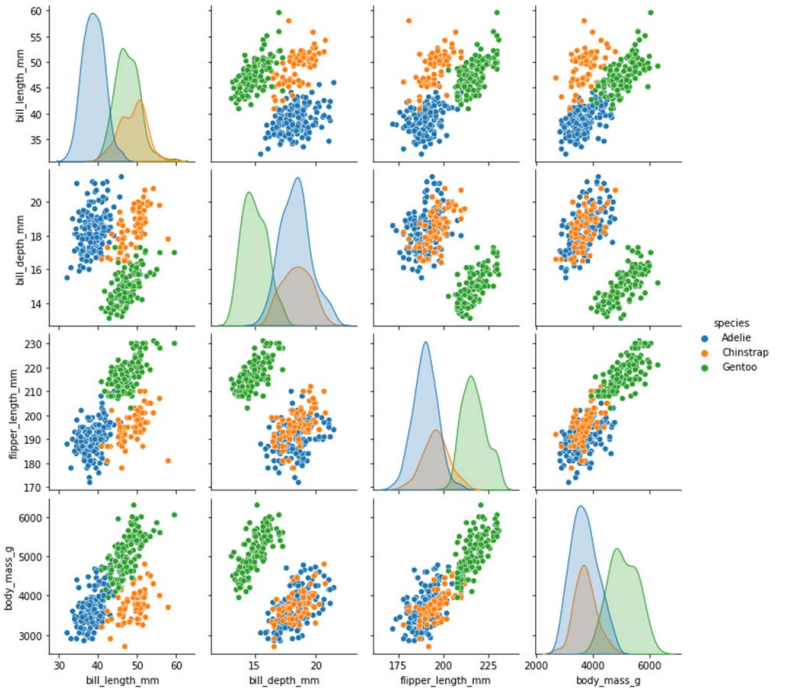


Рис. 43. Использование библиотеки Seaborn

Среда визуализации Colab позволяет использовать для анализа и визуализации данных все возможности популярных библиотек Python. Например, библиотека numpy используется для генерации случайных данных, а библиотека matplotlib – для их визуализации (рис. 44).

```
import numpy as np
from matplotlib import pyplot as plt
ys = 200 + np.random.randn(100)
x = [x for x in range(len(ys))]

plt.plot(x, ys, '-')
plt.fill_between(x, ys, 195, where=(ys > 195), facecolor='g',
alpha=0.6)
plt.title("Sample Visualization")
plt.show()
```

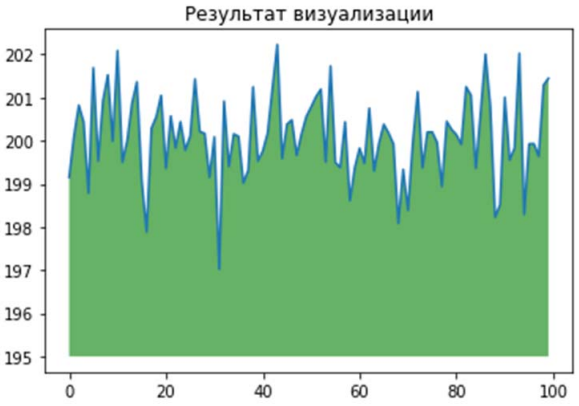


Рис. 44. Визуализация данных

Рассмотрим подробное описание диаграммы «Радар» с областями (рис. 45).

```
    ffrom math import pi
    import matplotlib.pyplot as plt
    # Вносим данные
    cat = [‘Область1’, ‘Область2’, ‘Область3’, ‘Область4’, ‘Об-
ласть5’]
    values = [70, 20, 80, 70, 65]
    N = len(cat)
    x_as = [n / float(N) * 2 * pi for n in range(N)]
    # Связываем последнее значение с первым, чтобы построить
радиальный график
    values += values[:1]
    x_as += x_as[:1]
    # Устанавливаем цвет и толщину линий
    plt.rc(‘axes’, linewidth=0.5, edgecolor=”#888888”)
    # Создаем диаграмму
    ax = plt.subplot(111, polar=True)
    # Устанавливаем стили для сетки
    ax.xaxis.grid(True,  color=”#888888”,  linestyle=’solid’,
linewidth=0.5)
    ax.yaxis.grid(True,  color=”#888888”,  linestyle=’solid’,
linewidth=0.5)
    ax.set_theta_offset(pi / 2)
    ax.set_theta_direction(-1)
    ax.set_rlabel_position(0)
    # Убираем стандартные метки
    plt.xticks(x_as[:-1], [])
    # Выводим шаг значения на график
```

```
    plt.yticks([30, 60], [«30», «60»])
    # Берем данные для диаграммы
    ax.plot(x_as,  values,  linewidth=0,  linestyle='solid',
zorder=3)
    # Заполняем область под значениями
    ax.fill(x_as, values, 'b', alpha=0.3)
    # Ограничиваем области
    plt.ylim(0, 100)
    # Отрисовываем все элементы
    for i in range(N):
        angle_rad = i / float(N) * 2 * pi
        if angle_rad == 0:
               ha, distance_ax = "center", 10
        elif 0 < angle_rad < pi:
               ha, distance_ax = "left", 1
        elif angle_rad == pi:
               ha, distance_ax = "center", 1
        else:
               ha, distance_ax = "right", 1
        ax.text(angle_rad, 100 + distance_ax, cat[i], size=10,
horizontalalignment=ha, verticalalignment="center")
    # Показываем итоговую диаграмму
    plt.show()
```

Рис. 45. Диаграмма «Радар»

На диаграмму (рис. 46) также имеется возможность добавить:

– произвольный текст (метод Axes.text());

– аннотацию (метод Axes.annotate()).

Текст в matplotlib поддерживает известный научный формат TeX, позволяющий записывать сложные математические выражения.

```
    import numpy as np
    import matplotlib
    import matplotlib.pyplot as plt
    if __name__ == “__main__”:
        x = np.arange(-8, 3, 0.1) # x – массив np.array
        y1 = abs(x**2 + 4*x – 5)
        y2 = [9] * len(x)
        fig, ax = plt.subplots()
        fig.canvas.set_window_title(“Графики функций»)
        # Настройки диаграммы и осей
        ax.set_title(«Графики функций: экстремум»)
        ax.set_xlabel(«Ось абсцисс»)
        ax.set_ylabel(«Ось ординат»)
        ax.grid(True)
        # 2 графика
        ax.plot(x, y1, ‘r’, linewidth=3, label=»Парабола»)
        ax.plot(x, y2, label=”Линия”)
        # Аннотации и текст
        ax.annotate(r”Экстремум функции = $\frac{-b}{2a} = \
frac{-4}{2} = -2$”,
                    xy=(-2, 9), xytext=(-4.8, 15.5),
                    arrowprops=dict(facecolor=”black”,
shrink=0.05))
        ax.text(-7, 24.5, «На диаграмме 2 графика:\nпарабола
и линия экстремума»)
        # Легенда
        ax.legend()
        plt.show()
```

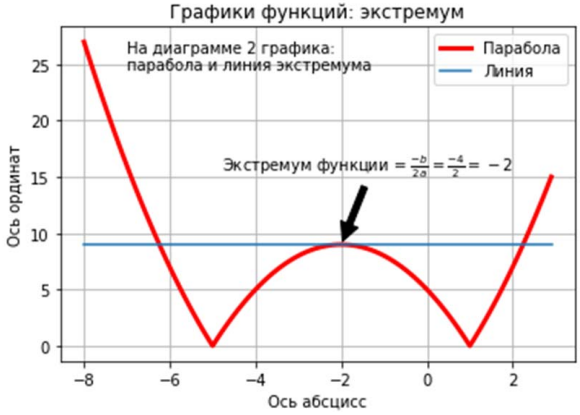


Рис. 46. График функции

Язык Python широко применяется для работы с данными. В экосистеме Python можно найти большое количество библиотек, предназначенных для визуализации данных, то есть построения графиков и диаграмм.

Контрольные вопросы

1. Что представляет собой анализ данных? Какие основные этапы он включает? Поясните каждый этап.
2. С чего начинается анализ данных? Что является ключевым фактором успешного анализа данных? Как можно определить проблему? Какое значение она имеет для анализа данных?
3. Какие процессы включает подготовка данных? Как они связаны между собой? Каковы особенности исследования данных? Какие шаги включает исследование данных?
4. Перечислите основные этапы анализа данных. Опишите каждый из этапов. Какие методы применяются на разных этапах анализа данных?
5. Какие шаги включает проект визуализации данных? Что необходимо учитывать в проекте визуализации?
6. Перечислите важные предпосылки успешного проекта визуализации данных.
7. Какую роль играет документирование в проекте визуализации?
8. Назовите основные подходы к выбору инструментов визуализации. Дайте им описание.
9. Какие моменты важно учитывать при планировании проекта визуализации данных?
10. Назовите шаги для обеспечения эффективного подхода к поиску идей на основе данных и визуализаций.
11. Выделите проблемные моменты в проектах визуализации данных. Обоснуйте их возможные решения.
12. Дайте характеристику R как инструментария, с помощью которого можно осуществлять анализ и визуализацию данных. Выделите преимущества применения R.
13. Какими элементами можно ограничить решения по визуализации в Python? Выделите преимущества применения Python.

14. Опишите библиотеки, которые используются для визуализации данных в Python?
15. Приведите примеры типов диаграмм, которые можно строить средствами R и Python.

Тесты для самоконтроля

1. Определите последовательность осуществления анализа данных. (*несколько вариантов ответа*)
1) определение проблемы
2) извлечение данных
3) подготовка данных
4) исследование и визуализация данных
5) проверка модели, тестирование

2. На каком этапе анализа данных осуществляется проверка модели, построенной на основе начальных данных? (*один вариант ответа*)
1) определение проблемы
2) подготовка данных
3) исследование и визуализация данных
4) тестирование

3. На каком этапе анализа данных предоставляются результаты, то есть выводы анализа? (*один вариант ответа*)
1) развертывание
2) подготовка данных
3) исследование и визуализация данных
4) тестирование

4. Какие этапы включает процесс развертывания? (*несколько вариантов ответа*)
1) очистка данных
2) преобразование данных
3) визуализация и интерпретация результатов
4) развертывание решения

5. С чего начинается процесс анализа данных? (*один вариант ответа*)

1) с определения проблемы
2) извлечения данных
3) подготовки данных
4) проверки модели

6. Какие задачи решаются во время планирования? (*несколько вариантов ответа*)

1) выбирается эффективная команда
2) определяются ресурсы
3) выбирается нужный инструментарий
4) осуществляется построение модели

7. Какой процесс позволяет определить, какой метод анализа данных лучше всего подойдет для определения модели? (*один вариант ответа*)

1) исследование данных
2) изучение данных
3) нормализация данных
4) превращение в оптимизированный набор данных

8. Как называется метод анализа данных, используемый для поиска групп, объединенных общими атрибутами? (*один вариант ответа*)

1) обобщение данных
2) группировка данных
3) кластерный анализ
4) регрессионный анализ

9. Как называется процесс в анализе данных, который нужен для создания или поиска подходящей статистической модели для предсказания вероятности результата? (*один вариант ответа*)

1) обобщение данных
2) группировка данных
3) кластерный анализ
4) предсказательная аналитика

10. Какие модели будут использоваться в ходе анализа данных, если полученный результат – качественная переменная? (*один вариант ответа*)

1) модели классификации
2) регрессионные модели
3) кластерные модели
4) предсказательные модели

11. Какие техники включают простые методы генерации моделей анализа данных? (*несколько вариантов ответа*)

1) линейная регрессия
2) классификация
3) кластеризация
4) дерево решений

12. На каком этапе анализа данных подбирается формат будущей визуализации? (*один вариант ответа*)

1) формулирование цели
2) сбор данных
3) подготовка данных
4) исследование данных

13. Какие существуют основные подходы к выбору инструментов визуализации? (*несколько вариантов ответа*)

1) без кодирования
2) немного кодирования
3) больше кодирования
4) меньше кодирования

14. К какому подходу к выбору инструментов визуализации относят MS Excel с функцией сводных таблиц для получения сложных диаграмм? (*один вариант ответа*)

1) без кодирования
2) немного кодирования
3) больше кодирования
4) меньше кодирования

15. К какому подходу к выбору инструментов визуализации относят Python, как мощный инструмент и популярный язык для анализа данных? (*один вариант ответа*)

1) без кодирования
2) немного кодирования
3) больше кодирования
4) меньше кодирования

16. Какие шаги нужно предпринять, чтобы минимизировать игнорирование конечных пользователей? (*несколько вариантов ответа*)

1) составить карту пользовательских историй
2) координировать свои действия с конечными пользователями и собирать их подробные бизнес-перспективы
3) создать бизнес-сценарии «как есть»
4) создать бизнес-сценарии «как будет»

17. Какой редактор используется для визуализации данных? (*один вариант ответа*)

1) R
2) Ar
3) VR
4) IR

18. Какие библиотеки предлагает R в дополнение к универсальным функциям? (*несколько вариантов ответа*)

1) ggplot2
2) lattice
3) plotly
4) petal

19. Какой пакет используется в R для построения графиков в полярной системе координат? (*один вариант ответа*)

1) ggplot2
2) lattice
3) plotly
4) plotrix

20. Какие функции используются в R для построения графиков поверхностей? (*несколько вариантов ответа*)
1) rgb()
2) wireframe()
3) expand.grid()
4) par()

21. Выберите свойства, которые характеризуют R. (*несколько вариантов ответа*)
1) имеет отличные возможности для всех этапов анализа
2) позволяет повторно использовать сценарий анализа
3) позволяет импортировать другие данные
4) легко описывает логику исследования

22. Что включает анализ данных в R? (*несколько вариантов ответа*)
1) импорт данных
2) выполнение функций для визуализации данных
3) выполнение функций для моделирования данных
4) вычислительный процесс

23. Для чего нужна в R интегрированная среда разработки? (*один вариант ответа*)
1) для создания сценариев
2) для создания программного кода
3) для создания пакетов
4) для создания визуализации

24. При помощи какой функции в R можно получить матрицу графиков разброса? (*один вариант ответа*)
1) plot()
2) points()
3) library()
4) expand()

25. Какой язык программирования применяют в области науки о данных? (*один вариант ответа*)
1) Python
2) Java
3) C#
4) R

26. Какой элемент позволяет определить размеры, в которых следует отображать соответствующие элементы? (*один вариант ответа*)
1) кодировка
2) геометрия
3) весы
4) масштаб

27. Какие библиотеки не используются в Python? (*несколько вариантов ответа*)
1) Matplotlib
2) Seaborn
3) ggplot2
4) plotly

28. Какая функция в Python позволяет получить первоначальный обзор данных в рамках их исследования? (*один вариант ответа*)
1) scatter()
2) pairplot()
3) range()
4) grades()

29. Какая функция в Python позволяет создать диаграмму рассеяния? (*один вариант ответа*)
1) scatter()
2) pairplot()
3) range()
4) grades()

Практическое задание

Подготовить данные и визуализировать результаты анализа данных с использованием MATLAB, Excel или Python.

Методические указания

1. Для выполнения задания подготовьте таблицу с данными. Допускается использование доступных встроенных в среду выборок данных.
2. Проведите анализ имеющихся данных, изучите структуру данных, зависимые переменные.
3. Создайте диаграмму размаха, чтобы показать: симметричны ли данные; насколько есть смещение и каково направление смещения; насколько плотно сгруппированы данные.
4. Постройте диаграмму рассеяния для определения зависимостей между основными данными в выборке данных.
5. Результаты исследования представьте доступными средствами визуализации и инфографики.

Пример выполнения задания на языке R

1. Необходимо подготовить исходные данные. В качестве исходных данных возьмем встроенную в базовую инсталляцию R выборку данных `mtcars` из американского журнала Motor Trend за 1974 год. Выборка содержит информацию о дизайне и технических характеристиках (число цилиндров, объем и мощность двигателя, расход топлива и т. д.) для 32 марок автомобилей.

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4  8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4 8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
```

```
Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1
```

Стандартным средством представления табличных данных в R являются фреймы данных. Посмотрим на структуру набора данных.

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Фрейм данных содержит 32 наблюдения с 11 характеристиками.

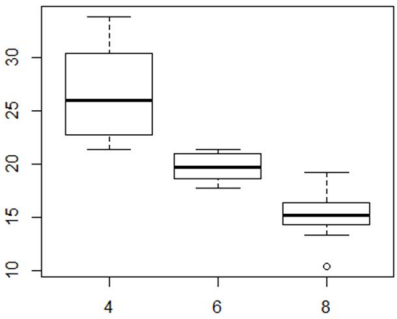
2. Для последующего анализа вызовем описательную статистику выборки данных `mtcars`.

```
> summary(mtcars)
      mpg             cyl             disp             hp             drat             wt
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424
      qsec             vs              am              gear            carb
 Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :17.71   Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
 Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

3. Диаграмма размаха часто используется в описательной статистике для быстрого исследования одного или нескольких наборов данных. Данная диаграмма позволяет увидеть, симметричны ли данные, смещение данных и направление смещения, насколько плотно сгруппированы данные.

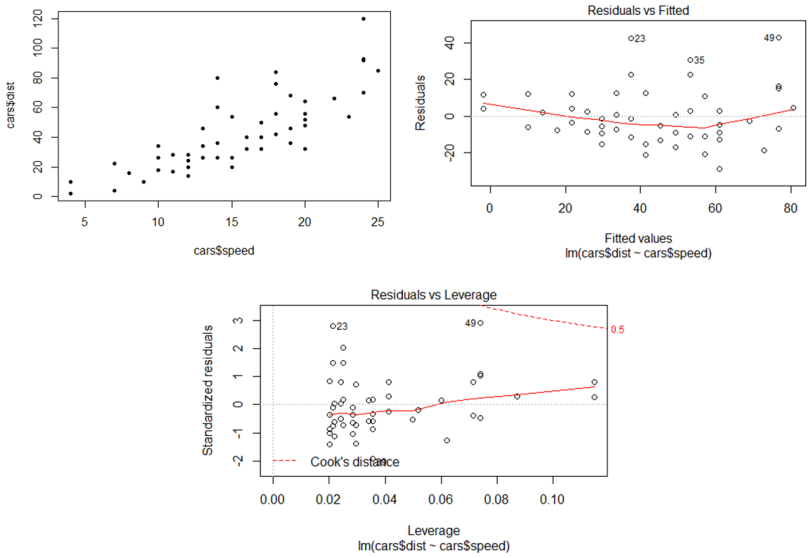
На рисунке представлена зависимость расхода топлива mpg от выбранной переменной (в данном случае – от количества цилиндров) в виде диаграммы размаха:

```
> boxplot(mpg ~ cyl, data = mtcars)
```



4. Диаграмма рассеяния позволяет установить, существует ли зависимость между переменными, а также понять, как объекты дифференцируются по значениям переменных. Например, для данной выборки можно проанализировать зависимость тормозного пути от скорости автомобиля, построив графики рассеяния.

```
> str(cars)
'data.frame':      50 obs. of 2 variables:
 $ speed: num 4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num 2 10 4 22 16 10 18 26 34 17 ...
> plot(cars)
```



Чем выше скорость, тем длиннее тормозной путь, а разброс зависит от множества факторов (авто, дороги, шин и т. п.).

ОТВЕТЫ НА ТЕСТЫ

№ вопроса	Вариант ответа	№ вопроса	Вариант ответа
Глава 1. Основы анализа данных			
1	1	16	1
2	1, 2	17	1
3	1	18	1
4	1, 2	19	1, 2, 3, 4
5	1	20	1
6	1	21	1
7	1	22	1
8	1	23	1, 2, 3
9	1	24	1, 2, 3
10	1	25	1
11	1	26	1
12	1, 2, 3, 4	27	1
13	1	28	1
14	1	29	1
15	1	30	1

№ вопроса	Вариант ответа	№ вопроса	Вариант ответа
Глава 2. Виды анализа данных			
1	1	16	1, 2, 3
2	1	17	1, 2
3	1	18	1
4	1	19	1
5	1	20	1
6	1	21	1, 2
7	1	22	1, 2
8	1	23	1, 2
9	1, 2	24	1
10	1	25	1
11	1	26	1
12	1	27	1
13	1	28	1, 2
14	1	29	1
15	1	30	1

№ вопроса	Вариант ответа	№ вопроса	Вариант ответа
Глава 3. Общее представление о визуализации данных			
1	2	16	3
2	2	17	3
3	1, 2, 3	18	2
4	1, 2, 3, 4	19	4
5	1, 3	20	2
6	1, 2, 3	21	3
7	1	22	3
8	3	23	3
9	1, 2, 3	24	1
10	1, 2, 3	25	1, 2, 3
11	1, 2, 3	26	1, 2, 3
12	1, 2, 3	27	4
13	1, 2, 3	28	1, 2, 3
14	1, 2, 3	29	2
15	1, 2, 3	30	1, 2, 3

№ вопроса	Вариант ответа	№ вопроса	Вариант ответа
Глава 4. Проект анализа и визуализации данных			
1	1, 2, 3, 4, 5	16	1, 2, 3
2	4	17	1
3	1	18	1, 2, 3
4	3, 4	19	4
5	1	20	2, 3
6	1, 2, 3	21	1, 2, 3
7	1	22	1, 2, 3
8	3	23	1
9	4	24	1
10	1	25	1
11	1, 2, 4	26	3
12	4	27	3, 4
13	1, 2, 3	28	2
14	1	29	1
15	3		

ЗАКЛЮЧЕНИЕ

В результате изучения учебно-методического пособия:

✓ *студент:*

– освоит основные понятия анализа и визуализации данных;
– разберет понятия и преимущества корреляционного анализа;
– выделит структурные элементы регрессионного анализа и факторного анализа;
– рассмотрит сферу применения и алгоритм проведения кластерного анализа;
– получит общее представление о визуализации данных;
– рассмотрит наиболее распространенные типы визуализации данных и инструменты;
– выяснит основные этапы анализа данных для их визуализации;
– рассмотрит процесс визуализации данных в R и Python;

✓ *преподаватель дисциплины:*

– пополнит методическую копилку для проведения курса, посвященного анализу и визуализации данных;
– получит снижение трудоемкости при подготовке к занятиям;
– разнообразит тестовый материал для проверки знаний у студентов;

✓ *специалист в предметной области:*

– систематизирует знания в области анализа и визуализации данных;
– получит методический набор инструментария, который можно применять в практической деятельности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Воскобойников, Ю. Е. Обработка и анализ экспериментальных данных в пакетах MathCAD и Excel : учеб. пособие / Ю. Е. Воскобойников. – Новосибирск : Новосибирский государственный архитектурно-строительный университет (Сибстрин), 2020. – 160 с. – URL: www.iprbookshop.ru/107639.html (дата обращения: 10.04.2022). – Режим доступа: по подписке. – ISBN 978-5-7795-0906-0.
2. Моделирование и визуализация экспериментальных данных : учеб. пособие (лабораторный практикум) / авт.-сост. Е. В. Крахоткина. – Ставрополь : Северо-Кавказский федеральный университет, 2018. – 124 с. –URL: www.iprbookshop.ru/92565.html (дата обращения: 10.04.2022). – Режим доступа: по подписке.
3. Кумратова, А. М. Методы хранения и анализа данных : учеб. пособие / А. М. Кумратова, И. И. Василенко ; Кубанский государственный аграрный университет имени И. Т. Трубилина. – Краснодар : КубГАУ, 2021. – 182 с. – URL: e.lanbook.com/book/254297 (дата обращения: 10.04.2022). – Режим доступа: по подписке. – ISBN 978-5-907474-28-4.
4. Маккинли, Уэс. Python и анализ данных / Уэс Маккинли ; пер. с англ. А. Слинкин. – 2-е изд. – Саратов : Профобразование, 2019. – 481 с. – URL: www.iprbookshop.ru/88752.html (дата обращения: 10.04.2022). – Режим доступа: по подписке. – ISBN 978-5-4488-0046-7.
5. Орлов, А. И. Прикладной статистический анализ : учебник / А. И. Орлов. – Москва : Ай Пи Ар Медиа, 2022. – 812 с. – URL: www.iprbookshop.ru/117038.html (дата обращения: 10.04.2022). – Режим доступа: по подписке. – ISBN 978-5-4497-1480-0.
6. Синева, И. С. Анализ данных в среде R. Учебное пособие. Часть 1 / И. С. Синева. – Москва : Московский технический университет связи и информатики, 2018. – 32 с. – URL: www.iprbookshop.ru/92422.html (дата обращения: 10.04.2022). – Режим доступа: по подписке.
7. Шнарева, Г. В. Анализ данных : учеб.-метод. пособие / Г. В. Шнарева, Ж. Г. Пономарева. – Симферополь : Университет экономики и управления, 2019. – 128 с. – URL: www.iprbookshop.ru/89482.html (дата обращения: 10.04.2022). – Режим доступа: по подписке.

ГЛОССАРИЙ

Анализ данных – процесс, состоящий из нескольких шагов, в которых сырые данные преобразуются и обрабатываются с целью создания визуализации и прогнозирования на основе математической модели.

Анализ текста – это метод анализа, применяемый для извлечения машиночитаемых фактов.

Визуализация данных – это метод преобразования необработанной информации: текста, чисел или символов – в графический формат; представление данных в виде, обеспечивающем максимально эффективную работу специалиста по их изучению.

Выборка (или выборочная совокупность) – конкретная группа, из которой собираются эмпирические данные.

Генеральная совокупность – вся группа, которую необходимо исследовать и о которой необходимо сделать выводы.

Гипотеза – это утверждение, которое может быть проверено научными исследованиями.

Диагностический анализ – это исследование, которое проводится для выявления причин возникновения определенных проблем или сбоев в работе системы. Он включает в себя изучение данных, собранных в ходе работы системы, и поиск возможных причин этих проблем.

Дискриминантный анализ – метод классификации в интеллектуальном анализе данных.

Дисперсия – это мера разброса значений случайной величины относительно ее математического ожидания. Она показывает, насколько сильно значения случайной величины отклоняются от среднего значения.

Искусственная нейронная сеть – это система, которая изменяет свою структуру на основе информации, которая проходит через сеть.

Кластеризация – это задача группировки набора объектов таким образом, чтобы объекты в одной группе (называемой кластером) были более похожи друг на друга, чем на объекты в других группах (кластерах).

Кластерный анализ – метод анализа данных, используемый для поиска групп, объединенных общими атрибутами (также называется группировкой).

Корреляционный анализ – метод, который используется для проверки прочности связи между переменными.

Коэффициент корреляции – число между –1 и 1, описывающее силу и направление взаимосвязи между переменными.

Линейная регрессия – это регрессионная модель, которая использует прямую линию для описания отношений между переменными.

Множественная линейная регрессия – метод, который используется для оценки взаимосвязи между двумя или более независимыми переменными и одной зависимой переменной.

Нейронная сеть – это парадигма программирования, вдохновленная биологическими факторами, которая представляет собой метафору мозга для обработки информации.

Нечеткая логика – это метод анализа данных, основанный на вероятности, который помогает справиться с неопределенностями в методах интеллектуального анализа данных.

Обобщение – процесс, при котором количество данных для интерпретации уменьшается без потери важной информации.

Описательная статистика – метод получения представления о «середине» и «распространении» данных с помощью мер центральной тенденции и изменчивости.

Регрессионная модель – метод описания взаимосвязи между переменными путем подгонки линии тренда к наблюдаемым данным.

Регрессионный анализ – это метод, который работает путем моделирования отношений между зависимой переменной и одной или несколькими независимыми переменными.

Статистический анализ – это метод выполнения нескольких статистических операций для количественной оценки данных.

Факторный анализ – это метод, который используется для преобразования большого количества переменных в меньшее количество факторов.

Оглавление

ВВЕДЕНИЕ 3

Глава 1. ОСНОВЫ АНАЛИЗА ДАННЫХ 6
Тема 1. Анализ данных: понятие, виды, способы реализации и сферы применения 6
Тема 2. Генеральная и выборочная совокупности, их значение в анализе данных 15
Тема 3. Описательная статистика и показатели изменчивости вариации 25
Контрольные вопросы 44
Тесты для самоконтроля 45
Практическое задание 52

Глава 2. ВИДЫ АНАЛИЗА ДАННЫХ 57
Тема 4. Понятие и процедура корреляционного анализа 57
Тема 5. Понятие и структурные элементы регрессионного анализа 67
Тема 6. Особенности и преимущества факторного анализа 72
Тема 7. Понятие, сфера применения и алгоритм проведения кластерного анализа 80
Тема 8. Методы проверки гипотез о взаимосвязи переменных 90
Контрольные вопросы 96
Тесты для самоконтроля 97
Практическое задание 104

Глава 3. ОБЩЕЕ ПРЕДСТАВЛЕНИЕ О ВИЗУАЛИЗАЦИИ ДАННЫХ 108
Тема 9. Визуализация данных и ее характеристики 108
Тема 10. Визуализация данных и визуализация информации 114
Тема 11. Визуализация данных в понимании и передаче аналитических данных 118
Тема 12. Наиболее распространенные типы визуализации данных 122

Тема 13. Инструменты визуализации данных 132
Контрольные вопросы 136
Тесты для самоконтроля 137
Практическое задание 144

Глава 4. ПРОЕКТ АНАЛИЗА И ВИЗУАЛИЗАЦИИ ДАННЫХ 149
Тема 14. Этапы анализа данных для их визуализации 149
Тема 15. Визуализация данных в R 165
Тема 16. Визуализация данных в Python 177
Контрольные вопросы 185
Тесты для самоконтроля 186
Практическое задание 192

ОТВЕТЫ НА ТЕСТЫ 195

ЗАКЛЮЧЕНИЕ 199

БИБЛИОГРАФИЧЕСКИЙ СПИСОК 200

ГЛОССАРИЙ 201

Учебное издание

Гущина Оксана Михайловна,
Аникина Оксана Владимировна,
Желнина Евгения Валерьевна

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Учебно-методическое пособие

В оформлении издания использованы иллюстрации с сайта freepik.com

Редактор *Е.В. Пилясова*
Технический редактор *Н.П. Крюкова*
Компьютерная верстка: *Л.В. Сызганцева*
Дизайн обложки: *Е.В. Веселова*

Подписано в печать 09.07.2025. Формат 60×84/16.
Печать оперативная. Усл. п. л. 11,97.
Тираж 100 экз. Заказ № 1-66-22.

Издательство Тольяттинского государственного университета
445020, г. Тольятти, ул. Белорусская, 14,
тел. 8 (8482) 44-91-47, www.tltsu.ru