МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра «Прикладная математика и информатика»
(наименование)

01.03.02 Прикладная математика и информатика
(код и наименование направления подготовки / специальности)

Компьютерные технологии и математическое моделирование
(направленность (профиль) /специализация)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(БАКАЛАВРСКАЯ РАБОТА)

на тему «Моделирование системы HR-аналитики на основе анализа данных сайтов вакансий»

Обучающийся В.С. Курочкин _____
(Инициалы Фамилия) (личная подпись)

Руководитель канд.пед.наук, доцент, О.М. Гущина
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Консультант канд.пед.наук, доцент, Т.С. Якушева
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2022

Аннотация

Темой данной бакалаврской работы является «Моделирование системы HR-аналитики на основе анализа данных сайтов вакансий».

Работа выполнена студентом Тольяттинского государственного университета, института математики, физики и информационных технологий, группы ПМИб-1802а, Курочкиным Владиславом Сергеевичем.

Объект исследования: анализ статистических совокупностей.

Предмет исследования: система HR – аналитик.

Цель работы: моделирование системы HR-аналитики на основе анализа статистической совокупности данных сайта вакансий, опираясь на выбор статистической стратегии.

Для достижения цели работы необходимо решить следующие задачи:

– Рассмотреть основные понятия математической статистики.

– Научиться пользоваться выбором статистического критерия по алгоритму.

– Собрать данные и агрегировать их с помощью Python.

– Выдвинуть гипотезы и проанализировать данные с помощью Python.

Отчет состоит из введения, трех глав и заключения.

В первой главе представлены основные определения и методы статистического анализа.

Во второй главе описан алгоритм выбора статистического критерия, а также их описания.

В третьей главе проводится разработка и тестирование системы HR - аналитики.

Бакалаврская работа выполнена на 61 страницах, состоит из введения, трех разделов, заключения, списка литературы, состоящего из 25 литературных источников и 46 рисунков.

Annotation

The title of the graduation work is: «Modeling of HR analytics system based on data analysis of job sites».

The graduation work consists of an introduction, three sections, a conclusion, one table, list of references, including foreign sources and 46 pictures.

The key issue of the thesis is the design of an HR analytics system based on the analysis of data from job sites, the method of analyzing statistical aggregates. We touch upon the problem of the external direction of HR analytics and model methods for analyzing the labor market.

The aim of the work is to model the HR analytics system based on the analysis of the statistical aggregate of the job site data, based on the choice of a statistical strategy.

The graduation work may be divided into several logically connected parts which are the following: descriptions of the main definitions and methods of statistical analysis; analysis of existing statistical strategies; algorithm for choosing a statistical strategy; modeling algorithms for analyzing statistical aggregates.

In conclusion, I would like to emphasize that this work is relevant for identifying the relationship between factors and the response in the labor market, the algorithms developed during the study are relevant both for a certain region and for all areas of the Russian Federation.

Содержание

Введение

HR-аналитика - относительно новый термин, который впервые появился в академической литературе в 2004 году. Это систематическая идентификация и количественная оценка факторов бизнес-результатов. Аналитика определяется как пересечение информатики, принятия решений и количественных методов для организации, анализа и объяснения растущего объема данных, генерируемых современным обществом.

HR-аналитика, также называемая аналитикой людей, рабочей силы или талантов, включает в себя сбор, анализ и отчетность кадровых данных. Это позволяет вашей организации измерять влияние ряда показателей на общую эффективность бизнеса и принимать решения на основе данных.

Существует два направления HR-аналитики:

– Внутренняя – управление процессами и персоналом для достижение максимального эффекта

– Внешняя – понимание общих аспектов рынка труда

На сегодняшний день именно внутреннее направление проходит скоростной путь эволюции, пренебрегая внешней аналитикой из-за чего компании сталкиваются с проблемой получения недостаточного количества данных, необходимых для анализа соотношения предложений и спроса на рынке труда, уровня зарплат, конкуренции и общей демографической ситуации.

Цель работы: моделирование системы HR-аналитики на основе анализа статистической совокупности данных сайта вакансий, опираясь на выбор статистической стратегии.

Объект исследования: анализ статистических совокупностей.

Предмет исследования: система HR – аналитик.

Для достижения цели работы необходимо решить следующие задачи:

– Рассмотреть основные понятия математической статистики.

– Научиться пользоваться выбором статистического критерия по

алгоритму.

– Собрать данные и агрегировать их с помощью Python.

– Выдвинуть гипотезы и проанализировать данные с помощью Python.

Данная работа содержит в себе введение, три раздела, заключение и список используемой литературы.

В первой главе представлены основные определения и методы статистического анализа.

Во второй главе описан алгоритм выбора статистического критерия, а также их описания.

В третьей главе проводится разработка и тестирование системы HR-аналитики.

1 Основные понятия статистического анализа

1.1 Описательная статистика

Статистика [17] — это дисциплина математики, которая занимается анализом данных и числами. Изучение сбора, анализа, интерпретации, представления и организации данных известно как статистика.

Статистика стала универсальным языком науки, а анализ данных может привести к важным результатам. Как ученые, исследователи и менеджеры, работающие в секторе природных ресурсов, мы все полагаемся на статистический анализ, который помогает нам ответить на вопросы, возникающие у населения, которым мы управляем. Например:

– Увеличилось ли количество конкурентноспособных предприятий, обнаруженных в городе N?

– Произошло ли существенное изменение в экономике за нынешний год?

– Какая доля безработных людей?

Это типичные вопросы, ответы на которые требуют статистического анализа. Чтобы ответить на эти дилеммы, необходимо собрать хорошую случайную выборку из интересующей нас совокупности. Затем мы используем описательную статистику для организации и обобщения наших выборочных данных. Следующим шагом является статистика вывода, которая позволяет нам использовать нашу выборочную статистику и распространять результаты на популяцию, измеряя при этом надежность результата. Но прежде, чем мы приступим к рассмотрению статистических методов и критериев статистического анализа, необходимо сделать краткий обзор описательной статистики.

Выборка (выборочная совокупность) [5] — это метод отбора отдельных членов или подмножества для получения статистических выводов из них и оценки характеристик.

Выборочное среднее арифметическое (mean) — это среднее значение набора данных. Среднее значение выборки можно использовать для расчета центральной тенденции, стандартного отклонения и дисперсии набора данных. Выборочное среднее может применяться для различных целей, включая расчет средних значений генеральной совокупности. Рассчитывается по формуле (1).

$$\hat{x} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad (1)$$

где

N – объем выборки;

x_i – i-й элемент выборки;

$\hat{x}$ – выборочное среднее.

Медиана (median) [14] — это среднее число в отсортированном, восходящем или нисходящем списке чисел, и оно может быть более информативным для этого набора данных, чем среднее значение.

Размах — это разброс данных от самого низкого до самого высокого значения в распределении, вычисляется путем вычитания наименьшего значения из наибольшего.

Формула для расчета размаха (2) [13]:

$$R = x_{max} - x_{min}\ , \quad (2)$$

где

R – размах;

x_{max} – максимальное значение;

x_{min} – минимальное значение.

Размах является самой простой мерой изменчивости для расчета. Чтобы найти размах, требуются следующие действия:

– Упорядочить все значения в вашем наборе данных от меньшего к большему.

– Вычесть наименьшее значение из наибольшего значения.

Этот процесс одинаков независимо от того, являются ли значения положительными или отрицательными, целыми числами или дробями.

Выборочная дисперсия (variance) [2] - ожидание квадрата разности точек данных от среднего значения набора данных. Это абсолютная мера дисперсии, которая используется для проверки отклонения точек данных по отношению к среднему значению данных, она вычисляется по формуле (3).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N-1}, \quad (3)$$

где

σ^2 – выборочная дисперсия;

x_i – i-й элемент выборки;

$\hat{x}$ – выборочное среднее;

N – количество значений в выборке.

Для выборок мы используем $N - 1$ в формуле, потому что использование N дало бы нам смещенную оценку, которая последовательно занижает изменчивость.

Выборочное стандартное отклонение (Std. Deviation) - среднеквадратичное различие между наблюдениями и средним значением выборки, рассчитывается по формуле (4), указанной ниже:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N-1}}, \quad (4)$$

где

σ – выборочное стандартное отклонение;

σ^2 – выборочная дисперсия;

x_i – i-й элемент выборки;

$\hat{x}$ – выборочное среднее;

N – количество значений в выборке.

Математическое ожидание, также известное как ожидаемое значение - суммирование или интегрирование возможных значений случайной величины [18].

Например, можно рассчитать ожидаемую стоимость инвестиции в определённый момент в будущем. Рассчитывая математическое ожидание перед тем, как инвестировать, можно выбрать наилучший сценарий, который по мнению инвестора, даст наилучший результат.

Случайная величина делится на несколько типов:

– Дискретной: число возможных значений X — это числимое конечное или бесконечное множество точек

– Непрерывной: X может принимать любое значение в заданном диапазоне

Математическое ожидание случайного дискретного выражения рассчитывается по данной формуле (5):

$$\mu = \sum_{i=1}^{n} x_i \times p_i \,, \tag{5}$$

где

μ – математическое ожидание;

x – случайная величина;

p – вероятность появления случайной величины.

Для непрерывной случайной величины X вычисляется следующим образом (6):

$$\mu = \int_{-\infty}^{\infty} xf(x)dx \,, \tag{6}$$

где $f(x)$ - плотность распределения случайной величины X.

В данном пункте были рассмотрены основные понятия описательной статистики.

1.2 Типы данных в описательной статистике

Данные играют важную роль в области науки о данных. Необработанные данные подвергаются масштабным экспериментам для получения осмысленной информации, которая помогает достичь многих бизнес-целей. Понимание различных типов данных в статистике, приведенных на рисунке 1, сделает вас на один шаг ближе к выбору типа данных, соответствующего вашим бизнес-требованиям. Знание типов данных поможет точно применять статистические измерения к необработанным данным и делать важные выводы. Перед тем как переходить к определению типов данных, мы должны их собрать и только после этого определять их тип.

Рисунок 1 - Типы данных в статистике

Категориальные данные представляют характеристики. Их также называют качественными данными, это означает, что мы не можем выразить их числовым значением, а следовательно, измерить. Он включает в себя такие переменные, как слова, символы, изображения, которые помогают

сортировать информацию по категориям, например, по местам отдыха, полу, языку и т. д.

Качественный тип данных включает в себя:

– Номинальный тип данных – переменные, которые не имеют естественного порядка. В качестве примера номинального типа данных можно привести пол, цвет, семейное положение.

– Порядковый означает то, что находится в порядке. Порядковый тип данных включает переменные, которые следуют естественному порядку. В качестве примера порядкового типа данных можно привести рейтинг, время суток.

Числовые данные (количественные) — это один из самых простых типов данных для понимания. Как следует из названия, он представляет числовое значение и помогает ответить на такие вопросы, как сколько, сколько, как долго и так далее.

Он пытается количественно определить элементы, измеряя числовые переменные, которые заставляют их учитываться в природе. Ключевым моментом здесь является то, что числовая переменная может принимать бесконечное количество значений.

Например, рост человека может варьироваться от *x* см до *y* см и может быть дополнительно разбит на дробные значения.

Количественные данные классифицируются на две следующие категории:

– Дискретный тип данных включает только целые числа или дискретные значения. Он содержит конечное число значений, которые нельзя разделить на более мелкие части. Он учитывает только те элементы, которые мы не можем измерить, а только сосчитать. В качестве примера дискретного типа данных можно привести количество сотрудников в организации.

– Непрерывный тип данных — это данные, которые вы можете измерить. Он содержит бесконечное количество значений, которые можно разделить на более мелкие части. Он учитывает только те элементы, которые

вы не можете посчитать, а только измерить. В качестве примера дискретного типа данных можно привести время, затраченное на обучение.

Тип данных в описательной статистике является одним из первых шагов выбора статистического критерия, несомненно очень важено понимать и определять типы данных.

1.3 Анализ нормальности распределения данных

Нормальное распределение [4], [7] часто используется для приближенного описания случайных явлений, в которых на интересующий нас признак оказывает воздействие большое количество независимых и случайных факторов, среди которых нет резко выделяющихся.

В математической статистике исключительно большую роль играет нормальное распределение (открыто Муавром в 1733 г. и затем детально изучалось Лапласом и Гауссом). В честь Гаусса нормальное распределение часто называют гауссовским. Оно является непрерывным распределением с плотностью вероятности, вычислить которую можно по формуле (7).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (7)$$

где μ и σ^2 – параметры распределения.

Определим математическое ожидание M(x) и дисперсию D(x) случайной величины, распределенной по закону Гаусса, указанную в формуле (8):

$$M(x) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (8)$$

Произведем замену переменных (9), (10), (11) и получим уравнение 12:

$$\frac{x-\mu}{\sigma} = t \tag{9}$$

$$x = \sigma t + \mu \tag{10}$$

$$dx = \sigma dt \tag{11}$$

$$M(x) = \int_{-\infty}^{\infty} \frac{\sigma t}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{12}$$

Первый член уравнения (12) равен нулю, а интеграл второго – единице. Тогда математическое ожидание будет равно первому параметру нормального закона распределения: $M(x) = \mu$.

Согласно определению, дисперсия есть центральный момент второго порядка (13):

$$D(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{13}$$

Заменяя $\frac{x-\mu}{\sigma} = t$, получим следующие выражение (14):

$$D(x) = \sigma^2 \int_{-\infty}^{\infty} \frac{t^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{14}$$

На рисунке 2 приведены кривые плотности нормального распределения случайной величины с различными значениями σ. Видно, что дифференциальная кривая нормального распределения (7) имеет колоколообразную форму.

К косвенным методам нормальности распределения относится оценка коэффициента асимметрии и оценка эксцесса.

Коэффициент асимметрии (Skewness) [15] — это один из способов измерения асимметрии распределения. Асимметрию можно определить как меру асимметрии распределения вероятностей. Если кривая нормального распределения искривлена влево или вправо, то такое распределение называется асимметричным.

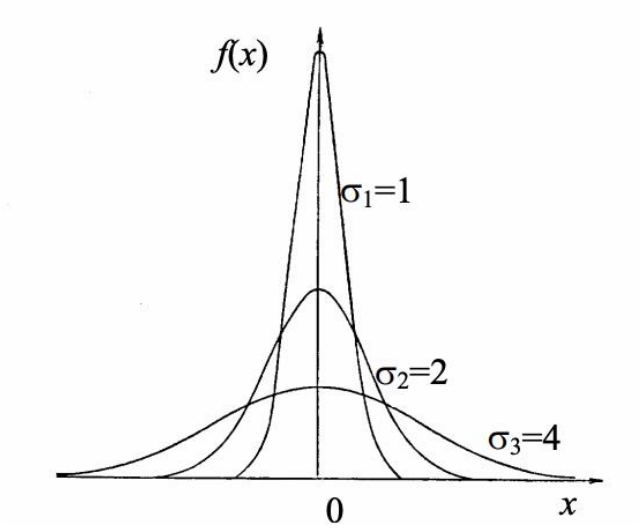


Рисунок 2 – Кривая нормального распределения с параметрами $\mu = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_3 = 4$

Важнейшей мерой асимметрии является коэффициент асимметрии, данный Карлом Пирсоном. Он также известен как коэффициент асимметрии Пирсона.

Коэффициент асимметрии, представленный на рисунке 3 можно определить как меру, которая используется для определения силы и направления асимметрии выборочного распределения с использованием описательной статистики, такой как среднее значение, медиана или мода [12]. Коэффициент асимметрии используется для сравнения выборочного распределения с нормальным. Если значение очень велико, это означает, что существует большая разница между выборочным распределением по сравнению с нормальным распределением.

В зависимости от значения коэффициента асимметрии можно сделать следующие выводы о распределении:

– Если среднее значение превышает моду и медиану, то распределение имеет положительную асимметрию. Другими словами, если коэффициент асимметрии положительный, то распределение скошено

вправо.

– Если мода превышает медиану и среднее значение, то распределение имеет отрицательную асимметрию. Таким образом, коэффициент асимметрии будет отрицательным и распределение будет скошено влево.

– Если значения среднего, медианы и моды равны, то распределение является нормальным и коэффициент асимметрии будет равен 0.

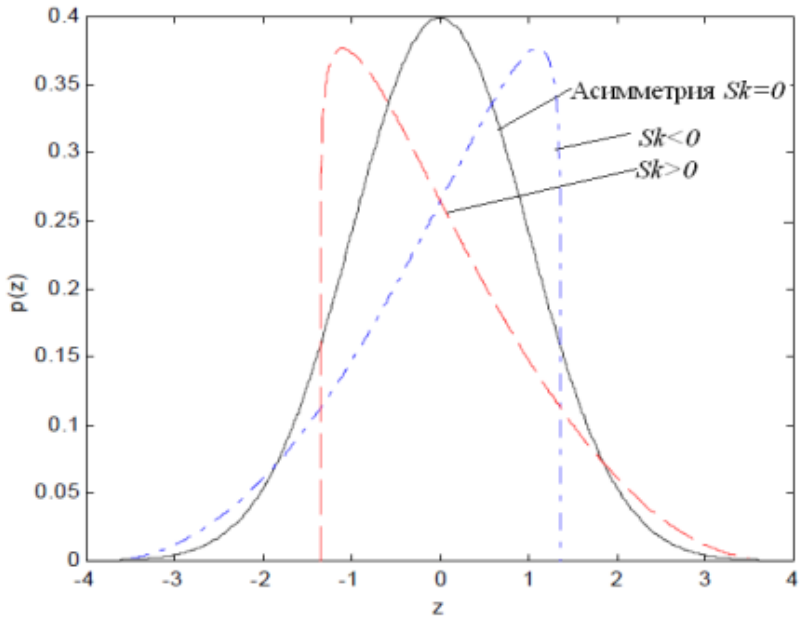


Рисунок 3 - Коэффициент асимметрии

Формула асимметрии Пирсона (15):

$$Sk = \frac{\bar{X} - Mo}{\sigma}, \quad (15)$$

где σ – среднее квадратическое отклонение статистической совокупности.

Рассмотрим следующий немаловажный метод нормальности распределения – Эксцесс [8]. Это мера совокупного веса хвостов распределения по отношению к центру распределения. Когда набор приблизительно нормальных данных отображается с помощью гистограммы, он показывает пик колокола и большинство данных в пределах трех стандартных отклонений (плюс или минус) от среднего значения. Однако, когда присутствует высокий эксцесс, хвосты простираются дальше, чем три стандартных отклонения нормального колоколообразного распределения.

Эксцесс иногда путают с мерой пикообразности распределения. Однако эксцесс — это мера, описывающая форму хвостов распределения по отношению к его общей форме. Распределение может иметь бесконечный пик с небольшим эксцессом, а распределение может быть идеально плоским с бесконечным эксцессом. Таким образом, эксцесс измеряет «хвостость», а не «заостренность».

Есть три категории эксцесса, которые могут отображаться набором данных. Все показатели эксцесса сравниваются со стандартным нормальным распределением или кривой нормального распределения. Рассмотрим каждый из них:

– Данные, следующие за мезокуртическим распределением, показывают избыточный эксцесс, равный нулю или близкий к нулю. Это означает что, если данные следуют нормальному распределению, они следуют мезокуртическому распределению. Пример мезокуртского распределение рисунок 4.

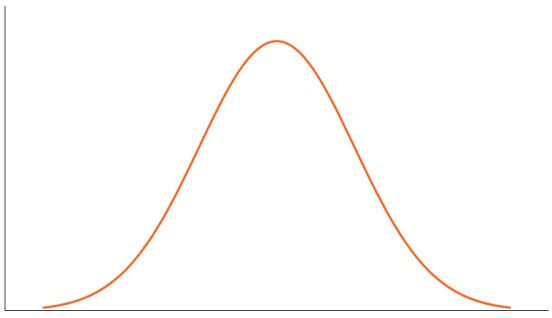


Рисунок 4 – Мезокуртическое распределение

– Лептокуртик приведенное на рисунке 5, указывает на положительный избыточный эксцесс. Лептокуртическое распределение показывает тяжелые хвосты с обеих сторон, что указывает на большие выбросы. В финансах лептокуртическое распределение показывает, что доход от инвестиций может быть склонен к экстремальным значениям с обеих сторон. Следовательно, инвестиции, доходность которых соответствует лептокуртному распределению, считаются рискованными.

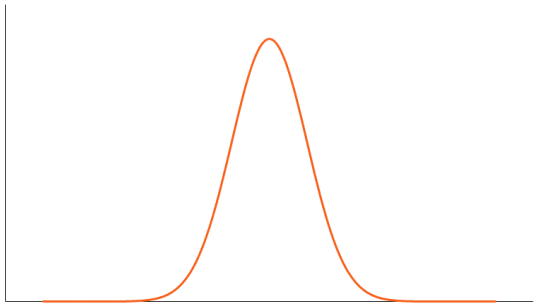


Рисунок 5 – Лептокуртическое распределение

– Платикуртическое распределение показывает отрицательный избыточный эксцесс, это отлично изображено на рисунке 6. Эксцесс показывает распределение с плоскими хвостами. Плоские хвосты указывают на небольшие выбросы в распределении. В финансовом контексте плоскокуртное распределение доходов от инвестиций желательно для инвесторов, потому что существует небольшая вероятность того, что инвестиции принесут экстремальную прибыль.

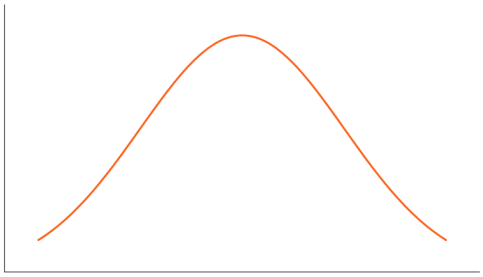


Рисунок 6 – Платикуртическое распределение

Стандартная ошибка асимметрии (16):

$$S_{sk} = \sqrt{\frac{6 \times N \times (N-1)}{(N-2) \times (N+1) \times (N+3)}} \quad (16)$$

Стандартная ошибка эксцесса (17):

$$S_k = \sqrt{\frac{4 \times (N^2-1) \times S_{sk}^2}{(N-3) \times (N-5)}} \quad (17)$$

Для того чтобы оценить является ли распределение нормальным или нет, нужно сравнить его значение асимметрии и эксцесса по модулю с их стандартными ошибками формула (16) и (17), если абсолютная величина асимметрии и эксцесса, меньше их стандартных ошибок в 3 и более раз, то распределение считают нормальным.

К расчетным методам анализа распределения данных относятся Критерий Колмогорова Смирнова и Шапиро – Уилка.

p (Significance – наблюдаемый уровень значимости) – вероятность допустить ошибку, утверждая, что фактор влияет на отклик.

Тест качества подгонки Колмогорова-Смирнова сравнивает данные с известным распределением и позволяет узнать, имеют ли они одинаковое распределение. Хотя тест является непараметрическим — он не предполагает какого-либо конкретного базового распределения — он обычно используется

в качестве теста на нормальность, чтобы увидеть, нормально ли распределены данные. Он также используется для проверки предположения о нормальности в дисперсионном анализе. Если $p < 0{,}2$ распределение ненормально.

Тест Шапиро-Уилка [16] — это способ определить, происходит ли случайная выборка из нормального распределения. Тест дает вам значение W; маленькие значения указывают на то, что ваша выборка не имеет нормального распределения (вы можете отклонить нулевую гипотезу о том, что ваша популяция имеет нормальное распределение, если ваши значения ниже определенного порога). Формула для значения W (18):

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (18)$$

где

x_i - упорядоченные значения случайной выборки;

a_i - константы, полученные из ковариаций, дисперсий и средних значений выборки (размера n) из нормально распределенной выборки.

Тест имеет систематическую ошибку по размеру выборки. Если $p < 0{,}5$, то распределение ненормально.

Вывод по разделу

Проверка нормальности распределения количественных данных является вторым важным шагом, на основе данного анализа мы определим нормально ли распределены данные, и на основе распределения выберем критерий статического анализа.

В данном разделе были рассмотрены основы статистического анализа данных, а именно:

– Описательная статистика и ее основные термины;

– Типы данных;

– Тесты для анализа колличественных шкал на нормальность распределения.

2 Анализ статистических совокупностей и выбор статического критерия

2.1 Алгоритм выбора статистического критерия для проверки гипотез

Проверка гипотез [10] — это формальная процедура исследования наших представлений о мире с использованием статистики. Чаще всего используется учеными для проверки конкретных предсказаний, называемых гипотезами, которые вытекают из теорий.

Проверка гипотез состоит из 4 основных этапов:

– Сформулировать свою исследовательскую гипотезу как нулевую гипотезу и альтернативную гипотезу (H_0) и (H_1);

– Собирать данные, чтобы проверить гипотезу;

– Выполнить соответствующий статистический тест;

– На основе полученных результатов решить, отвергать или не отвергать нулевую гипотезу.

Статистический критерий (статистические тесты) используются для проверки гипотез. Чтобы выбрать правильную статическую стратегию нужно выполнить следующий алгоритм (рисунок 7):

– Определить тип данных (количественный или качественный);

– Если тип данных является количественным определить нормальность распределения;

– Определить количество сравниваемых групп;

– Определить связанны ли сравниваемые группы между собой, то есть являются ли единицы наблюдения в группах разными носителями признака.

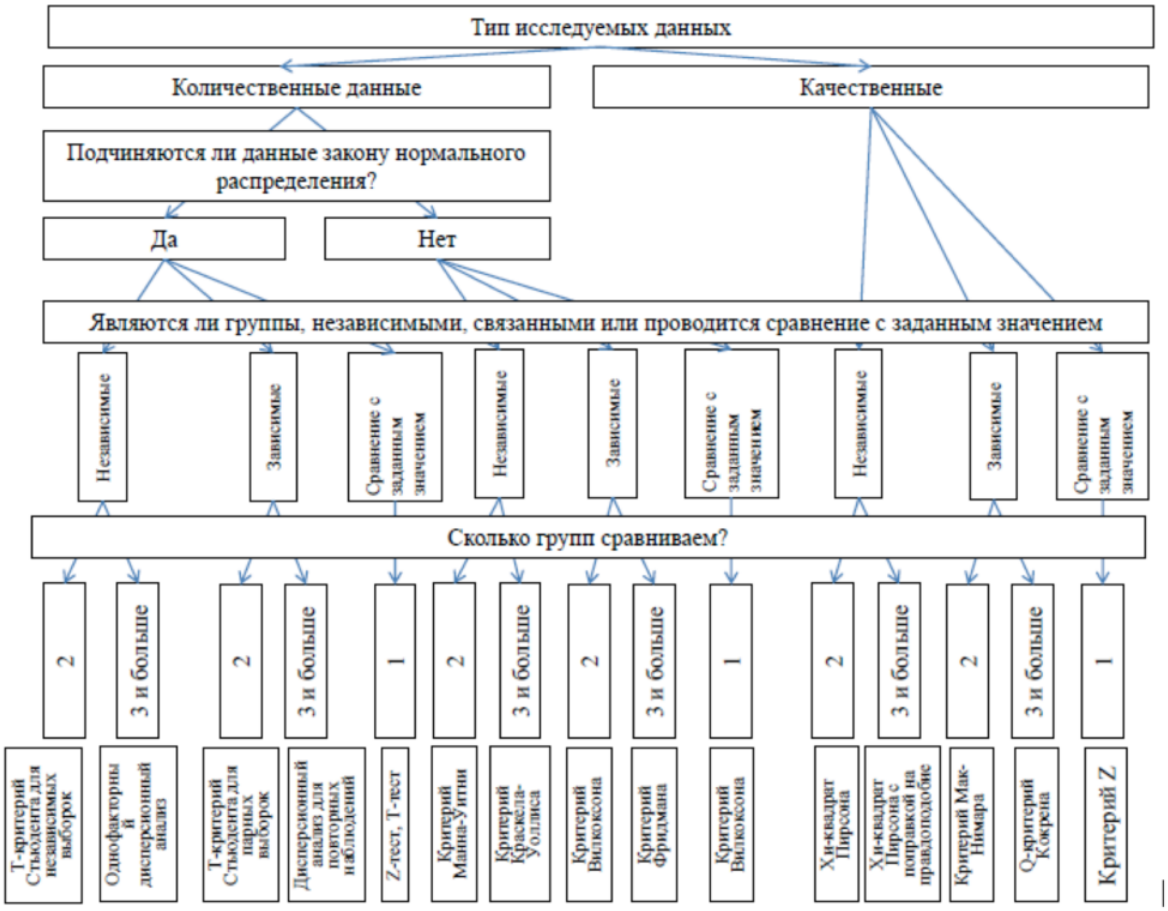


Рисунок 7 – Алгоритм выбора статистического критерия

После определения алгоритма рассмотрим существующие статистические стратегии.

2.2 Корреляция Пирсона, Спирмена

Коэффициенты корреляций позволяют ответить на вопросб, влияет ли колличественный фактор на колличественный отклик рисунок 8.

Рисунок 8 – Схема влияния коэффицикнта корреляции на данные

Коэффициент корреляции (r) – мера линейной связи между количественными фактором и откликом.

Коэффициент корреляции изменяется в пределах $-1 \leq r \leq 1$, чем ближе его абсолютное значение к 1, тем сильнее связь.

Для словестной интерпритации исполбзуются следующие диапозоны, которые нужно рассматривать по модулю:

– $r < 0{,}25$ – связь слабая;

– $0{,}25 \leq r < 0{,}75$ – связь умерянная;

– $r \geq 0{,}75$ – связь сильная;

– Если $r > 0$, то связь прямая;

– Если $r < 0$, то связь обратная.

Рассмотрим схему выбора корреляционного исследования указана на рисунке 9.

Рисунок 9 – Схема выбора корреляционного исследования.

Корреляция Пирсона является наиболее широко используемой статистикой корреляции для измерения степени взаимосвязи между линейно связанными переменными. Например, на фондовом рынке, если мы хотим измерить, как две акции связаны друг с другом, корреляция Пирсона используется для измерения степени взаимосвязи между ними. Точечно-двурядная корреляция проводится по формуле корреляции Пирсона, за исключением того, что одна из переменных является дихотомической. Формула (19) используется для расчета корреляции Пирсона r :

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i - \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \quad (19)$$

где

r_{xy} – коэффициент корреляции Пирсона r между x и y;

n – количество наблюдений;

x_i – значение x (для i-го наблюдения);

y_i – значение y (для i-го наблюдения).

Для корреляции Пирсона [9] обе переменные должны быть нормально распределены. Другие предположения включают линейность и гомоскедастичность . Линейность предполагает прямолинейную связь между каждой из двух переменных, а гомоскедастичность предполагает, что данные равномерно распределены по линии регрессии.

Ранговая корреляция Спирмена [1] — это непараметрический тест, который используется для измерения степени связи между двумя переменными. Критерий ранговой корреляции Спирмена не содержит никаких предположений о распределении данных и является подходящим корреляционным анализом, когда переменные измеряются по шкале не ниже порядковой.

Формула (20) используется для расчета ранговой корреляции Спирмена:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (20)$$

где

ρ – ранговая корреляция Спирмена;

n – количество наблюдений;

d_i – разница между рангами соответствующих переменных.

Предположения корреляции Спирмена заключаются в том, что данные должны быть как минимум порядковыми, а оценки одной переменной должны быть монотонно связаны с другой переменной.

В данном подразделе были рассмотрены корреляции Пирсона и Спирмена, их основные формулы и отличия. Также был определен коэффициент корреляции и его интерпритация диапозонов.

2.3 U-критерий Манна-Уитни для независимых выборок

Популярным непараметрическим тестом для сравнения результатов между двумя независимыми группами является U-критерий Манна-Уитни [20], который изображен на рисунке 10. U-критерий Манна-Уитни, иногда называемый критерием Манна-Уитни-Уилкоксона или критерием суммы рангов Уилкоксона, используется для проверки того, могут ли две выборки быть получены из одной и той же совокупности (т. е. две совокупности имеют одинаковую форму). Некоторые исследователи интерпретируют этот тест как сравнение медианы между двумя популяциями. Параметрический тест сравнивает средние значения ($H_0: \mu_1 = \mu_2$) между независимыми группами.

Рисунок 10 – Схема U-критерий Манна-Уитни

Нулевая и двусторонняя исследовательские гипотезы для непараметрического теста формулируются следующим образом:

H_0: Две популяции равны по сравнению с

H_1: Две популяции не равны

Этот тест часто выполняется как двусторонний тест, и, таким образом, исследовательская гипотеза указывает на то, что популяции не равны, в отличие от определения направленности. Односторонняя исследовательская гипотеза используется, если интерес заключается в обнаружении положительного или отрицательного сдвига в одной популяции по сравнению с другой. Процедура теста включает в себя объединение наблюдений из двух выборок в одну комбинированную выборку, отслеживание того, из какой выборки взято каждое наблюдение, а затем ранжирование от низшего к высшему от 1 до $n_1 + n_2$ соответственно.

Статистический показатель для U-теста Манна-Уитни обозначается U и является меньшим из U_1 назодится по формуле (21) и U_2 формула (22), определенных ниже:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 , \quad (21)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \quad (22)$$

где

R_1 - сумма рангов для группы 1;

R_2 - сумма рангов для группы 2.

В данном подразделе был рассмотрен U-критерий Манна-Уитни для независимых выборок, его основные поняти, какие параметрические значения сравнивает тест, для каких типов данных применяется.

2.4 Критерий Хи-квадрат Пирсона

Хи-квадрат Пирсона [22] — это непараметрический метод, который позволяет оценить значимость различий между фактическим (рисунок 11) количеством исходов или качественных характеристик выборки,

попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей).

Рисунок 11 – Схема критерия Хи-квадрат Пирсона

Ограничение применения критерия:

– Сопоставляемые показатели должны быть измерены в номинальной шкале

– Данный метод позволяет проводить анализ не только четырехпольных таблиц, когда и фактор, и исход являются бинарными переменными, то есть имеют только два возможных значения (например, мужской или женский пол)

– При анализе четырехпольных таблиц ожидаемые значения в каждой из ячеек должны быть не менее 10.

Рассчитываем ожидаемое количество наблюдений для каждой из ячеек таблицы сопряженности (при условии справедливости нулевой гипотезы об отсутствии взаимосвязи) путем перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений. Общий вид ожидаемых значений представлен в Таблице 1.

Таблица 1 – Расчет ожидаемого колличества наблюдений

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	$\frac{(A+B)\times(A+C)}{(A+B+C+D)}$	$\frac{(A+B)\times(B+D)}{(A+B+C+D)}$	$A+B$
Фактор риска отсутствует (0)	$\frac{(C+D)\times(A+C)}{(A+B+C+D)}$	$\frac{(C+D)\times(B+D)}{(A+B+C+D)}$	$C+D$

Всего	$A + C$	$B + D$	$A + B + C + D$

Находим значение критерия x^2 по следующей формуле (23):

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (23)$$

где

где i – номер строки (от 1 до r);

j – номер столбца (от 1 до c);

O_{ij} – фактическое количество наблюдений в ячейке ij;

E_{ij} – ожидаемое число наблюдений в ячейке ij.

В данном подразделе был рассмотрен Критерий Хи-квадрат Пирсона, его основные поняти, основные формулы для его нахождения, расчет ожидаемого количества наблюдений, а также его ограничения.

Вывод по разделу

В этом разделе был рассмотрен алгоритм выбора статистического критерия, виды статистических стратегий такие как:

– Корреляция Пирсона, Спирмена;

– U-критерий Манна-Уитни для независимых выборок;

– Критерий Хи-квадрат Пирсона.

А также их основные определения, расчет, формулы, влияние одних типов данных на другие.

3 Программная разработка и тестирование системы HR-аналитики

3.1 Обзор существующих инструментов и распределение данных для реализации модели

Для реализации системы будем использовать объектно-ориентированный язык программирования Python [3], [6], [24] версии 3.7 и интерактивную вычислительная среду Jupyter Notebook. Данный язык программирования был выбран из-за ряда следующих причин:

– Python — лучший и наиболее часто используемый язык для машинного обучения и науки о данных. Благодаря большому сообществу и большому количеству библиотек, Python очень помогает в разработке приложений в области науки о данных. В нем есть библиотеки для манипулирования данными, визуализации данных, очистки данных и ряда других связанных функций;

– Python — очень продуктивный язык. Благодаря простоте Python разработчики могут сосредоточиться на решении проблемы;

– Python не знает тип переменной, пока мы не запустим код. Он автоматически назначает тип данных во время выполнения. Не нужно беспокоиться об объявлении переменных и их типов данных;

– Отлично подходит для группировки и агрегации данных.

Исходя из всех вышеперечисленных плюсов, было принято использовать Python в качестве основного языка для моделирования системы HR-аналитики, а также для обработки, группировки и агрегирования данных, для этого будут использоваться следующие библиотеки:

Dask - это библиотека Python с открытым исходным кодом, которая позволяет работать с произвольно большими наборами данных и значительно увеличивает скорость вычислений;

NumPy [21] — это библиотека с открытым исходным кодом на Python, которая помогает в математических, числовых расчетах и вычислениях,

научном, инженерном программировании и программировании данных. NumPy — наиболее важная библиотека для выполнения математических и статистических операций. Отлично работает для многомерных массивов и умножения матриц.

Pandas [23], [25] - это библиотека Python с открытым исходным кодом, который наиболее широко используется для обработки данных/анализа данных и задач машинного обучения. Она построена поверх вышеупомянутой библиотеки под названием NumPy

Matplotlib [19] — это кроссплатформенная библиотека для визуализации данных и графического построения графиков для Python и его числового расширения NumPy. Таким образом, он предлагает жизнеспособную альтернативу MATLAB с открытым исходным кодом.

Выбрав подходящие инструменты, распределим данные по таблицам с помощью библиотек Pandas и Dask.

Таблица Workexpf содержит информацию об опыте работы соискателя, и имеет следующие поля:

- achievements - достижения на предыдущем месте работы (номинальный тип данных);
- date_from - дата первого трудоустройства (номинальный тип данных);
- date_to - дата окончания работы (номинальный тип данных);
- id_cv - идентификатор резюме, используется как внешний ключ для связи между таблицами (номинальный тип данных);
- job_title - название должности (номинальный тип данных).

В таблице представлено 58766 полей, пример данной таблицы представлен на рисунке 12.

	achievements	date_from	date_to	id_cv	job_title
58762	За период работы увеличен объем производства к...	2008-03-01	2017-03-01	8cc2102c-fd0d-11e7-a659-e37b4be0b9ed	Генеральный директор
58763	Неоднократно поощрялся высшим руководством, на...	1994-06-01	2017-08-01	ff48c2b0-2157-11e8-a5e0-037acc02728d	Начальник, заместитель начальника
58764	И меются публикации в российских изданиях	2012-07-01	2013-11-01	bbdbb110-f94b-11e7-b311-736ab11edb0c	Иммунолог
58765	Самостоятельно внедрила методы количественного...	2014-01-01	NaN	bbdbb110-f94b-11e7-b311-736ab11edb0c	Биолог
58766	NaN	2015-12-01	2016-12-01	67f7d339-2ca9-11e8-9855-0f468c90bfa7	Санитарный инструктор роты

Рисунок 12 – Таблица Workexpf

Таблица Invitations хранит приглашения на собеседование, отправленные работодателями, и имеет следующие поля:

– id_candidate - идентификатор кандидата (номинальный тип данных);
– date_creation - дата создания резюме (номинальный тип данных);
– id_hiring_organization - идентификатор нанимающей организации (номинальный тип данных);
– id_cv - идентификатор резюме (номинальный тип данных);
– id_vacancy - идентификатор вакансии (номинальный тип данных);
– id_invitation - идентификатор приглашения (номинальный тип данных);
– region_code - код региона в формате КЛАДР (номинальный тип данных).

В таблице представлено 212881 полей, пример данной таблицы представлен на рисунке 13.

	id_candidate	date_creation	id_hiring_organization	id_cv	id_vacancy	id_invitation	region_code
212877	9b9383b0-83a4-11e6-8fec-736ab11edb0c	2016-11-24	1097536008382	0d3cbcd7-85e8-11e6-8fec-736ab11edb0c	aeeb5df2-1c54-11e6-82b1-d1490b16fa29	ed381190-b23e-11e6-ab05-037acc02728d	7500000000000
212878	ab9b6f00-2af6-11e5-85ad-9370902ea73e	2018-11-21	1031400355073	0731b6bc-2af9-11e5-85ad-9370902ea73e	a5ee8d82-102c-11e8-bf8f-bf2cfe8c828d	682bf1a0-ed7f-11e8-a68c-e37b4be0b9ed	1400000000000
212879	9caa3d90-f373-11e6-a86f-037acc02728d	2017-06-18	1147746084133	a959fa67-3ef7-11e7-9f39-4376a32b3f45	601a7d92-3bf3-11e7-9247-bf2cfe8c828d	836ae4b0-5435-11e7-b7b8-037acc02728d	7700000000000
212880	a2fe7380-6bca-11e7-a400-037acc02728d	2017-08-09	1115658007728	7b281979-6bce-11e7-bb9b-4376a32b3f45	51fc7b61-7982-11e7-aa2e-bf2cfe8c828d	ce8d1140-7cf6-11e7-b45a-ef76bd2a03c1	5600000000000
212881	c742f690-4e0c-11ea-a184-9122a281f90e	2020-04-03	1185029013663	15706ea5-4e0e-11ea-a184-9122a281f90e	49ae3e32-4987-11ea-af61-bf2cfe8c828d	6275e5d0-753d-11ea-a201-0f468c90bfa7	5000000000000

Рисунок 13 – Таблица Invitations

Таблица Citizens – аналитические данные по субъектам Российской Федерации. Содержит следующие поля:

– region_code - код региона в формате КЛАДР (номинальный тип данных);
– region_name – название субъекта (номинальный тип данных);
– cvs_count - количество резюме по субъектам (номинальный тип данных);
– medium_salary - средняя заработная плата (номинальный тип данных).

В таблице представлено 89 полей, пример данной таблицы представлен на рисунке 14.

	region_code	region_name	cvs_count	medium_salary
85	8700000000000	Чукотский автономный округ	242	56174.65
86	8900000000000	Ямало-Ненецкий автономный округ	21171	46049.60
87	9100000000000	Республика Крым	95178	27194.39
88	9200000000000	г. Севастополь	9145	37272.06
89	9900000000000	г. Байконур	289	37425.67

Рисунок 14 – Таблица Citizens

Таблица Edu хранит информацию об образовании, и имеет следующие поля:

– id_cv - идентификатор резюме (номинальный тип данных);
– faculty – название специальности (номинальный тип данных);
– graduate_year – год окончания обучения (номинальный тип данных);
– legal_name – название учебного заведения (номинальный тип данных).

В таблице представлено 152046 полей, пример данной таблицы представлен на рисунке 15.

	faculty	graduate_year	id_cv	legal_name
152042	NaN	2018.0	c2249810-4150-11ec-8e6c-1b29d3b53cbb	«Алтайский государственный технический универс...
152043	Экономики и управления в пищевой отрасли	2017.0	d3872c30-097e-11e7-b313-4376a32b3f45	«Московский государственный университет пищевы...
152044	NaN	1986.0	931aba80-42e1-11ec-a63f-1b29d3b53cbb	ШИХИКЕНТСКАЯ СОШ
152045	NaN	NaN	d9224ff0-4397-11ec-85d8-839f0d9a4379	NaN
152046	Аивт	2021.0	b8ae53b0-e3aa-11eb-98be-ab5d2eb93a75	Курганский технологический колледж

Рисунок 15 – Таблица Edu

Таблица Stat_Companies - количестве компаний по субъектам Российской Федерации. Имеет следующие поля:

- region_code – код региона в формате КЛАДР (номинальный тип данных);
- region_name – название субъекта (номинальный тип данных);
- company_count – всего компаний (количественная шкала);
- micro_company – компании с численностью менее 50 сотрудников (количественная шкала);
- small_company – компании с численность от 51 до 100 сотрудников (количественная шкала);
- midle_company – компании с численностью от 101 до 250 сотрудников (количественная шкала);
- big_company – компании с численностью от 251 до 500 сотрудников (количественная шкала);
- large_company – компании с численностью более 500 сотрудников (количественная шкала).

В таблице представлено 86 полей, пример данной таблицы представлен на рисунке 16.

	region_code	region_name	company_count	micro_company	small_company	midle_company	big_company	large_company
82	8700000000000	Чукотский автономный округ	388	6	322	8	8	44
83	8900000000000	Ямало-Ненецкий автономный округ	4268	142	3261	143	257	465
84	9100000000000	Республика Крым	10021	483	8312	371	342	513
85	9200000000000	г. Севастополь	1528	71	1247	44	43	123
86	9900000000000	г. Байконур	107	6	89	5	1	6

Рисунок 16 – Таблица Stat_Companies

Таблица Adedu содержит информацию о дополнительном образовании соискателя, и имеет следующие поля:

– course_name – наименование дополнительного образования (номинальный тип данных);
– description – описание обучения (номинальный тип данных);
– graduate_year – год окончания обучения (номинальный тип данных);
– id_cv – идентификатор резюме (номинальный тип данных).

В таблице представлено 173819 полей, пример данной таблицы представлен на рисунке 17.

	course_name	description	id_cv	graduate_year
173815	Пользователь ПК, Изучение бухгалтерских програ...	NaN	f855b1f0-4144-11ec-8e6c-1b29d3b53cbb	2002.0
173816	Секретарь делопроизводитель	NaN	c61135d0-45d5-11ec-b2a7-1b29d3b53cbb	2009.0
173817	"Инвестиционный менеджмент", 144 а.ч.	NaN	7920c880-226e-11ea-a17c-ef76bd2a03c1	2021.0
173818	1с: Зарплата. Управление персоналом	NaN	c2249810-4150-11ec-8e6c-1b29d3b53cbb	2019.0
173819	Вопросы профилактики, диагностики и лечения ко...	NaN	6c9892b0-42cd-11ec-85d8-839f0d9a4379	2021.0

Рисунок 17 – Таблица Adedu

Таблица Organizations хранит информацию об организациях работодателей, и имеет следующие поля:

– id_organization – идентификатор работодателя (номинальный тип данных);
– business_size – размер организации работодателя (номинальный тип данных);
– date_creation – дата создания сущности организации на сайте (номинальный тип данных);
– hr_agency – имеется ли кадровое агентство HR в компании (бинарный тип данных);
– inner_info_deleted – удалена ли компания на сегодняшний момент с портала (бинарный тип данных).

В таблице представлено 81140 полей, пример данной таблицы представлен на рисунке 18.

	id_organization	business_size	date_creation	hr_agency	inner_info_deleted
81136	309741134400014	SMALL	2017-05-24	NaN	f
81137	1152443000456	SMALL	2016-03-16	NaN	f
81138	1054900270105	MICRO	2015-08-29	NaN	f
81139	1053819020133	SMALL	2016-03-30	NaN	f
81140	1162375032346	SMALL	2019-07-18	NaN	f

Рисунок 18 – Таблица Orgsnizations

Таблица Curricula_Vitae хранит резюме по регионам, в нее входят следующие поля:

- birthday – дата рождения указанное в резюме (номинальный тип данных);
- business_trips – готовность к командировкам (бинарный тип данных);
- busy_type – тип занятости соискателя указанное в резюме (бинарный тип данных);
- country – гражданство соискателя (номинальный тип данных);
- date_creation – дата создания резюме (бинарный тип данных);
- experience – опыт работы, указанный в резюме (количественный тип данных);
- gender – пол соискателя (номинальный тип данных);
- id_candidate – идентификатор соискателя (номинальный тип данных);
- id_cv - идентификатор резюме (номинальный тип данных).
- retraining_capability – готовность соискателя к переобучению (бинарный тип данных);

– salary – зарплата, желаемая соискателям, указанная в резюме.

В таблице представлено 9070 полей, пример данной таблицы представлен на рисунке 19.

	birthday	business_trips	busy_type	country	date_creation	experience	gender	id_candidate	id_cv	retraining_capability	salary
9066	NaN	0.0	Полная занятость	Российская Федерация	2017-03-25	5.0	NaN	19aff750-115e-11e7-a63e-037acc02728d	0566c1d0-116c-11e7-a63e-037acc02728d	1.0	12000
9067	1989.0	0.0	Полная занятость	Российская Федерация	2020-08-06	0.0	Женский	f3e2fe80-a3fb-11ea-b2aa-7bf9d8e248ac	54433070-d7ea-11ea-be51-ab5d2eb93a75	NaN	30000
9068	NaN	0.0	Полная занятость	Российская Федерация	2017-01-09	5.0	NaN	ae044f50-d64e-11e6-98e3-736ab11edb0c	297ff4ab-d653-11e6-98e3-736ab11edb0c	1.0	20000
9069	NaN	0.0	Полная занятость	Российская Федерация	2016-12-01	0.0	NaN	d83beeb0-b74d-11e6-9cf5-736ab11edb0c	320c11d7-b74f-11e6-9cf5-736ab11edb0c	1.0	12000
9070	NaN	0.0	Полная занятость	Российская Федерация	2018-05-16	7.0	NaN	ef2f2820-5752-11e8-ae23-ef76bd2a03c1	728a8ae0-5909-11e8-b616-736ab11edb0c	1.0	26000

Рисунок 19 – Таблица Curricula_Vitae

После того, как мы собрали все данные и определились с основными инструментами перейдем к анализу данных.

3.2 Моделирование функции разреза по вакансиям

Перед тем как переходить к описательной статистике создадим функцию разреза по определенной вакансии, чтобы рассмотреть, как менялась ее актуальность с течением времени. Блок – схема алгоритма показана на рисунке 20, реализация на Python рисунок 21.

Алгоритм работы:

– Принимает наименование профессии (name);

– Фильтрует таблицу Workexp по заданному значению и присваивает данный DataFrame переменной bigdata1;

– Переменная bigdata2 принимает значения таблицы Curricula_Vitae и использует поля id_cv и date_creation;

– Используется merge двух DataFrame (bigdata1 и bigdata2);

– Drop поля id_cv, так как данная колонка более не нужна;

– Объединяем и сортируем DataFrame по времени;

– Возвращаем полученный DataFrame bigdata.

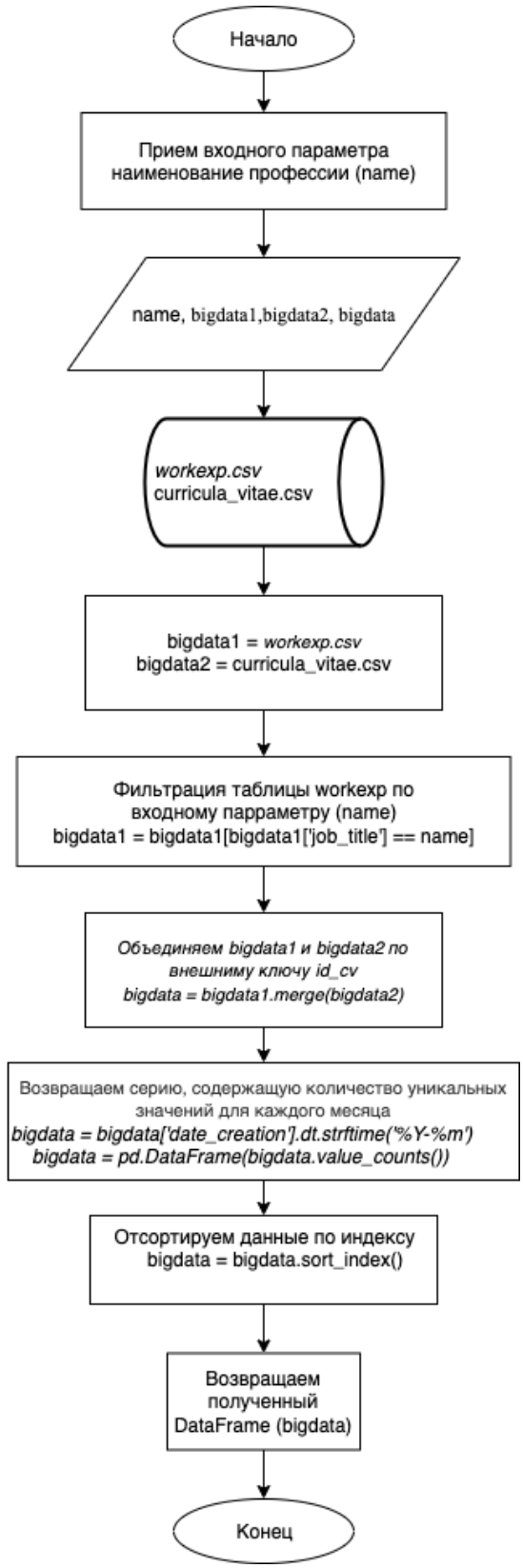


Рисунок 20 – Блок – схема алгоритма

```
def applicationsByYear(name):
    bigdata1 = pd.read_csv('/Users/vladkurockin/Desktop/Diplom/my.csv/workexp_split/workexp.csv',
                           sep = ';',
                          usecols = ['job_title','id_cv'])
    bigdata1 = bigdata1[bigdata1['job_title'] == name]
    bigdata1 = bigdata1.reset_index(drop = True)
    bigdata2 = pd.read_csv('/Users/vladkurockin/Desktop/Diplom/data/cur_vi_dc_idcv.csv',
                           sep = ',')
    bigdata = bigdata1.merge(bigdata2)
    bigdata = bigdata.drop(['id_cv'],axis=1)
    bigdata['date_creation'] = pd.to_datetime(bigdata['date_creation'], errors='coerce')
    bigdata = bigdata['date_creation'].dt.strftime('%Y-%m')
    bigdata = pd.DataFrame(bigdata.value_counts())
    bigdata.columns = ['N']
    bigdata = bigdata.sort_index()
    return bigdata
```

Рисунок 21 – Разрез по вакансиям на рынке труда по годам

Создадим функцию для построения графика на основе полученного DataFrame (рисунок 22).

```
def applicationsByYearGraph(data):
    plt.axes([0,0,2,1])
    plt.xticks(rotation=90, fontsize=11)
    for i in range(0,len(data)):
        plt.annotate(data['N'].iloc[i], xy = (data.index[i],data['N'].iloc[i]))
    plt.grid()
    plt.plot(data.index, data['N'])
    plt.xlabel('Месяцы')
    plt.ylabel('Колличество заявок')
    plt.title('Динамика заявок по годам', fontsize=20)
    return plt.show()
```

Рисунок 22 – Визуализация разреза по вакансиям на рынке труда по годам

Протестируем наш алгоритм на вакансии инженер. Произведем вызов функции applicationsByYear('Инженер'), посмотрим на результат (рисунок 23).

	N
2016-10	2306
2016-11	1368
2016-12	1000
2017-01	1510
2017-02	901
2017-03	1196
2017-04	806
2017-05	881

Рисунок 23 – Выборка вакансий инженер по годам

Данная таблица является огромной, для этого распределим всю генеральную совокупность на графике (рисунок 24), вызовем функцию applicationsByYearGraph() и поместим в нее полученный выше DataFrame.

Рисунок 24 – Динамика заявок по годам

Из данного графика можно сделать вывод, что инженеры на сегодняшний день, все меньше ищут работу на сайтах вакансий, это является отличным показателем заинтересованности государства в данной профессии,

многие технические высшие учебные заведения предоставляют рабочие места после их окончания.

3.3 Моделирование и тестирование функции разреза по регионам

Разрез по регионам предоставит очень полезную информацию о выбранном регионе:

- Преобладание размера компаний;
- Среднюю зароботную плату в регионе;
- Код региона;
- Уровень цен.

Данный разрез и его информация поможет нам в дальнейшей сортировке данных и построения новых таблиц.

Функция regionInfo (рисунок 25), принимает в качестве параметра названия региона – region, на основании входного параметра вытаскивает информацию из таблиц stat_companies и regions, а именно:

- Всего компаний в регионе;
- Компаний с численностью сотрудников менее 50 человек;
- Компаний с численностью сотрудников от 51 до 100 человек;
- Компаний с численностью сотрудников от 101 до 250 человек;
- Компаний с численностью сотрудников от 251 до 500 человек;
- Компаний с численностью сотрудников более 500 человек;
- Средняя заработная плата;
- Уровень цен.

Поиск реализован с помощью метода loc, данный метод основывается на получении доступа к данным на основе значения индекса, который был передан функции.

```python
def regionInfo(region):
    data = stat_companies[stat_companies['region_name']==region]
    data2 = regions[regions['region_name']==region]
    regionInfo = pd.DataFrame({'Код региона': data2['region_code'],
                   'Компаний Всего': data['company_count'].loc[data.index[0]],
                   'Компаний с численностью сотрудников менее 50 человек': data['micro_company'].loc[data.index[0]],
                   'Компаний с численностью сотрудников от 51 до 100 человек': data['small_company'].loc[data.index[0]],
                   'Компаний с численностью сотрудников от 101 до 250 человек': data['midle_company'].loc[data.index[0]],
                   'Компаний с численностью сотрудников от 251 до 500 человек': data['big_company'].loc[data.index[0]],
                   'Компаний с численностью сотрудников более 500 человек': data['large_company'].loc[data.index[0]],
                   'Средняя заработная плата': data2['medium_salary_difference'],
                   'Уровень цен': data2['price_level']})
    microC = regionInfo['Компаний с численностью сотрудников менее 50 человек'].loc[regionInfo.index[0]]
    smallC = regionInfo['Компаний с численностью сотрудников от 51 до 100 человек'].loc[regionInfo.index[0]]
    midleC = regionInfo['Компаний с численностью сотрудников от 101 до 250 человек'].loc[regionInfo.index[0]]
    bigC = regionInfo['Компаний с численностью сотрудников от 251 до 500 человек'].loc[regionInfo.index[0]]
    largeC = regionInfo['Компаний с численностью сотрудников более 500 человек'].loc[regionInfo.index[0]]
    allC = regionInfo['Компаний Всего'].loc[regionInfo.index[0]]
    priceLevel = regionInfo['Уровень цен'].loc[regionInfo.index[0]]
    mediumSal = regionInfo['Средняя заработная плата'].loc[regionInfo.index[0]]
    regionCode = regionInfo['Код региона'].loc[regionInfo.index[0]]

    fig, ax = plt.subplots(figsize=(7, 6), subplot_kw=dict(aspect="equal"))
    labels = ['Micro', 'Small', 'Midle', 'Big', 'Large']
    vals = [microC, smallC, midleC, bigC, largeC]
    exp = (0.1, 0.2, 0, 0, 0)
    ax.pie(vals, labels=labels, autopct='%.2f', explode=exp, shadow=True)
    ax.set_title('Соотношение компаний в {0}\n Средняя заработная плата = {1}; Уровень цен = {2}; Код региона (КЛАДР) = {3}

    plt.show()
    return regionInfo
```

Рисунок 25 – Реализация функции regionInfo

Протестируем функцию regionInfo на примере Чувашской Республики (рисунок 26). Вызовим regionInfo("Чувашская Республика")

Рисунок 26 – Разрез компаний в Чувашской Республике

В данном подразделе мы смоделировали систему разреза региона и протестировали ее на Чувашской республике. Исходя из полученных данных можно сказать, что в данном регионе преобладают компании с численностью сотрудников от 51 до 100 человек, а среднии ежимесячные траты занимают практически половину месячного дохода.

3.4 Моделирование функций нормальности распределения для колличественных величин

Для работы с колличественными шкалами и дальнейшим анализом данных разработаем функции по определению нормальности распределения данных:

– Распределение Гаусса;

– Распределение Q-Q Plot;

– Тест Шапиро-Уилка.

Реализуем функцию нормальности распреднлнния Гаусса.

Функцию расчитывает основные статистические метрики для каждого поля DataFrame:

– Mini – минимальное значение среди набора переданных значений;

– Maxi – максимальное значение среди набора переданных значений;

– Ran – диапозон значений, расчитывается как разница между максимальным и минимальным значением;

– Mean – среднее значение элементов DataFrame;

– Median – вычислет медиану DataFrame;

– St_dev – вычисляет стандартное отклонение, меру разброса распределения DataFrame;

– Skew – вычисляет асимметрию набора данных;

– Kurt – возвращает несмещенный эксцесс по запрошенной оси.

Эксцесс получен с использованием определения эксцесса Фишера;

– Points – расчет точек стандартного отклонения.

Основываясь на полученных данных реализованная функция Gauss моделирует график случайного отклонения.

Реализация представлена на рисунке 27.

```
def Gauss(data):
    group = data.columns # Список столбцов
    size = len(group)
    plt.figure(figsize = (7 * size, 3), dpi = 600) # Параметры графика

    # Применяем расчеты к каждому столбцу
    for j,i in enumerate(group):

        # Рассчитываем основные статистические метрики
        mini = data[i].min()
        maxi = data[i].max()
        ran = data[i].max()-data[i].min() # Диапазон значений
        mean = data[i].mean()
        median = data[i].median()
        st_dev = data[i].std() # Стандартное отклонение
        skew = data[i].skew() # Скошенность
        kurt = data[i].kurtosis() # Эксцесс

        # Расчет точек стандартного отклонения
        points = mean - st_dev, mean + st_dev

        # Строим график с каждым набором данных
        plt.subplot(1, size, j+1)
        sns.distplot(data[i], hist = True, kde=  True)

        sns.lineplot(points, [0,0], color = 'black', label = "std_dev")
        sns.scatterplot([mini,maxi], [0,0], color = 'orange', label = "min/max")
        sns.scatterplot([mean], [0], color = 'red', label = "mean")
        sns.scatterplot([median], [0], color = 'blue', label = "median")
        plt.xlabel('{}'.format(i), fontsize = 20)
        plt.ylabel('density')
        plt.title('Стандартное отклонение = {}; Эксцесс = {};\n Скошенность = {}; Разброс, шаг гистограммы = {}\n Среднее = {}; Ме
                                                                    round(points[0],2),
                                                                    round(points[1],2)),
                                                                    round(kurt,2),
                                                                    round(skew,2),
                                                                (round(mini,2),round(maxi,2),round(ran,2)),
                                                                    round(mean,2),
                                                                    round(median,2)))
```

Рисунок 27 – Реализация функции Gauss

При анализе колличественных шкал на нормальность распределения мы не можем пологаться на один тест, смоделируем тест Q – Q Plot представленный на рисунке 28 и тест Шапиро – Уилка (рисунок 29).

```
def qqTestDF (data):
    qqplot(data, line='s')
    return pyplot.show
```

Рисунок 28 – Реализация теста Q – Q Plot

Реализуем тест Шапиро – Уилка, данный тест считается одним из самых надежных тестов на нормальность. Функция принимает параметр DataFrame, на его основе вычисляет наблюдаемый уровень значимости – p и сравнивает его с уровнем значимости 0.05.

```
def testShapiroU (data):
    stat, p = scipy.stats.shapiro(data)
    print('Statistics=%.3f, p-value=%.3f' % (stat, p))

    alpha = 0.05
    if p > alpha:
        return print('Принять гипотезу о нормальности')
    else:
        return print('Отклонить гипотезу о нормальности')
```

Рисунок 29 – Реализация теста Шапиро – Уилка

Проанализируем нормальность распределение зарплат Юристов в Чувашской Республике на примере маленьких компаний с численностью сотрудников менее 50 человек.

Объединим таблицы curricula_vitae, workexp, organizations и responses, отфильтруем данные job_title = «Юрист», region_code = « 2100000000000», business_size = « SMALL», удалим колонки id_organization и id_cv так как они служили foreign key для объединения таблиц. Получим DataFrame, изображенный на рисунке 30.

	business_size	region_code	job_title	salary
0	SMALL	2100000000000	Юрист	17,000
1	SMALL	2100000000000	Юрист	17,000
2	SMALL	2100000000000	Юрист	17,000
3	SMALL	2100000000000	Юрист	17,000
4	SMALL	2100000000000	Юрист	17,000
...	...	...	...	...
3020	SMALL	2100000000000	Юрист	20,000
3021	SMALL	2100000000000	Юрист	30,000
3022	SMALL	2100000000000	Юрист	30,000
3023	SMALL	2100000000000	Юрист	30,000
3024	SMALL	2100000000000	Юрист	30,000

Рисунок 30 – Зарплаты Юристов по Чувашской Республике в маленьких компаниях

Проанализируем полученный DataFrame на нормальность распределения (рисунок 31) с помощью функции Gauss, которая была создана нами ранее перед проверкой распределения разделим зарплаты на 1000 для более удобного анализа.

Рисунок 31 – Анализ нормальности распределения зарплат по кривой Гаусса

Из данного графика можно сделать вывод, что данные распределены ненормально, данный вывод влияет на выбор статистического критерия, средняя зароботная плата по данной профессии составляет 30 тысяч рублей. Проанализируем нормальность распределения с помощью теста Q – Q Plot (рисунок 32).

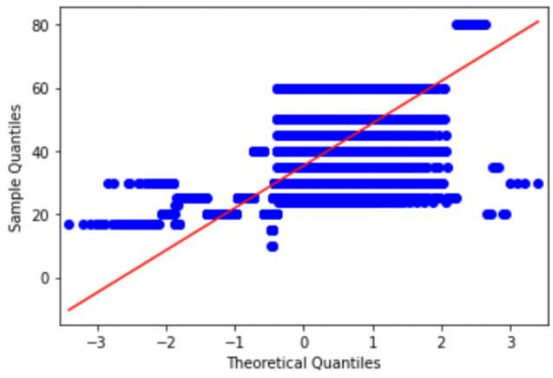


Рисунок 32 – Анализ нормальности распределения зарплат тест Q – Q Plot

Анализ нормальности распределения зарплат по тесту Q – Q Plot показал ненормальность распределения, так как данные распределены неравермерно вдоль моды и имеют множество выбросов. Проведем последний тест Шапиро – Уилка, чтобы окончательно убедиться в ненормальности распределения (рисунок 33).

```
Statistics=0.913, p-value=0.000
Отклонить гипотезу о нормальности
```

Рисунок 33 – Нормальности распределения зарплат тест Шапиро-Уилка

Тест Шапиро-Уилка отклонил гипотезу о нормальности, как и остальные тесты, теперь мы можем сделать окончательный вывод, что данные имеют ненормальное распределение.

В данном подразделе мы смоделировали тесты для анализа нормальности распределения данных с подробным тестированием на примере зарплат Юристов в Чувашской Республике в маленьких компаниях с численностью сотрудников менее 50 человек.

3.5 Моделирование и тестирование U-критерий Манна-Уитни на примере влияния гендерного признака на опыт работы

В данном подразделе мы проведем U-теста Манна-Уитни с использованием Pandas и SciPy [11] на примере влияния гендернего признака на опыт работы в выбраном регионе. На первом этапе мы получим наши данные, с помощью группировки таблиц. После того, как данные будут сохранены, мы проведем непараметрический тест.

Отсортируем данные для анализа и поместим их в новый DataFrame «bigdata», алгоритм позволит фильтровать данные по профессии в нужным регионе алгоритм указан на рисунке 34, блок схема алгоритма рисунок 35.

```
def filterByJT_RC(job_title, region_code):
    bigdata1 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/workexp.csv',
                           sep = ';',
                           usecols = ['job_title','id_cv'])
    bigdata1 = bigdata1[bigdata1['job_title'] == job_title]
    bigdata1 = bigdata1.reset_index(drop = True)
    bigdata2 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/curricula_vitae.csv',
                           sep = ';',
                           usecols = ['experience','id_cv','gender'])
    bigdataWorkCur = bigdata1.merge(bigdata2)
    bigdata3 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/responses.csv',
                       sep = ';',
                       usecols = ['region_code','id_cv'])
    bigdata3 = bigdata3[bigdata3['region_code'] == region_code]
    bigdata = bigdataWorkCur.merge(bigdata3)
    bigdata = bigdata.drop(['id_cv'],axis=1)
    bigdata = bigdata[bigdata['gender'].notna()]
    return bigdata.to_csv(r'/Users/vladkurockin/Desktop/Diplom/data/filterByJT_RC.csv',index=False)
```

Рисунок 34 – Реализация алгоритма фильтрации данных по названию вакансии и номеру региона

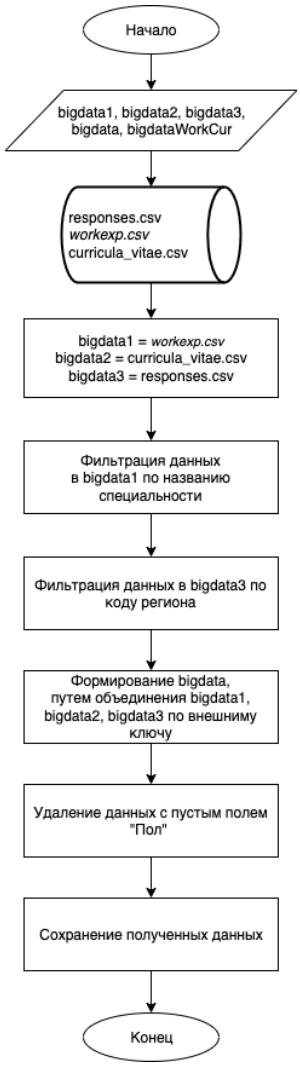


Рисунок 35 – Блок – схема алгоритма фильтрации данных по названию вакансии и номеру региона

Отфильтровав данные рисунок 36 с помощью алгоритма, который представлен на рисунке 34, имеются два гендорных типа (Мужчины и Женщины) по профессии «Юрист» в регионе «Чувашская республика» , различающиеся по опыту работы. U-критерий Манна-Уитни подходит для сравнения опыта работы двух гендерных типов, предположем следующую гипотезу:

H_0: Можно ли утверждать что опыт работы у Мужчин в сфере «Юрист» по региону «Чувашская республика», выше чем у Женщен.

```
filterByJT_RC('Юрист', 2100000000000)
```

	job_title	experience	gender	region_code
0	Юрист	6	Женский	2100000000000
1	Юрист	6	Женский	2100000000000
2	Юрист	6	Женский	2100000000000
3	Юрист	6	Женский	2100000000000
4	Юрист	6	Женский	2100000000000
...	...	...	...	...
2276	Юрист	7	Женский	2100000000000
2277	Юрист	7	Женский	2100000000000
2278	Юрист	0	Женский	2100000000000
2279	Юрист	0	Женский	2100000000000
2280	Юрист	0	Женский	2100000000000

Рисунок 36 – Полученные данные

У нас имеются две независимые группы с наблюдениями $x_1,\ x_2, \dots, x_m$ и $y_1,\ y_2, \dots, y_n$, выбранные из «М» и «Ж», U-критерий Манна-Уитни сравнивает каждое наблюдение x_i из выборки «М» с каждым наблюдением y_i из выборки «Ж».

Перед тем как приступить к тесту Манна – Уитни проверим нормальность распределения, опыта работы у мужчин и женщин, используя тест Шапиро – Уилка и Гистограмму рисунок 37

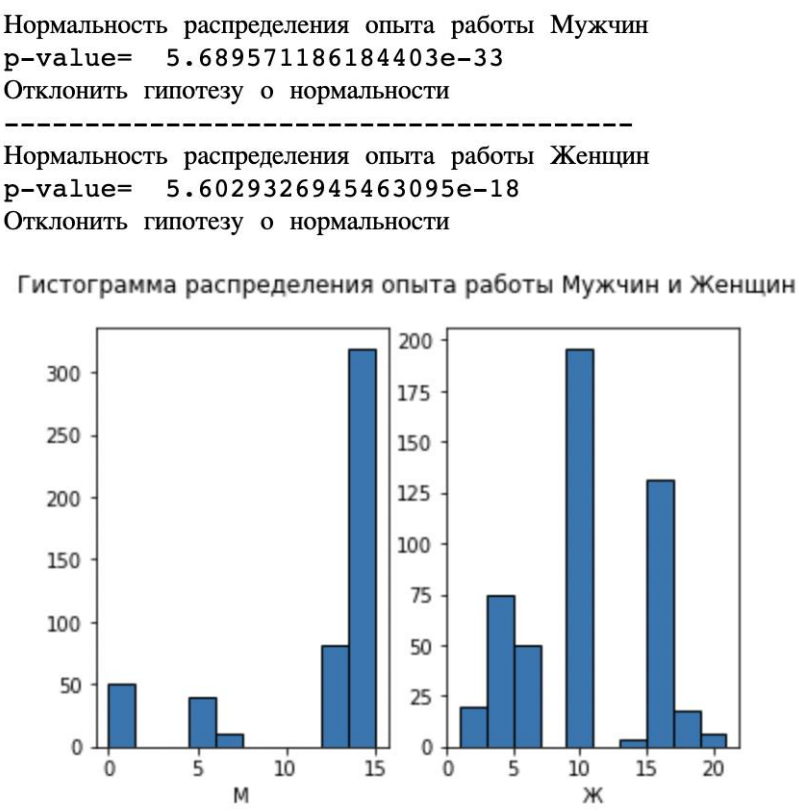


Рисунок 37 – Нормальность распределения опыта работы

Поскольку значение p , полученное из теста Шапиро-Уилка, является значимым ($p < 0{,}05$), мы делаем вывод, что данные опыта работы по Мужчинам и Женщинам не распределены нормально. На гистограмме форма распределения данных не выглядит нормальной. Следовательно, U-критерий Манна-Уитни подходит для анализа двух гендерных типов.

Выполним U-критерий Манна-Уитни рисунок 38.

```
stats.mannwhitneyu(x=dataMale['experience'], y=dataFemale['experience'], alternative = 'greater')
print('p-value=%.12f' % (p))
alpha = 0.05

if p < alpha:
        print('Фактор влияет на отклик принять H0')
        print('Различие опыта работы между М и Ж в сфере ' + dataJT_RC['job_title'].loc[dataJT_RC.index[0]]
              + str(dataMale['experience'].median() - dataFemale['experience'].median()) + ' лет')

else:
    print('Фактор не влияет на отклик отклонить H0')
```

Рисунок 38 – Реализация U – критерия Манна-Уитни

В приведенном выше примере рисунок 38, значение p полученное из mannwhitneyu, основано на нормальном приближении, поскольку размер выборки велик ($n > 20$). Если размер выборки мал, нормальная аппроксимация не подходит. Функция mannwhitneyu автоматически вычисляет точное значение p. Как точное, так и нормальное приближение значений p должны быть примерно одинаковыми.

Поскольку значение p, полученное из U-критерия Манна-Уитни, является значимым (U = 0.000043796767, $p < 0{,}05$), мы заключаем, что опыт гендерного типа Мужчин значительно выше, чем опыт Жещин. Вычислим разничу двух медиан и узнаем отличие рисунок 39.

```
p-value=0.000043796767
Фактор влияет на отклик принять H0
Различие опыта работы между М и Ж в сфере Юрист по данному региону составляет 5.0 лет
```

Рисунок 39 – Вывод реализация U – критерия Манна-Уитни

В данном подразделе мы смоделирование и протестирование U-критерий Манна-Уитни на примере влияния гендерного признака на опыт работы, а также смоделировали алгоритм для фильтрации данных и сбора нужной информации из общего набора данных.

3.6 Моделирование и тестирование критерия Хи-квадрат Пирсона на примере влияния отклика работодателя на образование.

Тест хи-квадрат — это статистический тест, используемый для определения взаимосвязи между категориальными переменными /столбцами в наборе данных. Он исследует корреляцию между переменными, которые не содержат непрерывных данных.

Чтобы использовать тест хи-квадрат, мы можем предпринять следующие шаги:

– Сформулировать нулевую (H_0) и альтернативную (H_1) гипотезы;

– Определите значение альфы (α) для в соответствии с доменом, с которым вы работаете. В идеале $\alpha = 0{,}05$, это означает, что мы готовы взять на себя 0,5% риска/погрешности;

– Сформировать данные;

– Проверьте данные на наличие Nans или других ошибок;

– Выполнить тест и сделайте вывод, следует ли отклонить или принять нулевую гипотезу (H_0).

Формула для теста хи-квадрат выглядит следующим образом (23)

Перед выполнением теста, создадим алгоритм формирования таблицы по названию профессии из всей генеральной совокупности рисунок 40.

```
def applications(name):
    bigdata1 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/workexp.csv',
                           sep = ';',
                          usecols = ['job_title','id_cv'])
    bigdata1 = bigdata1[bigdata1['job_title'] == name]
    bigdata1 = bigdata1.reset_index(drop = True)

    bigdata2 = pd.read_csv('/Users/vladkurockin/Desktop/Diplom/data/Faculty_InInfo.csv',
                           sep = ',')

    bigdata = bigdata1.merge(bigdata2)
    bigdata = bigdata.drop(['id_cv'],axis=1)
    bigdata['faculty'] = bigdata['faculty'].fillna('Образования нет')
    bigdata['faculty'].mask(bigdata['faculty'] != 'Образования нет', 'Образования есть' , inplace=True )
    bigdata['inner_info_status'].mask(bigdata['inner_info_status'] == 'Ожидает подтверждения',
                                      'Не одобрено' , inplace=True )

    bigdata = bigdata.drop(['job_title'],axis=1)
    bigdata = pd.crosstab(bigdata['faculty'], bigdata['inner_info_status'])

    return bigdata
```

Рисунок 40 – Функция формирования кростаблицы частот по названию профессии из всей генеральной совокупности

Мы выполним следующие шаги, чтобы создать таблицу непредвиденных обстоятельств:

– Объединим таблицы «workexp» и «Faculty_Info» по внешнему ключу «id_cv»;

– Проверьте строки на наличие Nans и заменим их на «Образования нет»;

– Сгрупируем данные и получим частоту их появления, путем создания кростаблицы.

Смоделировав функцию получим данные для исследования на примере «Юристов» рисунок 41.

```
applications('Юрист')
```

inner_info_status faculty	Не одобрено	Одобрено
Образования есть	6494	214396
Образования нет	1843	51060

Рисунок 41 – Исходные данные для анализа

В приведенной выше таблице рисунок 41, непредвиденных обстоятельств показаны различные типы откликов работодателей, на заявки соискателей с образованием и без него в качестве индекса.

Теперь, когда у нас сформирован DataFrame, выдвиним следующую гипотезы:

– H_0: нет никакой взамосвязи между откликом работодателя и высшим образованием;

– H_1: взаимосвязь существует.

Если p-значение значимо, вы можете отклонить нулевую гипотезу и заявить, что результаты подтверждают альтернативную гипотезу.

Смоделируем функцию для расчета основных значений Хи – квадрат, с построением графической визуализации, и передадим в нее исходные данные для анализа, чтобы принять или опровергнуть гипотезу рисунок 42.

Первое значение (42,569) — это значение хи-квадрат, за которым следует значение p (6,82e-07), затем идут степени свободы (1) Мы можем отклонить нулевую гипотезу, поскольку значение $p < 0{,}05$. Таким образом, результаты показывают, что в профессии Юрист отклик работодателя влияет на образование.

Рисунок 42 – Расчет критерия Хи – квадрат на полученных данных

В данном подпунке была создана функция формирования кросс таблицы частот по названию профессии из всей генеральной совокупности.

На ее основе был проведен тест Хи – квадрат, для выяснения влияния образования на отклик соискателя.

3.7 Моделирование и тестирование корреляции Пирсона, Спирмена на примере влияния уровня зарплат на опыт работы

В данном подразделе мы смоделируем и проведем корреляционны анализ Пирсона и Спирмена с использованием Pandas, SciPy, Matplotlib на примере влияния уровня зарплат на опыт работы. На первом этапе мы получим наши данные, с помощью группировки таблиц. После того, как данные будут сохранены, мы проведем моделирование и анализ.

Для получения данных из генеральной совокупности. Смоделируем функцию, которая принимает параметры «Название должности» и «Код региона». Объединим таблицы curricula_vitae, workexp и responses, отфильтруем данные job_title = «Юрист», region_code = «2100000000000», удалим колонку id_cv так как она служила foreign key для объединения таблиц. Полученный DataFrame сохраним по указонному пути, реализация функции рисунок 43, полученный DataFrame рисунок 44.

```
def filterByJT_RCEXpSal(job_title, region_code):
    bigdata1 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/workexp.csv',
                           sep = ';',
                           usecols = ['job_title','id_cv'])
    bigdata1 = bigdata1[bigdata1['job_title'] == job_title]
    bigdata1 = bigdata1.reset_index(drop = True)
    bigdata2 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/curricula_vitae.csv',
                           sep = ';',
                           usecols = ['experience','id_cv', 'salary'])
    bigdataWorkCur = bigdata1.merge(bigdata2)
    bigdata3 = pd.read_csv('/Users/vladkurockin/Downloads/_Rabota_v_RF_trudvsem.rf_186_02.12.21/responses.csv',
                       sep = ';',
                       usecols = ['region_code','id_cv'])
    bigdata3 = bigdata3[bigdata3['region_code'] == region_code]
    bigdata = bigdataWorkCur.merge(bigdata3)
    bigdata = bigdata.drop(['id_cv'],axis=1)
    bigdata = bigdata[bigdata['salary'].notna()]
    return bigdata.to_csv(r'/Users/vladkurockin/Desktop/Diplom/data/filterByJT_RCEXpSal.csv',index=False)
```

Рисунок 43 – Реализация функции для получения данных

	job_title	experience	salary	region_code
0	Юрист	4	25,000	2100000000000
1	Юрист	4	25,000	2100000000000
2	Юрист	4	25,000	2100000000000
3	Юрист	4	25,000	2100000000000
4	Юрист	4	25,000	2100000000000
...	...	...	...	...
3405	Юрист	0	25,000	2100000000000
3406	Юрист	0	25,000	2100000000000
3407	Юрист	0	25,000	2100000000000
3408	Юрист	0	25,000	2100000000000
3409	Юрист	0	25,000	2100000000000

Рисунок 44 – Полученные данные по Юристам

Смоделируем функцию для корреляционного анализа полученных данных рисунок 45, данная функция работает следующим образом:

– Принимает параметр DataFrame;

– Проверяет нормальность распределения «Зарплаты» и «Опыта» тестом Шапиро – Уилка;

– Исходя из нормальности распределения выбирает корреляционный анализ, если все данные распределенны нормально используем корреляцию Пирсона, в противном случае Спирмана;

– Строит линию регрессии и график;

– Делает вывод о связи и достоверности.

```
def correlation (data):
    alpha = 0.05
    print('Проверка нормальности распределения:')
    print('\nУровень значимости p для зарплат')
    sal = testShapiroU(data['salary'])
    print('------------------------------')
    print('Уровень значимости p для опыта работы')
    exp = testShapiroU(data['experience'])

    slope, intercept, r, p, stderr = scipy.stats.linregress(bigdata1['salary'], bigdata1['experience'])
    line = f'Regression line: y={intercept:.2f}+{slope:.2f}x, r={r:.2f}'

    fig, ax = pyplot.subplots(figsize = (14,8))
    ax.plot(data['salary'], data['experience'], linewidth=0, marker='s', label='Data points')
    ax.plot(data['salary'], intercept + slope * data['salary'], label=line)
    ax.set_xlabel('Зарплата')
    ax.set_ylabel('Опыт работы')
    ax.legend(facecolor='white')
    ax.grid()
    pyplot.show()

    if (sal > alpha and exp > alpha):
        print('Данные распределены нормально используем корреляцию Пирсона')
        corr, pvalue = scipy.stats.pearsonr(data['salary'], data['experience'])
        print('Коэффициент корреляции = ', corr)
        print('p-value = ', pvalue)
        if (corr < 0.25):
            print('Связь слабая и прямая')
        if (corr >= 0.25 and corr < 0.75):
            print('Связь умерянная и прямая')
        if (corr > 0.75):
            print('Связь сильная и прямая')
        if (p < 0.001):
            print ('Высокая достоверность')
        if (p > 0.001):
            print ('Низкая достоверность')
    else:
        print('Данные распределены ненормально используем корреляцию Спирмена')
        coef, p = spearmanr(data['salary'], data['experience'])
        if (coef < 0.25):
            print('Связь слабая и прямая')
        if (coef >= 0.25 and coef < 0.75):
            print('Связь умерянная и прямая')
        if (coef > 0.75):
            print('Связь сильная и прямая')
        print('Коэффициент корреляции = ', coef)
        print('p-value = ', p)
        if (p < 0.001):
            print ('Высокая достоверность')
        if (p > 0.001):
            print ('Низкая достоверность')
```

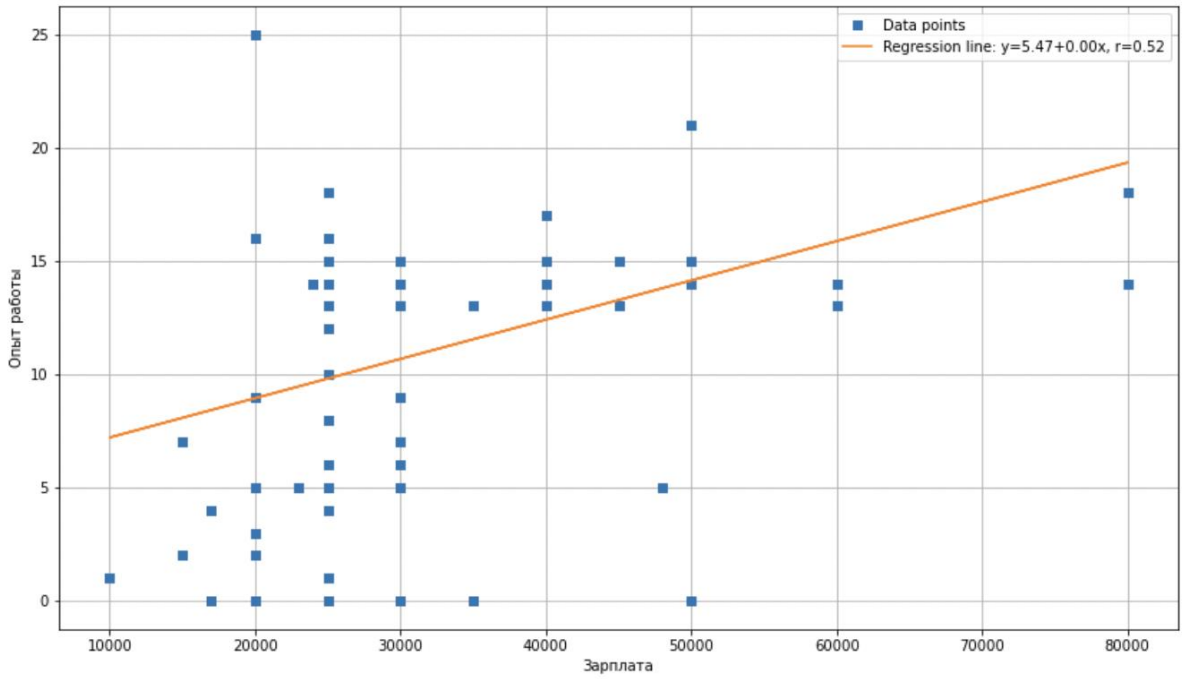
Рисунок 45 – Функция корреляционного анализа

Проведем корреляционный анализ и выясним влияние опыта работы на зарплату в сфере Юрист по Чувашской Республике рисунок 46.

Проверка нормальности распределения:

Уровень значимости p для зарплат
p-value = 2.7941891378636852e-42

Уровень значимости p для опыта работы
p-value = 0.0



Данные распределены ненормально используем корреляцию Спирмена
Связь умерянная и прямая
Коэффициент корреляции = 0.5342931129914894
p-value = 4.5639191655065195e-251
Высокая достоверность

Рисунок 46 – Корреляционный анализ влияния зарплаты на опыт работы

Из полученных данных можно сделать вывод, что данные «опыт работы» и «зарплата» являются ненормально распределенными, используем корреляцию Спирмена. Коэффициент корреляции $\rho = 0.53$, из этого следует что связь умеренная и прямая. Возвращаемое значение $p < 0{,}001$, что подтверждает высокую достоверность результата.

В данном подразделе мы реализовали и протестировали корреляционный анализ на примере влияния опыта работы на зарплату в сфере «Юрист» в Чувашской Республике. А также реализовали функцию сбора данных.

Вывод по разделу

В этом разделе мы определились с инструментами разработки, разбили данные по таблицам и описали каждую из них, разработали алгоритмы по объединению данных в отдельные DataFrame, привели смоделировали и протестировали:

– Функцию разреза по вакансиям;

– Функцию разреза по регионам;

– Функции нормальности распределения для количественных величин;

– Тест U-критерий Манна-Уитни на примере влияния гендерного признака на опыт работы;

– Тест Хи-квадрат Пирсона на примере влияния отклика работодателя на образование;

– Корреляционный анализ Пирсона, Спирмена на примере влияния уровня зарплат на опыт работы.

Для алгоритмов были приведены блок – схемы и участки кода.

Заключение

В данной пояснительной записке к выпускной квалификационной работе, были рассмотрены основные определения описательной статистики, был смоделирован анализ статистических совокупностей путем выбора статического критерия на примере собранных данных сайтов вакансий. Также были разработанны алгоритмы формирования агрегированных таблиц по требуемым аргументам. Реализация производилось на языке программирования Python с использованием библиотек указанных в подразделе 3.1.

В ходе исследования предметной области были указаны материалы, позволяющие ознакомиться с теоретическими данными и реализацией системы HR – аналитики, а также приложены иные материалы, способствующие пониманию того, как работает программная часть модуля, такие как блок-схема и рисунки, отображающий возможности взаимодействия программных компонентов продукта.

Одним из достоинств данной программы является наглядность (все вычисления представляются на графиках). Среди минусов можно отметить, что программа содержит не все статистические стратегии. Данные недостатки можно исправить в последующих версиях программы. Кроме того, в перспективах развития программы можно воссоздать общую базу данных и реализовать Web - приложение.

Подводя итог, можно сделать вывод, что разработанное программное обеспечение успешно функционирует и позволяет анализировать данные сайтов вакансий путем постановки гипотез.

Список используемой литературы

1. Афанасьев, В. В. Теория вероятностей [Текст] / В. В. Афанасьев. – М.: ВЛАДОС, 2007. – 350 с.

2. Валеев С.Г., Клячкин В.Н. Практикум по прикладной статистике. Учебное пособие. – Ульяновск: УлГТУ, 2008. – 130 с.: ил.

3. Васильев, А. Н. Python на примерах. Практический курс по программированию / А.Н. Васильев. - М.: Наука и техника, 2016. - 432 с.

4. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998.

5. Горяинова, Е. Р., Панков, А. Р., Платонов, Е. Н. Прикладные методы анализа статистических данных [Текст] : учеб. пособие / Е. Р. Горяинова, А. Р. Панков, Е. Н. Платонов ; Нац. исслед. ун-т «Высшая школа экономики». — М.: Изд. дом Высшей школы экономики, 2012. — 310, [2] с. — 1000 экз. — 978-5 7598-0866-4 (в обл.).

6. Изучаем Python. Программирование игр, визуализация данных, веб-прило – жения. — СПб.: Питер, 2017. — 496 с.: ил. — (Серия «Библиотека программиста»).

7 . Крамер Г. Математические методы статистики. – М.: Мир, 1975.

8. Математическая статистика: учеб. пособие /Д.К. Агишева, С.А. Зотова, Т.А. Матвеева, В.Б. Светличная; ВПИ (филиал) ВолгГТУ. – Волгоград, 2010. – 159 с.: ил.

9. Наследов, А. Д. Математические методы психологического исследования. Анализ и интерпретация данных [Текст] / А. Д. Наследов. – СПб.: Речь, 2004. – 392 с.

10. Никитин, О. Р. Методы измерения статистических параметров радиосигналов : учеб. пособие / О. Р. Никитин, Н. Н. Корнеева ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ. – Владимир, 2020. – 227 с.

11. Нуньес-Иглесиас Х., Уолт ван дер Ш., Дэшноу Х. Элегантный SciPy/ пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2018. – 266 с.: ил.

12. Общая теория статистики: учебное пособие: в 2 частях - Ч.1. Описательная статистика / Татьяна Борисовна Ершова. – Комсомольск-наАмуре: Изд-во АмГПГУ, 2012.–120 с.

13. Основы теории статистики : [учеб. пособие] / В. В. Полякова, Н. В. Шаброва ; М-во образования и науки Рос. Федерации, Урал. федер. ун-т. – 2-е изд., испр. и доп. – Екатеринбург : Изд-во Урал. ун-та, 2015. – 148 с.

14. Практикум по общей теории статистики: учеб. пособие / И.И. Елисеева, Н.А. Флуд, М.М. Юзбашев; под ред. И.И. Ели- сеевой. – М.: Финансы и статистика, 2008. – 512 с.: ил.

15. Справочник по прикладной статистике. – М.: Финансы и статистика. 1989. Т.1.

16. Тест Шапиро-Уилка URL:https://translated.turbopages.org/proxy_u/enru.ru.46c1ad93629661100584c65674722d776562/https/en.wikipedia.org/wiki/Wilk%E2%80%93Shapiro_test.

17. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере/ Под ред. В.Э.Фигурнова. - М.: ИНФРА-М, 1998. - 528 с.

18. Чубинский А.Н. Методы и средства научных исследований. Методы планирования и обработки результатов экспериментов: учебное пособие для студентов, обучающихся по направлениям 35.03.02 и 35.04.02 «Технология лесозаготовительных и деревоперерабатывающих производств», профиль «Технология деревообработки» / А.Н. Чубинский, Д.С. Русаков, И.М. Батырева, Г.С. Варанкина – СПб.: СПбГЛТУ, 2018.– 109 с.

19. Devpractice Team. Библиотека Matplotlib. - devpractice.ru. 2019. – 100 с.: ил.

20. Gregory W. Corder, Dale I. Foreman. Nonparametric Statistics. URL: https://faculty.ksu.edu.sa/sites/default/files/nonparametric_statistics_a_step-by-step_approach.pdf

21. NumPy User Guide. URL: https://numpy.org/doc/stable/numpy-user.pdf

22. Pearson's chi-squared test. URL: https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

23. Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил. — (Серия «Бестселлеры O’Reilly»).

24. Timothy J.Barth, Michael Griebel, David E. Keyes, Risto M. Nieminen, Dirk Roose, Tamar Schlick. Programming for Computations – Python. URL: https://library.oapen.org/bitstream/handle/20.500.12657/27997/1002000.pdf?sequence=1

25. Wes McKinney and the Pandas Development Team. pandas: powerful Python data analysis toolkit. URL: https://pandas.pydata.org/docs/pandas.pdf