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)

09.04.03 Прикладная информатика
(код и наименование направления подготовки)

Управление корпоративными информационными процессами
(направленность (профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

на тему «Модели и алгоритмы системы анализа качества аудитории корпоративного сайта»

Обучающийся Т.Б. Холдаров
(И.О. Фамилия) (личная подпись)

Научный руководитель д.т.н., доцент, С.В. Мкртычев
(ученая степень, звание, И.О. Фамилия)

Тольятти 2023

Оглавление

Введение

Одной из важнейших критериев сайта является его посещаемость.

Посещаемость сайта отражает количество желаний и проблем, которые хотят получить или решить пользователи.

Чтобы понять какие задачи нужно решать на сайте для увеличения его посещаемости, необходимо сформировать портрет пользователя.

«Портрет пользователя – это набор разных характеристик и информацнных данных о пользователях, с помощью которого можно выявить проблему для отдельных групп лиц» [13].

Так же портрет пользователя можно представить, как комплексную характеристику и описание типичного пользователя продукта или услуги, основанное на анализе данных о поведении пользователей. Он включает в себя информацию о демографических характеристиках пользователей, их интересах, потребностях, привычках и предпочтениях. Портрет пользователя помогает лучше понимать целевую аудиторию и создавать более эффективные маркетинговые стратегии и продукты, а также оптимизировать пользовательский опыт.

Для создания портрета пользователя проводится анализ данных, которые могут включать в себя:

Данные о демографии: возраст, пол, место жительства, образование, доходы и другие характеристики, которые могут быть важны для понимания аудитории.

Данные о поведении: данные о частоте использования продукта, продолжительности сеансов, популярных функциях и действиях пользователей, которые позволяют понять, как пользователи взаимодействуют с продуктом.

Данные о предпочтениях: данные о том, какие функции или продукты наиболее популярны у пользователей, что они ищут, какие задачи они решают, и какие проблемы они хотят решить.

Данные о сегментации: данные о том, как различные группы

пользователей взаимодействуют с продуктом, какие проблемы они решают, какие функции наиболее важны для каждой группы, и как можно оптимизировать пользовательский опыт для каждой группы.

После анализа данных и создания портрета пользователя, можно использовать его для создания более эффективных стратегий маркетинга и продуктов, которые лучше соответствуют потребностям и предпочтениям целевой аудитории. Также можно использовать портрет пользователя для оптимизации пользовательского опыта, понимая, как пользователи взаимодействуют с продуктом и что наиболее важно для них.

На данный момент существует множество различных систем анализов целевой аудитории сайта.

Построение эффективной системы анализа целевой аудитории сайта представляет актуальность и научно-практический интерес.

Объектом исследования является процесс сегментирования целевой аудитории сайта.

Предметом исследования являются модели и алгоритмы системы анализа целевой аудитории сайта.

Целью работы является создание моделей и алгоритмов для быстрой и эффективной сегментации пользователей.

Гипотеза исследования заключается в том, что применение методов интеллектуального анализа данных для сегментации целевой аудитории сайта повысит эффективность и гибкость сегментации.

В ходе работы применялись следующие методы научного исследования: системный анализ, эмпирические исследования, эксперименты, статистический анализ, визуализация данных, компьютерное моделирование и машинной обучение, а также методы валидации и верификации.

При выполнении диссертации последовательно пройдены следующие этапы:

– проведён обзор современного состояния исследований в области систем анализа целевой аудитории сайта;

– проведён анализ наиболее эффективных методов и алгоритмов для сегментации целевой аудитории;
– выбраны методы, на основании которых будет разработать алгоритм для сегментации целевой аудитории сайта;
– проведена апробацию моделей, полученных в ходе сравнение моделей с целью определения более эффективной;
– проверена эффективность примененной модели, сравнена ее работа с существующей моделью;
– оценён полученный результат и сделан вывод.

Научная новизна диссертационной работы, заключается в:

– улучшении модели системы анализа целевой аудитории сайта;
– применении альтернативных методов для анализа целевой аудитории.

Практическая значимость диссертационной работы заключается в том, что совершенствование анализа целевой аудитории способствует увеличению общего числа пользователей веб-сайта.

При проведении научного исследования принято участие в научном журнале «Вестник научных конференций». Результатом участия в данном журнале является статья (принята к публикации).

На защиту выносится:

– алгоритм сегментирования целевой аудитории сайта;
– результаты апробации и оценки эффективности разработанных методов.

Диссертация состоит из введения, 4 глав, заключения, списка используемой литературы и используемых источников и приложения.

Работа изложена на 74 страницах, содержит 66 рисунков, 3 таблицы и 30 источников.

Использованные в диссертации изображения находятся в свободном доступе, являются общественным достоянием или собственными изображениями.

Глава 1 Системы анализа целевой аудитории сайта

1.1 Анализ современного состояния исследований в области систем анализа целевой аудитории

«Целевая аудитория - это группа людей, определяемая различными общими атрибутами и характеристиками. Это группы людей, которые с наибольшей вероятностью купят продукт или услугу. Определение этой аудитории - важная часть создания успешной контентной стратегии на сайте и медиа» [24].

Также стоит добавить, что целевая аудитория может делится на группы.

Данный метод называется сегментацией целевой аудитории.

«Сегментация целевой аудитории – это разделение потенциальных покупателей на группы, объединенные общими потребностями» [15].

Проблематики в области систем анализа целевой аудитории сайта посвятили свои работы Номейн Алексей, Шевченко. Д, Krzysztof Kubacki, Sharyn Rundle-Thiele и другие.

Методы анализа целевой аудитории сайта:

- анализ статистики - этот метод основывается на анализе данных, полученных от пользователей сайта, таких как число посетителей, время пребывания на сайте, просмотры страниц и другие параметры. Эти данные могут помочь понять, какие страницы и функции сайта наиболее популярны среди пользователей и как пользователи взаимодействуют с сайтом;
- массовые опросы - этот метод включает в себя опрос большого количества пользователей, которые могут представлять целевую аудиторию сайта. Опросы могут проводиться онлайн или офлайн, и могут включать в себя вопросы о предпочтениях, потребностях, проблемах и мнениях пользователей;

– интервью с пользователями - этот метод включает в себя глубокое интервью с отдельными пользователями сайта, которые могут представлять целевую аудиторию. это может помочь понять, как пользователи используют сайт, какие проблемы они испытывают при его использовании, а также получить более детальную информацию о их потребностях и предпочтениях;
– опрос представителей бизнеса - этот метод включает в себя опрос представителей бизнеса, которые могут использовать сайт в качестве инструмента маркетинга или продаж. Это может помочь понять, как бизнес использует сайт, какие проблемы он испытывает при его использовании и какие функции на сайте могут быть улучшены;
– экспертные гипотезы - этот метод включает в себя использование опыта и знаний экспертов, работающих в отрасли, связанной с сайтом. Это может помочь понять, какие функции на сайте могут быть наиболее полезны для пользователей, а также определить тенденции и вызовы в отрасли.

К методам анализа целевой аудитории сайта относят разбивку целевой аудитории на категории или типы целевой аудитории, а именно [23]:

– демография - этот тип целевой аудитории определяется по различным демографическим характеристикам, таким как возраст, пол, семейное положение, образование, доход и т.д. Эта информация позволяет понять, какие категории пользователей наиболее заинтересованы в продукте или услуге, и настроить маркетинговые кампании соответствующим образом;
– интересы - этот тип целевой аудитории определяется на основе интересов и предпочтений пользователей. Интересы могут быть связаны с конкретными темами, например, спортом, культурой, модой и т.д., или с определенными видами продуктов или услуг, например, электроникой, путешествиями, здоровьем и т.д. Эта информация

помогает понять, какие товары или услуги наиболее востребованы пользователем;

– стили поведения - этот тип целевой аудитории определяется на основе стилей поведения пользователей в сети, например, с каких устройств они заходят на сайт, как часто они посещают сайт, какое время суток предпочитают для посещения сайта, как долго они проводят время на сайте и т.д. Эта информация позволяет оптимизировать сайт и контент для удобства использования и повышения уровня вовлеченности пользователей [25];
– частота потребления (покупка) конкретных продуктов - этот тип целевой аудитории определяется на основе частоты покупки или потребления конкретных продуктов или услуг. Эта информация помогает определить, какие категории пользователей наиболее вероятно будут совершать покупки на сайте, и настроить маркетинговые кампании соответствующим образом.

Таким образом, разбивка целевой аудитории на категории или типы позволяет получить более детальное представление о пользователе сайта и его потребностях, что в свою очередь помогает оптимизировать сайт и маркетинговые кампании, увеличить конверсию и улучшить пользовательский опыт.

Кроме того, при анализе целевой аудитории необходимо учитывать ее изменчивость, поскольку пользователи могут изменять свои потребности и предпочтения со временем [2], [30].

Самый основной тип целевой аудитории определяется демографией.

Демографические факторы включают:

– возраст — это один из наиболее важных факторов, поскольку потребности и предпочтения пользователей могут изменяться в зависимости от возраста. Например, молодые пользователи могут быть заинтересованы в продуктах, связанных с развлечениями и технологиями, в то время как более старшие пользователи могут

искать более традиционные продукты или услуги;

- пол - является одним из важных факторов, который влияет на потребности и предпочтения пользователей. Например, продукты и услуги, связанные с красотой и здоровьем, могут быть более интересны женщинам, в то время как продукты, связанные с технологиями и спортом, могут быть более интересны мужчинам;
- социоэкономический статус — это фактор, который связан с уровнем дохода и социальным статусом пользователей. Он может влиять на выбор продуктов и услуг, которые представляют для них наибольший интерес. Например, пользователи с высоким доходом могут быть более заинтересованы в продуктах и услугах премиум-класса;
- доход - уровень дохода пользователя может влиять на его потребности и возможность приобретения определенных продуктов и услуг. Например, пользователи с низким доходом могут искать продукты и услуги, которые представляют хорошее соотношение цены и качества, в то время как пользователи с высоким доходом могут быть более заинтересованы в продуктах и услугах, которые представляют уникальный дизайн или высокую производительность;
- образование - уровень образования пользователя также может влиять на его потребности и предпочтения. Например, пользователи с высшим образованием могут быть более заинтересованы в продуктах и услугах, которые представляют высокий уровень интеллектуальной сложности или предназначены для профессионального использования.

На рисунке 1 изображена модель демографии.

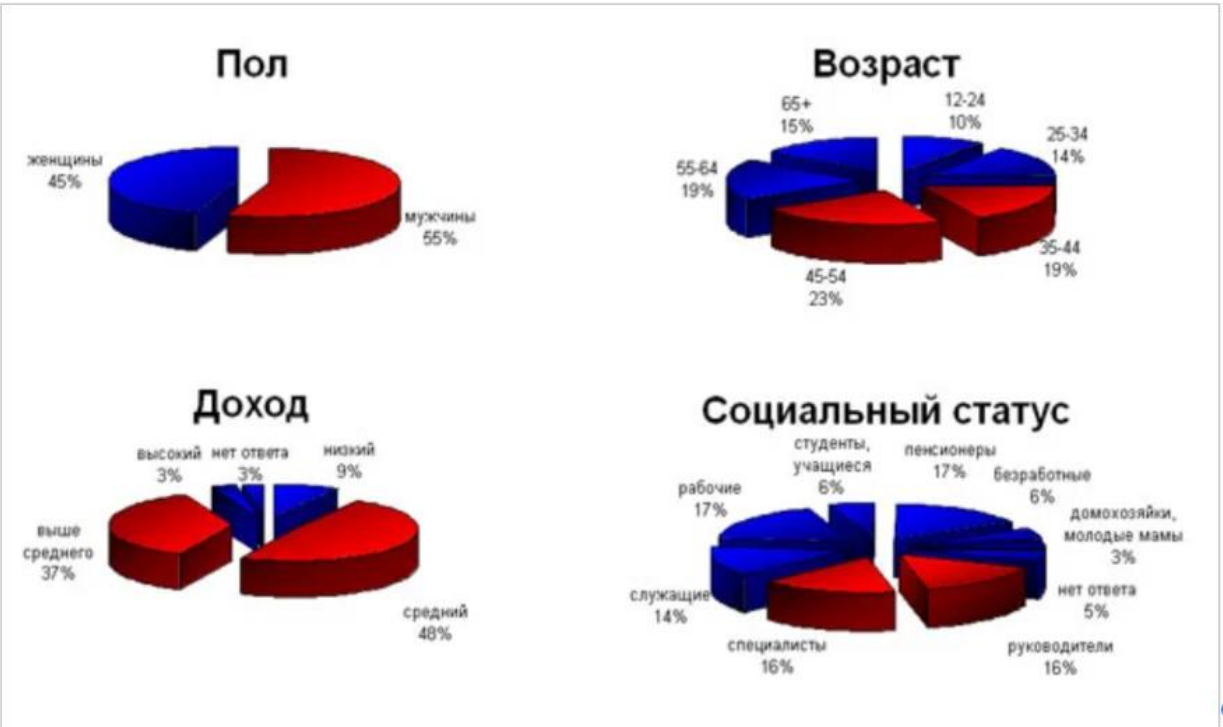

Рисунок 1 – Модель демографии

Основные цели каждого метода анализа:

– выявление ключевых характеристик основных групп ЦА;

– формирование персон, представляющих каждую из групп;

– описание их потребностей, контекста и поведенческих стереотипов.

В данном разделе был произведён анализ современного состояния в области систем анализа целевой аудитории сайта [3], [26].

1.2 Обзор и анализ источников по теме исследований

Рассмотрим работы, в которых представлены методы и алгоритмы для построения систем анализа целевой аудитории.

«Сегментирование целевой аудитории — это разделение аудитории на группы, где они объединены по признаку схожих потребностей (запросов).» [15]

На рисунке 2 изображена модель целевой группы, разделённой на

подгруппы – сегменты.

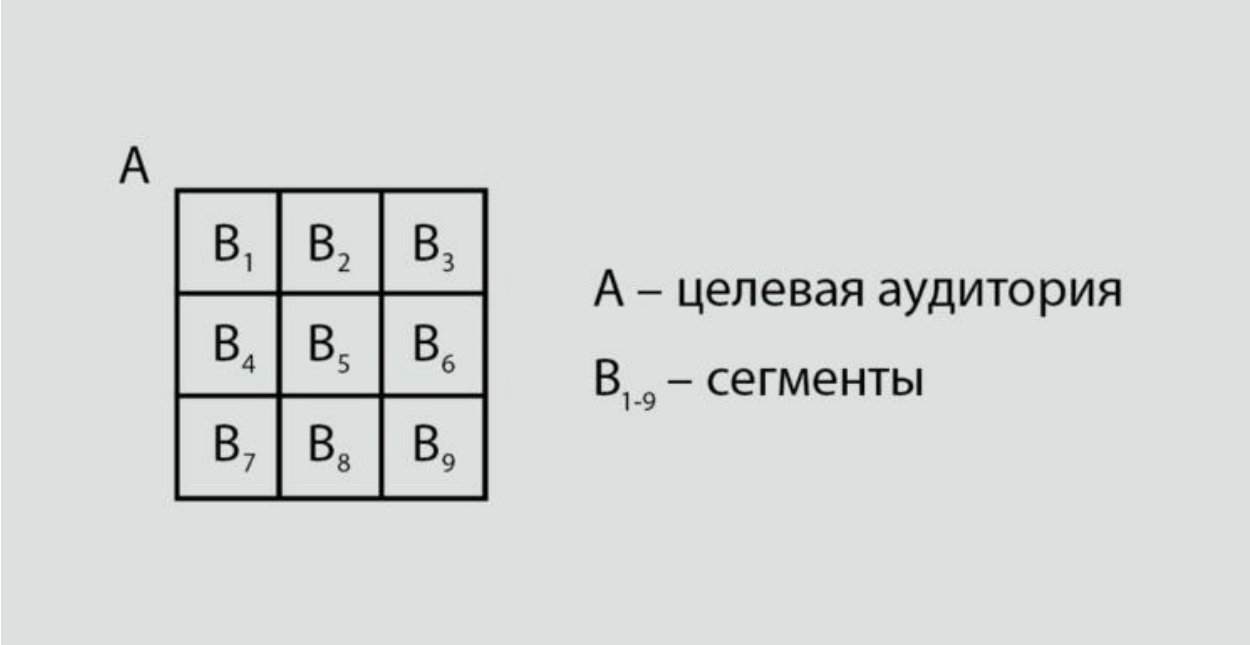

Рисунок 2 – Модель целевой аудитории

Существует множество принципов разделения ЦА, в зависимости от ситуации можно выбрать один из параметров главным при сегментации целевой аудитории.

Так же существуют популярные методы сегментации, которые на сегодняшний день являются устойчивыми [21].

Метод сегментирования «5W» Марка Шеррингтона - это метод, который позволяет разбить целевую аудиторию на группы или сегменты на основе ответов на пять вопросов, начинающихся на букву W и изображённых на рисунке 3.

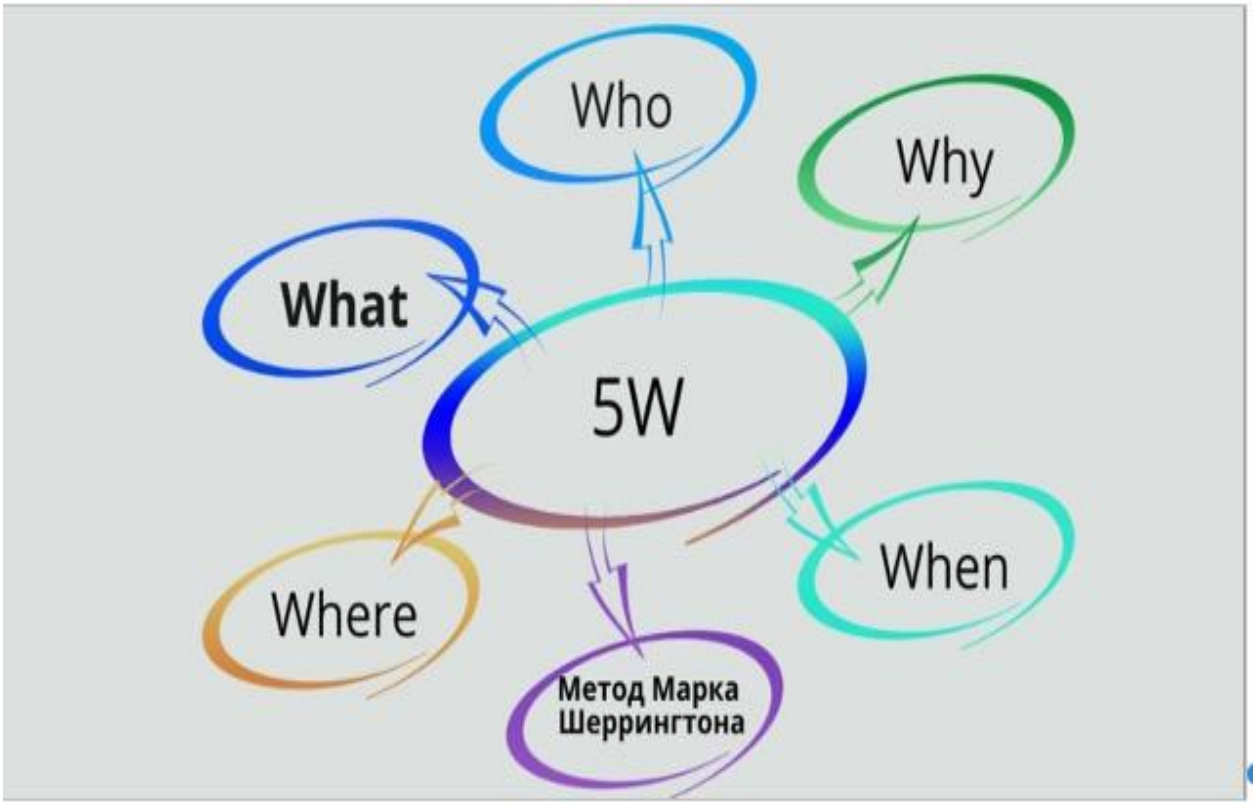

Рисунок 3 – Метод сегментирования «5W» Марка Шеррингтона

Этот метод является эффективным инструментом для разработки маркетинговых стратегий и продвижения продуктов на рынке.

Метод основан на следующих пяти вопросах:

– who (кто) - кто является вашей целевой аудиторией? Кто является вашим идеальным покупателем? Этот вопрос помогает определить демографические, социальные и психографические характеристики вашей аудитории, такие как возраст, пол, доход, образование, интересы, мнения и т.д;

– what (что) - что именно вы предлагаете? Какие продукты или услуги вы предоставляете? Этот вопрос помогает понять, какие конкретные потребности или проблемы решает ваш продукт или услуга;

– when (когда) - когда ваша целевая аудитория обычно покупает ваш продукт или услугу? В какие моменты времени это происходит? Этот вопрос помогает понять, какие факторы могут повлиять на решение

покупателя о приобретении вашего продукта [27];

– where (где) - где обычно происходят покупки вашего продукта или услуги? Этот вопрос помогает определить каналы распространения продукта или услуги, а также места, где они могут быть наиболее эффективно продвинуты;
– why (почему) - почему ваша целевая аудитория должна купить ваш продукт или услугу? Какие преимущества вы предоставляете по сравнению с конкурентами? Этот вопрос помогает понять, какой уникальный продуктовый пропозишн (USP) может быть использован для привлечения целевой аудитории и увеличения конверсии.

Таким образом, метод сегментирования «5W» Марка Шеррингтона позволяет получить детальную информацию о целевой аудитории и использовать ее для настройки маркетинговых кампаний и продвижения продуктов на рынке [4], [22].

Эта метод является очень простой, потому что не требует специальных навыков. Возьмём, например, целевую аудиторию по направлениям студии танца и поделим её на возможное количество частей.

В таблице 1 отображена сегментация данных в виде опроса для танцевальной студии.

Таблица 1 – Сегментация в виде опроса «5W»

Вид танца			
Вопросы	Фитнес на пилоне	Свадебный вальс	Детские танцы
Что?	Фитнес	Вальс	Детский танец
Кто?	Девушки 26-36 лет	Пары, готовящиеся к свадьбе 21-31 лет	Родители детей 7-11 лет
Почему?	Желание хорошо выглядеть	Танец жениха и невесты	Физическое развитие ребёнка

Продолжение таблицы 1

Вид танца			
Вопросы	Фитнес на пилоне	Свадебный вальс	Детские танцы
Когда?	Каждый год	Каждый год	Преимущественно в учебное время года, с сентября по май
Где?	Тематические сообщества в соцсетях, фитнес-залы, салоны красоты,	Тематические сообщества в соцсетях, банкетные залы, салоны свадебной одежды	Тематические сообщества в соцесетях, секции, поликлиники, школы

Также существует ещё один метод сегментации, который называется «KHRAMATRIX».

Методика «KHRAMATRIX» является развитием метода «5W», который помогает более глубоко разобраться в поведении клиентов. Этот метод был разработан Евгением Храмовым и позволяет более точно определить характеристики целевой аудитории. Он состоит из четырех основных пунктов:

- описание ЦА по географическим и демографическим критериям - этот пункт включает в себя описание целевой аудитории по местоположению и демографическим характеристикам, таким как возраст, пол, социальный статус, образование и доход;
- поведенческие особенности - этот пункт описывает особенности поведения целевой аудитории, такие как интересы, привычки, стиль жизни и ценности;
- готовность к покупке: этот пункт определяет, насколько целевая аудитория готова к совершению покупок. В рамках этого пункта можно выделить несколько уровней готовности к покупке, такие как готовность приобрести, сбор информации, сравнение вариантов и желание попробовать;
- целевое действие, которое подталкивает к покупке: этот пункт

определяет, какие конкретные действия приводят к совершению покупки целевой аудиторией. Например, это может быть посещение сайта, регистрация на нем, оформление заказа или оплата товара.

В таблице 2 отражен метод сегментации в виде опроса «KHRAMATRIX».

Таблица 2 – Метод KHRAMATRIX

Вид танца			
Вопросы	Фитнес на пилоне	Свадебный вальс	Детские танцы
Что?	Фитнес	Вальс	Детский танец
Кто?	Девушки 26-36 лет	Пары, готовящиеся к свадьбе 21-31 лет	Родители детей 7-11 лет
Почему?	Желание хорошо выглядеть	Танец жениха и невесты	Физическое развитие ребёнка
Когда?	Каждый год	Каждый год	Преимущественно в учебное время года, с сентября по май
Где?	Тематические сообщества в соцсетях, фитнесзалы, салоны красоты,	Тематические сообщества в соцсетях, банкетные залы, салоны свадебной одежды	Тематические сообщества в соцесетях, секции, поликлиники, школы
Описание ЦА	Девушки, которые, часто посещают салоны красоты, много времени проводят в соцсетях, следят за питанием и фигурой. Интересуются, но не могут решиться. Уровень дохода средний	Пара, которая готовится к свадьбе. Уровень дохода средний и выше.	Родители работающие. Уровень дохода средний.

Продолжение таблицы 2

Вид танца			
Вопросы	Фитнес на пилоне	Свадебный вальс	Детские танцы
Поведенческие особенности	Посещает разные места для молодежи: бары и кафе.	Пара в отношениях, посещают соответствующие места. Кинотеатры, рестораны, парки и т.д.	Жизнь в быстром темпе: ребёнок, дом, работа, школа.
Готовность купить	Сравнивает разные студии	Готовы к покупке	Анализируют данные, сравнивают
Целевое действие	Пригласить вступить в сообщество в соцсети.	Совершить продажу и предложить скидку друзьям.	Пригласить на пробное занятие

Также существует ещё один популярный метод для сегментации целевой аудитории - этот метод называется LTV.

LTV (Lifetime Value) – это метод сегментации целевой аудитории, который основывается на уровне пожизненной ценности клиента и на том, сколько дохода он принесет. При использовании этого метода, клиенты делятся на три категории: эконом-, средний и VIP-класс. Определяются также другие характеристики клиентов, такие как уровень дохода, объемы покупок, приверженность к марке и лояльность в целом [6], [28].

Этот метод позволяет оптимизировать расходы на рекламные кампании, планируя бюджет на основе того, сколько дохода принесет каждый клиент, и будет ли это выгодно. Например, на привлечение VIP-клиентов можно потратить больше финансов, чем на эконом-класс.

Для расчета методики LTV можно использовать несколько показателей. В первом случае, LTV рассчитывается как разница между доходом за год и

затратами на рекламу, деленная на количество клиентов. Во втором случае, LTV определяется как средний чек клиента, умноженный на количество покупок за определенный период времени. В третьем случае, LTV рассчитывается как средний чек, умноженный на число покупок клиента в месяц.

Методика LTV является популярным и эффективным методом для сегментации целевой аудитории, который позволяет компаниям оптимизировать расходы на рекламные кампании и лучше понимать, какие категории клиентов приносят больше прибыли [29].

Сегментация по методу LTV представлена в таблице 3.

Таблица 3 – Сегментация по методу LTV

Показатели	Сегмент 1	Сегмент 2	Сегмент 3
Средний чек	4150	1200	2600
Число продаж клиенту на месяц	3	4	3
Время удержания клиента (в месяцах)	10	4	1

Группа 1. 4 150 * 3 * 10 = 124 500 руб.

Группа 2. 1 200 * 4 * 4 = 19 200 руб.

Группа 3. 2 600 * 3 * 1 = 7 800 руб.

«В конце все данные перемножаются. Берутся средние показатели за год и рассчитываем на человека, усреднено.

После обработки данных видно, что первая группа наиболее интересна для бизнеса. Именно в этот сегмент нужно вкладывать максимум усилий» [5].

Таким образом, в данном разделе были рассмотрены различные методы и подходы к анализу целевой аудитории сайта, такие как разбивка на категории по демографическим, интересным и поведенческим факторам, а

также методы LTV, KHRAMATRIX и 5W.

Анализ целевой аудитории является важным этапом в разработке и оптимизации сайта, так как он позволяет лучше понимать потребности и предпочтения пользователей, что, в свою очередь, может привести к улучшению пользовательского опыта и увеличению конверсии.

Каждый из этих методов имеет свои преимущества и недостатки, а также может быть использован в комбинации с другими методами для получения более точных результатов и глубокого понимания характеристик целевой аудитории [7],[8].

Так же стоит добавить, что почти для такого вида сегментации необходимо проводить опрос или же иметь какие-то конкретные данные что усложняет процесс деления на группы пользователей.

Выводы по главе 1

В рамках первой главы было проведено исследование современного состояния систем анализа целевой аудитории, а именно были рассмотрены такие методы сегментации как LTV (Customer Lifetime Value), KHRAMATRIX и 5W, а также был дан обзор и анализ источников по данной теме.

LTV (Customer Lifetime Value) - это метрика, оценивающая общую прибыль, которую компания может получить от одного клиента на протяжении всего периода сотрудничества с ним. Она позволяет определить самых ценных клиентов и разработать стратегии, направленные на их удержание и максимизацию доходов.

KHRAMATRIX - это метод сегментации аудитории на основе комплексного анализа поведенческих, психографических и демографических характеристик пользователей. Он позволяет создавать более точные и детализированные портреты целевых сегментов, что упрощает разработку индивидуальных маркетинговых стратегий.

5W - это метод сегментации, основанный на анализе пяти основных вопросов: Who (Кто?), What (Что?), When (Когда?), Where (Где?) и Why

(Почему?). Он позволяет выявить общие закономерности и различия между пользователями, а также определить ключевые параметры, влияющие на их решения и поведение.

В процессе исследования данных методов сегментации пользователей, таких как KHRAMATRIX, LTV и 5W, было выявлено, что текущие методы имеют ряд недостатков. Одним из таких недостатков является ручное разделение на группы пользователей, что может быть трудоемким и подвержено ошибкам. Кроме того, эти методы предполагают наличие определенных обязательных параметров для сегментации, что может быть неудобным для быстрого разделения пользователей на группы. В связи с этим, возникает потребность в разработке и применении более автоматизированных и гибких методов сегментации. Такие методы должны позволять учитывать индивидуальные особенности пользователей и быстро адаптироваться к изменениям в их поведении и предпочтениях.

Автоматизация процесса сегментации позволит упростить и ускорить этот процесс, а гибкие методы сегментации смогут работать с разнообразными данными и учитывать разные потребности пользователей. Разработка новых методов может привести к повышению эффективности и точности сегментации, что, в свою очередь, способствует оптимизации маркетинговых стратегий и улучшению обслуживания клиентов.

Глава 2 Выбор алгоритма для сегментации целевой аудитории сайта

2.1 Технология интеллектуального анализа данных

В процессе развития информационных технологий, а также систем сбора и хранения данных - баз данных (databases), хранилищ данных (data warehousing), и с недавних пор, облачных репозиториев, возникла проблема анализа больших объемов данных, когда аналитик или управленец не в состоянии вручную обработать большие массивы данных и принять решение. Понятно, что аналитику необходимо каким-то образом представить исходную информацию в более компактном виде, с которой может справится человеческий мозг за приемлемое время.

Выделим несколько уровней информации:

- исходные данные (сырые данные, исторические данные или просто данные) – необработанные массивы данных, получаемые в результате наблюдения за некой динамической системой или объекта и отображающие его состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год);
- информация – обработанные данные, которые несут в себе некую информационную ценность для пользователя; сырые данные, представленные в более компактном виде (например, результаты поиска);
- знания — несут в себе некое ноу-хау, отображают скрытые взаимосвязи межу объектами, которые не являются общедоступными (в противном случае, это будет просто информация); данные с большой энтропией (или мерой неопределенности);
- исторически сложилось, что у термина Data Mining есть несколько вариантов перевода (и значений);
- извлечение, сбор данных, добыча данных (еще используют Information Retrieval или IR);

– извлечение знаний, интеллектуальный анализ данных (Knowledge Data Discovery или KDD, Business Intelligence).

IR оперирует первыми двумя уровнями информации, соответственно, KDD работает с третьим уровнем. Если же говорить о способах реализации, то первый вариант относится к прикладной области, где главной целью являются сами данные, второй — к математике и аналитике, где важно получить новое знание из большого объема уже имеющихся данных. Чаще всего извлечение данных (сбор) является подготовительным этапом для извлечения знаний (анализ).

С помощью методов Data minig можно найти ранее не известные знания, например найти неочевидных закономерностей в большом массиве данных. Рассмотрим Data minig в области извлечении знаний подробнее, а также рассмотрим сферы, где они используются.

Сферы применения Data minig:

– анализ покупательской корзины;

– сегментация клиентов;

– электронная коммерция.

«Data minig часто используется в банковских сферах, например, для выявления мошенничества с кредитными картами. Для выявления подозрительных операций с кредитными картами применяются подозрительные стереотипы поведения определяемы в результате анализа банковских транзакции, которые в последствие оказались мошенническими.

Также банк может разбивать клиентов при помощи инструментов Data Mining на различные группы, банк имеет возможность сделать свою маркетинговую политику более целенаправленной, а потому - эффективной, предлагая различным группам клиентов именно те виды услуг, в которых они нуждаются»[5].

«Методы интеллектуального анализа можно использовать и для определения клиентов, которые один раз воспользовались услугами данного кампании с большой доли вероятности останутся и верными. В итоге средств

на маркетинг нужно тратить там, где отдача больше всего» [15].

«По некоторым экспертным оценкам стоимость приобретение нового клиента в 5-10 раз превышает удержание нынешнего, по тем же оценкам стоимость возврата ушедшего клиента превышает 50-100 раз удержанию нынешнего клиента»[15].

Поэтому методы Data Mining помогают определять тех прибыльных клиентов, которые собираются уйти, а значит своевременно вести на тех клиентов компанию предназначенную для их удержания.

Подводя итог можно описать основные задачи, решаемые Data Mining:

- классификация - отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов;
- кластеризация - разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга;
- сокращение описания - для визуализации данных, упрощения счета и интерпретации, сжатия объемов собираемой и хранимой информации.
- ассоциация - поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя»;
- прогнозирование - нахождение будущих состояний объекта на основании предыдущих состояний (исторических данных);
- анализ отклонений - например, выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы;
- визуализация данных - представления информации в виде графиков, диаграмм, графов и других графических элементов, для упрошенного нахождение основных закономерностей.

В этой разделе была описана технология интеллектуального анализа данных, а именно Data Minig. Проведём обзор и анализ её методологий в следующем разделе.

2.2 Обзор и анализ методологий технологий Data Mining

В данном разделе приведём обзор и анализ алгоритмов, которые используются в технологии Data Mining, которые в дальнейшем будут использоваться для сегментации целевой аудитории сайта.

«К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы; методы поиска ассоциативных правил, в том числе алгоритм Apriori; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов»[1].

Рассмотрим метод кластеризации по подробнее. Метод кластеризации - это метод, обнаруживающий естественную группировку набора шаблонов, точек или объектов.

Цель кластеризации данных, также известной как кластерный анализ, является разделение набора данных на кластеры (также называемые группами) сходных объектов на основе меры сходства (или несходства), тем самым сводя к минимуму сходство между объектами, принадлежащими к разным кластерам и максимизация сходства между объектами, принадлежащими к одному кластеру, т. е. объекты внутри группы больше похожи друг на друга (высокое внутри кластерное сходство), чем объекты, принадлежащие к разным группам (низкое меж кластерное сходство), пример изображён на рисунке 4.

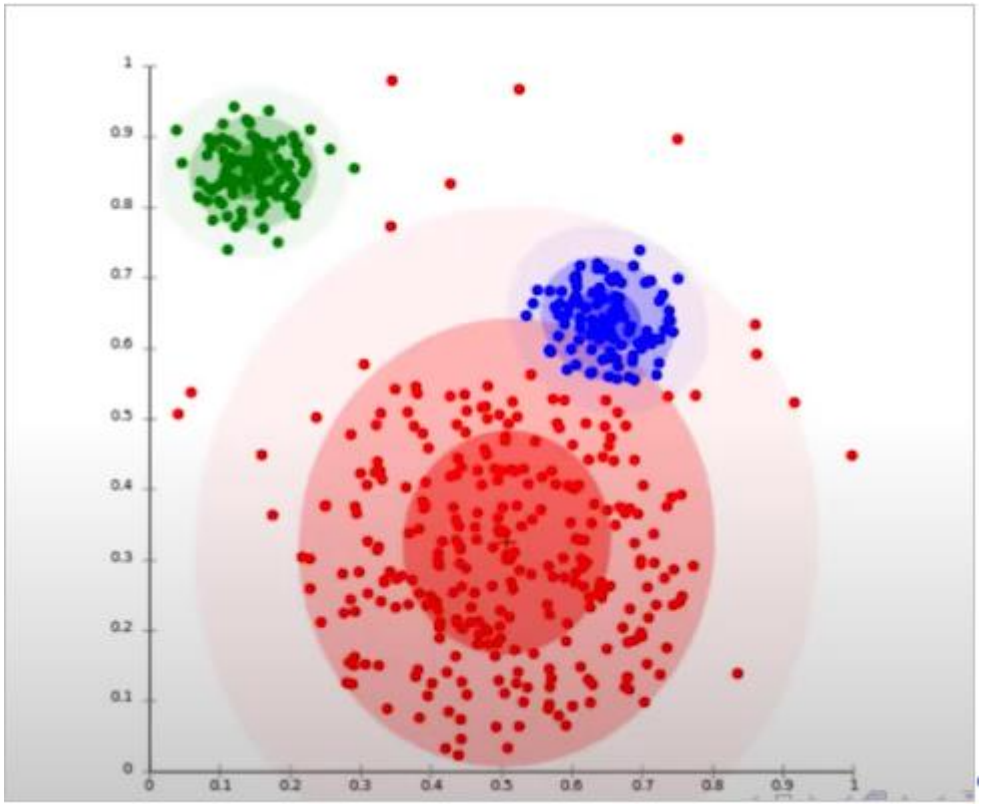

Рисунок 4 – Модель демографии

Существует несколько мер сходства (или несходства), и выбор подходящей меры зависит от анализируемых данных и цели анализа. Кластеризация успешно применяется в самых разных приложениях, например, при сегментации изображений, распознавании объектов и символов, поиске документов, дистанционном зондировании, сжатии данных и т. д.

Методы кластеризации можно разделить на две категории в зависимости от структуры абстракции, а именно, иерархическая кластеризация и групповая кластеризация.

Групповая кластеризация изображена на рисунке выше, а иерархическая – на рисунке 5.

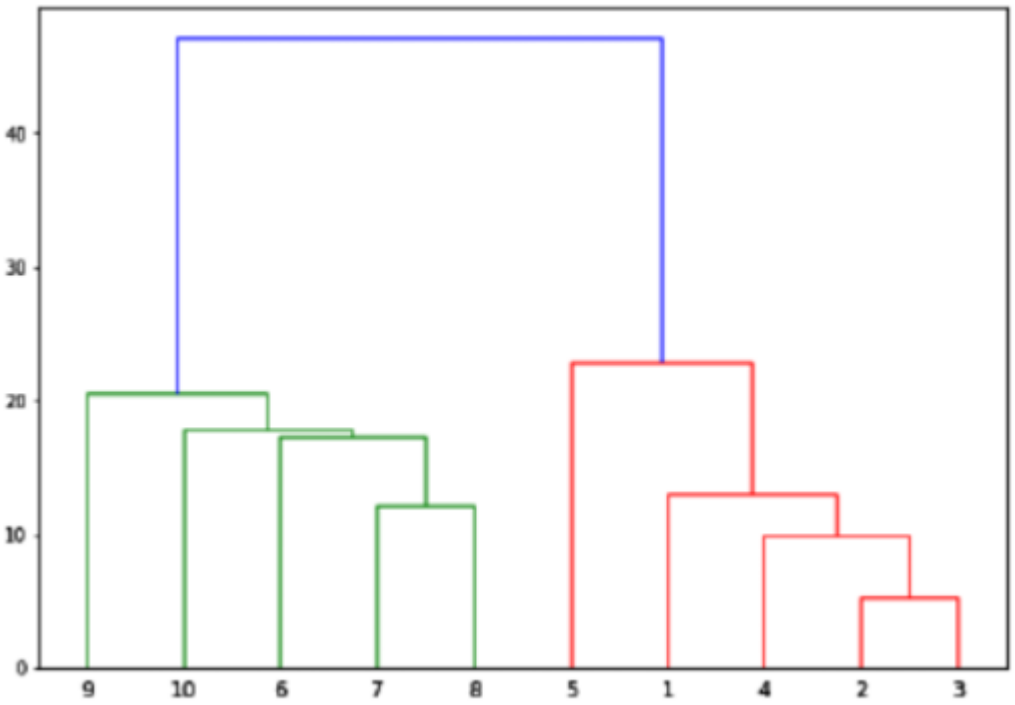

Рисунок 5 – Иерархическая кластеризация

Методы иерархической кластеризации группируют объекты данных с помощью последовательности разделов либо от одноэлементных кластеров к кластеру, содержащему все индивидуумы, либо наоборот. В результате получается иерархическая структура разделов, известная как дендрограмма, в которой каждый раздел вложен в раздел на следующем уровне иерархии. Иерархические методы могут быть как агломеративными, так и разделительными.

Алгоритмы агломерации вначале помещают каждый шаблон в отдельный кластер, а затем последовательно объединяют (или агломерируют) пары кластеров, пока все кластеры не будут объединены в один кластер, содержащий все шаблоны, или не будет достигнута желаемая цель.

Алгоритмы разделения начинаются со всех шаблонов, помещенных в один кластер, а затем продолжаются путем рекурсивного разделения кластера на более мелкие кластеры, пока не будут достигнуты отдельные шаблоны или не будет выполнен критерий остановки.

В отличие от иерархической кластеризации, групповая кластеризация распределяет набор шаблонов в заданное или предполагаемое число кластеров без иерархической структуры. Одной из важных проблем при секционной кластеризации является нахождение разделения заданных данных с заданным числом кластеров, которое минимизирует (или максимизирует) некоторую критериальную функцию. Сумма квадратов функций ошибок является одним из наиболее широко используемых критериев.

Рассмотрим применение алгоритмов групповой кластеризации. Кластеризация представляет из себя деление на классы например с помощью неё можно сжимать данные, разбивать пользователей на группы, обнаруживать аномалии, так же стоит заметить что наиболее обширное применение кластеризации это в сегментации пользователей, а именно разбиение пользователей на группы в которых будет содержатся свой качественный приближенный признак возможный отличный от остальных групп, а может и не содержатся, во всём этом поможет использование алгоритмов кластеризации.

Рассмотрим популярные и распространённые алгоритмы кластеризации.

– k-means;
– affinity propagation;
– алгомеративная кластеризация.

K-Means - один из самых простых и наиболее известных алгоритмов методов разбиения. Алгоритм K-Means основан на расстоянии, потому что сходство шаблонов вычисляется с помощью евклидова расстояния или косинусного расстояния. Алгоритмы кластеризации на основе расстояния эффективны для данных с эллипсоидальным или гиперсферическим распределением. Если границы разделения между кластерами нелинейны, алгоритмы не сработают. Один из подходов к решению этой проблемы заключается в нелинейном преобразовании входных данных в многомерное пространство признаков (то есть в едином отображении) и последующем выполнении кластеризации в пределах этого пространства признаков [9].

Также существует MiniBatchKMeans он работает по такому же принципу как и K-Means но для выборки из данных он берет случайную подвыборку данных что и сокращает нахождение центров кластеров но из-за этого уменьшается точность.

Affinity Propagation (метод распространения близости) – получает на вход матрицу схожести между элементами дата сета и возвращает набор меток, присвоенных этим элементам. График этого метода изображен на рисунке 6.

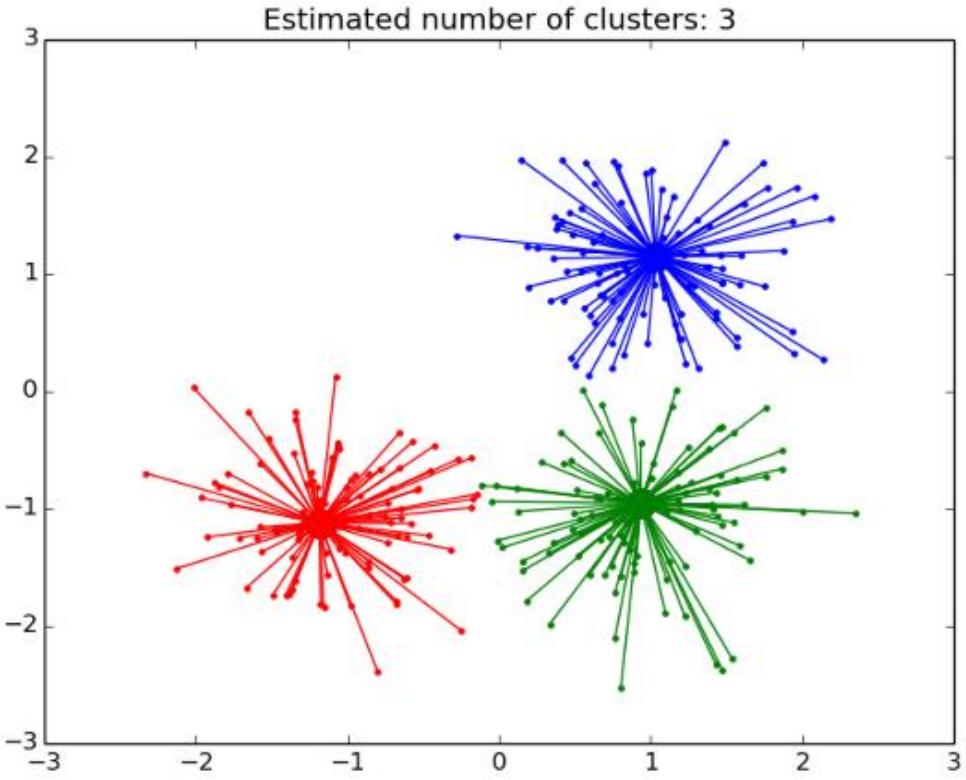

Рисунок 6 – Метод распространения близости

По данным сайта scikit-learn.org были получены данные кластеризации по разным алгоритмам, они изображены на рисунке 7.

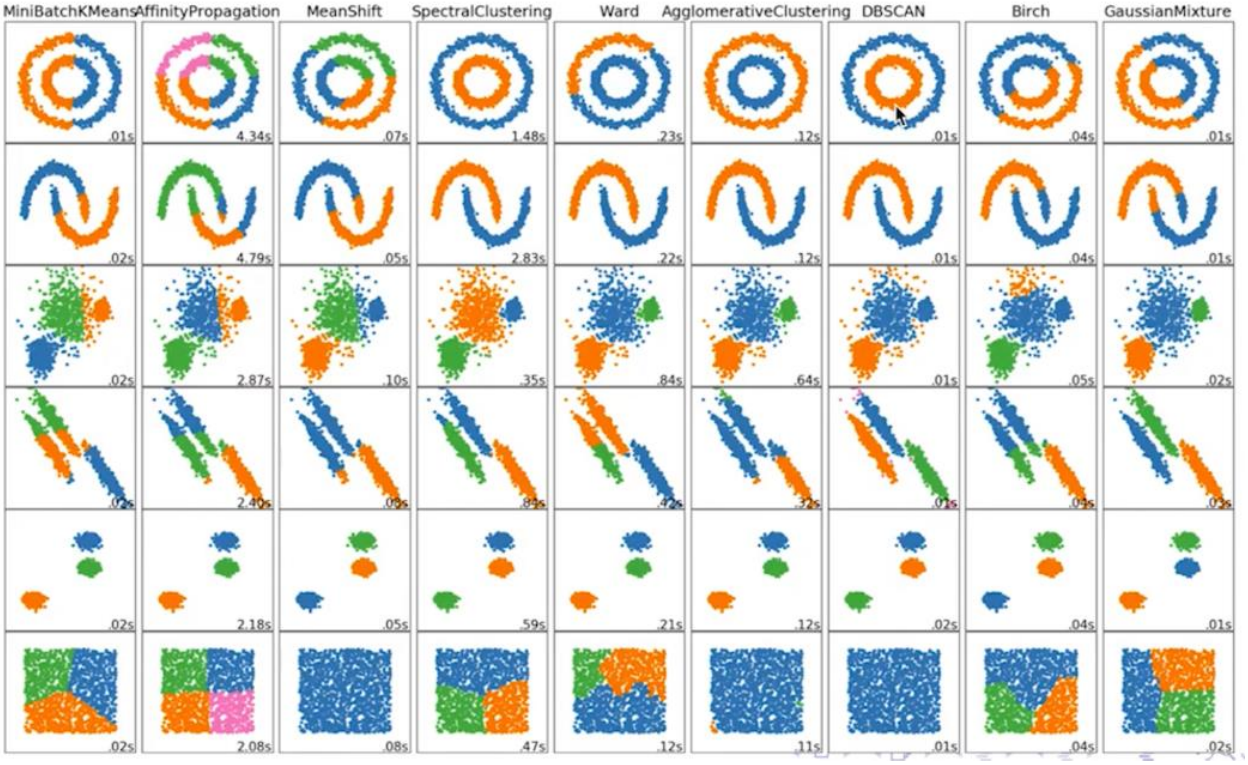

Рисунок 7 – Данные методов кластеризации

На рисунке 7 изображены данные кластеризации различных алгоритмов с одними и теме же данными. Стоит заметить, что алгоритм K-means справился хорошо только с разделение небольших групп точек, а в остальных вариантах представлен не совсем верное разделение если сравнивать визуально.

Хочется отметить, что алгоритм DBSCAN и OPTICS справились со всеми выборками почти на отлично, они даже могут отличить помехи.

Вывод по главе 2

В данном разделе был произведён обзор, анализ и сравнение разновидностей методов и алгоритмов DATA MINIG. Были определены фавориты среди алгоритмов кластеризации, которые в дальнейшем будут использоваться для сегментации целевой аудитории сайта.

Глава 3 Применение алгоритма сегментации и выбор оптимального алгоритма поиска кластеров для набора данных пользователей

3.1 Считывание набора данных

Сегментация клиентов — одно из наиболее важных применений неконтролируемого обучения. Используя методы кластеризации, компании могут идентифицировать несколько сегментов клиентов, что позволяет им ориентироваться на потенциальную пользовательскую базу она изображена на рисунке 8.

В дипломной работе будет использована кластеризацию K-Means, которая является основным алгоритмом кластеризации немаркированного набора данных [10], [11].

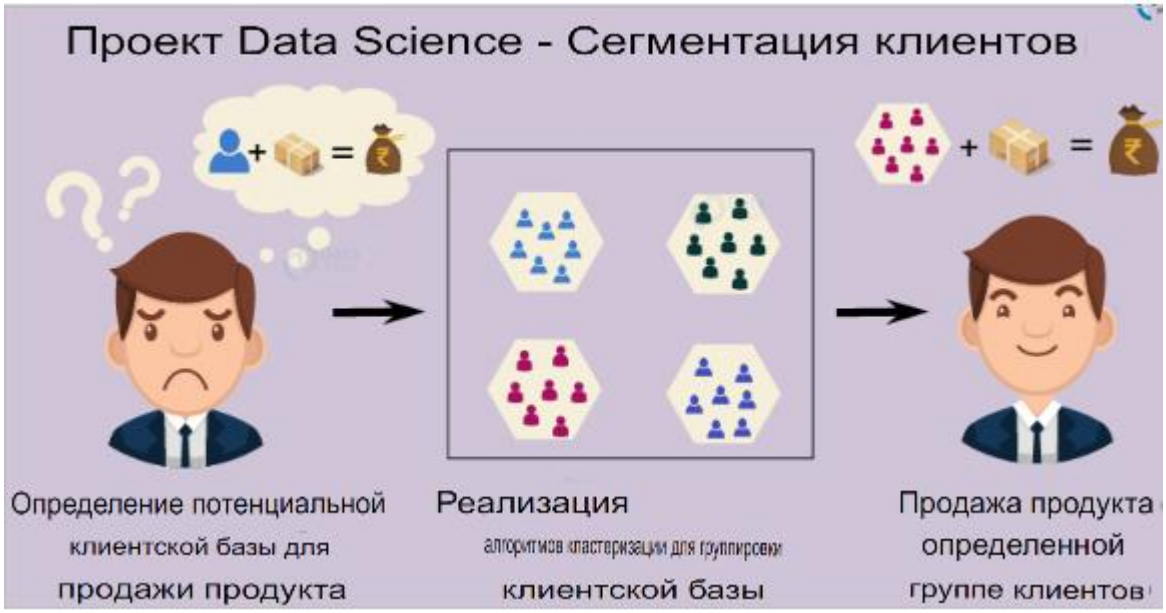

Рисунок 8 – Сегментация клиентов

Сегментация клиентов - это процесс разделения клиентской базы на несколько групп лиц, которые имеют сходство в различных отношениях, имеющих отношение к маркетингу, таких как пол, возраст, интересы и

различные привычки в отношении расходов.

Компании, внедряющие сегментацию клиентов, исходят из того, что каждый клиент имеет разные требования и требует особых маркетинговых усилий для их надлежащего удовлетворения.

Компании стремятся получить более глубокий подход к клиенту, на которого они ориентируются. Поэтому их цель должна быть конкретной и должна быть адаптирована для удовлетворения требований каждого отдельного клиента.

Кроме того, с помощью собранных данных компании могут лучше понять предпочтения клиентов, а также требования для выявления ценных сегментов, которые принесут им максимальную прибыль.

Таким образом, они могут более эффективно разрабатывать свои маркетинговые методы и минимизировать возможность риска для своих инвестиций [12].

Техника сегментации клиентов зависит от нескольких ключевых дифференциаторов, которые делят клиентов на группы, на которые нужно ориентироваться.

Данные, относящиеся к демографии, географии, экономическому статусу, а также поведенческим моделям, играют решающую роль в определении направления компании по работе с различными сегментами.

На первом этапе проведем исследование данных. Импортируем основные пакеты, а затем считываем данные, а также рассмотрим входные данные, чтобы получить о них необходимые сведения, так же выведем название столбцов, операции изображены на рисунке 9.

```
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")
str(customer_data)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1
...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

```
names(customer_data)
```

```
## [1] "CustomerID"             "Gender"
## [3] "Age"                    "Annual.Income..k.."
## [5] "Spending.Score..1.100."
```

Рисунок 9 – Считывание данных из набора

Теперь отобразим первые шесть строк нашего набора данных с помощью функции head() и используем функцию summary() для вывода его статистических данных, данные изображены на рисунке 10.

```
head(customer_data)
```

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                     39
## 2          2   Male  21                 15                     81
## 3          3 Female  20                 16                      6
## 4          4 Female  23                 16                     77
## 5          5 Female  31                 17                     40
## 6          6 Female  22                 17                     76
```

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

Рисунок 10 – Вывод данных и получение статистических данных столбца

На рисунке отражено среднее минимальное и максимально минимальное значение столбца age.

3.2 Визуализация пола клиента

Далее создадим гистограмму и круговую диаграмму, чтобы показать гендерное распределение в наборе данных customer_data. Рисунок 11 и 12.

```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

Рисунок 11 – Программный код создание диаграммы

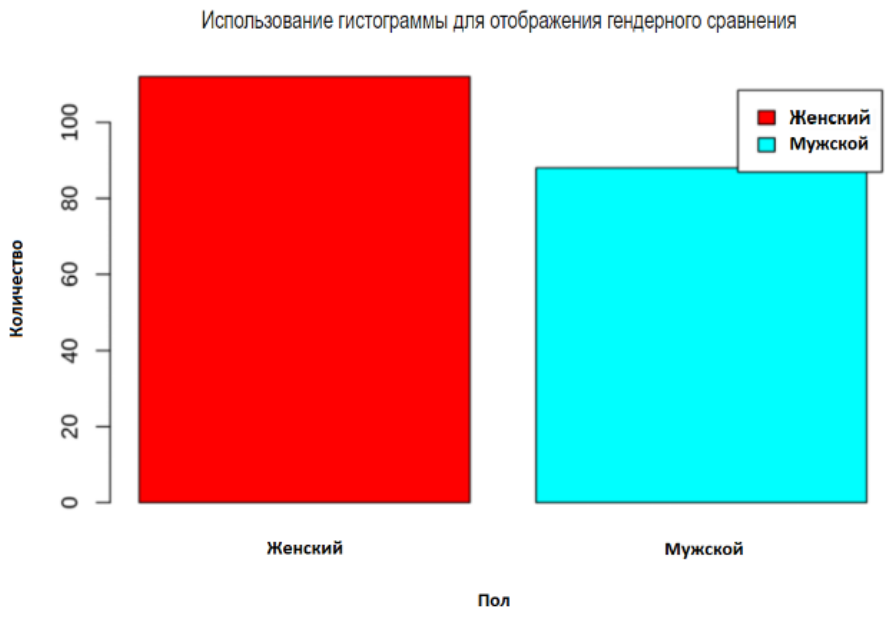

Рисунок 12 – Диаграмма пола

Из приведенной выше гистограммы видим, что количество женщин

выше, чем мужчин [14].

Гистограмма — это графическое представление распределения частоты встречаемости значений в наборе данных. Она состоит из столбцов, высота которых соответствует количеству наблюдений, попадающих в каждый из интервалов значений. Гистограмма позволяет быстро оценить форму распределения данных, выявить выбросы, аномалии и неравномерность распределения [16], [17].

Теперь давайте визуализируем круговую диаграмму, чтобы наблюдать соотношение распределения мужчин и женщин (рисунки 13 и 14).

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
     main="Pie Chart Depicting Ratio of Female and Male")
```

Рисунок 13 – Программный код круговой диаграммы

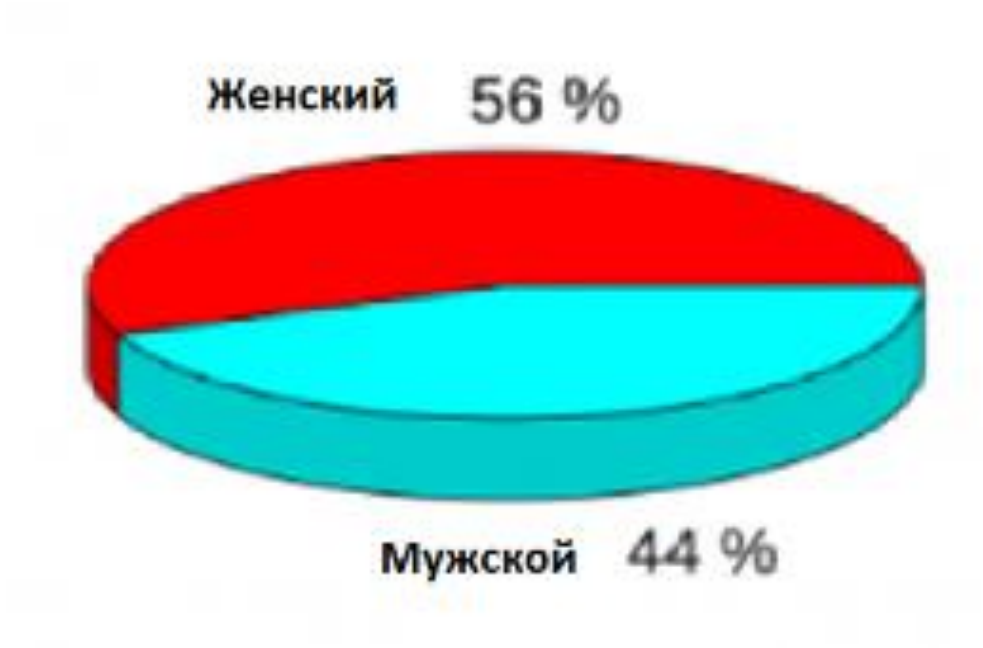

Рисунок 14 – Круговая диаграмма

Круговая диаграмма - это графическое представление данных, которое используется для отображения соотношения различных категорий или долей в целом. Она представляет собой круг, разбитый на секторы, размер каждого сектора соответствует относительному размеру категории или доли, которую она представляет.

Из приведенного выше графика можно сделать вывод, что процент женщин составляет 56% , тогда как процент мужчин в наборе данных клиентов составляет 44%.

3.3 Визуализация возрастного распределения

Построим гистограмму возрастного распределения, чтобы просмотреть распределение частоты возрастов клиентов.

Возьмём выборку по переменной Age (рисунки 15-18).

```
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
```

Рисунок 15 – Программный код гистограммы возрастного распределения

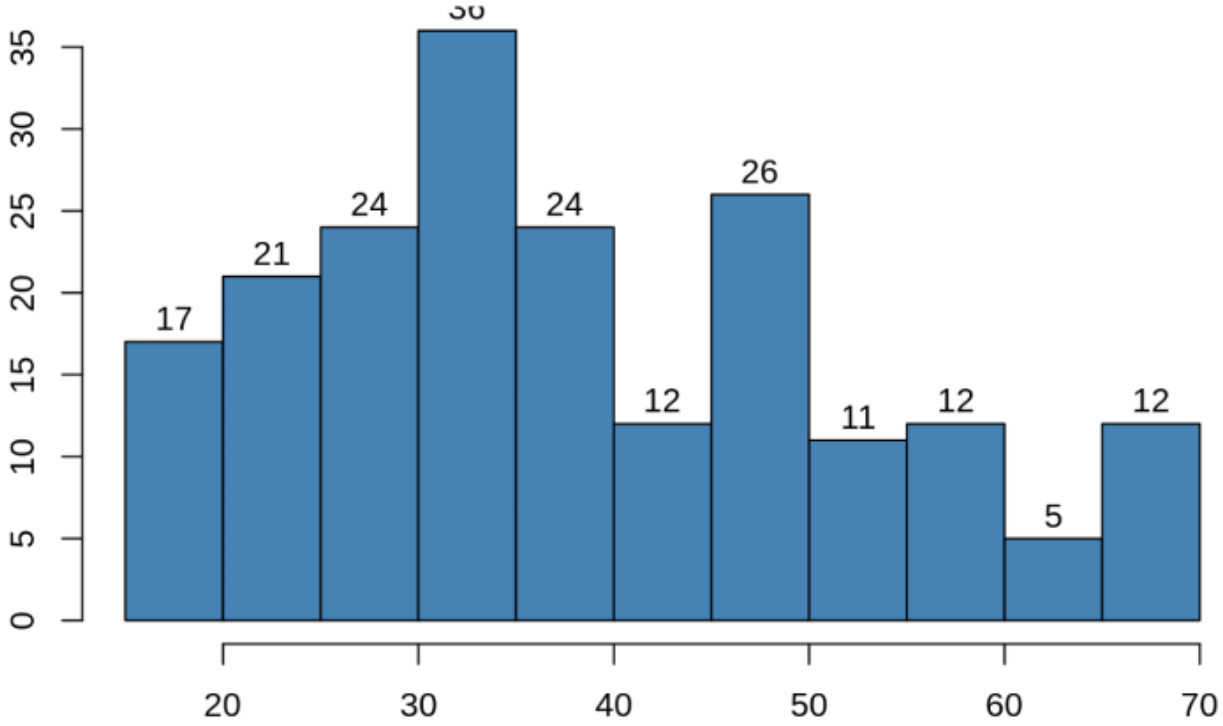

Рисунок 16 – Гистограмма возрастного распределения

На рисунке 16 видно, что набольшее количество пользователей в возрасте от 30 до 40 лет.

```
boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")
```

Рисунок 17 – Программный код коробочной диаграммы «Ящик с усами»

Коробочная диаграмма (Box plot) – это графическое представление распределения вероятностей набора данных, которое позволяет оценить основные статистические характеристики, такие как медиана, квартили, минимальное и максимальное значение и выбросы [18], [19].

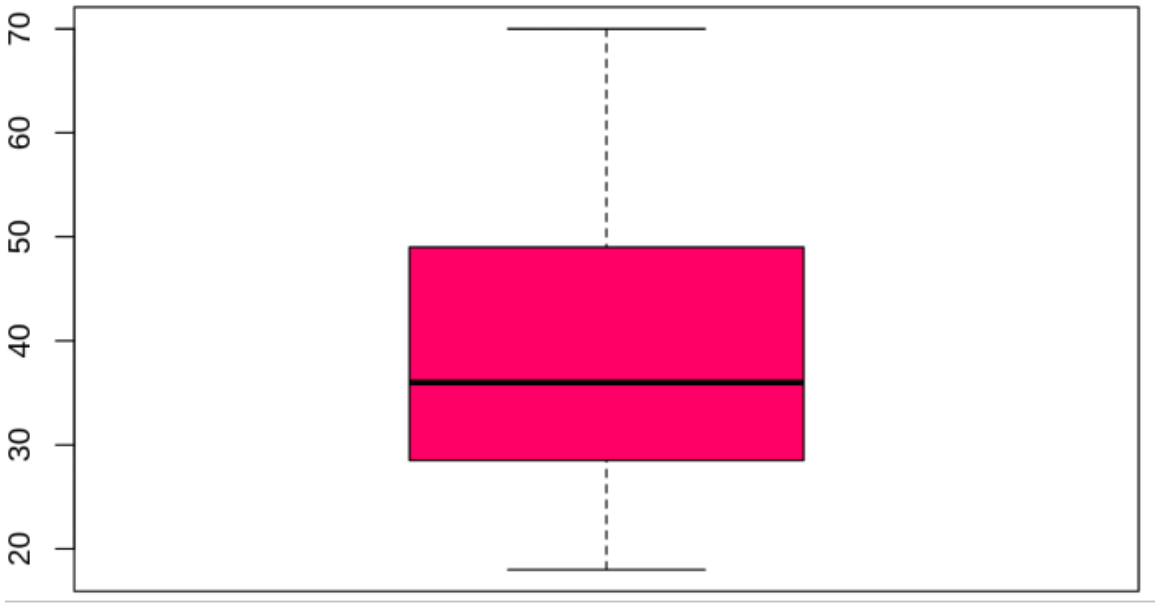

Рисунок 18 – Коробочная диаграмма «Ящик с усами» по возрасту пользователей

Коробочная диаграмма включает в себя горизонтальный прямоугольник, который представляет интерквартильный размах (между первым и третьим квартилями), вертикальную линию внутри прямоугольника, которая обозначает медиану, и две линии-усы, которые идут вверх и вниз от прямоугольника. Усы могут быть максимальным и минимальным значениями данных, которые не являются выбросами, или они могут быть определены с помощью формулы.

Из приведенных выше двух графиков можно сделать вывод, что максимальный возраст клиентов составляет от 30 до 35 лет. Минимальный возраст клиентов составляет 18 лет, тогда как максимальный возраст составляет 70 лет.

3.4 Анализ годового дохода клиентов

В этом разделе будет создана визуализации для анализа годового дохода клиентов, а также будет построена гистограмма, для анализа этих данных, используя график плотности.

Программный код и график изображены на рисунках 19 и 20 соответственно [20].

```
summary(customer_data$Annual.Income..k..)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00

hist(customer_data$Annual.Income..k..,
    col="#660033",
    main="Histogram for Annual Income",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)
```

Рисунок 19 – Программный код готового дохода клиента

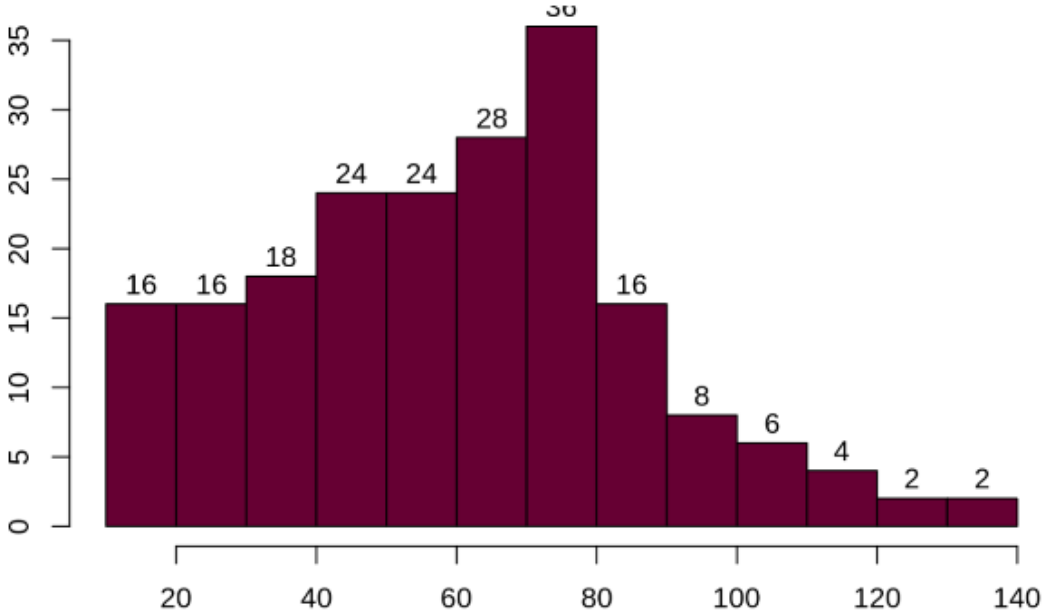

Рисунок 20 – Гистограмма готового дохода

На рисунке 21 изображён программный код графика оценки плотности по годовому доходу.

```
plot(density(customer_data$Annual.Income..k..),
     col="yellow",
     main="Density Plot for Annual Income",
     xlab="Annual Income Class",
     ylab="Density")
polygon(density(customer_data$Annual.Income..k..),
        col="#ccff66")
```

Рисунок 21 – Программный код графика оценки плотности

Ниже на рисунке 22 изображён график оценки плотности по годовому доходу пользователей.

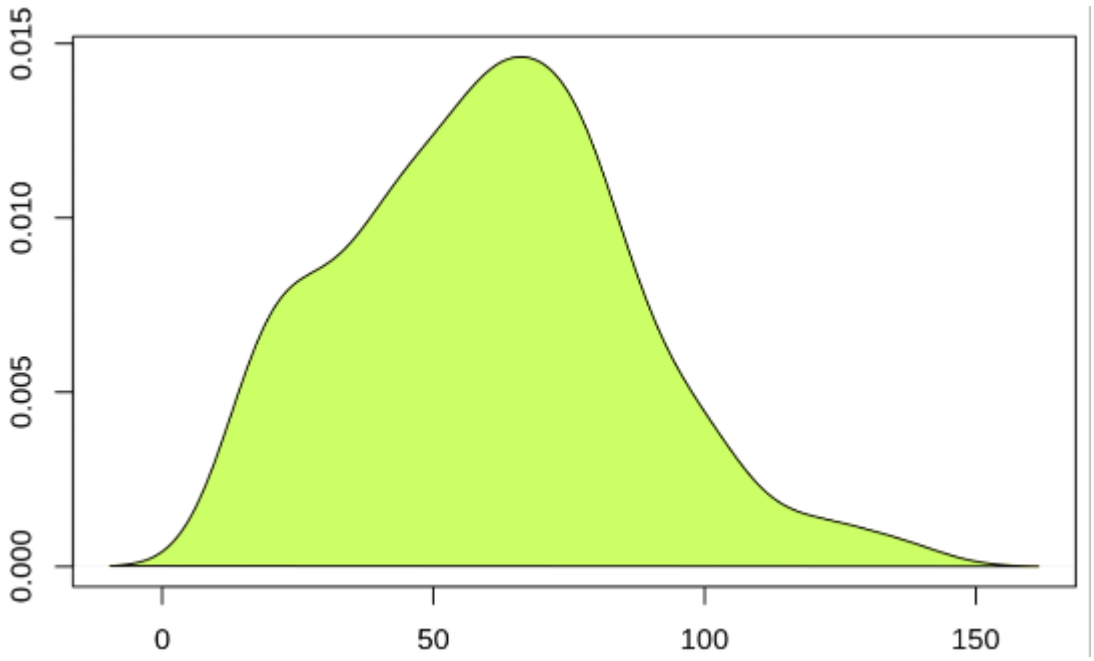

Рисунок 22 – Графике оценки плотности по готовому доходу

Анализируя графики выше, можно прийти к выводу к выводу, что минимальный годовой доход клиентов составляет 15, а максимальный доход - 137. Люди со средним доходом 70 имеют самую высокую частоту в нашем распределении гистограммы. Средняя зарплата всех клиентов 60,56. На графике оценки плотности изображенного на рисунке 22, который показан

выше, видно, что годовой доход имеет нормальное распределение.

3.5 Анализ расходов клиентов

Построим коробочную диаграмму расходов на клиентов, так называемую «ящик с усами», и также построим гистограмму, рисунок 24 и 26 соответственно. Программный код к ним представлен на рисунках 23-26.

```
summary(customer_data$Spending.Score..1.100.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.75   50.00   50.20   73.00   99.00
```

```
boxplot(customer_data$Spending.Score..1.100.,
        horizontal=TRUE,
        col="#990000",
        main="BoxPlot for Descriptive Analysis of Spending Score")
```

Рисунок 23 – Программный код коробочной диаграммы по расходам

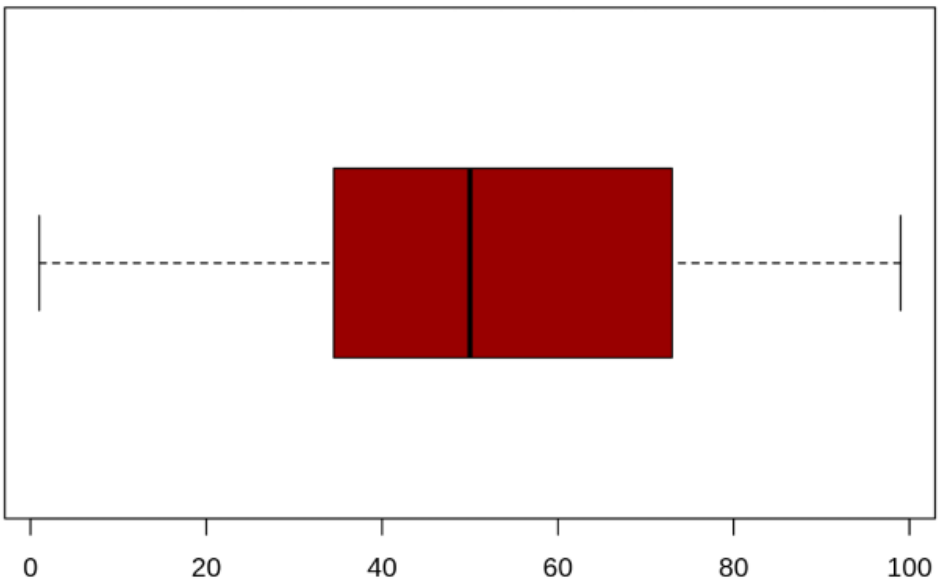

Рисунок 24 – Коробочная диаграмма по зарплате пользователей

```
hist(customer_data$Spending.Score..1.100.,
     main="HistoGram for Spending Score",
     xlab="Spending Score Class",
     ylab="Frequency",
     col="#6600cc",
     labels=TRUE)
```

Рисунок 25 – Программный код «Гистограммы» по расходам

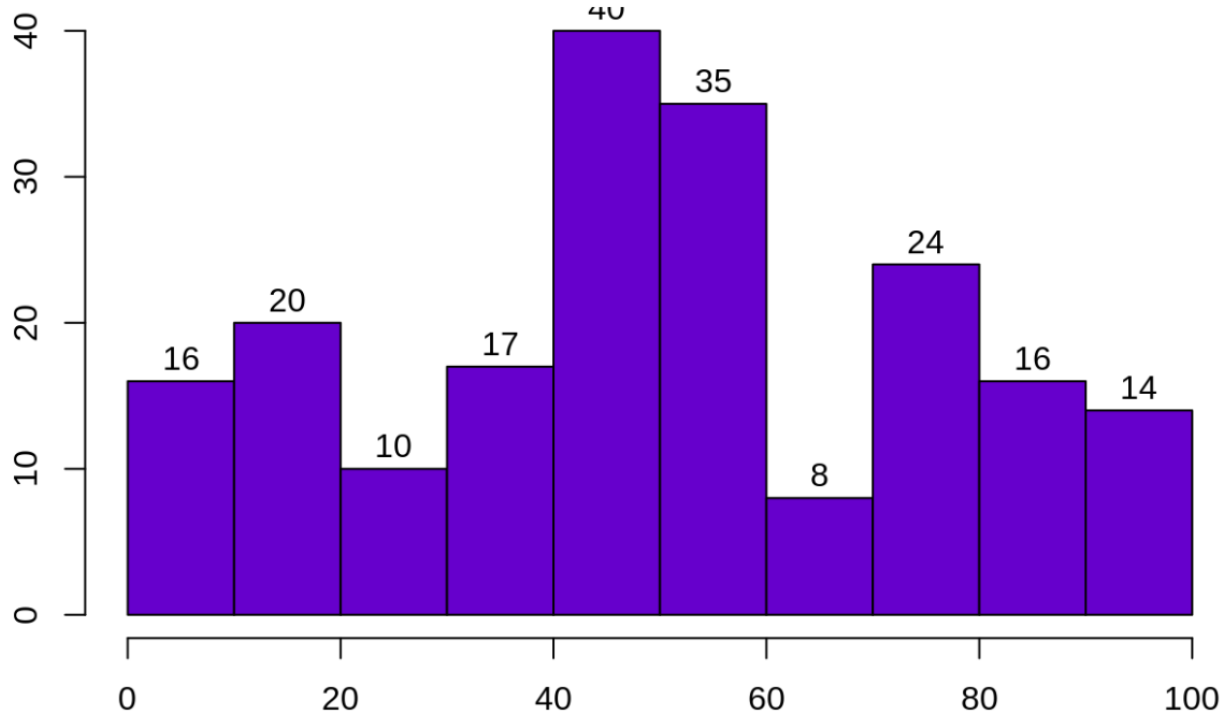

Рисунок 26 – Гистограмма по расходам

Минимальный балл расходов — 1, максимальный — 99, а средний — 50, 20. Можно заметить, что описательный анализ оценки расходов показывает, что минимальное значение равно 1, максимальное значение равно 99, а среднее значение равно 1. составляет 50, 20. Из гистограммы можно сделать вывод, что клиенты между классами 40 и 50 имеют самый высокий показатель расходов среди всех классов.

3.6 Алгоритм К-средних и поиск количества кластеров

При использовании алгоритма кластеризации k-средних первым шагом является указание количества кластеров (k), которое нужно получить в конечном результате. Алгоритм начинается со случайного выбора k объектов из набора данных, которые будут служить начальными центрами кластеров. Эти выбранные объекты являются кластерными средствами, также известными как центроиды. Затем оставшимся объектам присваивается ближайший центр тяжести. Этот центр тяжести определяется евклидовым расстоянием между объектом и средним значением кластера. Назовём этот шаг «назначением кластера». Когда назначение завершено, алгоритм переходит к вычислению нового среднего значения каждого кластера, присутствующего в данных. После пересчета центров проверяются наблюдения, не находятся ли они ближе к другому кластеру. С использованием обновленного кластерного среднего происходит переназначение объектов. Это повторяется многократно через несколько итераций, пока назначения кластера не перестанут изменяться. Кластеры, присутствующие в текущей итерации, такие же, как и кластеры, полученные в предыдущей итерации.

Подводя итог кластеризации K-Means:

- указываем количество кластеров, которые необходимо создать;
- алгоритм случайным образом выбирает k объектов из набора данных. Этот объект является начальным кластером или средним значением;
- ближайший центроид получает назначение нового наблюдения. Находим это назначение на евклидовом расстоянии между объектом и центроидом;
- у k кластеров в точках данных обновляем центр тяжести путем вычисления новых средних значений, присутствующих во всех точках данных кластера. Центроид k-го кластера имеет длину p, которая содержит средние значения всех переменных для наблюдений в k-м

кластере. Обозначим число переменных через p;

– итерационная минимизация суммы в сумме квадратов. Затем посредством итеративной минимизации общей суммы квадрата задание перестанет колебаться, когда достигнет максимальной итерации. Значение по умолчанию — 10, которое программное обеспечение R использует для максимального количества итераций.

При работе с кластерами необходимо указать количество используемых кластеров. Так же существуют три популярных метода для поиска количества кластеров:

– локтевой метод;
– метод среднего силуэта;
– статистический метод разрыва.

Метод локтя - Основная цель методов разбиения кластеров, таких как k-means, состоит в том, чтобы определить кластеры таким образом, чтобы внутри кластерная вариация оставалась минимальной.

Формула алгоритма (1):

$$minimize\left(\sum W\ (Ck)\right), k = 1 \dots k, \quad (1)$$

где Ck представляет k-й кластер, W(Ck) обозначает внутри кластерную вариацию.

Измерив общую внутри кластерную вариацию, можно оценить компактность границы кластеризации. Затем можно перейти к определению оптимальных кластеров следующим образом:

Сначала вычислить алгоритм кластеризации для нескольких значений k. Это можно сделать, создав вариацию в пределах k от 1 до 10 кластеров. Затем вычислить общую внутри кластерную сумму квадратов (iss). Затем приступить к построению iss на основе количества k кластеров. На графике расположение изгиба или колена указывает на оптимальное количество

кластеров.

На рисунке 27 описан программный код метода локтевого метода, а на рисунке 28 построен его график.

```
library(purrr)
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withins
s
}

k.values <- 1:10


iss_values <- map_dbl(k.values, iss)

plot(k.values, iss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total intra-clusters sum of squares")
```

Рисунок 27 – Программный код внутрикластерной суммы квадратов

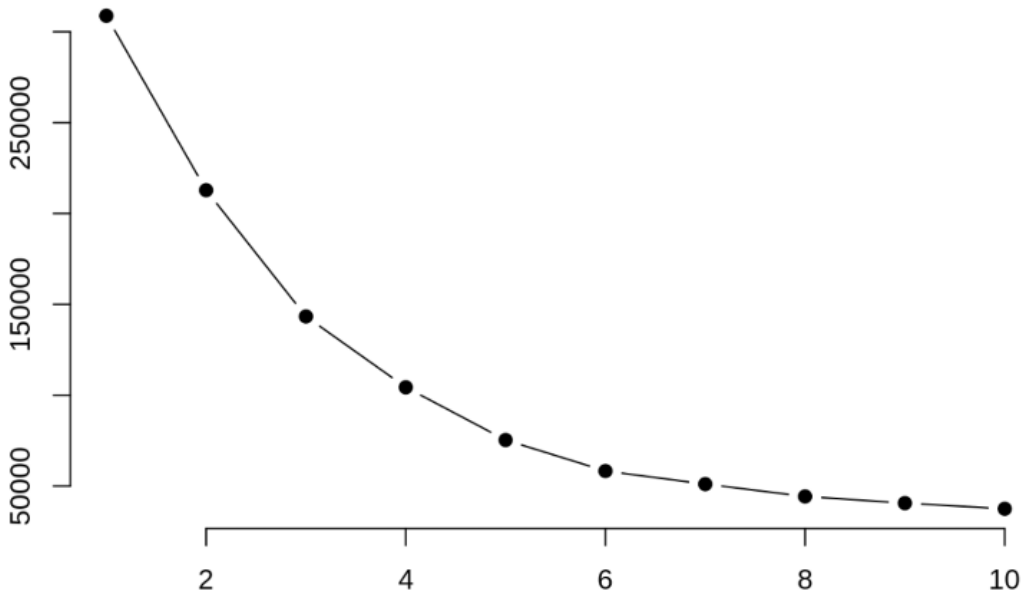

Рисунок 28 – Общая внутрикластерная сумма квадратов

Из приведенного выше графика можно сделать вывод, что 4 — это

подходящее количество кластеров, так как на графике видно, что количество кластеров появляются на изгибе локтевого графика.

Метод среднего силуэта - с помощью метода среднего силуэта можно измерить качество операции кластеризации. Благодаря этому можно определить, насколько хорошо внутри кластера находится объект данных. Если получаем высокую среднюю ширину силуэта, это означает, что получилась хорошая кластеризация. Метод среднего силуэта вычисляет среднее значение наблюдений силуэта для различных значений k. При оптимальном количестве k-кластеров можно максимизировать средний силуэт по значимым значениям для k кластеров.

Используя функцию силуэта в пакете кластера, можно вычислить среднюю ширину силуэта, используя функцию kmean. Здесь оптимальный кластер будет иметь наивысшее среднее значение.

Получим значения средней ширины силуэта от двух до десяти кластеров (рисунки 29-38).

```
library(cluster)
library(gridExtra)
library(grid)


k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))
```

Рисунок 29 – Программный для нахождения средней ширины силуэта

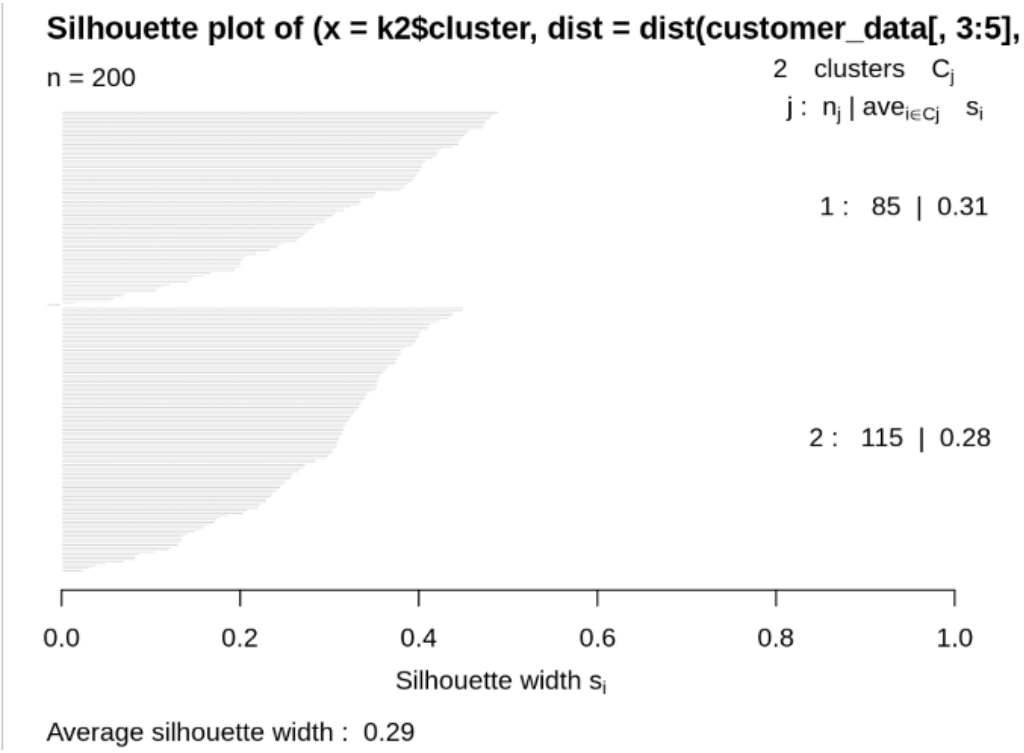

Рисунок 30 – График средней ширины для 2 кластеров

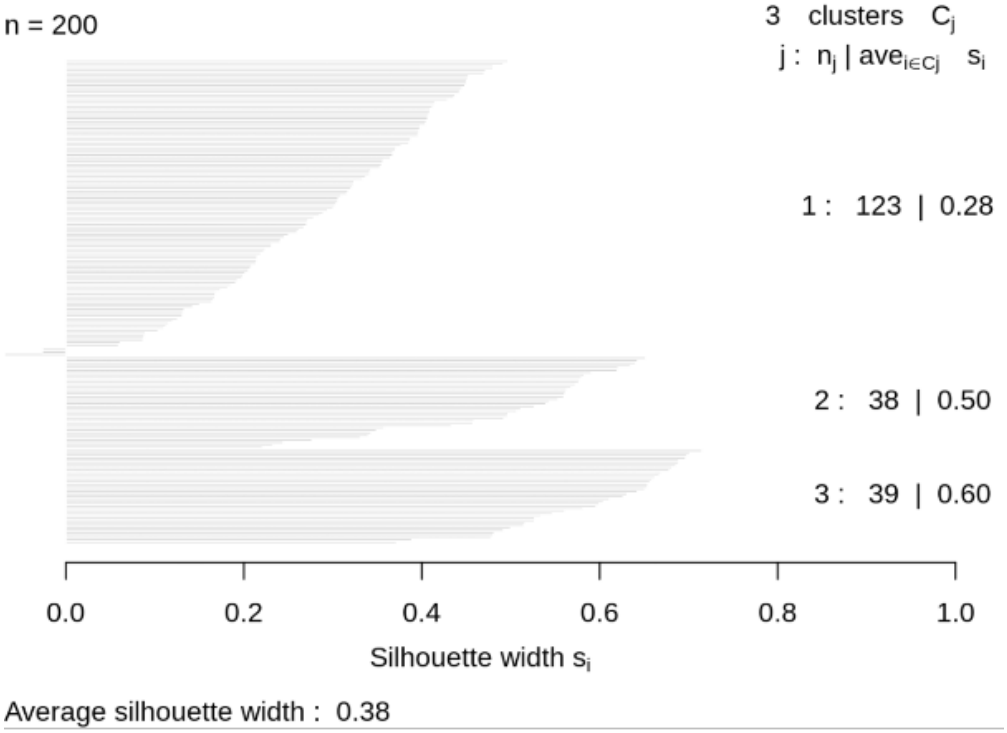

Рисунок 31 – График средней ширины для 3 кластеров

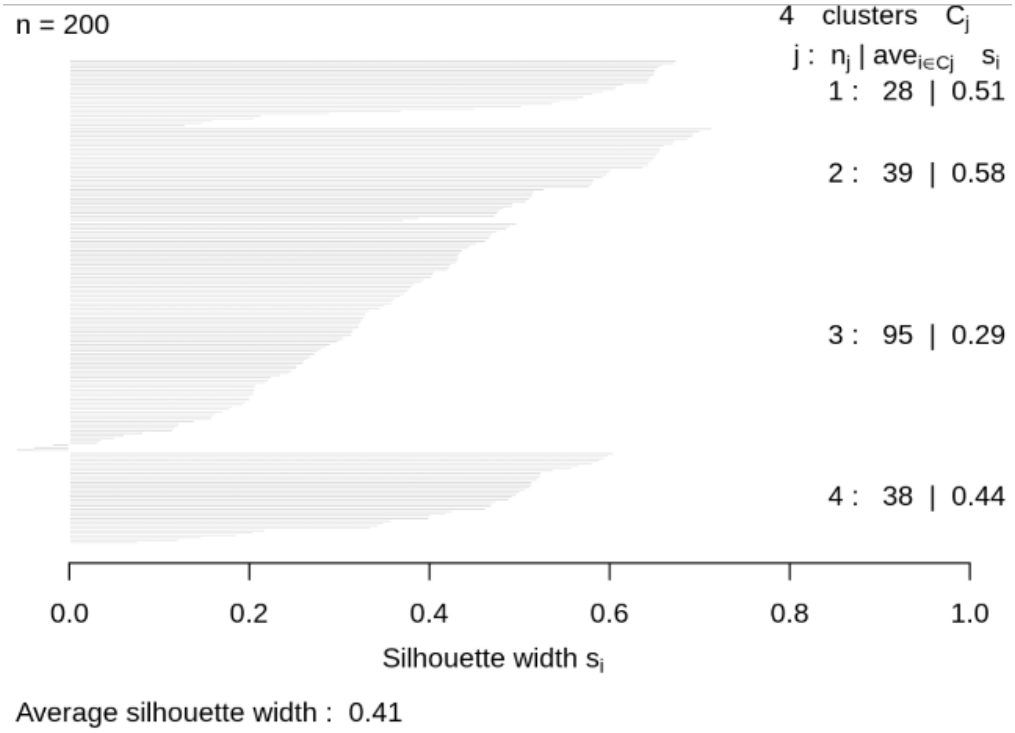

Рисунок 32 – График средней ширины для 4 кластеров

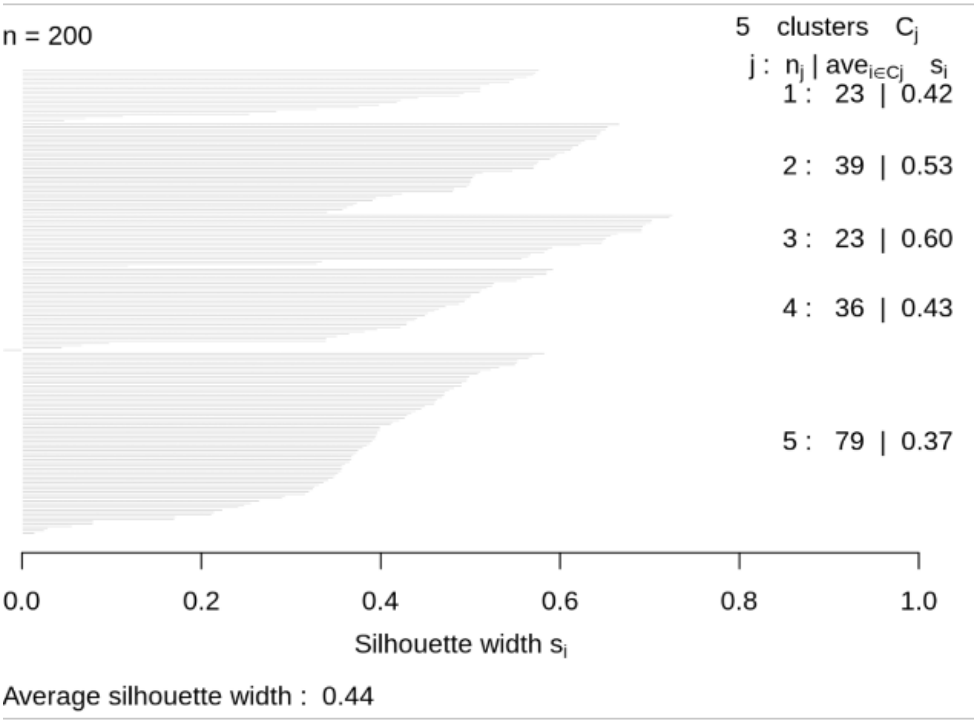

Рисунок 33 – График средней ширины для 5 кластеров

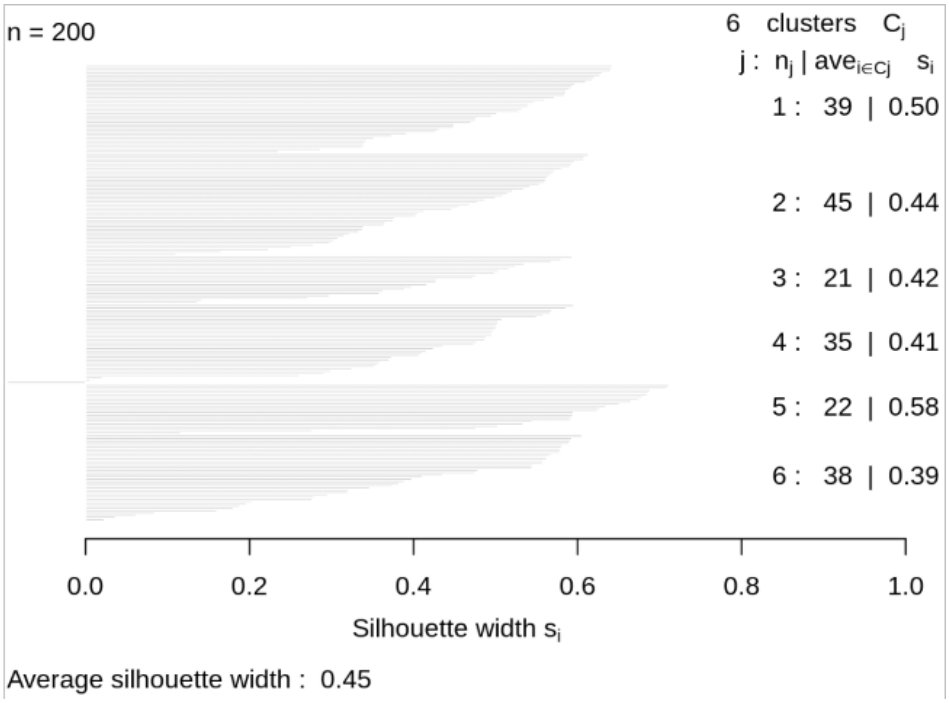

Рисунок 34 – График средней ширины для 6 кластеров

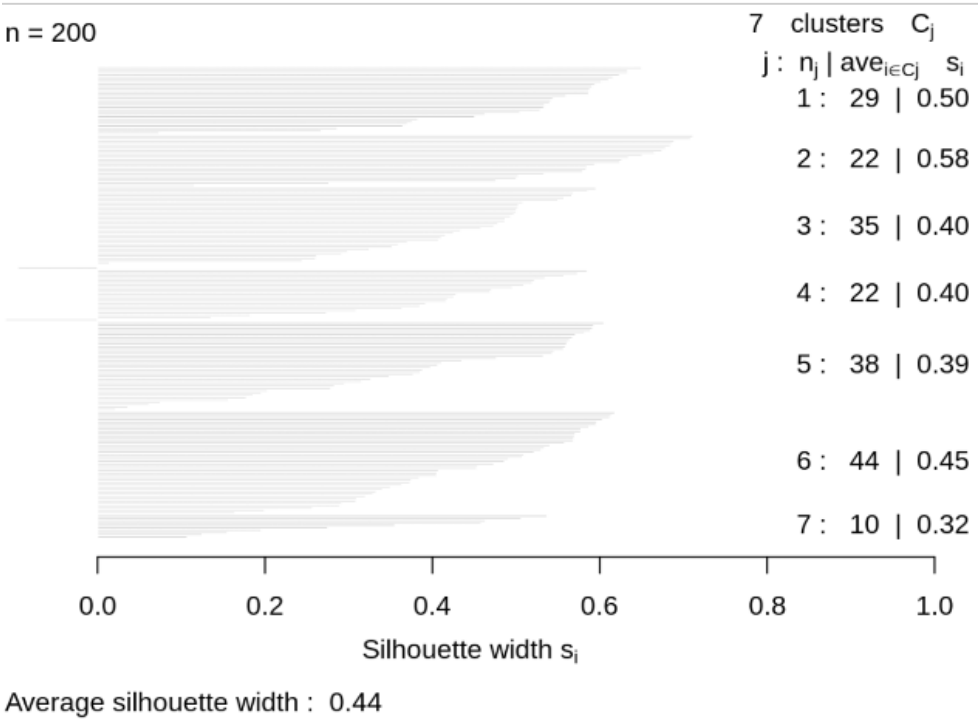

Рисунок 35 – График средней ширины для 7 кластеров

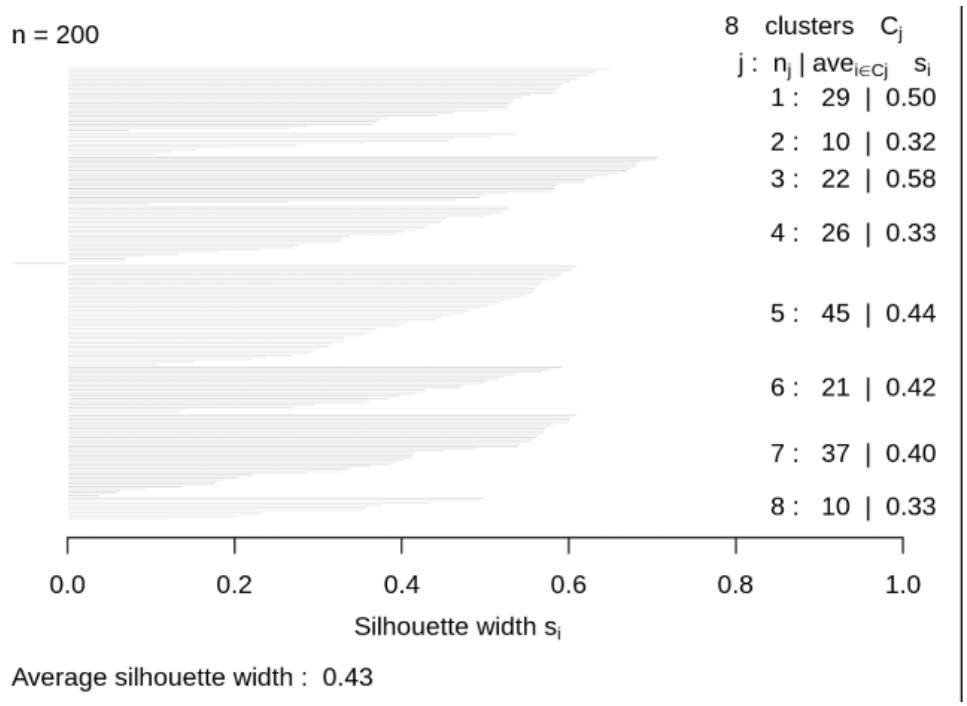

Рисунок 36 – График средней ширины для 8 кластеров

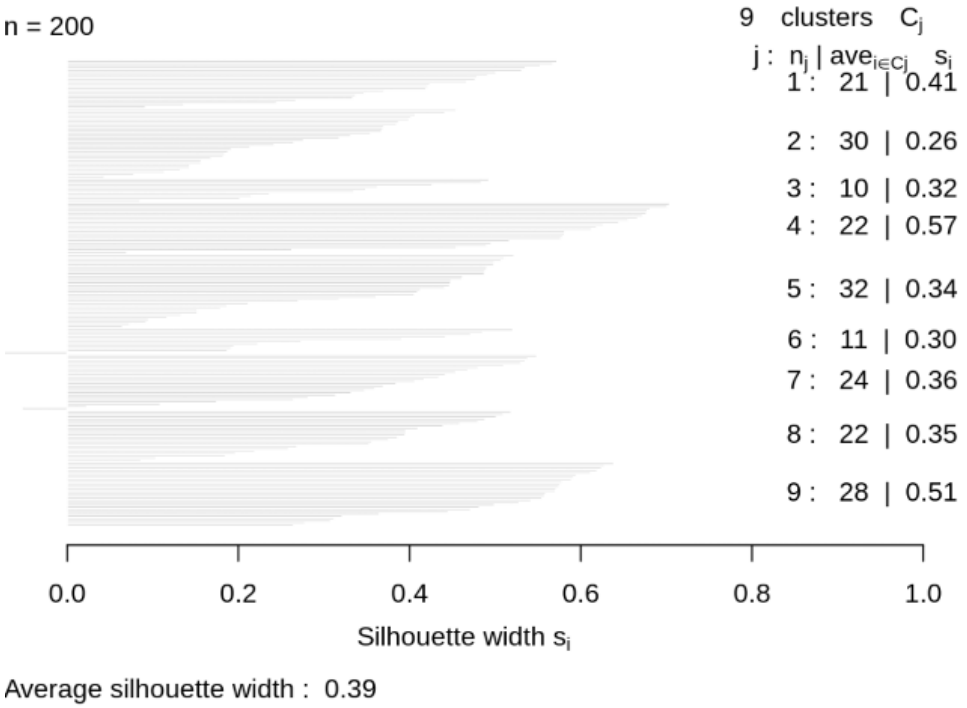

Рисунок 37 – График средней ширины для 9 кластеров

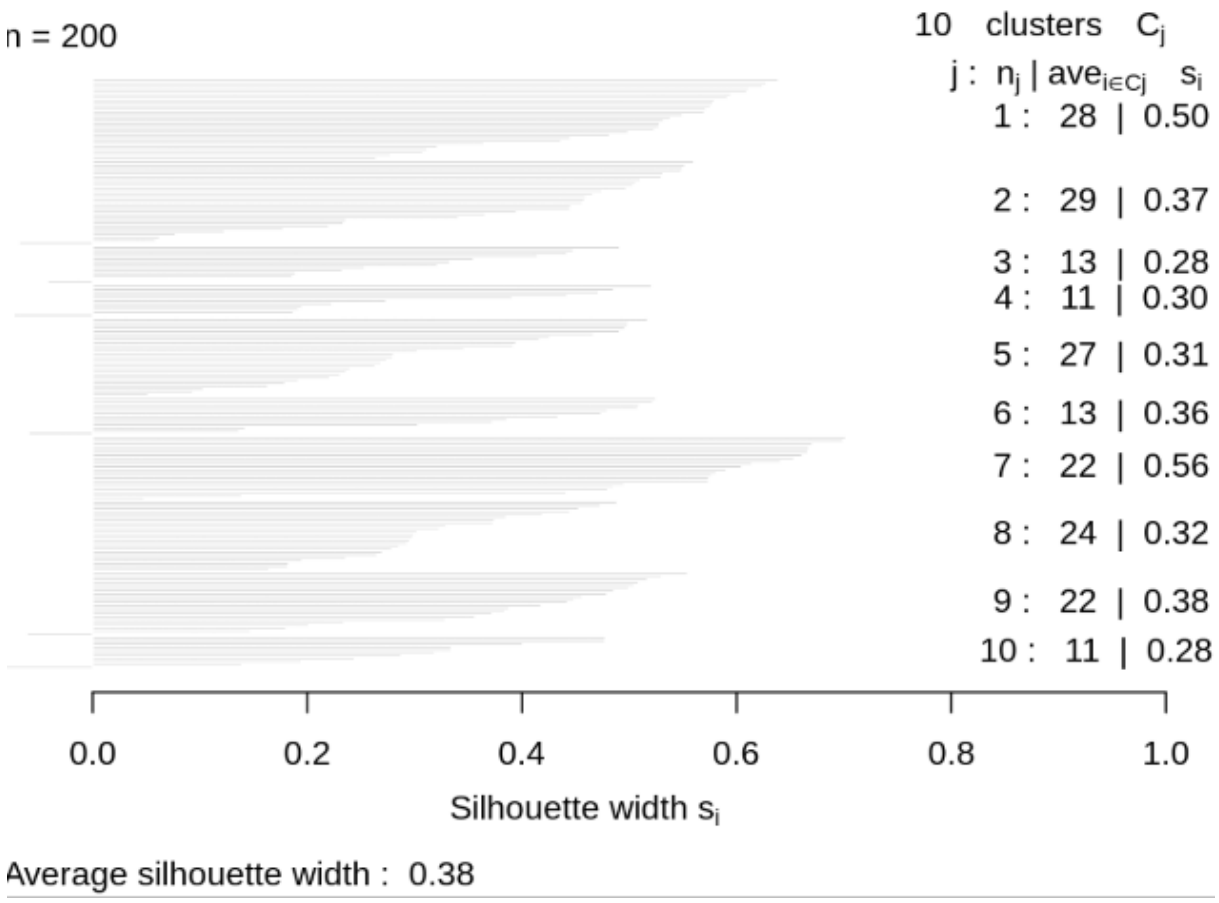

Рисунок 38 – График средней ширины для 10 кластеров

После получение всех ширины всех силуэтов используем функцию fviz_nbclust() для определения и визуализации оптимального количества кластеров.

На рисунках 39 и 40 изображены программный код и график полученных результатов, соответственно.

Наивысшей точкой в данном графике по оси ширины занимает значение 6 для кластера. Используя данный метод можно прийти к выводу что оптимальным значением классификаторов является значение 6.

```
library(NbClust)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.g
l/13EFCZ
```

```
fviz_nbclust(customer_data[,3:5], kmeans, method = "silhouette")
```

Рисунок 39 – Программный код

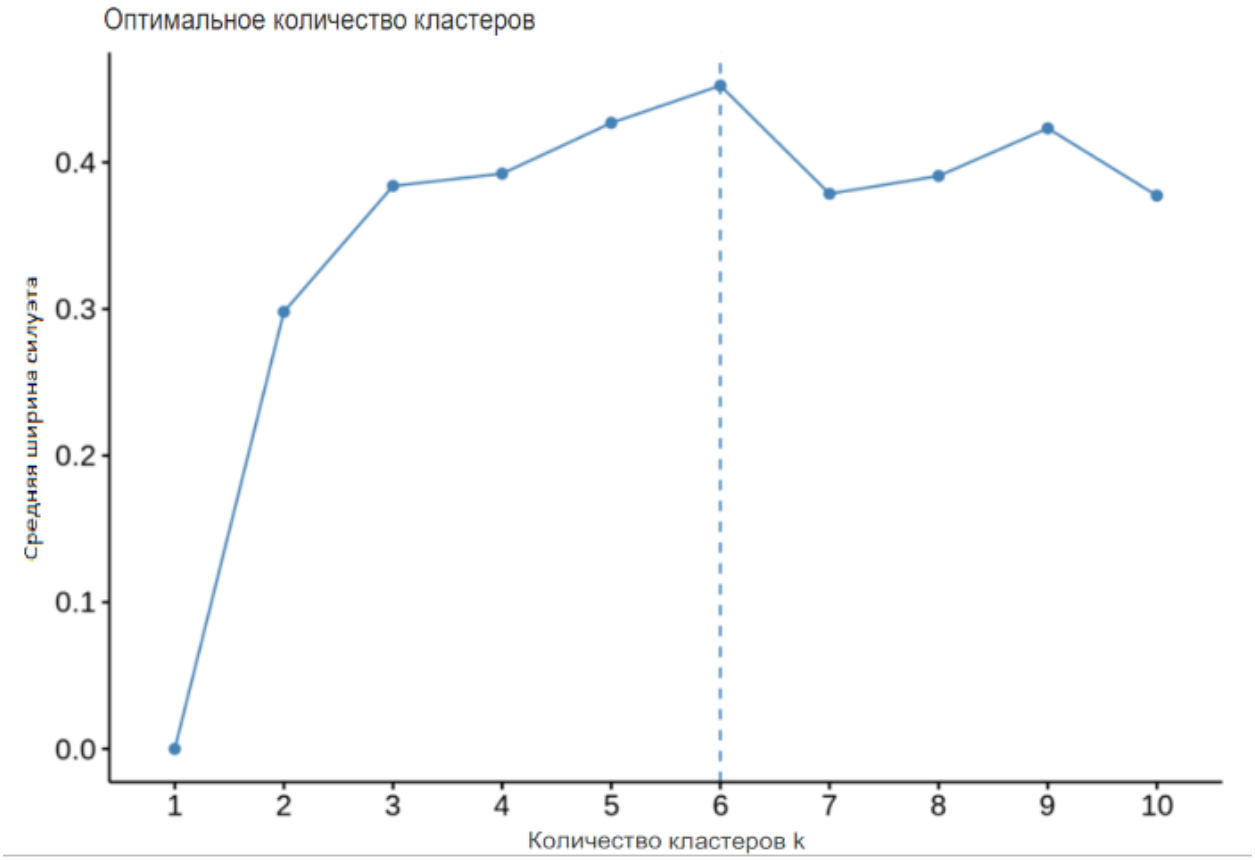

Рисунок 40 – График полученных результатов

«Статистический метод разрыва» - В 2001 г. исследователи Стэнфордского университета Р. Тибширани, Г. Уолтер и Т. Хасти опубликовали метод «Статистический метод разрыва». Можно использовать этот метод для любого метода кластеризации, такого как K-средние,

иерархическая кластеризация и т.д. Используя «Статистический метод разрыва», можно сравнить общую внутри кластерную вариацию для разных значений k вместе с их ожидаемыми значениями при нулевом эталонном распределении данных. С помощью моделирования Монте-Карло можно создать образец набора данных. Для каждой переменной в наборе данных можем вычислить диапазон между $\min(x_i)$ и $\max(x_j)$, в котором можно равномерно получать значения от нижней границы интервала до верхней границы. Статистический метод разрыва изображён на рисунке 41, а его программный код на рисунке 42.

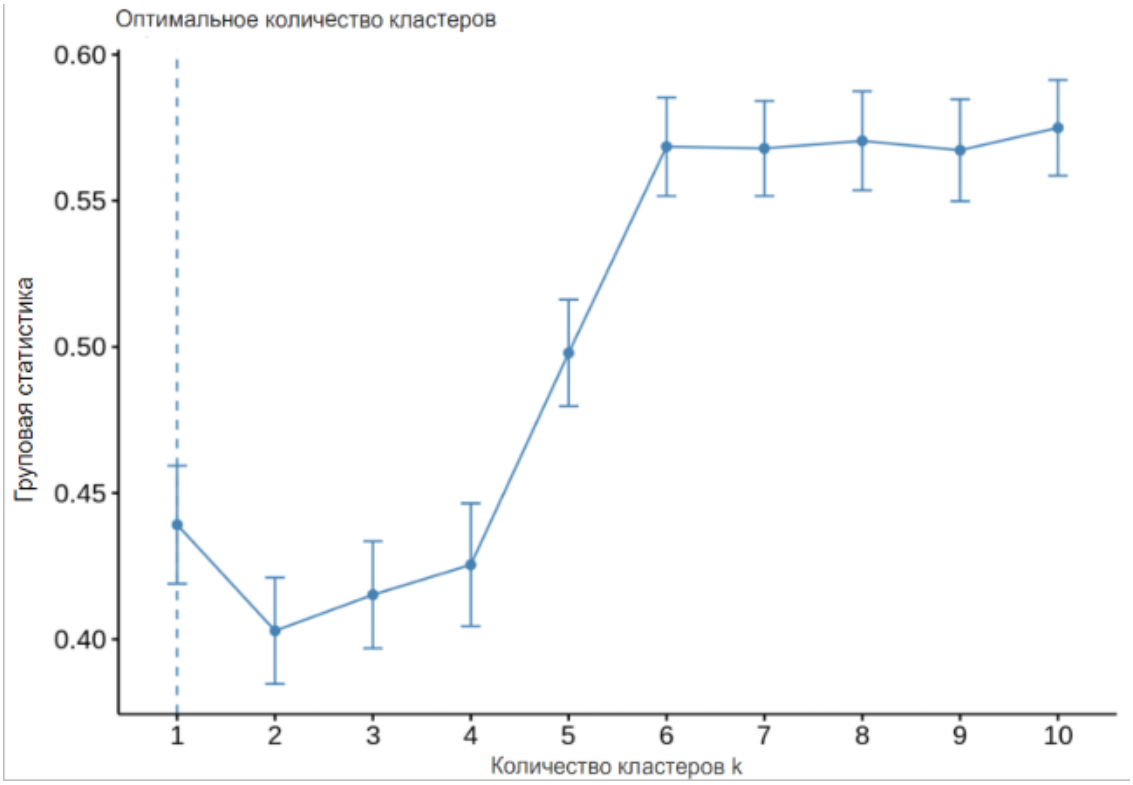

Рисунок 41 – Статистический метод разрыва

```
# compute gap statistic
set.seed(123)
gap_stat <- clusGap(customer_data[,3:5], FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)
fviz_gap_stat(gap_stat)
```

Рисунок 42 – Программный код статического метода разрыва

Различные методов дали следующие результаты:

– метод статистического разрыва = 6 кластеров;

– метод среднего силуэта = 6 кластеров;

– локтевой метод = 4 кластера.

Можно прийти к выводу что оптимальным для сегментации целевой аудитории является значение 6 кластеров, а так же, как метод среднего силуэта более простой в расчётах, можно сделать вывод что данный метод является эффективным для данного набора данных.

3.7 Сегментация целевой аудитории

В этом разделе выполним сегментацию целевой аудитории, на рисунке 44 выполнен программный код кластеризации по типу доход расход, а на рисунке 43 изображен его программный код.

```
## VISULIASE THE CLUSTERS
set.seed(1)
ggplot(customer_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +
    geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
    scale_color_discrete(name=" ",
                         breaks=c("1", "2", "3", "4", "5","6"),
                         labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4",
"Cluster 5","Cluster 6")) +
    ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```

Рисунок 43 – Программный K-means

На рисунке 44 изображёна K-means кластеризация по типам «доход и расход».

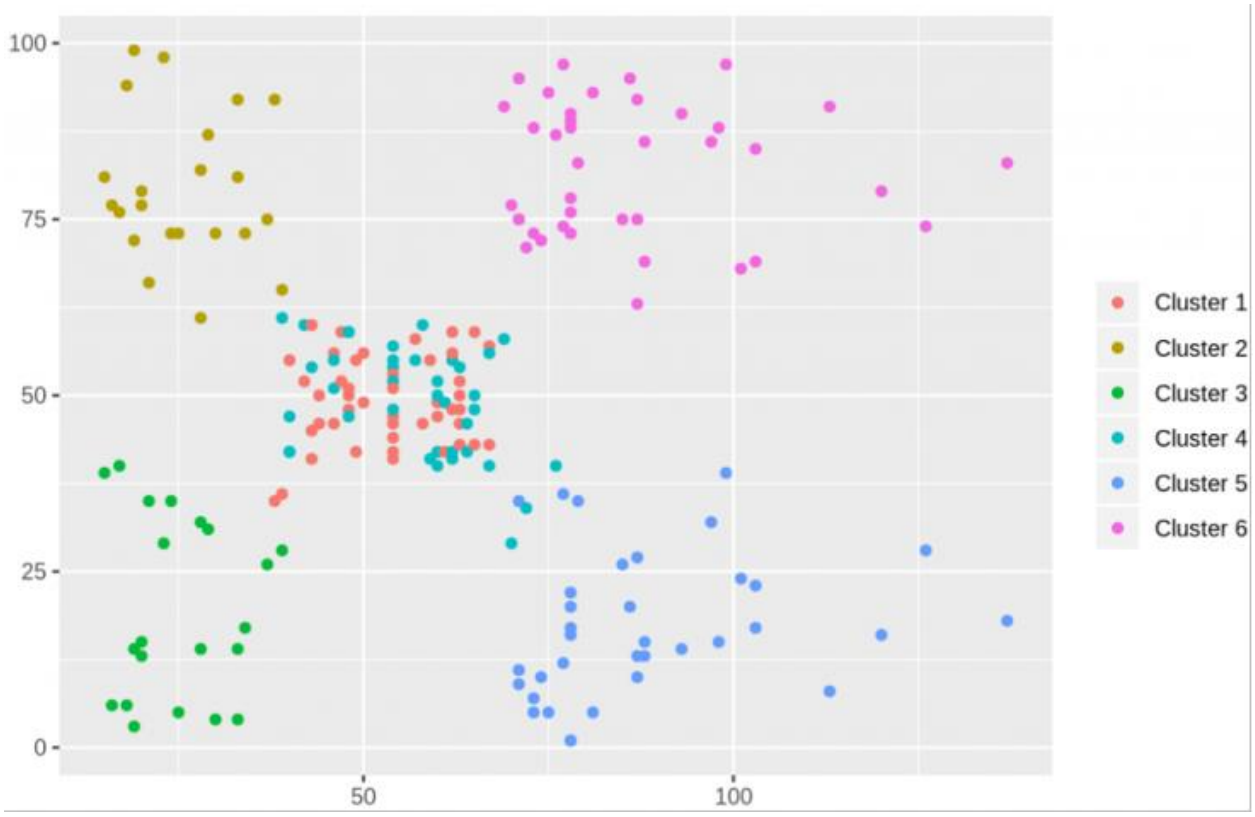

Рисунок 44 – K-means кластеризация по типам «доход и расход»

Из приведенной выше визуализации видим, что существует следующее распределение 6 кластеров, описанных ниже.

Кластер 1. Этот кластер представляет данные о клиентах с высоким годовым доходом, а также с высокими годовыми расходами.

Кластер 3. Этот кластер обозначает данные о клиентах с низким годовым доходом, а также с низким годовым расходом дохода.

Кластер 2. Этот кластер обозначает высокий годовой доход и низкие годовые расходы.

Кластер 6 и 4. Эти кластеры представляют данные о клиентах со средней зарплатой, а также среднегодовые расходы по зарплате.

Кластер 5. Этот кластер представляет собой низкий годовой доход, но высокие годовые расходы.

Уменьшим размерность полученных данных, программный код и метод K-means с уменьшением размерности представлены на рисунке 45 и 46

соответственно.

```
kCols=function(vec){cols=rainbow (length (unique (vec)))
return (cols[as.numeric(as.factor(vec))])}

digCluster<-k6$cluster; dignm<-as.character(digCluster);    # K-means clusters

plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K-means",ylab="classes")
legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))
```

Рисунок 45 – Программный код K-means с уменьшением размерности данных

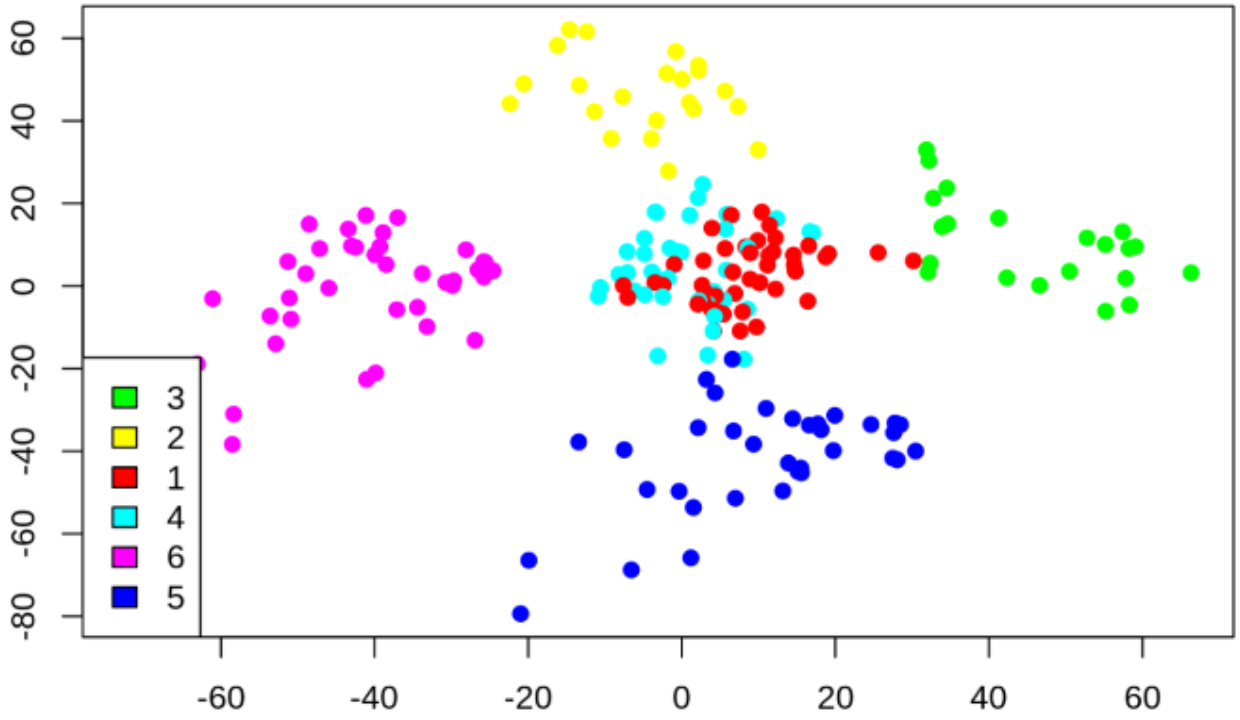

Рисунок 46 – K-means с уменьшением размерности данных

Кластер 4 и 1. Эти два кластера состоят из клиентов со средним баллом PCA1 и средним баллом PCA2.

Кластер 6. Этот кластер представляет клиентов с высоким PCA2 и низким PCA1.

Кластер 5. В этом кластере есть клиенты со средним значением PCA1 и низким значением PCA2.

Кластер 3. Этот кластер состоит из клиентов с высоким доходом PCA1 и высоким PCA2.

Кластер 2. В него входят клиенты с высоким PCA2 и средними годовыми расходами дохода.

PCA (Principal Component Analysis, или главные компоненты) - это статистический метод, который используется для уменьшения размерности данных, сохраняя при этом как можно больше информации. Это достигается путем преобразования исходных переменных в новый набор ортогональных переменных, называемых главными компонентами. Главные компоненты упорядочиваются таким образом, что первая компонента объясняет наибольшую долю общей дисперсии данных, вторая - следующую по величине долю, и так далее.

PCA широко используется в машинном обучении и анализе данных для визуализации, сжатия данных, удаления шума и т.д. Это особенно полезно, когда исходные данные имеют множество коррелированных переменных, и существует необходимость упростить анализ, сокращая размерность данных.

3.8 Сравнение алгоритмов

В данном подразделе сравним выбранные ранее методы сегментации такие как DBSCAN и OPTICS с методом k-means.

Алгоритмом DBSCAN основан на плотности данных и может обрабатывать данные с шумом и выбросами, а также может определять кластеры произвольной формы. В контексте сегментации целевой аудитории это может быть особенно полезным, так как целевая аудитория может иметь различные характеристики и не всегда легко выделить ее в явном виде, один из сильных недостатков данного метода является сложность правильного выбора параметров радиуса соседства и минимальное количество точек в окрестности, так как эти параметры подбирается вручную. Результат применения алгоритма на выбранных данных представлен на рисунке 56.

На рисунке 47 изображено импортирование библиотеки pandas, numpy и matplotlib, а также класс DBSCAN и функцию StandardScaler из модуля sklearn для кластеризации и масштабирования данных.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import OPTICS
from sklearn.preprocessing import StandardScaler
from matplotlib.lines import Line2D
```

Рисунок 47 – Импортирование необходимых библиотек и модулей

На рисунке 48 указывается имя файла, содержащего набор данных, и считывает его содержимое с помощью функции pd.read_csv().

```
# Загрузка данных
filename = 'Mall_Customers.csv'
data = pd.read_csv(filename)
```

Рисунок 48 – Загрузка данных из файла

На рисунке 49 выбираем два столбца из нашего набора данных для использования в кластеризации: "Annual Income (k$)" и "Spending Score (1-100)". Затем получаем массив значений из этих столбцов и сохраняем его в переменной X.

```
# Подготовка данных: используем столбцы "Annual Income (k$)" и "Spending Score (1-100)"
X = data[['Annual Income (k$)', 'Spending Score (1-100)']].values
```

Рисунок 49 – Подготовка данных

Так как алгоритм DBSCAN чувствителен к масштабу данных,

используем функцию StandardScaler() для масштабирования значений признаков в X, на рисунке 50. Это преобразование приведет данные к нулевому среднему значению и единичному стандартному отклонению.

```
# Масштабирование данных
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

Рисунок 50 – Масштабирование данных

На рисунке 51 создаем экземпляр класса DBSCAN с заданными значениями параметров eps и min_samples. Затем обучаем алгоритм на данных X и получаем метки кластеров с помощью функции fit_predict().

```
# Кластеризация DBSCAN
dbscan = DBSCAN(eps=0.35, min_samples=3)
cluster_labels = dbscan.fit_predict(X)
```

Рисунок 51 – Кластеризация DBSCAN

На рисунке 52, помечаем индексы точек, отмеченных как помехи (шум), и удаляем их из массива данных X и меток кластеров cluster_labels. Полученные отфильтрованные массивы сохраняются в переменных X_filtered и cluster_labels_filtered.

```
# Удаление помех (шума)
noise_indices = np.where(cluster_labels == -1)
X_filtered = np.delete(X, noise_indices, axis=0)
cluster_labels_filtered = np.delete(cluster_labels, noise_indices)
```

Рисунок 52 – Удаление помех (шума)

На рисунке 53, используем функцию plt.scatter() для создания графика

точек, раскрашенных в соответствии с метками кластеров cluster_labels_filtered. Также задаем метки осей и название графика.

```
# Визуализация кластеров без помех
scatter = plt.scatter(X_filtered[:, 0], X_filtered[:, 1], c=cluster_labels_filtered, cmap='viridis')
plt.xlabel('Annual Income (k$) (scaled)')
plt.ylabel('Spending Score (1-100) (scaled)')
plt.title('DBSCAN Clustering without Noise')
```

Рисунок 53– Визуализация кластеров без помех

На рисунке 54 в этом фрагменте кода создаем список `legend_elements`, который содержит элементы легенды для каждого уникального значения метки кластера. Затем добавляем легенду с помощью функции `plt.legend()`, передавая список элементов легенды и указывая положение легенды на графике.

```
# Создание легенды с информацией о цветах кластеров
legend_elements = [Line2D([0], [0], marker='o', color='w', label=f'Cluster {i}',
                          markerfacecolor=c, markersize=8) for i, c in enumerate(scatter.cmap(scatter.norm(np.unique(cluster_labels_filtered))))]

plt.legend(handles=legend_elements, loc='upper right')
```

Рисунок 54– Создание легенды с информацией о цветах кластеров

Наконец, на рисунке 55, используем функцию plt.show() для отображения графика с визуализированными кластерами без помех и легендой, показывающей, к какому кластеру принадлежит какой цвет.

```
plt.show()
```

Рисунок 55 – Отображение графика

Для визуализации алгоритма DBSCAN были переданы такие параметры

такие параметры как радиус соседства и минимальное количество точек в окрестности а так же были удалены помехе с выборки, результат работы изображён на рисунке 56.

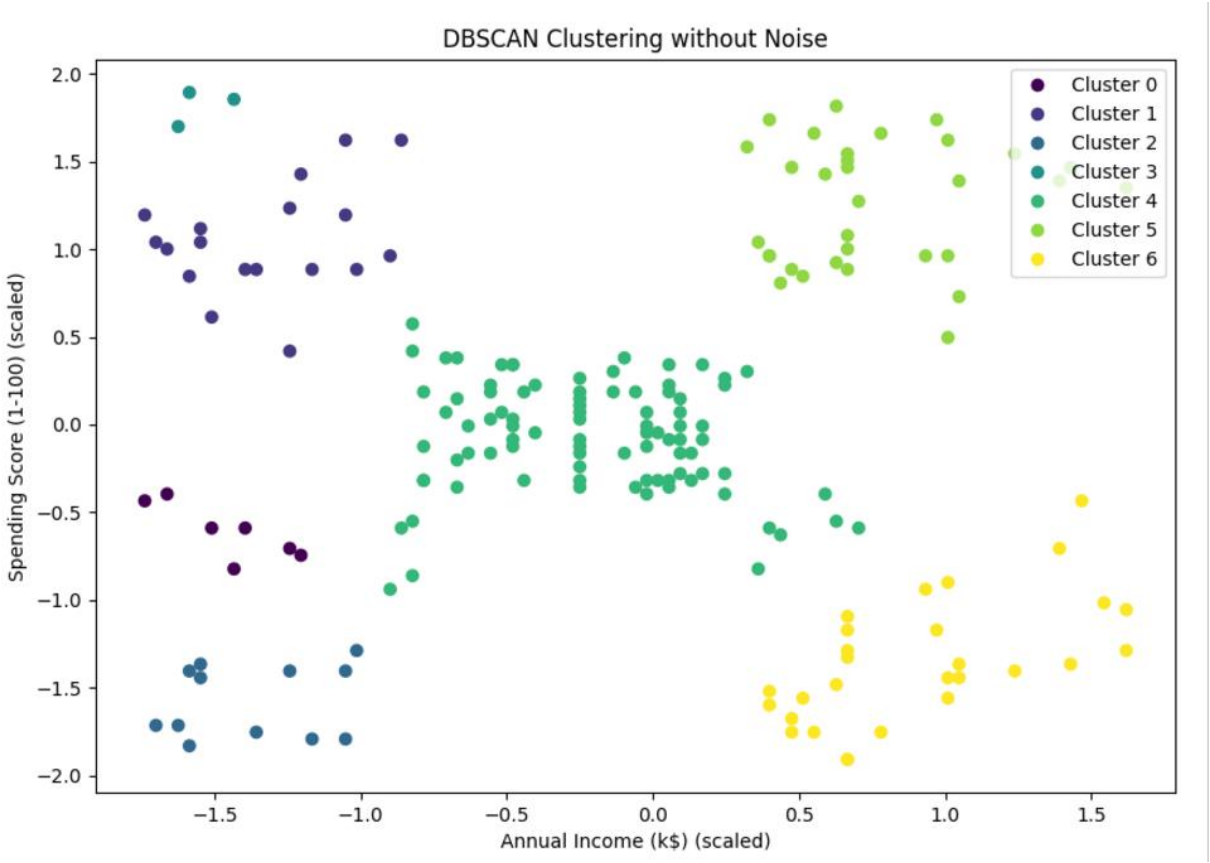

Рисунок 56 – Метод DBSCAN

На рисунке 56 можно наблюдать, что групп, кластеров по сравнение с k-means стало больше, визуально группы разделены не четко. Можно сделать вывод что данный метод для данного набора данных не подходит.

Сравним ещё один вариант кластеризации OPTICS, для данного набора данных.

Метод OPTICS (Ordering Points To Identify the Clustering Structure) представляет собой плотностной алгоритм кластеризации, разработанный как улучшение DBSCAN. Вот ключевые отличия OPTICS от DBSCAN:

Отличия от DBSCAN:

– параметры: В то время как DBSCAN требует выбора двух параметров - радиуса (eps) и минимального числа точек (minPts), OPTICS требует только значения minPts. OPTICS автоматически определяет локальную плотность для каждой точки, устраняя необходимость задавать глобальный радиус;

– разнообразие плотности: DBSCAN может испытывать трудности при обработке данных с кластерами разной плотности. В то время как OPTICS адаптируется к разнообразным плотностям кластеров, что позволяет обрабатывать более сложные структуры данных;

– упорядоченное представление данных: OPTICS генерирует упорядоченное представление данных, основанное на плотности, которое позволяет выявлять структуру кластеров на различных уровнях плотности. Это делает OPTICS более гибким для анализа кластеров по сравнению с DBSCAN.

Алгоритм OPTICS (Ordering Points To Identify the Clustering Structure) принимает следующие ключевые параметры:

– min_samples (minPts): Минимальное количество точек, требуемое для формирования кластера. Этот параметр определяет порог плотности для классификации точек как составляющих кластер или как шум;

– xi (ξ): Параметр, используемый для определения кластеров из сгенерированной упорядоченной последовательности точек. Значение xi находится в диапазоне от 0 до 1, и определяет степень плотности, которая требуется для слияния двух кластеров. Более высокие значения xi приводят к большему количеству, но меньшим по размеру кластерам, а более низкие значения приводят к меньшему количеству, но большим по размеру кластерам.

Программный код и результат выполнения представлены на рисунке 57-65 и 66, соответственно.

На рисунке 57, импортируем библиотеки pandas, numpy и matplotlib, а также класс OPTICS и функцию StandardScaler из модуля sklearn для кластеризации и масштабирования данных.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import OPTICS
from sklearn.preprocessing import StandardScaler
from matplotlib.lines import Line2D
```

Рисунок 57 – Импортирование необходимых библиотек и модулей

На рисунке 58, указывается имя файла, содержащего набор данных, и считывается его содержимое с помощью функции pd.read_csv().

```
# Загрузка данных
filename = 'Mall_Customers.csv'
data = pd.read_csv(filename)
```

Рисунок 58 – Загрузка данных из файла

На рисунке 59, выбираем два столбца из нашего набора данных для использования в кластеризации: "Annual Income (k$)" и "Spending Score (1-100)". Затем получаем массив значений из этих столбцов и сохраняем его в переменной X.

```
# Подготовка данных: используем столбцы "Annual Income (k$)" и "Spending Score (1-100)"
X = data[['Annual Income (k$)', 'Spending Score (1-100)']].values
```

Рисунок 59 – Подготовка данных

Так как алгоритм OPTICS чувствителен к масштабу данных, используем функцию StandardScaler() для масштабирования значений признаков в X, изображено на рисунке 60. Это преобразование приведет данные к нулевому среднему значению и единичному стандартному отклонению.

```
# Масштабирование данных
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

Рисунок 60 – Масштабирование данных

На рисунке 61 создаем экземпляр класса OPTICS с заданными значениями параметров min_samples и xi. Затем обучаем алгоритм на данных X и получаем метки кластеров с помощью функции fit_predict().

```
# Кластеризация OPTICS
optics = OPTICS(min_samples=6, xi=0.003)
cluster_labels = optics.fit_predict(X)
```

Рисунок 61 – Кластеризация OPTICS

На рисунке 62, находим индексы точек, отмеченных как помехи (шум), и удаляем их из массива данных X и меток кластеров cluster_labels. Полученные отфильтрованные массивы сохраняются в переменных X_filtered и cluster_labels_filtered.

```
# Удаление помех (шума)
noise_indices = np.where(cluster_labels == -1)
X_filtered = np.delete(X, noise_indices, axis=0)
cluster_labels_filtered = np.delete(cluster_labels, noise_indices)
```

Рисунок 62 – Удаление помех (шума)

На рисунке 63, используем функцию plt.scatter() для создания графика точек, раскрашенных в соответствии с метками кластеров cluster_labels_filtered. Также задаем метки осей и название графика.

```
# Визуализация кластеров без помех
scatter = plt.scatter(X_filtered[:, 0], X_filtered[:, 1], c=cluster_labels_filtered, cmap='viridis')
plt.xlabel('Annual Income (k$) (scaled)')
plt.ylabel('Spending Score (1-100) (scaled)')
plt.title('OPTICS Clustering without Noise')
```

Рисунок 63 – Визуализация кластеров без помех

На рисунке 64, в этом фрагменте кода создаем список legend_elements, который содержит элементы легенды для каждого уникального значения метки кластера. Затем добавляем легенду с помощью функции plt.legend(), передавая список элементов легенды и указывая положение легенды на графике.

```
# Создание легенды с информацией о цветах кластеров
legend_elements = [Line2D([0], [0], marker='o', color='w', label=f'Cluster {i}',
                          markerfacecolor=c, markersize=8) for i, c in enumerate(scatter.cmap(scatter.norm(np.unique(cluster_labels_filtered))))]

plt.legend(handles=legend_elements, loc='upper right')
plt.show()
```

Рисунок 64 – Создание легенды с информацией о цветах кластеров

Наконец, на рисунке 65, используем функцию plt.show() для отображения графика с визуализированными кластерами без помех и легендой, показывающей, к какому кластеру принадлежит какой цвет.

```
plt.show()
```

Рисунок 65 – Отображение графика

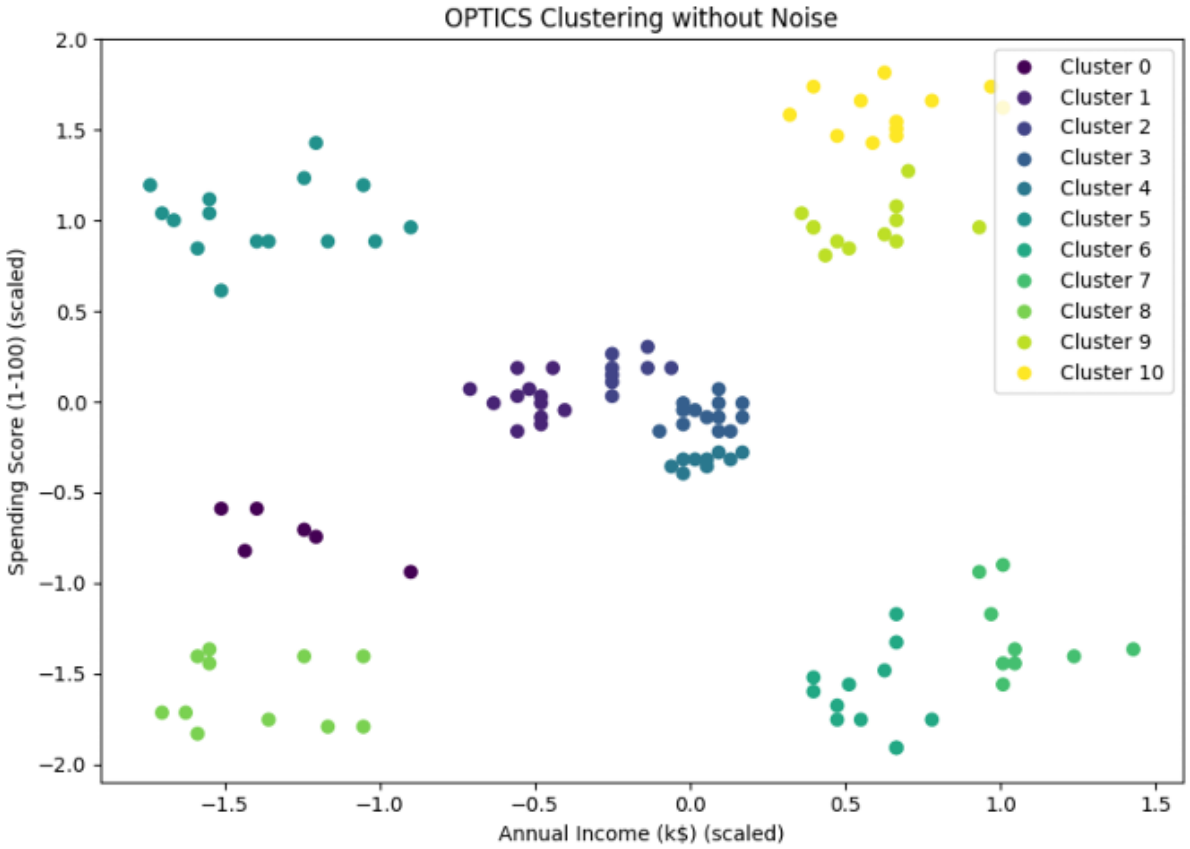

Рисунок 66 – Метод OPTICS

На рисунке 66 можно наблюдать что, количество кластеров стало больше, чем методе DBSCAN так же визуально разделение на группы по сравнению с начальными данными стало более невыраженным для данного

набора. Можно прийти к выводу что данный алгоритм так же не подходит для данного набора данных.

Стоит заметить, что метод кластеризации k-means для данного набора данных визуально справился отлично, так как группы были разделены так как они делятся визуально. Стоит добавить так же помогло правильное нахождения количества кластеров. Благодаря возможности указанию кластеров можно объединить в ещё более крупные группы, которые возможно могут потребоваться аналитикам.

Выводы по главе 3

В данном разделе бы воссоздан алгоритм машинного обучения, известный как неконтролируемое обучение. В частности, был использован алгоритм кластеризации, называемый кластеризацией K-средних, DBSCAN и OPTICS. Был проведён сравнительный анализ в процессе выбран лучший алгоритм для системы анализа целевой аудитории сайта, а так же был выбран оптимальный алгоритм поиска кластеров для сегментации k-means из набора данных пользователей который был разделён на целевые субъекты.

Глава 4 Оценка модели и анализ результатов

4.1 Анализ полученных результатов

Сравнительный анализ алгоритмов DBSCAN, OPTICS и k-means для сегментации аудитории сайта проведен на основе следующих критериев:

- разделение на кластеры: K-means требует заранее определить число кластеров, но количество кластеров можно вычислить. DBSCAN и OPTICS требует заранее определить радиус и плотность, найти эти параметры можно только по методу проб и ошибок;
- скорость и масштабируемость: K-means обычно быстрее и легче масштабировать для больших наборов данных. Для сегментации аудитории сайта, где количество пользователей может быть очень большим, k-means может быть более подходящим с точки зрения вычислительных затрат. DBSCAN и OPTICS обычно требуют больше времени и ресурсов для выполнения, особенно на больших наборах данных. Однако, с применением оптимизированных алгоритмов и техник индексации, их производительность может быть улучшена;
- преимущества K-means: быстрее и легче масштабировать для больших наборов данных. Простота реализации и интерпретации. Можно вычислить количество кластеров или задать нужно количество – что может быть необходимо для сегментации групп;
- преимущества DBSCAN и OPTICS: могут обнаруживать кластеры произвольной формы и разной плотности. Устойчивы к выбросам. Определяют количество кластеров автоматически на основе плотности данных;
- недостатки DBSCAN и OPTICS: обычно требуют больше времени и ресурсов для выполнения, особенно на больших наборах данных. Требуется правильный выбор параметров (например, радиуса для

DBSCAN) который нельзя вычислить с помочью формул и методов только с помощью подбора что ухудшает их возможности.

При выборе алгоритма для сегментации аудитории сайта важно учесть специфику данных, размеры набора данных и требования к точности кластеризации. K-means может быть предпочтительным для больших наборов данных с простой структурой кластеров, в то время как DBSCAN и OPTICS могут быть более подходящими для обработки сложных аудиторий сайта с разнообразными формами и плотностями кластеров. Для выбранного набора данных алгоритм k-means визуально справился лучше чем другие, стоит так же добавить что было найден лучший алгоритм для поиска кластеров.

Так же если сравнивать метод k-means со стандартными решениями а именно методами опроса, такие как "5W" Марка Шеррингтона и метод сегментирования KHRAMATRIX то можно выделить основные преимущества:

Метод сегментирования KHRAMATRIX является еще более сложным методом, который учитывает не только поведение пользователей, но и данные о социальных связях между ними. Этот метод помогает выявить группы пользователей, которые связаны между собой более тесно, чем с пользователями из других групп.

Выбор метода кластеризации зависит от конкретных целей и задач, а также от количества и качества данных. Метод k-means хорошо подходит для большинства разновидностей а так же количества данных. Метод сегментирования "5W" Марка Шеррингтона хорошо подходит для выявления общих черт и потребностей пользователей, а метод сегментирования KHRAMATRIX - для более глубокого анализа социальных связей между пользователями.

Метод k-means имеет свои преимущества перед другими методами кластеризации, такими как метод сегментирования "5W" Марка Шеррингтона и метод сегментирования KHRAMATRIX.

Одним из главных преимуществ метода k-means является его простота,

быстрота а так же что данный метод не зависит от определённых типов данных может производить сегментацию при любых параметрах выборки. Так же K-means является одним из самых быстрых методов кластеризации, который может обрабатывать большие объемы данных в короткие сроки. Это позволяет быстро получить представление о данных и провести первичный анализ, что может быть полезно для бизнес-аналитики.

Кроме того, метод k-means является легко интерпретируемым, что означает, что результаты кластеризации могут быть легко объяснены и представлены заказчику. Кластеры, полученные методом k-means, представляют собой группы объектов, которые ближе всего к центру кластера, что делает их понятными и легко интерпретируемыми.

Также следует отметить, что метод k-means может быть использован для разных типов данных, включая числовые, бинарные и категориальные данные. Это расширяет возможности использования метода k-means и позволяет его применять для различных задач.

В целом, метод k-means может быть полезным инструментом для быстрого и простого анализа данных. Он прост в использовании, легко интерпретируем и может быть применен к различным типам данных. Однако, стоит помнить, что необходимо применять выбранный оптимальный алгоритм с алгоритмом поиском кластеров по ширине для сегментации набора данных пользователей на целевые субъекты, тогда процесс нахождения будет проще и точнее.

Вывод из данного под заголовка заключается в том что сегментация с методом k-means работает эффективнее среди методов data Mining таких как DBSCAN и OPTICS в сочетании с поиском кластеров по ширине, так же в сравнении со стандартными методами опроса, преимущества метода k-means заключается в том что он работает для любых типов данных, а значит это освобождает от проведения определённых опросов что бы получить конкретные параметры.

4.2 Оценка возможностей и применимости полученного решения

Полученное решение можно применять для любых типов данных пользователей сайта, для объединения их в группы и дальнейшего анализа в плане маркетинга и других видах деятельности, таких как:

- электронная коммерция: K-means может использоваться для сегментации клиентов на основе их покупательского поведения, истории покупок, предпочтений и интересов. Это позволяет предоставлять персонализированные рекомендации, разрабатывать акции и предложения, нацеленные на определенные группы пользователей;
- контент-сайты и блоги: Сегментация аудитории может помочь определить группы пользователей с похожими интересами, что облегчает создание и предоставление контента, наиболее релевантного для различных сегментов аудитории;
- социальные сети: Метод K-means может быть применен для группировки пользователей по интересам, активности, географическому расположению и демографическим характеристикам, что помогает улучшить рекомендации друзей, групп и мероприятий;
- образовательные платформы: K-means может использоваться для сегментации студентов и пользователей по различным характеристикам, таким как уровень знаний, интересы и стиль обучения, что облегчает создание персонализированных учебных планов и материалов;
- онлайн-игры и платформы развлечений: Метод K-means может быть применен для сегментации игроков и пользователей на основе их игрового поведения, предпочтений и уровня мастерства. Это позволяет предоставлять персонализированный контент и опыт, а

также улучшить баланс и дизайн игр;

- туристические сайты и платформы: Сегментация пользователей с помощью K-means может помочь определить различные группы туристов и путешественников, что позволяет предлагать им наиболее подходящие туры, предложения и услуги;
- финансовые и банковские услуги: K-means может использоваться для сегментации клиентов на основе их финансового поведения, рисков и потребностей. Это позволяет предоставлять персонализированные продукты и услуги, а также улучшать стратегии управления рисками и мониторинга;
- онлайн-здравоохранение: K-means может быть использован для сегментации пациентов и пользователей по различным характеристикам, таким как состояние здоровья, возраст, медицинская история и предпочтения. Это облегчает предоставление персонализированных решений здравоохранения, образовательных материалов и программ поддержки;
- работа и карьера: Сайты по поиску работы и развитию карьеры могут использовать K-means для сегментации пользователей по опыту, образованию, навыкам и интересам. Это помогает предоставлять более точные и релевантные предложения работы, курсов и программ обучения;
- новостные порталы и медиа: Метод K-means может использоваться для сегментации пользователей на основе их интересов и предпочтений в отношении новостей и медиаконтента. Это позволяет предоставлять персонализированные новостные ленты и рекомендации статей, учитывающие интересы различных групп пользователей.

В целом, метод K-means для сегментации пользователей на группы может быть применен в широком спектре отраслей и видов деятельности.

Выводы по главе 4

В данной главе научной работы был проведен сравнительный анализ методов и алгоритмов сегментации целевой аудитории, используя подходы Data Mining и стандартные методы сегментации в виде словесных опросов. Сравнительный анализ показал, что новая модель обладает более высокой эффективностью по сравнению со старой.

Это означает, что новая модель способна лучше определять группы аудитории с определенными характеристиками, чем старая. Кроме того, новая модель является простой в обработке больших объемов данных, что делает ее более удобной и эффективной для использования в условиях больших объемов данных.

Таким образом, вывод из проведенного сравнительного анализа заключается в том, что новая модель является более эффективной и простой в обработке больших данных, чем старая модель. Это подтверждает преимущества новой модели в анализе и сегментации целевой аудитории, что может повысить эффективность маркетинговых стратегий и улучшить взаимодействие с клиентами.

Заключение

Целью данной работы заключалась в создание моделей и алгоритмов для быстрой и эффективной сегментации пользователей, а также опровержение или принятия гипотезы о том что методы data mining для сегментации целевой аудитории сайта является более эффективным и гибкими, чем традиционные методы сегментации на основе опросов.

Цель данной работы заключается в разработке алгоритма сегментации клиентов целевой аудитории сайта. В рамках работы был использован алгоритм машинного обучения, называемый неконтролируемым обучением. Для решения задачи был применен алгоритм кластеризации K-средних.

В ходе работы проанализирован ряд алгоритмов для поиска кластеров, выбран лучший алгоритм для выявления кластеров в классификации k-means, построен алгоритм кластеризации K-средних. Он основан на выявлении сходства между объектами и группировке их в кластеры. Были определены параметры алгоритма, такие как количество кластеров и метод определения центров кластеров, а так же в ходе работы были построены такие алгоритмы как DBSCAN и OPTICS которые имеют свои преимущества и недостатки по сравнению с кластеризации K-средних. Так же был произведён их сравнительный анализ.

Результаты работы были успешными: была разработана эффективная система сегментации клиентов целевой аудитории сайта, в отличии от словесных опросов. Алгоритм позволяет автоматически выделять кластеры клиентов на основе общих характеристик, что поможет оптимизировать маркетинговые кампании и улучшить взаимодействие с клиентами.

Дальнейшие исследования в этой области могут быть направлены на улучшение алгоритма сегментации, например, путем добавления дополнительных параметров и анализа других алгоритмов машинного обучения.

Список используемых источников

1. Аль-Ватар А. Поток клиентов из Facebook и Instagram, 2020.
2. Бослаф С. Статистика для всех, 2022.
3. Бретт Л. Машинное обучение на R: экспертные техники для прогностического анализа, 2020.
4. Брюс П. Практическая статистика для специалистов Data Science, 2018.
5. Буявец И. Сегментирование целевой аудитории [Электронный ресурс]. URL: https://checkroi.ru/blog/segmentaciya-celevoy-auditorii/ (дата обращения 05.03.2023).
6. Гудфеллоу Я. Глубокое обучение, 2022.
7. Джеффри У. Анализ больших наборов данных, 2022.
8. Джоэл Грас. Data Science. Наука о данных с нуля, 2017.
9. Дэвид Фримэн. Машинное обучение и безопасность, 2022.
10. Клетте Р. Компьютерное зрение. Теория и алгоритмы, 2022.
11. Коэльо Л. Построение систем машинного обучения на языке Python, 2022.
12. ЛитРес, Практическое применение методов кластеризации, классификации и аппроксимации на основе нейронных сетей . [Электронный ресурс]. URL: https://www.litres.ru/book/raznoe-4340152/prakticheskoe-primenenie-metodov-klasterizacii-klassifikaci-64241736/ (дата обращения 05.03.2023).
13. Любанович Б. Простой Python. Современный стиль программирования. 2-е изд, 2020.
14. Миркин Б. Введение в анализ данных. Учебник и практикум, 2022.
15. Номейн А. Анализ целевой аудитории. Как составить портрет целевой аудитории, 2018.
16. Пащенко М. Как составить портрет целевой аудитории [Электронный ресурс]. URL: https://smmplanner.com/blog/kak-sostavit-portriet-tsielievoi-auditorii-chto-eto-zachiem-nuzhno-i-ghdie-iskat-informatsiiu/ (дата

обращения 05.03.2023).

17. Петер Флах, Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных, 2022.
18. Подробный алгоритм работы с ЦА [Электронный ресурс]. URL: https://zen.yandex.ru/media/algrigo/kto-tvoia-celevaia-auditoriia-podrobnyi-algoritm-raboty-s-ca-dlia-marketologa-6001ac7bf8b1af50bb713c81 (дата обращения 05.03.2023).
19. Потапов А. Распознавание образов и машинное восприятие, 2022.
20. Себастьян Рашка, Python и машинное обучение, 2022.
21. Сумарокова Е. В. Цифровые маркетинговые коммуникации: введение в профессию. Учебник для вузов, 2021.
22. Шай Шалев-Шварц. Идеи машинного обучения. От теории к алгоритмам, 2022.
23. Шакла Н. Машинное обучение и TensorFlow, 2018.
24. Шевченко. Д. Цифровой маркетинг-микс, 2021.
25. Ын Анналин. Теоретический минимум по Big Data. Всё что нужно знать о больших данных, 2018г.
26. Art Weinstein, Dennis J. Cahill. Lifestyle Market Segmentation, 2014.
27. Art Weinstein. Handbook of Market Segmentation, 2013.
28. Emereo Pty Limited. Target Market Segmentation a Complete Guide, 2019.
29. Krzysztof Kubacki , Sharyn Rundle-Thiele, Timo Dietrich. Segmentation in Social Marketing, 2016.
30. Malcolm McDonald, Ian Dunbar. Market Segmentation, 2012.