МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
Satbayev University
Институт кибернетики и информационных технологий
Кафедра кибербезопасность, обработка и хранение информации

Досман Әлижан Махамедқалиұлы

Применение методов кластеризации для анализа данных

ДИПЛОМНАЯ РАБОТА

Специальность 5В070300 – Информационные системы

Алматы 2021

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

Казахский национальный исследовательский технический университет имени К.И. Сатпаева
Институт кибернетики и информационных технологий
Кафедра кибербезопасность, обработка и хранение информации

«Допущен к защите»
Заведующий кафедрой КОиХИ
_____Н.А.Сеилова

ДИПЛОМНАЯ РАБОТА

На тему: Применение методов кластеризации для анализа данных

Специальность 5В070300 – Информационные системы

Выполнил**:** Досман Ә. М.

Научный руководитель
к.т.н, доцент
_____Сейлова Н. А.

«27» 05 2021г.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

Satbayev University

Институт кибернетики и информационных технологий

Кафедра кибербезопасность, обработка и хранение

информации

5В070300 – Информационные системы

УТВЕРЖДАЮ
Заведующий кафедрой КБОиХО
канд. техн. наук, ассистент-профессор
_____Н. А. Сейлова
«_31_» _05_ 2021г.

ЗАДАНИЕ
на выполнение дипломной работы

Обучающемуся: Досман Әлижан Махамедқалиұлы
Тема: Применение методов кластеризации для анализа данных
Утверждена *приказом Ректора Университета № 762-б от 24.11.2020г.*

Срок сдачи законченной работы *27.05.2021г.*

Исходные данные к дипломному проекту: результаты преддипломной практики, результат обзора современного состояния по данной теме, сбор теоретического материала.
Краткое содержание дипломной работы:
а) Обзор методов кластерного анализа;
б) Описание методики кластерного анализа;
в) Применение метода кластерного анализа
при обработке данных;
Рекомендуемая основная литература: *из 16 наименований*

ГРАФИК

подготовки дипломной работы (проекта)

Наименование разделов, перечень разрабатываемых вопросов	**Сроки представления научному руководителю**	**Примечание**
Обзор методов кластерного анализа	02.04.2021г.	
Описание методики кластерного анализа	16.05.2021г.	
Применение метода кластерного анализа при обработке данных	26.05.2021г.	

Подписи

консультантов и нормоконтролера на законченную дипломную работу (проект) с указанием относящихся к ним разделов работы (проекта)

Наименование разделов	**Консультанты, Ф.И.О. (уч.степень, звание)**	**Дата подписания**	**Подпись**
Применение методов кластеризации для анализа данных	Сейлова Н. А., кандидат технических наук	27.05.2021	
Нормоконтролер	Кабдуллин М.А., ассистент		
Программная часть			

Научный руководитель:_____ Сейлова Н. А.

Задание принял к исполнению обучающийся:_____ Досман Ә. М.

Дата "24" ноября 2020

АННОТАЦИЯ

Данная работа посвящена вопросам исследования методик кластерного анализа. В ходе выполнения работы был проведен обзор методов кластеризации; исследованы методы кластерного анализа для сегментации клиентов банка; проведена интерпертация результатов с использованием метода кластеризации. Анализ проведен на примере набора данных о клиентах банка, которые являются держателями кредитных карт. Проведена классификация групп клиентов, являющихся держателями кредитных карт. В ходе проведения работы были использованы методы многомерного статистического анализа и математической статистики.

Для выполнения данной работы используется язык программирования Python с использованием библиотек pandas, numpy, matplotlib и sklearn.

АҢДАТПА

Бұл жұмыс кластерлік талдау әдістемелерін зерттеу мәселелеріне арналған. Жұмысты орындау барысында кластерлеу әдістеріне шолу жүргізілді; банк клиенттерін сегменттеу үшін кластерлік талдау әдістері зерттелді; кластерлеу әдісін пайдалана отырып нәтижелерге интерпертация жүргізілді. Талдау несие карталарын ұстаушылар болып табылатын банк клиенттері туралы мәліметтер жиынтығы мысалында жүргізілді. Несие карталарын ұстаушылар болып табылатын клиенттер топтарының жіктелуі жүргізілді. Жұмыс барысында көп өлшемді статистикалық талдау және математикалық статистика әдістері қолданылды.

Бұл жұмысты орындау үшін pandas, numpy, matplotlib және sklearn кітапханаларын қолдана отырып, Python бағдарламалау тілі қолданылады.

THE SUMMARY

This paper is devoted to the study of cluster analysis methods. In the course of the work, a review of clustering methods was carried out; cluster analysis methods for bank customers segmentation were investigated; results were interperted using the clustering method. The analysis is carried out on the example of a set of data on bank customers who are credit card holders. The classification of groups of customers who are credit card holders is carried out. In the course of the work, the methods of multivariate statistical analysis and mathematical statistics were used.

To perform this work, the Python programming language is used using the pandas, numpy, matplotlib, and sklearn libraries.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Кластерный анализ одни из статистических методов, используемый для группировки похожих объектов в соответствующие категории. Его также можно назвать анализом сегментации, анализом таксономии или кластеризацией.

Цель кластерного анализа – сортировка различных объектов или точек данных в различные группы кластеров таким образом, чтобы степень связи между двумя объектами была высокой, если они принадлежат к одной группе, и низкой, если они принадлежат к разным группам.

Кластерный анализ отличается от многих других статистических методов тем, что он в основном используется, когда исследователи не имеют предполагаемого принципа или факта, который они используют в качестве основы своих исследований.

Этот метод анализа, как правило, выполняется во время экспериментальной фазы исследования, так как в отличие от таких методов, как факторный анализ, он не различает зависимые и независимые переменные. Вместо этого кластерный анализ применяется в основном для обнаружения структур в данных без объяснения или интерпретации.

Проще говоря, кластерный анализ обнаруживает структуры в данных, не объясняя, почему эти структуры существуют.

Например, когда кластерный анализ применяется в рамках исследования рынка, можно выделить определенные группы внутри клиентов. Анализ этих групп может затем определить, насколько вероятно, что кластер клиентов будет приобретать товары или услуги. Если эти группы четко определены, маркетинговая команда может затем ориентироваться на различные кластеры с помощью специально разработанной целевой коммуникации.

Цель дипломной работы – использование методов кластерного анализа, их сравнение и примерное определение к какой группе клиентов относится тот или иной кластер данных и выдача рекомендации для каждого кластера клиентов с помощью визуализации результатов кластеризации на различных графиках.

Набор данных для исследования содержит в себе данные о клиенте кредитной карточки, в котором содержатся информация о кредитном лимите клиента, и поведенческие переменные, описывающие способ их обращения в банк.

Основные задачи для выполнения дипломной работы:

- обработка данных;
- обучение модели;
- анализ данных;
- интерпретация результатов.

Задачи подробно разобраны в последующих главах

В первой главе рассмотрены возможности кластерного анализа, его значение и суть в машинном обучении, далее идет разбор применяемых в этой работе методов кластеризации. Сформированы основные задачи, которые необходимо выполнить в этой работе.

Во второй главе описаны используемые программные средства и описан алгоритм решения.

В третьей главе начинается техническая часть дипломной работы, которая включает в себя основную часть работы, обработку набора данных, обучение и построение модели, и выдача рекомендации.

1 Обзор методов кластерного анализа

1.1 Назначение кластерного анализа

Кластерный анализ может быть мощным инструментом сбора данных для различной организации, которые нуждаются в идентификации отдельных групп клиентов, сделки купли-продажи, или других типов поведении и вещей. Для примера можно взять страховую компанию, которая использует кластерный анализ чтобы определять требования мошенников, и банки, которые определяют кредитный рейтинг.

Определение кластерного анализа.

Кластерный анализ — это статистический метод анализа данных. Он выполняет работу путем организации элементов в группы или кластеры, на основе того, насколько они тесно связаны.

Кластерный анализ, как и факторный тип анализа связан матрицами данных в которых переменные не были заранее разделены на подмножество критериев и предикторов. Цель кластерного анализа найти похожие группы субъектов, где «сходство» между каждой парой субъектов может означать некоторую глобальную меру по всему набору характеристик.

Кластерный анализ — это алгоритм обучения без учителя, что означает что изначально неизвестно сколько кластеров существует в наборе данных до запуска модели. В отличии от многих статистических методов, кластерный анализ обычно используется, когда нет никаких предположений о вероятных отношениях в данных. Он предоставляет информацию о том где существуют ассоциации и закономерности в наборе данных, но не о том, что они могут означать, это уже работа аналитика.

Как используется кластерный анализ?

Во многом кластерный анализ применяется для классификации. Субъекты разделены на группы, так что каждый субъект больше похож на другие субъекты в своей группе, чем на субъекты вне группы.

В маркетинге кластерный анализ может использоваться для сегментации аудитории, так что различные группы клиентов могут быть нацелены на наиболее релевантные сообщения.

Каким бы ни было приложение, очистка данных – важный подготовительный шаг для успешного кластерного анализа. Кластеризация работает на уровне набора данных, где каждая точка оценивается относительно других, поэтому данные должны быть как можно более полными.

1.2 Обзор методов кластеризации

Кластеризация k-средних. На рисунке 1.1 можно увидеть визуализацию алгоритма кластеризации k-средних:

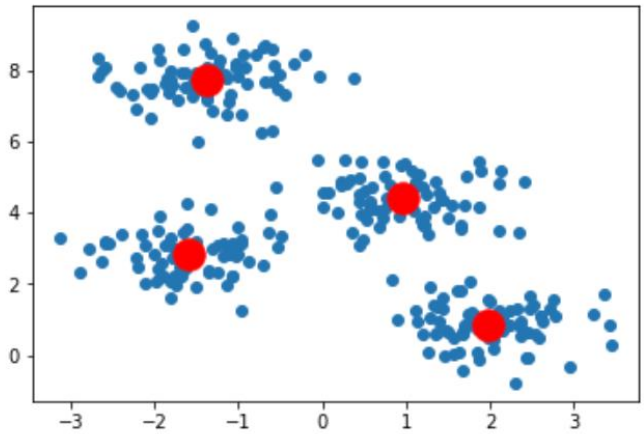

Рисунок 1.1 – Визуализация алгоритма кластеризации k-средних

Алгоритм k-средних присваивает каждую точку ближайшим к нему кластеру, центр которого также называется центроидом. Центр – это среднее значение всех точек в кластере, то есть его координаты являются средним арифметическим для каждого измерения отдельно по всем точкам в кластере.

Иерархическая кластеризация. На рисунке 1.2 приведен визуализация иерархической кластеризации:

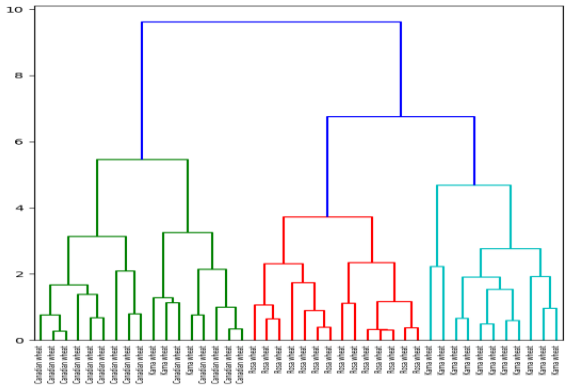

Рисунок 1.2 – Визуализация алгоритма иерархической кластеризации

Иерархическая кластеризация делится на два типа:

- agglomerative (агломеративный);
- divisive (разделительный).

Для обработки данных я использовал метод агломеративной иерархической кластеризации. В этом методе изначально все точки данных рассматриваются как отдельные кластеры. На каждой итерации аналогичные кластеры объединяются с другими кластерами, пока не будет сформирован один единственный кластер или K кластеров.

Кластеризацию так же можно разделить на две подгруппы:

- жесткая кластеризация;
- мягкая кластеризация.

До этого мы рассматривали кластеризации соответствующие к жесткому типу кластеризации, как k-средних и иерархическая.

В жесткой кластеризации каждая точка данных кластеризуются или группируются в любой один кластер. Каждая точка данных может либо полностью принадлежать кластеру, либо нет.

В мягкой кластеризации, вместо размещения каждой точки в отдельный кластер, назначается вероятность того, что эта точка будет находится в этом кластере. В мягкой кластеризации, так же называемой в нечеткой кластеризации, точка данных может принадлежать сразу нескольким кластерам со своей оценкой вероятности принадлежности.

Нечеткая кластеризация с-средних. На рисунке 1.3 приведен визуализация нечеткой кластеризации:

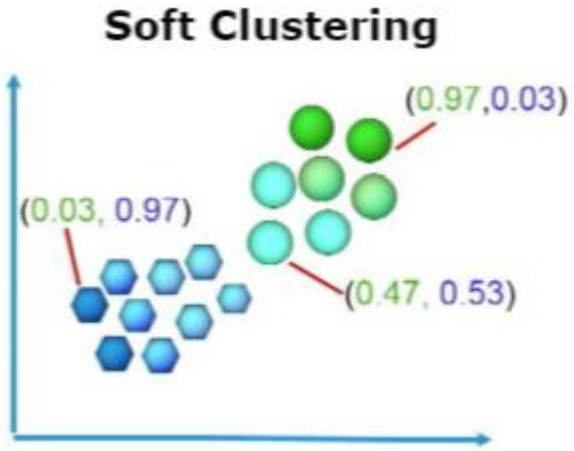

Рисунок 1.3 – Визуализация работы нечеткой кластеризации

Нечеткая кластеризация с-средних – это метод мягкой кластеризации, который каждой точке данных присваивает оценку вероятности принадлежности к тому или иному кластеру.

1.3 Постановка задачи

Цель дипломной работы – исследование методов кластеризации на примере набора данных клиентов банка, которые являются держателями кредитных карт.

Поставленная цель требует решения следующих задач:

- обзор методов кластеризации;
- выбор средств разработки;
- разведочный анализ данных;
- выбор модели;
- проведение кластерного анализа;
- интерпретация результатов.

В процессе выполнения анализа проводится стандартизация данных, определяется оптимальное количество кластеров с использованием выбранного

средства, проводится обучение модели, строятся диаграммы и проводится интерпретация результатов.

Работа выполняется с использованием языка программирования Phyton.

2 Описание методики кластерного анализа

2.1 Используемые средства

Для выполнения кластерного анализа я выбрал язык программирования Python 3, так как он является хорошим инструментом для машинного обучения и визуализации данных. Python – высокоуровневый язык программирования общего назначения с динамической строгой типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости кода и его качества, а также на обеспечение переносимости написанных на нём программ[4].

Среду для проведения работы я выбрал Jupyter Notebook. Jupyter Notebook является крайне удобным инструментом для аналитических отчетов, так как он позволяет хранить в себе код, графики а так же формулы, комментарии и изображения с помощью языка разметки Markdown.

При выполнении дипломной работы так же были необходимы такие библиотеки как pandas, NumPy, matplotlib, seaborn, scikit-learn, SciPy.

Библиотека NumPy используется для работы с многомерными массивами данных.

Pandas является библиотекой, которая дает возможность обрабатывать и анализировать данные, и строится она поверх библиотеки NumPy.

Библиотеки matplotlib и seaborn предназначены для визуализации данных. Они позволяют строить графики на заданных данных, что облегчает делать какие-либо выводы на их основе.

Scikit-learn библиотека для машинного обучения, она содержит в себе различные алгоритмы регрессии, кластеризации и классификации, взаимодействует с численным Python и научными библиотеками NumPy и SkiPy.

SciPy библиотека, используемая для научного и технического вычисления

2.2 Описание алгоритмов кластеризации

Алгоритмическая последовательность работы кластеризации k-средних выглядит следующим образом:

- указать количество кластеров k;
- инициализировать центроиды, сперва перемешивая набор данных, затем случайно выбирая k точек данных для центроидов без замены;
- повторять, пока центроиды не останутся без изменений. То есть назначение точек данных кластерам не изменяется;
- вычислить сумму квадрата промежутка между точками данных и всеми центроидами;
- определить каждую точку данных к ближайшему кластеру (центроиду);
- вычислить центроиды для каждого кластера, взяв среднее значение всех точек данных, принадлежащих каждому кластеру.

Блок-схема алгоритма k-средних приведен на рисунке 2.1:

Рисунок 2.1 – Блок-схема кластеризации k-средних

Алгоритм агломеративной иерархической кластеризации довольно прост:

- рассчитать матрицу близости;
- пусть каждая точка данных будет кластером;
- повторить: соединить два самых ближайших кластера и рассчитать новую матрицу близости;
- пока не останется больше одного кластера.

Блок-схема иерархической кластеризации на рисунке 2.2:

Рисунок 2.2 – Блок-схема иерархической кластеризации

Алгоритм нечеткой кластеризации с-средних:
- выполнить случайную инициализацию входных значении;
- вычислить центры масс каждого кластера используя:

$$c_i = \frac{\sum_{j=1}^{p} (u_{ij})^m x_j}{\sum_{j=1}^{p} (u_{ij})^m}, \quad (1)$$

где p – количество точек данных;

u_{ij} – принадлежность i-х данных к j-му центру кластера;

j – центр кластера;

c_i – представляет j-й центр кластера;

- вычислить значение функции погрешности используя:

$$E = \sum_{i=1}^{K} \sum_{j=1}^{p} u_{ij}^m \left\| c_i - x_j \right\|^2, \quad (2)$$

где $\|c_i - x_j\|$ – Евклидово расстояние между i-м данными и j-м центром кластера;

- возвращаться к предыдущему шагу, пока значение погрешности не будет ниже установленного предела или уменьшение погрешности относительно прошлой итерации не будет пренебрежимо мало используя:

$$u_{ij} = \frac{1}{\sum_{k=1}^{K}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}}}, \quad (3)$$

где k – шаг итерации;

d_{ij} – представляет собой евклидово расстояние между i-м данными и j-м центром кластера.

3 Применение метода кластерного анализа при обработке данных

3.1 Входные данные

Для выполнения дипломной работы был использован набор данных клиентов кредитных карт.

Названия столбцов данных:

- Sl_No (номер клиента);
- Customer Key (ключ клиента);
- Avg_Credit_Limit (средний кредитный лимит);
- Total_Credit_Cards (всего кредитных карт);
- Total_visits_bank (всего посещений банка);
- Total_visits_online (всего онлайн посещений);
- Total_calls_made (всего сделанных звонков).

Загружаем данные во фрейм данных. Прописываем путь к файлу, так как файл находится в директории, достаточно написать имя файла. Пример кода загрузки данных на рисунке 3.1:

```
B [2]: df = pd.read_csv('Credit Card Customer Data.csv')
```

Рисунок 3.1 – Чтение и загрузка набора данных

3.2 Исследовательский анализ данных

Проводим разведочный анализ данных для выведения основных закономерностей и аномалии данных.

Вывод первых 5 элементов на рисунке 3.2:

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1	87073	100000	2	1	1	0
1	2	38414	50000	3	0	10	9
2	3	17341	50000	7	1	3	4
3	4	40496	30000	5	1	1	4
4	5	47437	100000	6	0	12	3

Рисунок 3.2 – Первые 5 элементов набора данных

Вывод общей информации данных на рисунке 3.3:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sl_No                660 non-null    int64
 1   Customer Key         660 non-null    int64
 2   Avg_Credit_Limit     660 non-null    int64
 3   Total_Credit_Cards   660 non-null    int64
 4   Total_visits_bank    660 non-null    int64
 5   Total_visits_online  660 non-null    int64
 6   Total_calls_made     660 non-null    int64
dtypes: int64(7)
memory usage: 36.2 KB
```

Рисунок 3.3 – Общая информация по набору данных

Первые 2 столбца содержат в себе номер клиента и ключ клиента, в связи с чем они не имеют значимости для кластеризации данных, поэтому создаем копию набора данных без этих столбцов для дальнейшей работы с ним. Пример кода с функцией удаления столбцов на рисунке 3.4:

```
data = df.drop(['Sl_No','Customer Key'],axis=1).copy()
```

Рисунок 3.4 – Создание нового набора данных без двух столбцов

Строим парные диаграммы с участием каждого столбца, которые приведены на рисунке 3.5:

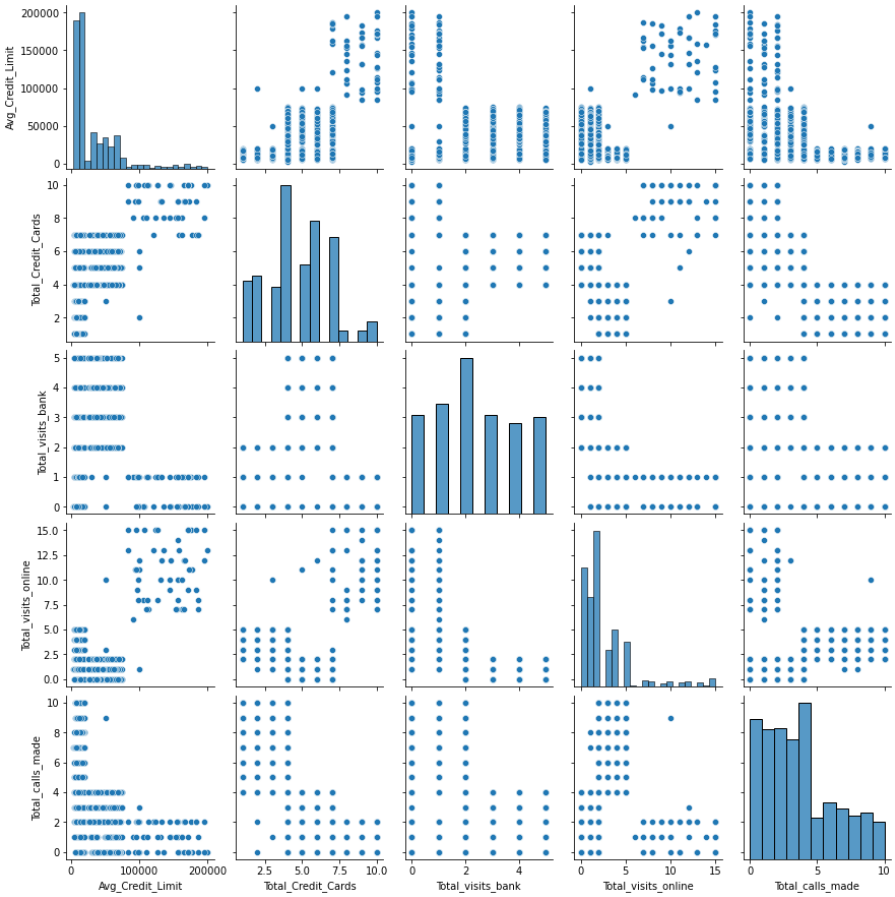

Рисунок 3.5 – Парные диаграммы всех столбцов

Наглядно можно увидеть некоторые зависимости между столбцами.

Далее строим «тепловую диаграмму» по корреляции столбцов, которая изображена на рисунке 3.6:

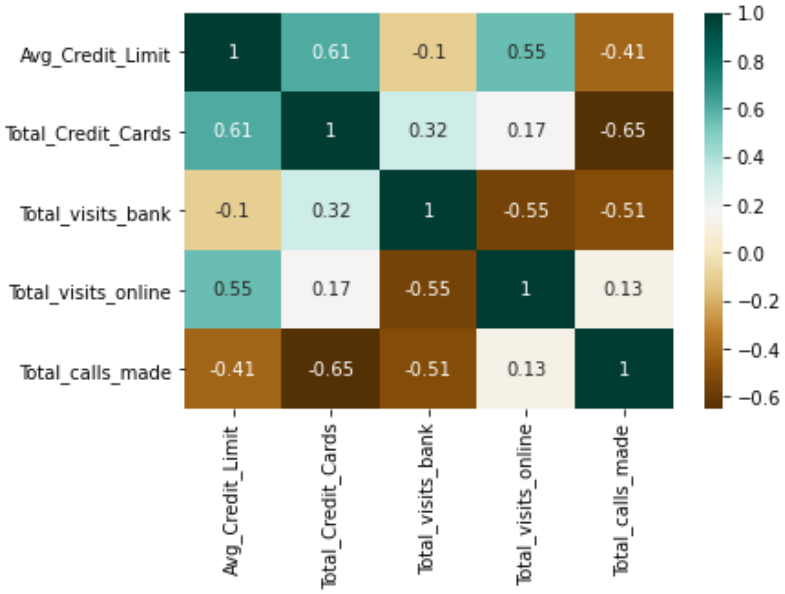

Рисунок 3.6 – Тепловая диаграмма корреляции столбцов

Можно заметить высокую корреляцию между столбцами среднего кредитного лимита и количества кредитных карточек, количества кредитных карточек и количества визитов в банк, среднего кредитного лимита и количества визитов онлайн. Так же можно заметить отрицательные корреляции.

3.3 Подготовка данных

Поскольку алгоритмы кластеризации, используют измерения на основе расстояния для выявления сходства между точками данных, рекомендуется выполнить стандартизацию данных, чтобы среднее значение равнялось нулю, а стандартное отклонение равнялось единице.

Создаем копию данных, чтобы оригинальные данные не подвергались изменению. Пример кода на рисунке 3.7:

```
X = data.copy()
```

Рисунок 3.7 – Создание копия набора данных

Проводим стандартизацию и преобразование файла. Пример кода на рисунке 3.8:

```
scaler = StandardScaler()
X = scaler.fit_transform(X)
X.shape
```

```
(660, 5)
```

Рисунок 3.8 – Стандартизация и преобразование набора данных

3.4 Определение оптимального числа кластеров

Чтобы выбрать оптимальное число кластеров в методе кластеризации k-средних можно использовать «Метод локтя».

Метод локтя – один из самых популярных способов найти оптимальное число кластеров. В этом методе используется концепция значения WCSS. WCSS означает «Сумма квадратов внутри кластера», который определяет общие вариации внутри кластера. Формула для расчета значения WCSS (для 3 кластеров) приведена ниже:

$$WCSS= \sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2 + \sum_{Pi\ in\ Cluster2} distance(P_i\ C_2)^2 + \sum_{Pi\ in\ CLuster3} distance(P_i\ C_3)^2. \quad (4)$$

В приведенной выше формуле WCSS, $\sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2$ – сумма квадратов промежутков между каждой точкой данных и ее центроидом в кластере 1 и того же значения для двух других членов.

Чтобы найти оптимальное значение кластеров, метод локтя включает следующие шаги:

- выполнить кластеризацию k-средних для данного набора данных для разных значений k (диапазоны от 1 до 10);
- для каждого значения K вычисляется значение WCSS;
- строит кривую между рассчитанными значениями WCSS и количеством кластеров k;
- острая точка изгиба или точка графика выглядит как рука, тогда эта точка считается лучшим значением k.

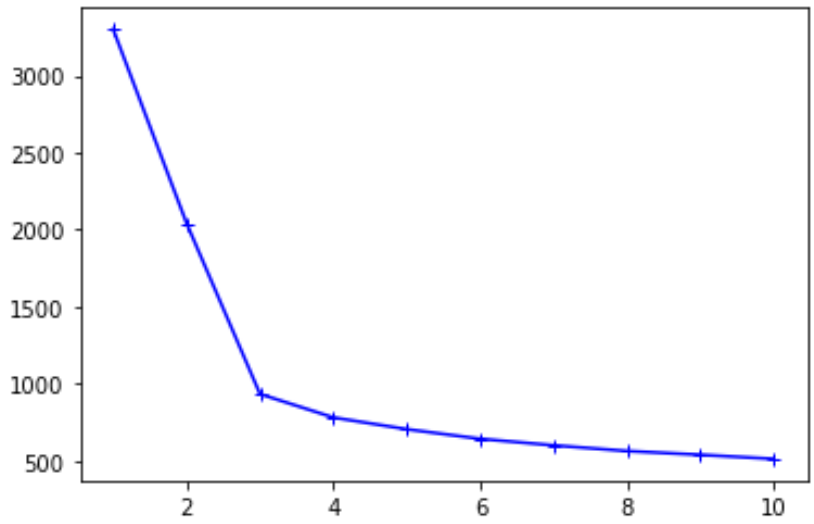

Рисунок 3.9 – График метода локтя

График на рисунке 3.9 показывает, что k = 3 – хороший выбор. Временами может быть сложно определить подходящее количество кластеров, так как кривая может монотонно убывать и не показывать резких изгибов или как в этом примере кривая может иметь очевидную точку, в которой она начинает сглаживаться.

Теперь рассмотрим вариант определения количества кластеров путем построения дендрограммы для иерархической кластеризации, которая изображена на рисунке 3.10:

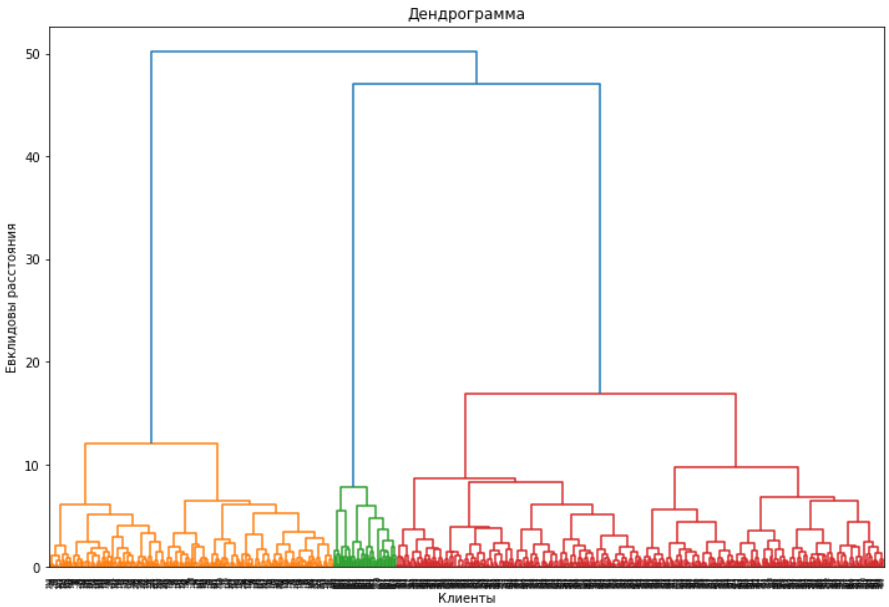

Рисунок 3.10 – Дендограмма иерархической кластеризации

На дендрограмме, которая изображена на рисунке 3.10 мы можем увидеть, что опстимальное число кластеров так же равняется 3.

Алгоритм начинается с поиска двух близлежащего друг к другу точек данных на основе евклидова расстояния.

После формирования одного большого кластера выбирается самое высокое вертикальное расстояние без проходящей через него горизонтальной линии, и по нему чертится одна горизонтальная линия. Количество вертикальных линий, которые проходят через проведенную горизонтальную линию, равно количеству кластеров.

3.5 Обучение модели

После того как выяснили, что оптимальное число кластеров 3 начинаем обучение модели.

Обучение модели k-средних. Пример кода показан на рисунке 3.11:

```
k_means = KMeans(n_clusters = 3, random_state = 42)
k_means.fit(X)
pred_km = k_means.labels_
```

Рисунок 3.11 – Обучение модели методом кластеризации k-средних

В pred_km сохраняем значения кластеров.

Далее создаем новый набор данных с дополнительным столбцом со значениями кластеров. Пример кода на рисунке 3.12:

```
data_km = pd.concat([data, pd.DataFrame({"cluster":pred_km})], axis=1)
```

```
data_km.head()
```

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	cluster
0	100000	2	1	1	0	0
1	50000	3	0	10	9	2
2	50000	7	1	3	4	0
3	30000	5	1	1	4	0
4	100000	6	0	12	3	1

Рисунок 3.12 – Новый набор данных с данными о кластеризации k-средних

По той же схеме обучаем модель и создаем новый набор данных с кластерами по алгоритму иерархической кластеризации. Пример кода на рисунке 3.13:

```
#Euclidean distance, and ward linkage (Евклидово расстояние и связь Уорда)
hc = AgglomerativeClustering(n_clusters = 3, affinity = 'euclidean', linkage = 'ward')
hc.fit(X)
pred_hc = hc.labels_
```

```
data_hc = pd.concat([data,pd.DataFrame({"cluster":pred_hc})], axis=1)
```

```
data_hc.head()
```

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	cluster
0	100000	2	1	1	0	0
1	50000	3	0	10	9	1
2	50000	7	1	3	4	0
3	30000	5	1	1	4	0
4	100000	6	0	12	3	2

Рисунок 3.13 – Обучение модели и создание нового набора данных с данными о иерархической кластеризации

3.6 Построение графиков

Построим диаграмму pairplot с результатами кластеризации. Диаграмма pairplot кластеризация k-средних приведена на рисунке 3.14:

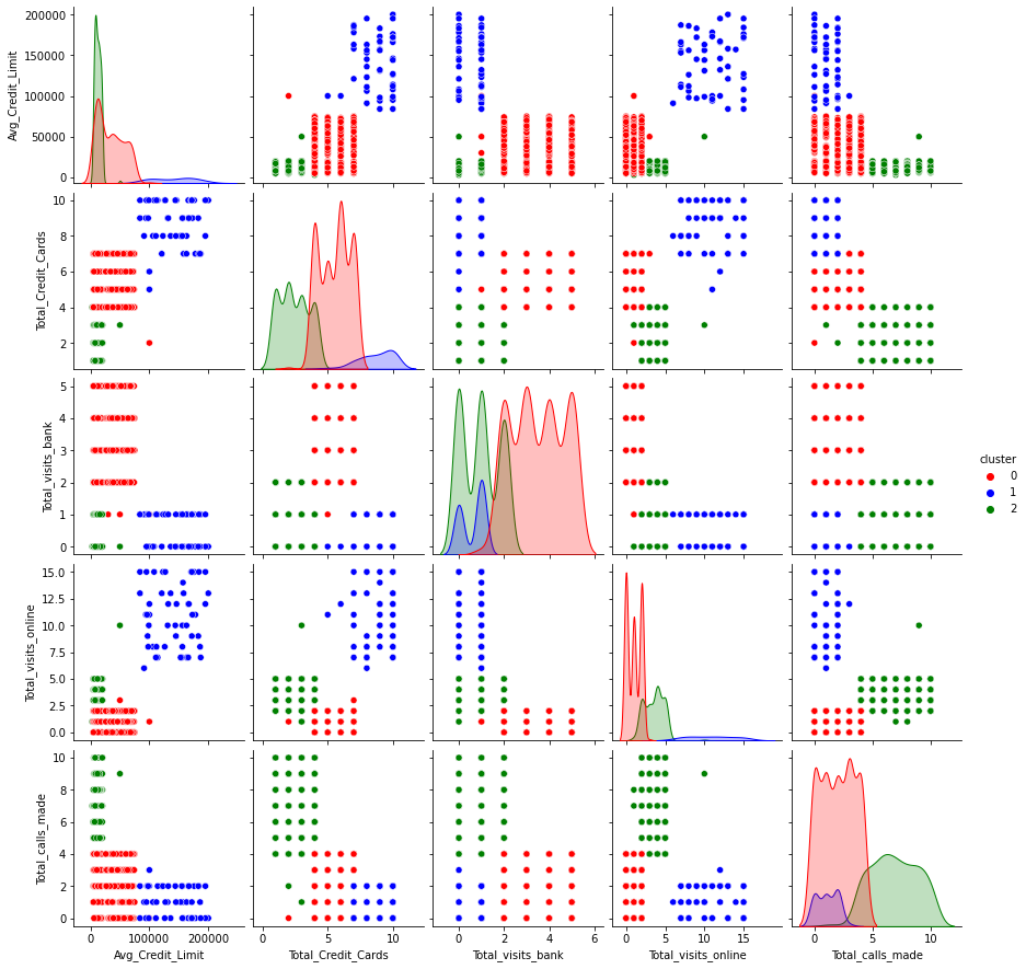

Рисунок 3.14 – Парные диаграммы всех столбцов с цветовой пометкой кластеров кластеризации k-средних

Диаграмма pairplot для иерархической кластеризации приведена на рисунке 3.15:

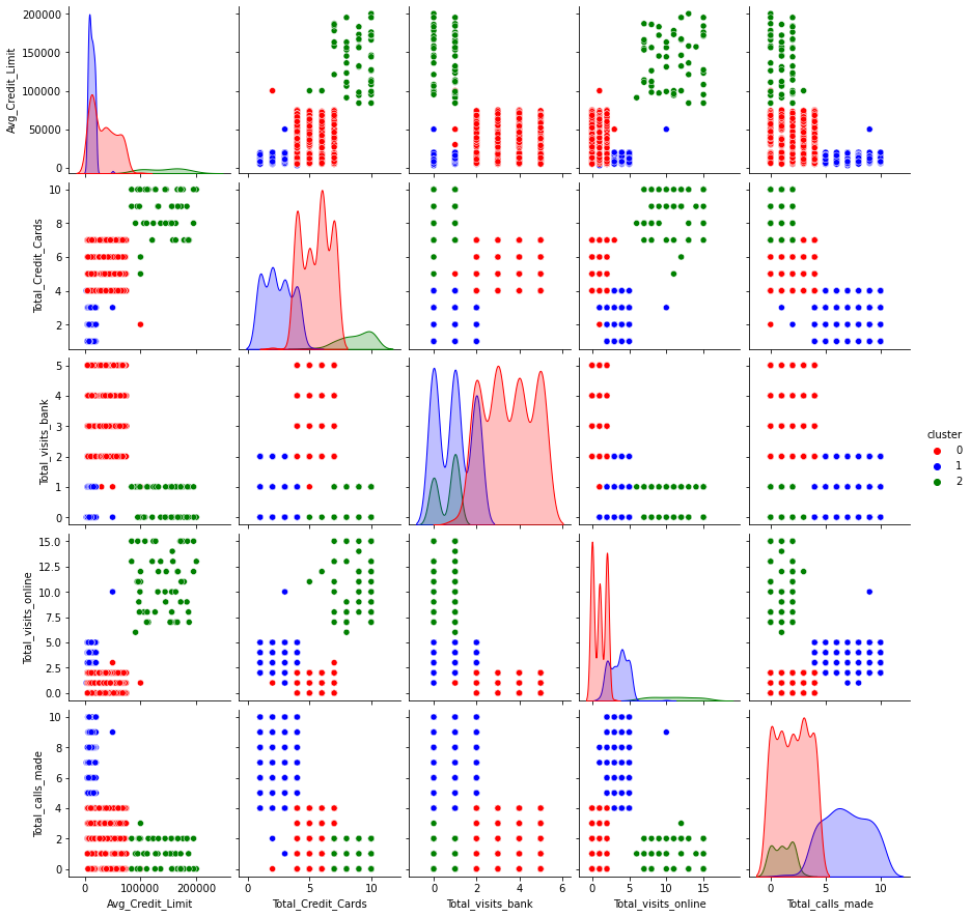

Рисунок 3.15 – Парные диаграммы всех столбцов с цветовой пометкой кластеров иерархической кластеризации

По графикам можно заметить, что оба метода кластеризации справились с задачей одинаково. На них можно заметить четкие разделения по кластерам.

Для еще одного примера визуализации данных используем метод анализа главных компонент (PCA - Principal Component Analysis).

Анализ главных компонентов (PCA) - это метод, используемый для подчеркивания вариативности и выявления сильных закономерностей в наборе данных. Его часто используют для упрощения изучения и визуализации данных.

Зная зависимости и их силу, мы можем выразить несколько признаков через один, слить воедино, так сказать, и работать уже с более простой моделью. Конечно, избежать потерь информации, скорее всего не удастся, но минимизировать ее нам поможет как раз метод PCA[16].

Визуализация кластеризации k-средних применением метода главных компонент на рисунке 3.16:

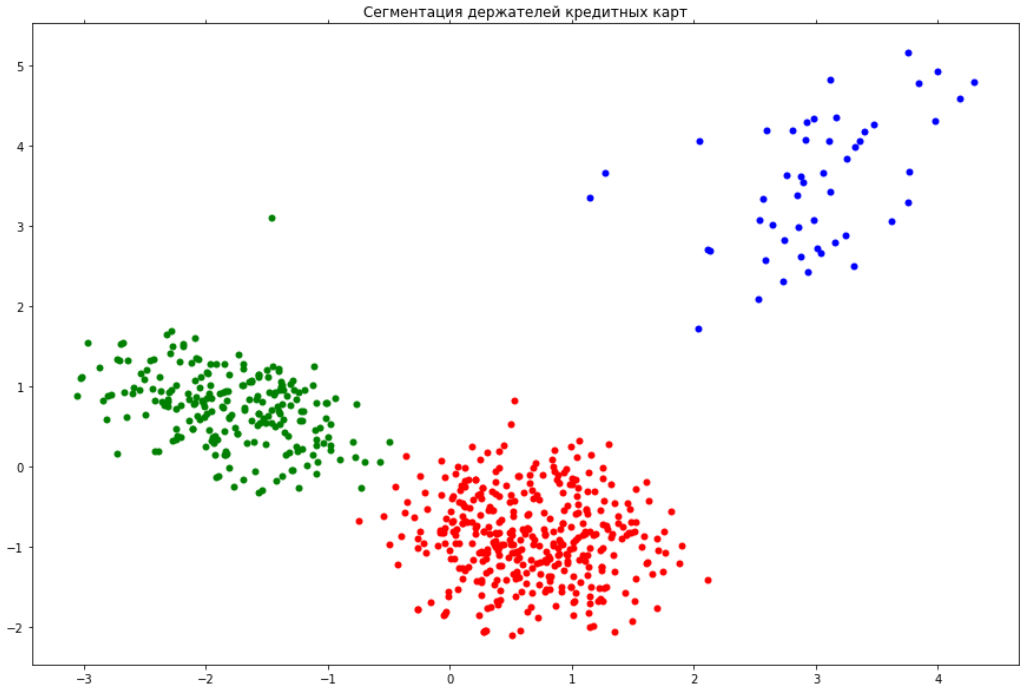

Рисунок 3.16 – Вид кластеризации k-средних применением метода главных компонент на наборе данных

Визуализация иерархической кластеризации применением метода главных компонент на рисунке 3.16:

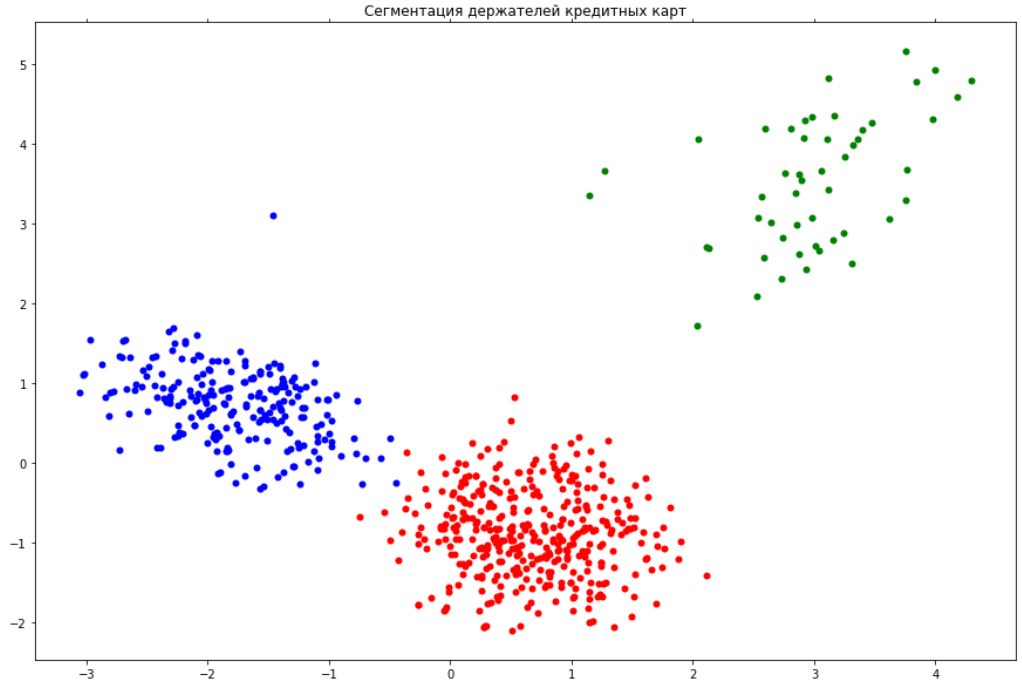

Рисунок 3.17 – Вид иерархической кластеризации применением метода главных компонент на наборе данных

Можно заметить совсем не большое различие между двумя графиками.

Построим диаграммы размахов (boxplot) по кластерам и по каждому параметру.

Диаграмма размахов по среднему кредитному лимиту на рисунке 3.18:

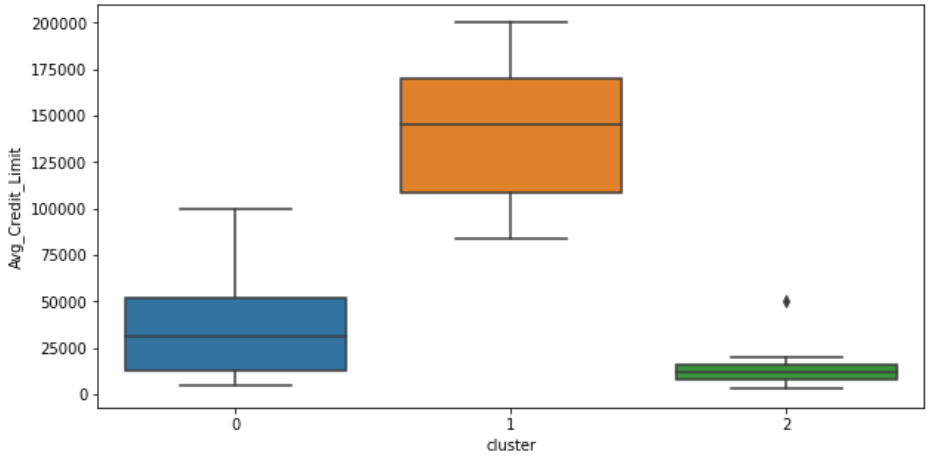

Рисунок 3.18 – Диаграмма размахов кредитного лимита по кластерам

Как видно из рисунка, распределение кредитного лимита в разных кластерах разные и они отличаются не только по значениям медиан, но и по разбросам значений.

Диаграмма размахов кластеров по количеству кредитных карт на рисунке 3.19:

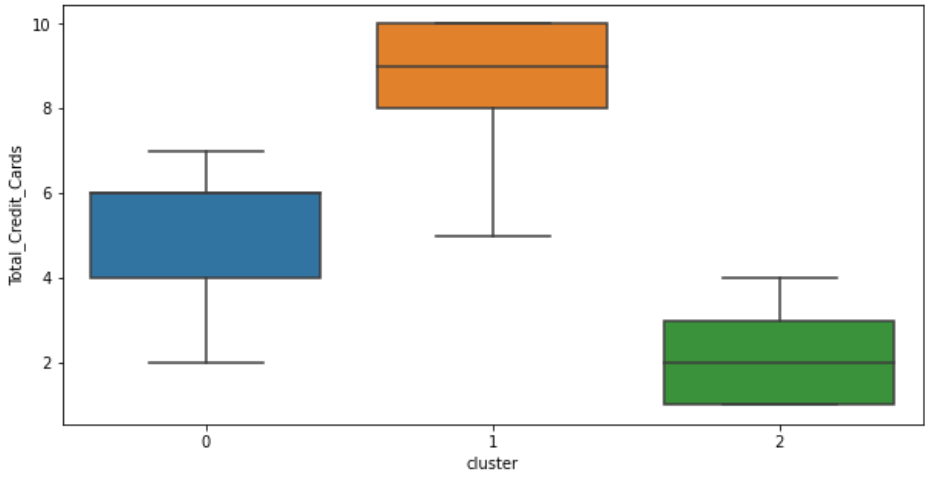

Рисунок 3.19 – Диаграмма размахов количества кредитных карт по кластерам

Диаграмма размахов кластеров по количеству визитов в банк на рисунке 3.20:

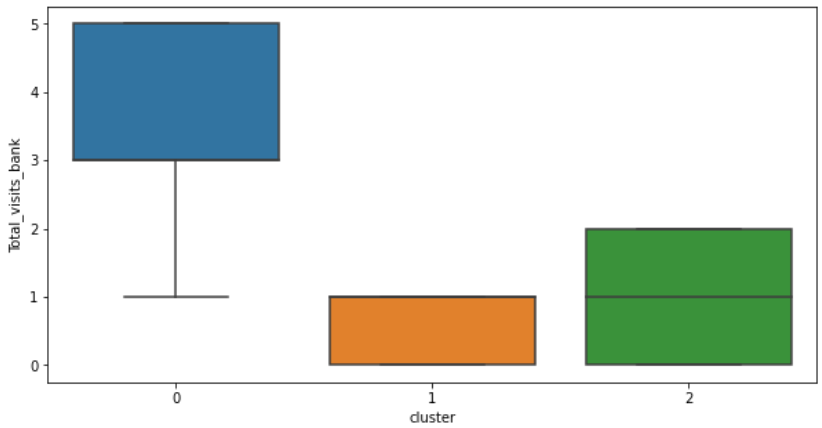

Рисунок 3.20 – Диаграмма размахов количества визитов в банк по кластерам

Диаграмма размахов кластеров по количеству визитов онлайн на рисунке 3.21:

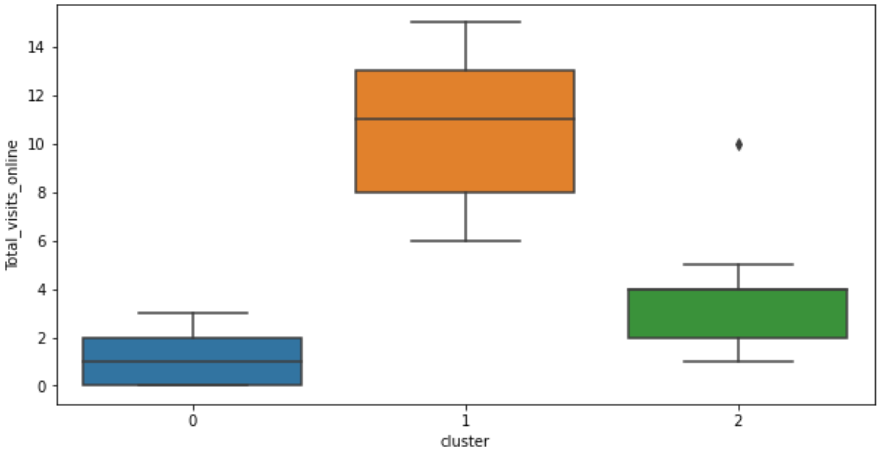

Рисунок 3.21 – Диаграмма размахов количества визитов онлайн по кластерам

Диаграмма размахов кластеров по количеству совершенных звонков на рисунке 3.22:

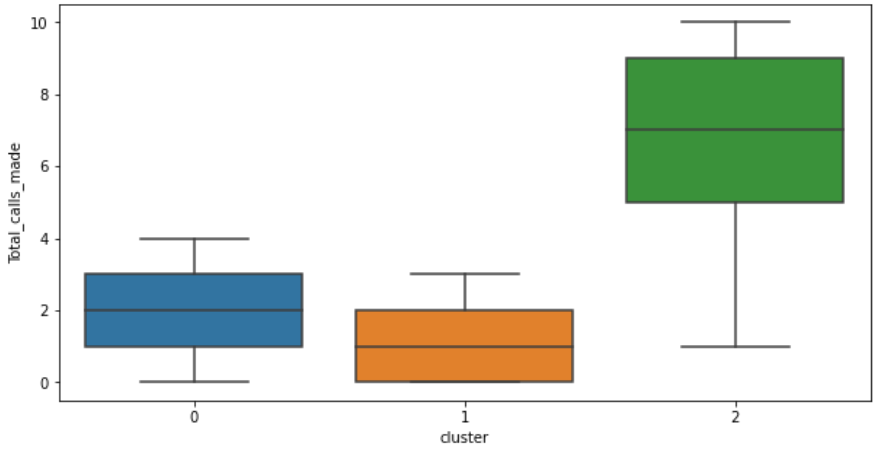

Рисунок 3.22 – Диаграмма размахов количества совершенных звонков по кластерам

Как видно из представленных рисунков в кластер 0 входят клиенты со средним кредитным лимитом, также клиенты этой категории часто ходят в отделения банка, редко пользуются онлайн услугами, количество звонков в среднем 1-3 раза.

Могут быть нацелены на перекрестные продажи через рекламные объявления в банке и через менеджеров банка

Кластер 1 имеет следующие характеристики:

- клиенты с высоким кредитным лимитом;
- часто пользуются онлайн услугами.

Предположительно современные клиенты с доходом выше среднего, которые хорошо разбираются в онлайн услугах. Могут быть привлечены посредством бонусов за покупки через интернет, предложениями на электронную почту. Предположительно наиболее являются наиболее прибыльными клиентами

Кластер 2 имеет следующие характеристки:

- клиенты с низким кредитным лимитом;
- редко посещают отделения банка;
- временами посещают банк онлайн;
- частые звонки.

Это могут быть клиенты с низкой заработанной платой, которые могут часто иметь проблемы с доходом, связи с чем часто обращаются к кредитным услугам банка. Возможно частые звонки связаны с тем, что у них могут быть проблемы с выплатой кредитов. Их можно использовать для перекрестных продаж посредством телефонных звонков.

После визуального исследования распределений применим формальный критерий Краскела-Уоллиса. Формальный критерий выполним на примере параметров количества кредитных карт (Total_Credit_Cards) и количества совершенных звонков (Total_calls_made). Так как значение p-value близко к нулю, то на уровне значимости 5% нулевая гипотеза отклоняется. Это означает, что распределения количества кредитных карт и количества совершенных звонков отличаются по кластерам. Таким образом, медианы показателей количество кредитных карт и количества совершенных звонков по разным кластерам не равны. Пример кода с выводом на рисунке 3.23:

```
groups = {}
for grp in data_hc['cluster'].unique():
    groups[grp] = data_hc['Total_Credit_Cards'][data_hc['cluster']==grp].values
#print(groups)
args = groups.values()
scipy.stats.kruskal(*args)
```

```
KruskalResult(statistic=459.4650897232965, pvalue=1.692089977152644e-100)
```

```
groups = {}
for grp in data_hc['cluster'].unique():
    groups[grp] = data_hc['Total_calls_made'][data_hc['cluster']==grp].values
#print(groups)
args = groups.values()
scipy.stats.kruskal(*args)
```

```
KruskalResult(statistic=428.5875993316468, pvalue=8.577985327534083e-94)
```

Рисунок 3.23 – Критерий Краскела-Уоллиса

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены вопросы изучения методов кластерного анализа для процессов в социально-экономических системах. Кластерный анализ выполнен на примере набора данных о клиентах банка – держателей кредитных карт. По поставленным в работе целям решены следующие задачи:

- проведен обзор существующих алгоритмов кластеризации;
- исследованы алгоритмы метода кластерного анализа;
- выбраны инструментальные средства для проведения кластерного анализа;
- проведено сравнение методов кластеризации;
- метод кластерного анализа применялся при классификации данных о клиентах банка;
- была проведена интерпретация полученных результатов.

Были использованы следующие методы:

- методы многомерного статистического анализа;
- алгоритмы кластерного анализа.

Полученные результаты работы могут быть использованы при сегментации клиентов банка – держателей кредитных карточек.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1 Современные тенденции в кластерном анализе. Бериков В. С., Лбов Г. С., 2013 г.

2 K-means and K-medoids applet. Mirkes E.M., University of Leicester, 2011.

3 Лекции по алгоритмам кластеризации и многомерного шкалирования. Воронцов К. В., 2010 г.

4 Python и машинное обучение. Рашка С., 2015 г.

5 Python и анализ данных. Маккинни У., 2015 г.

6 Изучаем pandas. Хейдт М., 2018 г.

7 Beginning Python Visualization: Crafting Visual Transformation Scripts. Shai Vaingast, 2009 г.

8 Прикладная математическая статистика. Кобзарь А. И., 2006 г.

9 Основные таблицы математической статистики. Ликеш И., Ляга Й., 1985 г.

10 Use of ranks in one-criterion variance analysis. Kruskal W. H., Wallis W. A., 1952 г.

11 Pattern Recognition with Fuzzy Objective Function Algoritms. Bezdek J. C., 1982 г.

12 Кластерный анализ. Мандель И. Д., 1988 г.

13 Применение кластерного анализа в государственном управлении. Хайдуков Д. С. 2009 г

14 Иерархический кластер-анализ и соответствия. Жамбю М., 1988 г.

15 Hierarchical grouping to optimize an objective function. Ward J.H., 1963 г.

16 A Tutorial on Principal Component Analysis. Jonathon Shlens, 2014 г.

ПРИЛОЖЕНИЕ

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

df = pd.read_csv('Credit Card Customer Data.csv')

df.head(5)

df.tail(5)

df.info()

data = df.drop(['Sl_No','Customer Key'],axis=1).copy()

data.describe().T

sns.pairplot(data)

sns.heatmap(data.corr(), cmap = "BrBG", annot = True)

from sklearn.preprocessing import StandardScaler, MinMaxScaler # Стандартизация
from sklearn.cluster import KMeans # Метод k-средних
from sklearn.decomposition import PCA # Метод главных компонентов
from sklearn.metrics.pairwise import cosine_similarity

X = data.copy()

scaler = StandardScaler()
X = scaler.fit_transform(X)
X.shape

clusters = 11
cost = []
for i in range(1,clusters):
   kmeans = KMeans(n_clusters = i, random_state = 42)
   kmeans.fit(X)
   cost.append(kmeans.inertia_)

plt.plot(range(1, 11), cost, "b+-")
```

```
plt.show()

k_means = KMeans(n_clusters = 3, random_state = 42)
k_means.fit(X)
pred_km = k_means.labels_

data_km = pd.concat([data, pd.DataFrame({"cluster":pred_km})], axis=1)

data_km.head()

pca = PCA()
principal_components = pca.fit_transform(X)
x, y = principal_components[:,0], principal_components[:,1]

print(principal_components.shape)

colors = {0:"red",1:"blue",2:"green"}

final_df = pd.DataFrame({'x': x, 'y':y, 'label':pred_km})
groups = final_df.groupby(pred_km)

fig, ax = plt.subplots(figsize=(15, 10))

for name, group in groups:
    ax.plot(group.x, group.y, marker='o', linestyle='', ms=6, color=colors[name],
mec='none')
    ax.set_aspect('auto')

ax.tick_params(axis='x',which='both',bottom='off',top='off',labelbottom='off')
    ax.tick_params(axis= 'y',which='both',left='off',top='off',labelleft='off')

ax.set_title("Сегментация держателей кредитных карт")
plt.show()

sns.pairplot(data_km, hue='cluster', palette = ["red","blue","green"])

import scipy.cluster.hierarchy as sch

plt.figure(figsize=(12,8))

dendrogram = sch.dendrogram(sch.linkage(X, method = 'ward'))

plt.title('Дендрограмма')
plt.xlabel('Клиенты')
```

```
        plt.ylabel('Евклидовы расстояния')
        plt.show()

        from sklearn.cluster import AgglomerativeClustering

        #Euclidean distance, and ward linkage (Евклидово расстояние и связь
Уорда)
        hc = AgglomerativeClustering(n_clusters = 3, affinity = 'euclidean', linkage =
'ward')
        hc.fit(X)
        pred_hc = hc.labels_

        data_hc = pd.concat([data,pd.DataFrame({"cluster":pred_hc})], axis=1)

        data_hc.head()

        pca = PCA()
        principal_components = pca.fit_transform(X)
        x, y = principal_components[:,0], principal_components[:,1]

        print(principal_components.shape)

        colors = {0:"red",1:"blue",2:"green"}

        final_df = pd.DataFrame({'x': x, 'y':y, 'label':pred_hc})
        groups = final_df.groupby(pred_hc)

        fig, ax = plt.subplots(figsize=(15, 10))

        for name, group in groups:
            ax.plot(group.x, group.y, marker='o', linestyle='', ms=6, color=colors[name],
mec='none')
            ax.set_aspect('auto')

ax.tick_params(axis='x',which='both',bottom='off',top='off',labelbottom='off')
            ax.tick_params(axis= 'y',which='both',left='off',top='off',labelleft='off')

        ax.set_title("Сегментация держателей кредитных карт")
        plt.show()

        sns.pairplot(data_hc, hue='cluster', palette = ["red","blue","green"])

        plt.subplots(figsize=(10,5))
        sns.boxplot(x = 'cluster', y = 'Avg_Credit_Limit', data = data_km)
```

```
plt.show()

plt.subplots(figsize=(10,5))
sns.boxplot(x = 'cluster', y = 'Total_Credit_Cards', data = data_km)
plt.show()

plt.subplots(figsize=(10,5))
sns.boxplot(x = 'cluster', y = 'Total_visits_bank', data = data_km)
plt.show()

plt.subplots(figsize=(10,5))
sns.boxplot(x = 'cluster', y = 'Total_visits_online', data = data_km)
plt.show()

plt.subplots(figsize=(10,5))
sns.boxplot(x = 'cluster', y = 'Total_calls_made', data = data_km)
plt.show()
```

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН
Сәтбаев университеті

ОТЗЫВ
НАУЧНОГО РУКОВОДИТЕЛЯ

на дипломный проект

Досман Әлижан Махамедкалиұлы

(Ф.И.О. обучающегося)

5В070300 – Информационные системы

Тема: «Применение методов кластеризации для анализа данных»

В социально-экономических исследованиях приходится оперировать данными с десятками признаков, зафиксированных по определенному объекту. Целью работы является исследование методов кластеризации, используемых в таком многомерном статистическом анализе на основе данных клиентов банка.

В результате выполнения работы были получены следующие результаты:

- проведен обзор существующих алгоритмов кластеризации;
- отобраны инструментальные средства для проведения кластерного анализа;
- исследованы методы кластерного анализа на примере данных клиентов банка – держателей кредитных карт;
- проведена интерпретация результатов.

Результаты работы могут быть использованы при создании групп клиентов банков – держателей кредитных карт.

В процессе выполнения работы Досман Ә. показал умение работать с технической литературой, проводить анализ предметной области информационных систем, навыки владения программными средствами. Все разделы дипломной работы выполнены автором самостоятельно.

В связи с вышеизложенным, считаю, что работа Досман Әлижан Махамедкалиұлы на тему «Применение методов кластеризации для анализа данных» может быть допущена к защите.

Научный руководитель,
ассоц. проф., к.т.н. Сейлова Н.А.

Ф КазНИТУ 706-16. Отзыв научного руководителя

Протокол анализа Отчета подобия

заведующего кафедрой / начальника структурного подразделения

Заведующий кафедрой / начальник структурного подразделения заявляет, что ознакомился(-ась) с Полным отчетом подобия, который был сгенерирован Системой выявления и предотвращения плагиата в отношении работы:

Автор: Лосина Ольгой Александровны

Название: Применение методов кластеризации для анализа данных

Коэффициент подобия 1: 0.00
Коэффициент подобия 2:0.0

Замена букв:0
Интервалы:0
Микропробелы:2
Белые знаки:0

После анализа отчета подобия заведующий кафедрой / начальник структурного подразделения констатирует следующее:

☐ обнаруженные в работе заимствования являются добросовестными и не обладают признаками плагиата. В связи с чем, работа признается самостоятельной и допускается к защите;
☐ обнаруженные в работе заимствования не обладают признаками плагиата, но их чрезмерное количество вызывает сомнения в отношении ценности работы по существу и отсутствием самостоятельности ее автора. В связи с чем, работа должна быть вновь отредактирована с целью ограничения заимствований;
☐ обнаруженные в работе заимствования являются недобросовестными и обладают признаками плагиата, или в ней содержатся преднамеренные искажения текста, указывающие на попытки сокрытия недобросовестных заимствований. В связи с чем, работа не допускается к защите.

Обоснование:
Заимствования являются добросовестными и не обладают признаками плагиата.

Дата 31.05.2021 Сейдаев Н.А., зав. кафедрой КВОиХИ

Ф КубГТУ 711-27. Протокол анализа Отчета подобия заведующего кафедрой

Протокол анализа Отчета подобия Научным руководителем

Заявляю, что я ознакомился(-ась) с Полным отчетом подобия, который был сгенерирован Системой выявления и предотвращения плагиата в отношении работы:

Автор: Досман Эдуард Михайлович

Название: Применение методов кластеризации для анализа данных

Коэффициент подобия 1: 0.00

Коэффициент подобия 2: 0.00

Замена букв: 0
Интервалы: 0
Микропробелы: 2
Белые знаки: 0

После анализа Отчета подобия констатирую следующее:

- ☐ обнаруженные в работе заимствования являются добросовестными и не обладают признаками плагиата. В связи с чем, признаю работу самостоятельной и допускаю ее к защите;
- ☐ обнаруженные в работе заимствования не обладают признаками плагиата, но их чрезмерное количество вызывает сомнения в отношении ценности работы по существу и отсутствием самостоятельности ее автора. В связи с чем, работа должна быть вновь отредактирована с целью ограничения заимствований;
- ☐ обнаруженные в работе заимствования являются недобросовестными и обладают признаками плагиата, или в ней содержатся преднамеренные искажения текста, указывающие на попытки сокрытия недобросовестных заимствований. В связи с чем, не допускаю работу к защите.

Обоснование:
Заимствования являются добросовестными и не обладают признаками плагиата.

31.05.2021
Дата *Подпись Научного руководителя*