**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод Кластеризации K-Mean

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Громова В.С. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи.**

1. Добавить к набору данных новый атрибут, вычислив его из имеющихся;
2. Очистить данные, удалив выбросы и дубли; обработать пропущенные значения; найти кривые данные и т.п.;
3. Построить графики зависимости одной переменной от другой, описать какие кластеры (группы) можно заметить. Построить несколько разных графиков, найти ярко выраженные группы.
4. Применить метод кластеризации K-mean к набору данных; проанализировать полученные результаты.

**Выполнение работы.**

Для данной лабораторной работы был выбран новый набор данных([Customer Segmentation (K-Means) | Analysis](https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis/input))

1. **Добавление нового атрибута**

Для добавления нового атрибута в датасет был использован атрибут, который является комбинацией двух других числовых признаков: **"Annual Income (k$)"** и **"Spending Score (1-100)"**. Создан новый столбец под названием Income\_Spending\_Score, который рассчитывается как произведение этих двух атрибутов:

Этот новый атрибут может быть полезен для дальнейшего анализа, поскольку он представляет собой сочетание уровня дохода и склонности к расходам. Это позволяет лучше понять поведение клиентов с учетом их финансовых возможностей и трат.

**2. Обработка данных**

Для подготовки данных к кластеризации были выполнены несколько операций очистки и предобработки:

**Удаление дубликатов**: Были удалены все строки с повторяющимися значениями, что позволяет избежать искажения результатов анализа.

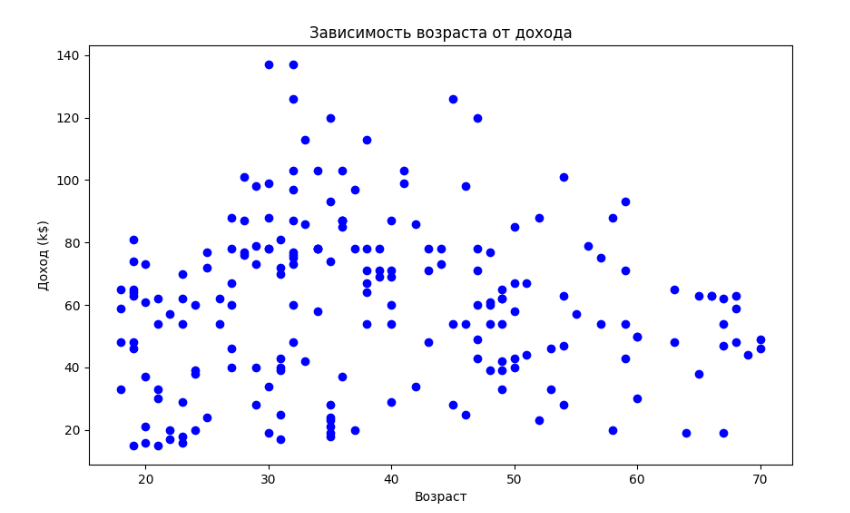
**Обработка пропущенных значений**: Для числовых столбцов были заполнены пропущенные значения средним значением по каждому из столбцов.

Таким образом, все числовые данные в датасете стали полными и готовыми для анализа.

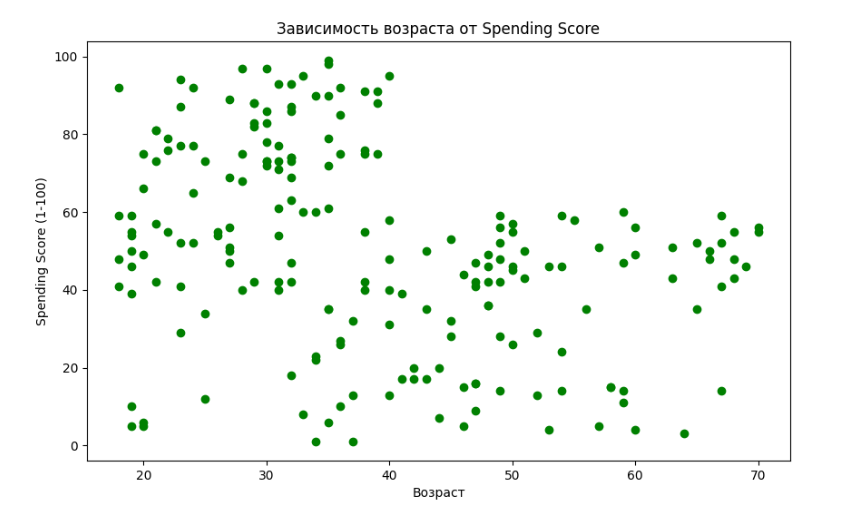
**3. Построение графиков зависимостей**

После предобработки данных были построены графики, отображающие зависимости между различными переменными, чтобы исследовать потенциальные кластеры.

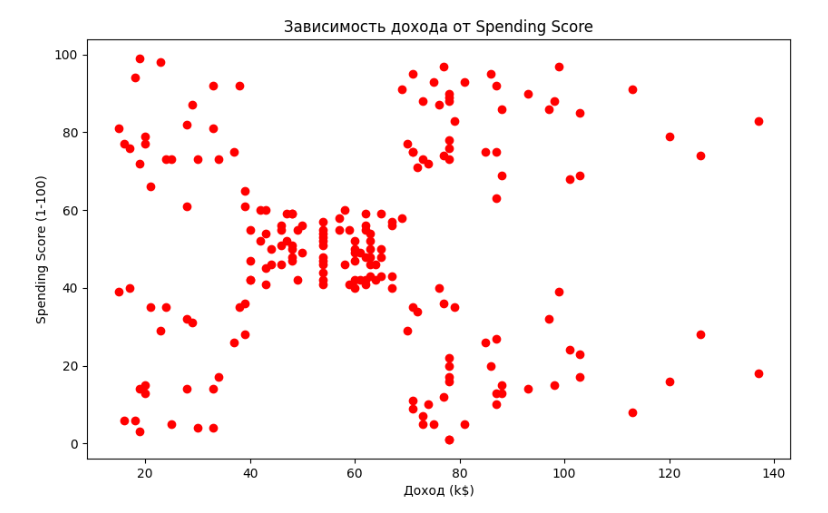
* **График зависимости возраста от дохода**



* **График зависимости возраста от Spending Score**:



* **График зависимости дохода от Spending Score**:



**4. Применение метода KMeans для кластеризации**

Метод **KMeans** был применен для разделения данных на несколько кластеров. Для этого было использовано 3 кластера

После выполнения кластеризации каждому клиенту был присвоен кластер, и эта информация была добавлена в новый столбец **'Cluster'**. Для того чтобы корректно выполнить кластеризацию, данные были предварительно стандартизированы с помощью **StandardScaler**.

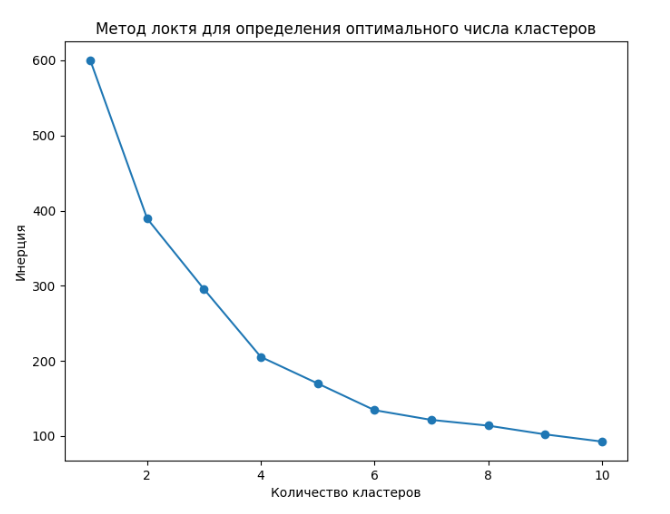
После того как кластеры были определены, была построена визуализация, отображающая распределение клиентов по кластерам:



На графике разные кластеры отображаются разными цветами, что позволяет наглядно увидеть, как разделяются клиенты в зависимости от их возраста и дохода.

**5. Оценка оптимального числа кластеров с использованием метода локтя**

Для того чтобы определить оптимальное количество кластеров, был использован метод локтя. Этот метод позволяет увидеть, при каком числе кластеров инерция (сумма квадратов расстояний внутри кластеров) начинает значительно снижаться, что указывает на оптимальное количество кластеров.



На графике видно, что инерция значительно снижается при переходе от одного кластера к двум, а затем уменьшается медленно. Этот момент на графике может указывать на оптимальное число кластеров для дальнейшего анализа.

**6. Выводы о числе кластеров**

После выполнения кластеризации и оценки оптимального числа кластеров, было определено, что оптимальное количество кластеров в данной задаче составляет **3**. Это количество кластеров было использовано в методе KMeans для разделения данных. Результаты кластеризации были визуализированы и проанализированы для дальнейшего использования.

Итак, на основе анализа было выделено три кластера, которые, возможно, отражают разные группы клиентов с различными характеристиками, такими как возраст, доход и поведение в плане расходов.

**Выводы.**

В ходе работы был проведен анализ данных о клиентах, содержащих информацию о возрасте, годовом доходе и индексе расходования. В процессе работы были добавлены новые атрибуты, очищены данные от пропущенных значений и дубликатов, а также выполнена стандартизация. Методом K-means была проведена кластеризация клиентов, выделены три группы, различающиеся по возрасту и уровню дохода. Кластеры позволили выявить разные типы потребительских групп, что может быть полезно для дальнейшего маркетингового анализа и таргетинга. В результате были получены навыки работы с методом кластеризации K-means и оценкой оптимального числа кластеров с использованием метода локтя.