Министерство науки и образования РФ

Федеральное государственное бюджетное учреждение

высшего образования

**«Тверской государственный технический университет»**

(ТвГТУ)

Кафедра программного обеспечения

**Отчет по лабораторной работе №5**

По дисциплине: «Анализ больших данных»

Тема: «Методы снижения размерности и задача кластеризации»

|  |
| --- |
| Выполнила:  студентка группы  Б.ПИН.РИС - 21.06  Александрова М.А. |
| Проверила:  старший преподаватель  кафедры ПО  Корнеева Е.И. |

Тверь, 2025

**Оглавление**

[**Введение** 3](#_Toc194746636)

[**Общее задание на работу** 3](#_Toc194746637)

[**База данных 1** 4](#_Toc194746638)

[**Описание базы данных – признаки, целевая переменная;** 4](#_Toc194746639)

[**Исследовательский анализ данных;** 4](#_Toc194746640)

[**Выводы по анализу:** 5](#_Toc194746641)

[**Снижение размерности. KernelPCA** 5](#_Toc194746642)

[**Вывод по метрикам:** 8](#_Toc194746643)

[**Оптимальное число кластеров. Метод локтя. Метод силуэта.** 9](#_Toc194746644)

[Алгоритм 1. 10](#_Toc194746645)

[Алгоритм 2. 12](#_Toc194746646)

[**Использованные материалы** 14](#_Toc194746647)

# **Введение**

# **Общее задание на работу**

Подробное описание:

1. Загрузить один из наборов данных. Загружаются все файлы по ссылке.
2. Провести разведочный анализ данных, ответив на следующие вопросы:
3. Сколько строк в датафрейме, сколько столбцов
4. Сколько места занимает датафрейм в оперативной памяти
5. Для каждой интервальной переменной подсчитать следующее - мин, медиана, среднее, макс и персентили 25, 75
6. Для каждой категориальной переменной рассчитать моду и сколько раз мода встречается в данных
7. Подготовка датасета к построению моделей ML
8. Провести анализ и обработку пропусков (либо заменить, либо удалить)
9. Провести анализ и обработку выбросов (либо заменить, либо удалить)
10. Провести анализ и обработку категориальных переменных (сколько таких переменных, закодируйте категориальные переменные одним из методов ( one hot encoding, mean target, frequence encoding)
11. Разделить датасет на трейн и тест
12. Построить следующие классификационные алгоритмы:
13. Knn
14. Logistic regression
15. SVM
16. Оценить качество алгоритмов, выбрать самый оптимальный алгоритм
17. Возможно ли улучшить алгоритм, предложите идеи?
18. База данных 1
    1. Описание базы данных – признаки, целевая переменная;
    2. Исследовательский анализ данных (вопросы и задания);
    3. Выводы по анализу;
    4. Алгоритм 1;
    5. Визуализация результата;
    6. Критерии качества;
    7. Алгоритм 2;
    8. Визуализация результата
    9. Критерии качества;
    10. Алгоритм 3;
    11. Визуализация результата
    12. Критерии качества;
    13. Вывод по результату и критериям.

**Ссылка на репозиторий**: <https://github.com/milana-cat/BigData.git>

# **База данных 1**

## **Описание базы данных – признаки, целевая переменная;**

В качестве первой базы данных была выбрана база данных качества вина.Она содержит данные о составе вина. Целевая переменная – качество (quality)

## **Исследовательский анализ данных;**

<class 'pandas.core.frame.DataFrame'>

Index: 1121 entries, 0 to 1142

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 fixed acidity 1121 non-null float64

1 volatile acidity 1121 non-null float64

2 citric acid 1121 non-null float64

3 residual sugar 1121 non-null float64

4 chlorides 1121 non-null float64

5 free sulfur dioxide 1121 non-null float64

6 total sulfur dioxide 1121 non-null float64

7 density 1121 non-null float64

8 pH 1121 non-null float64

9 sulphates 1121 non-null float64

10 alcohol 1121 non-null float64

11 quality 1121 non-null float64

dtypes: float64(12)

memory usage: 113.9 KB

None

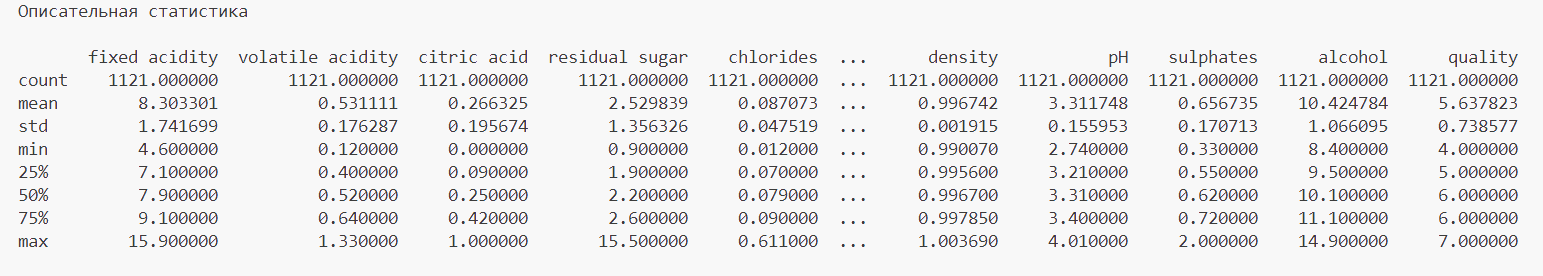
1.Данные загружены успешно.

2.A. Количество строк: 1121

Количество столбцов: 12

2.B. Объем памяти датафрейма: 113.9 KB

2.C. Статистики для интервальных переменных:



2.D. Анализ категориальных переменных:

Категориальных переменных не обнаружено.

3.A. Анализ пропусков:

Пропуски отсутствуют.

3.b. Категориальные переменные:

Категориальных переменных нет, кодирование не требуется.

3.C. Обработка выбросов:

Выбросы обработаны методом One-Hot Encoding.

## **Выводы по анализу:**

Был проведён исследовательский анализ данных. В ходе анализа были рассчитаны статистические характеристики количественных признаков, не было обнаружено пропусков и категориальных переменных.

Целевая переменная распределена биномиально. Преобладает группа качества 0.5.

## **Снижение размерности. KernelPCA**

Было проведено снижение размерности с помощью KernelPCA с различными ядрами: ['linear', 'poly', 'rbf', 'sigmoid', 'cosine']

Была построена визуализация для всех ядер:

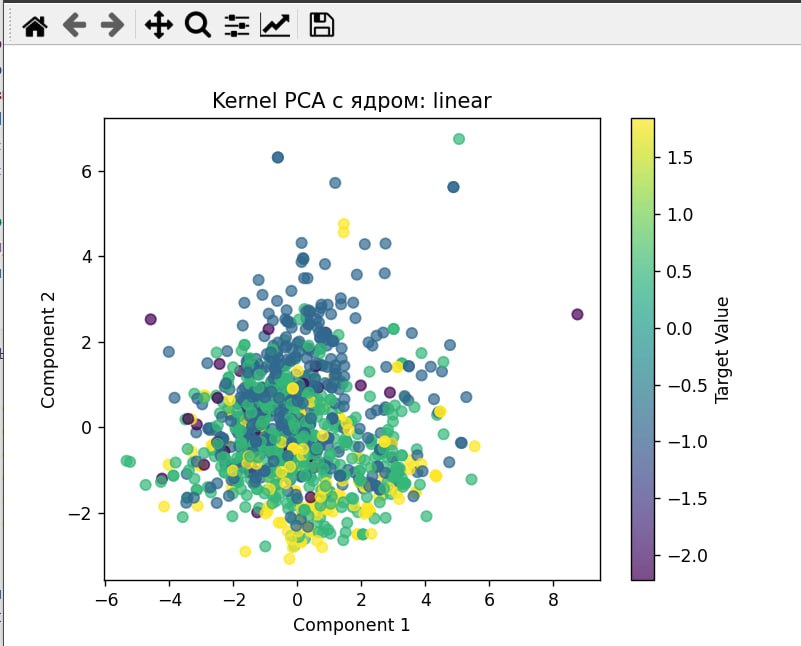


Рис. 1. Визуализация линейного ядра KernelPCA

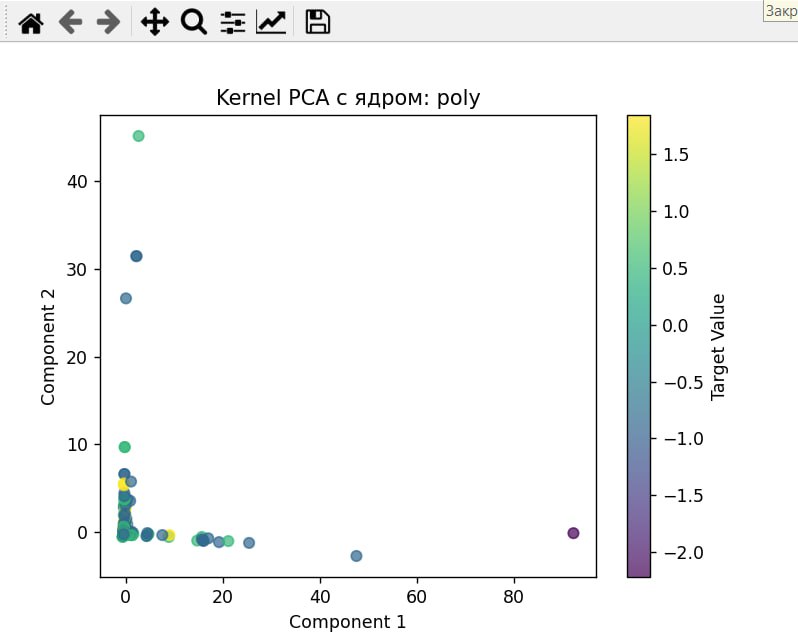


Рис. 2. Визуализация полиномиального ядра KernelPCA

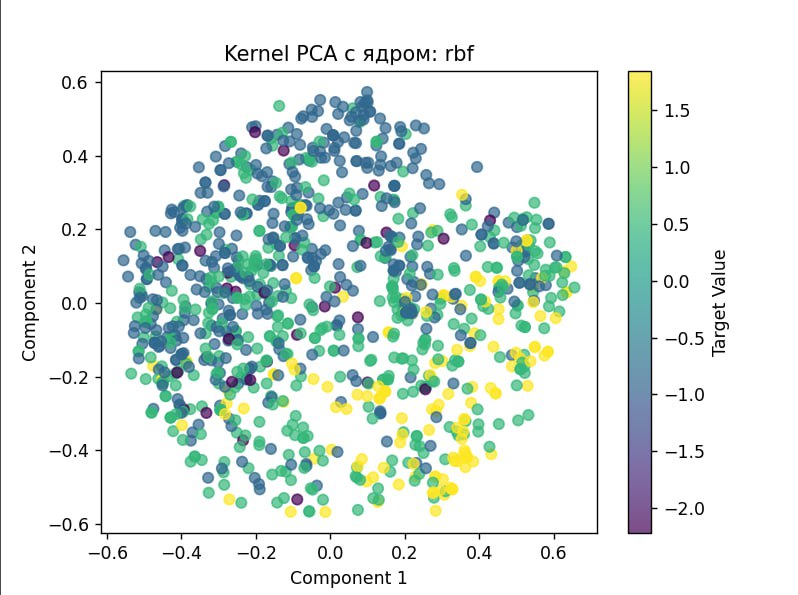


Рис. 3. Визуализация rbf ядра KernelPCA

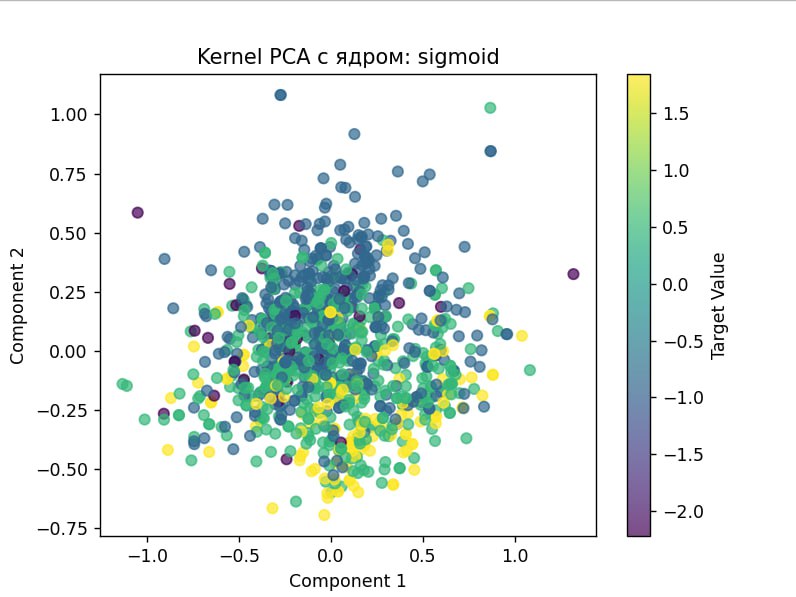


Рис. 4. Визуализация sigmoid ядра KernelPCA

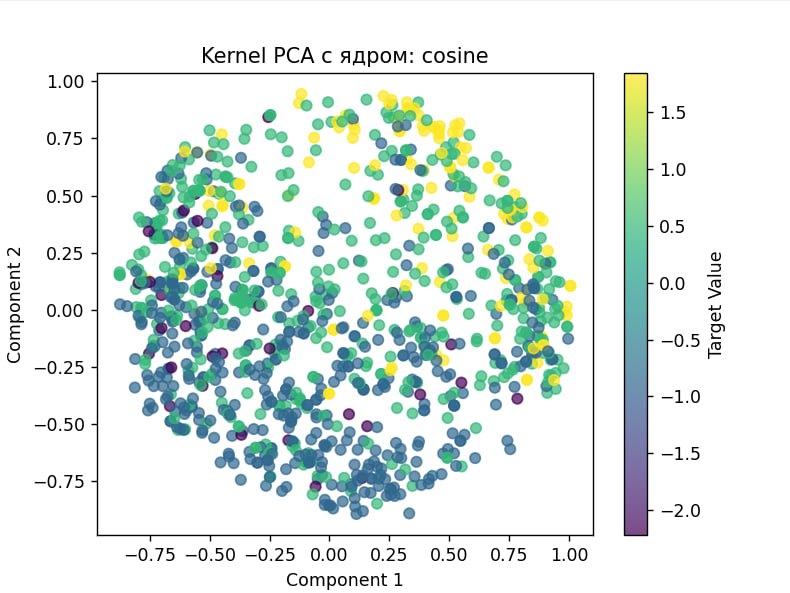


Рис. 5. Визуализация cosine ядра KernelPCA

Были вычислены метрики для линейного ядра KernelPCA:  
Общая дисперсия исходных данных: 1.0000

Общая дисперсия после Kernel PCA (linear): 1.0000

Потеря дисперсии (lost\_variance): 0.0000

### **Вывод по метрикам:**

Общая дисперсия до и после Kernel PCA (с линейным ядром) одинакова — 1.0000.

Это означает, что линейное ядро в Kernel PCA не изменило структуру данных с точки зрения общей дисперсии.

Потеря дисперсии (lost\_variance) равна 0.0000,

что указывает на то, что при понижении размерности с помощью Kernel PCA не потеряна ни одна доля информации, выраженной через дисперсию.

Так же было проведено снижение размерности методом t-SNE

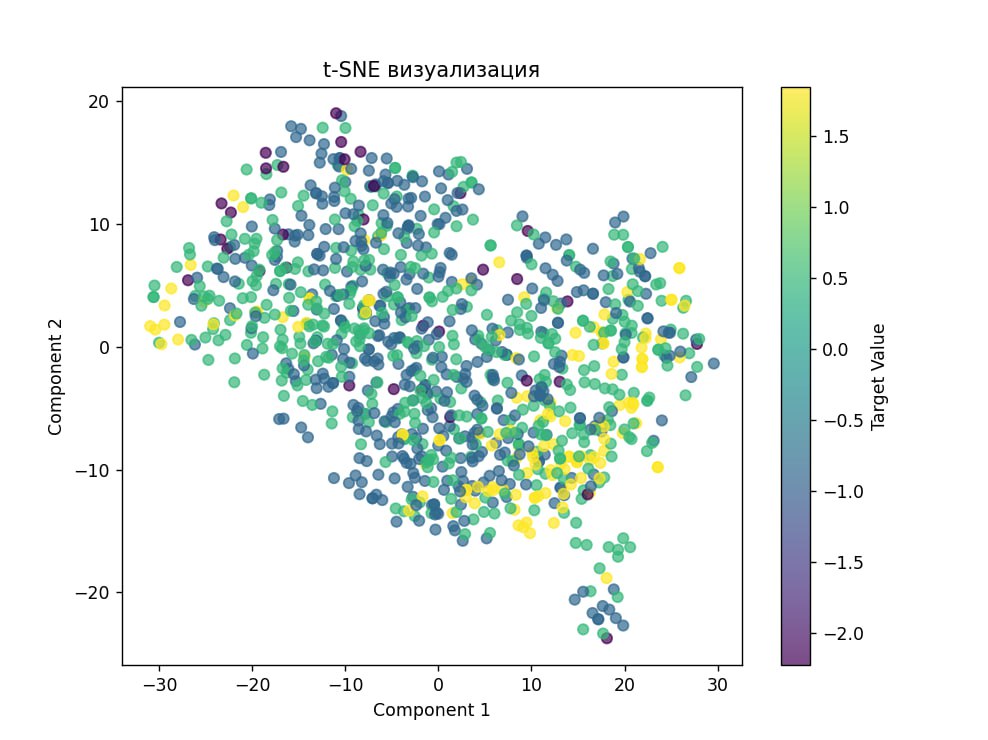


Рис. 6. Визуализация t-SNE

## **Оптимальное число кластеров. Метод локтя. Метод силуэта.**

В качестве масштабированного датасета будем использовать результат работы алгоритма Kernel PCA (cosine)

Было определено оптимальное число кластеров для кластеризации с помощью метода локтя и метода силуэта.

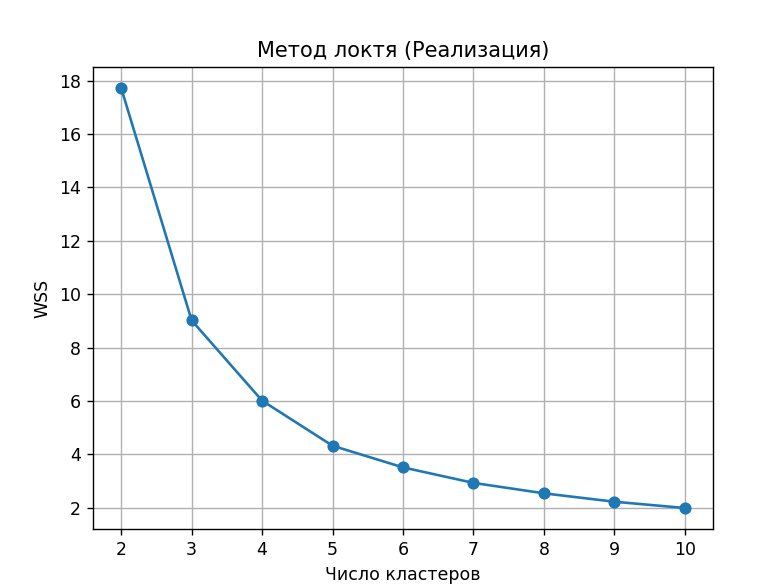


Рис. 8. Визуализация метода локтя

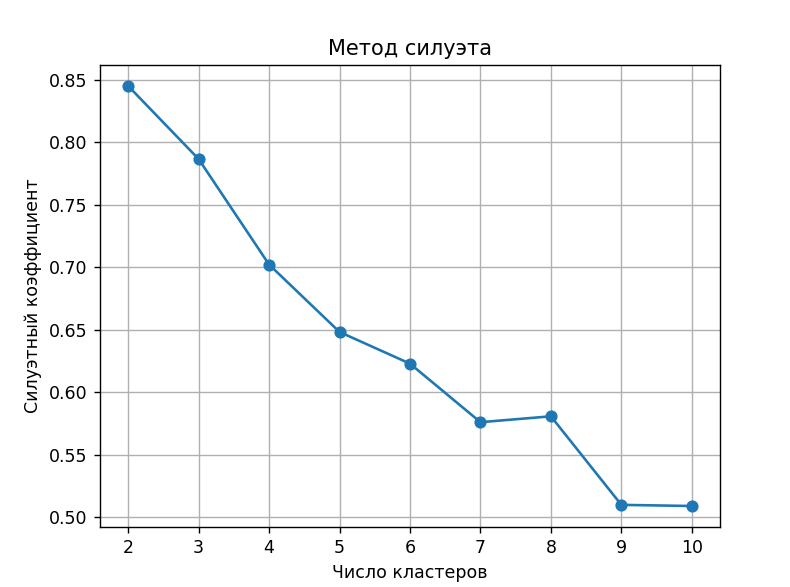


Рис. 9. Визуализация метода силуэта

**Вывод:**

Оптимальным числом кластеров является 2-3, т.к. с этого значения score методов начинает меняться незначительно.

## Алгоритм 1.

Была проведена кластеризация данных с помощью метода k-means

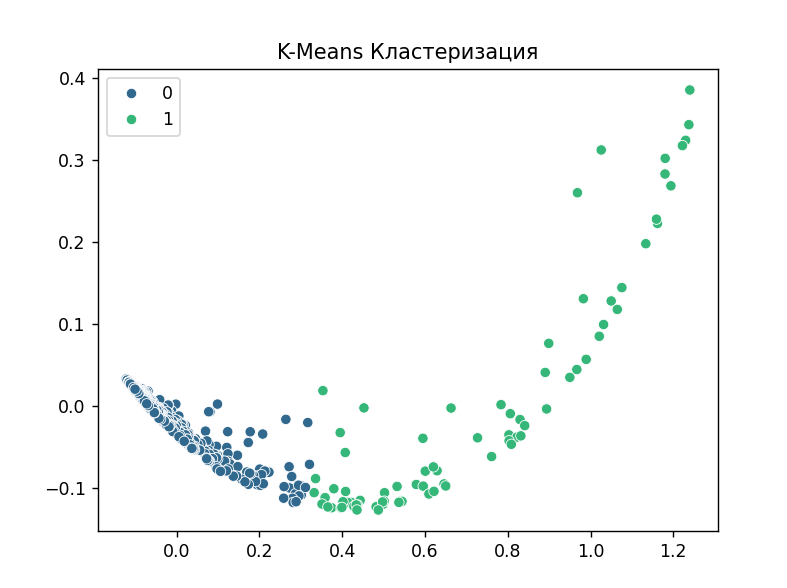


Рис. 10. Визуализация k-means кластеризации (число кластеров - 3)

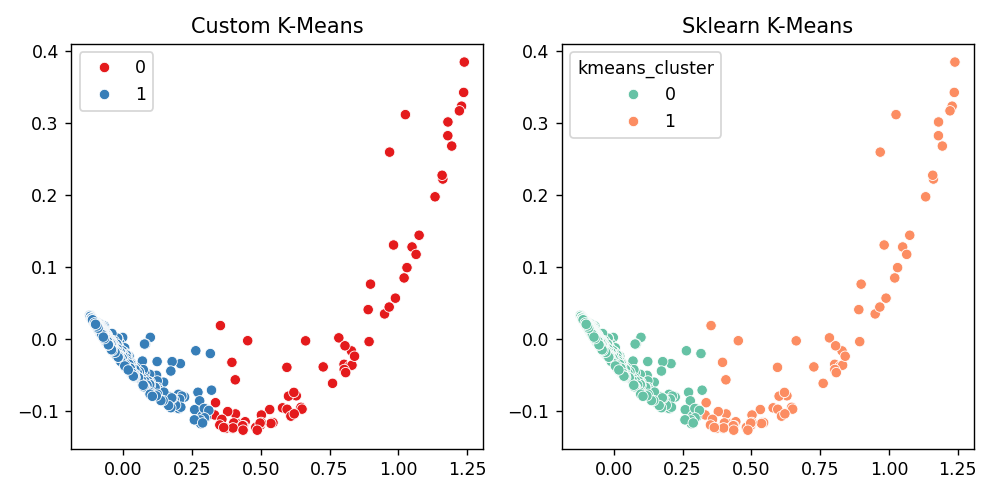


Рис. 11. Визуализация моей реализации k-means кластеризации (число кластеров - 2)

## Алгоритм 2.

Была проведена иерархическая кластеризация.

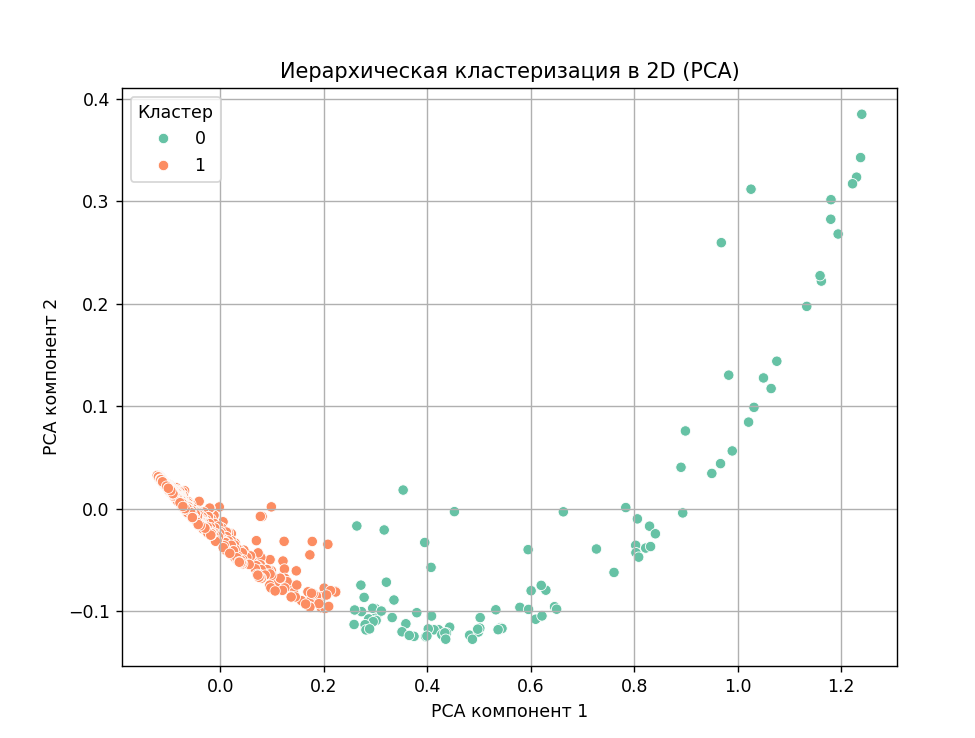


Рис. 12. Визуализация иерархической кластеризации (число кластеров - 2)

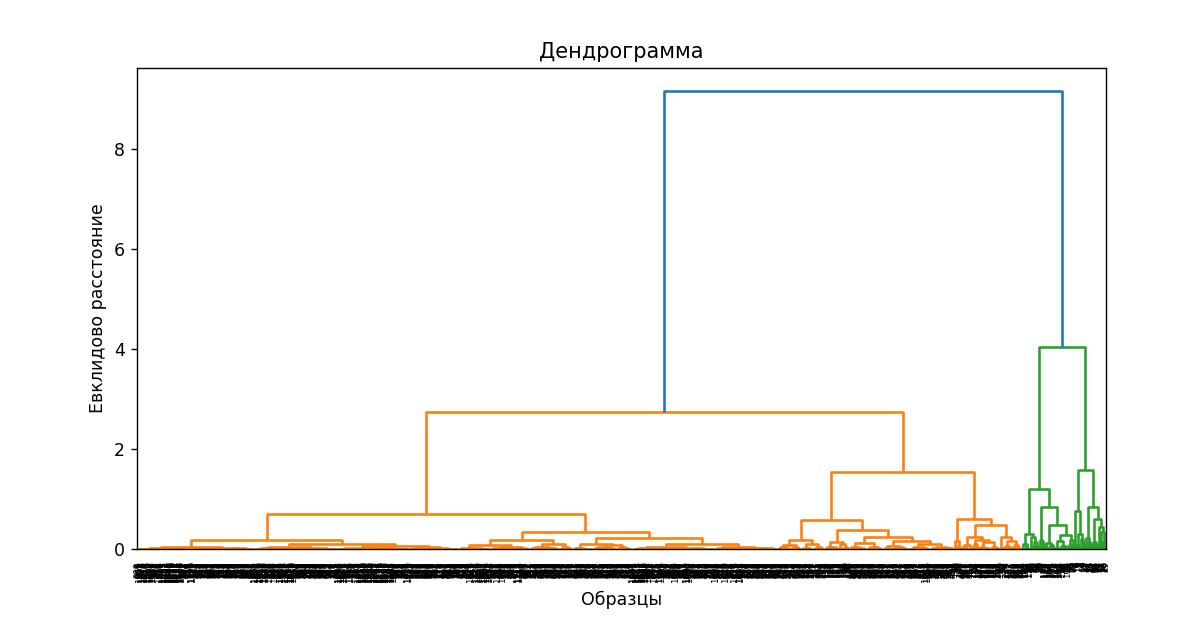
Также была построена дендрограмма иерархической кластеризации. 

Рис. 12. Дендрограмма иерархической кластеризации (число кластеров - 5)

# **Использованные материалы**

1. Документация библиотеки pandas // https://pandas.pydata.org/docs/ (дата обращения: 10.03.2025).
2. Документация библиотеки Seaborn // https://seaborn.pydata.org/ (дата обращения: 10.03.2025).
3. Документация библиотеки scikit-learn https://scikit-learn.org/stable/user\_guide.html/ (дата обращения: 10.03.2025).
4. Лекционный материал по дисциплине «Анализ больших данных»