Санкт-Петербургский Национальный Исследовательский Университет

Информационных Технологий, Механики и Оптики

Факультет Инфокоммуникационных Технологий

**Современные инструменты анализа данных**

**Лабораторная работа 1**

**Кластеризация данных.**

Выполнила

Улитина М. С.

Проверила

Добренко Н.В.

Санкт-Петербург, 2024

**Цель работы.** Освоить практическое применение методов кластеризации данных, в частности алгоритмов K-средних и DBSCAN, на различных наборах данных.

**Задачи:**

1. Изучить теоретический материал и примеры в Jupyter Notebook по [ссылке](https://colab.research.google.com/drive/17peHsZnxntneYG9ZPcuv4PDFufXfM7Bz?usp=sharing);
2. Использовать метод K-средних и метод DBSCAN на самостоятельно сгенерированной выборке;
3. Применить методы K-средних и DBSCAN к датасету Mall\_Customers:
4. Для каждого метода построить график и проанализировать результаты.

**Ход работы:**

Был изучен теоретический материал, с помощью make\_blobs, была создана выборка (рисунок 1).

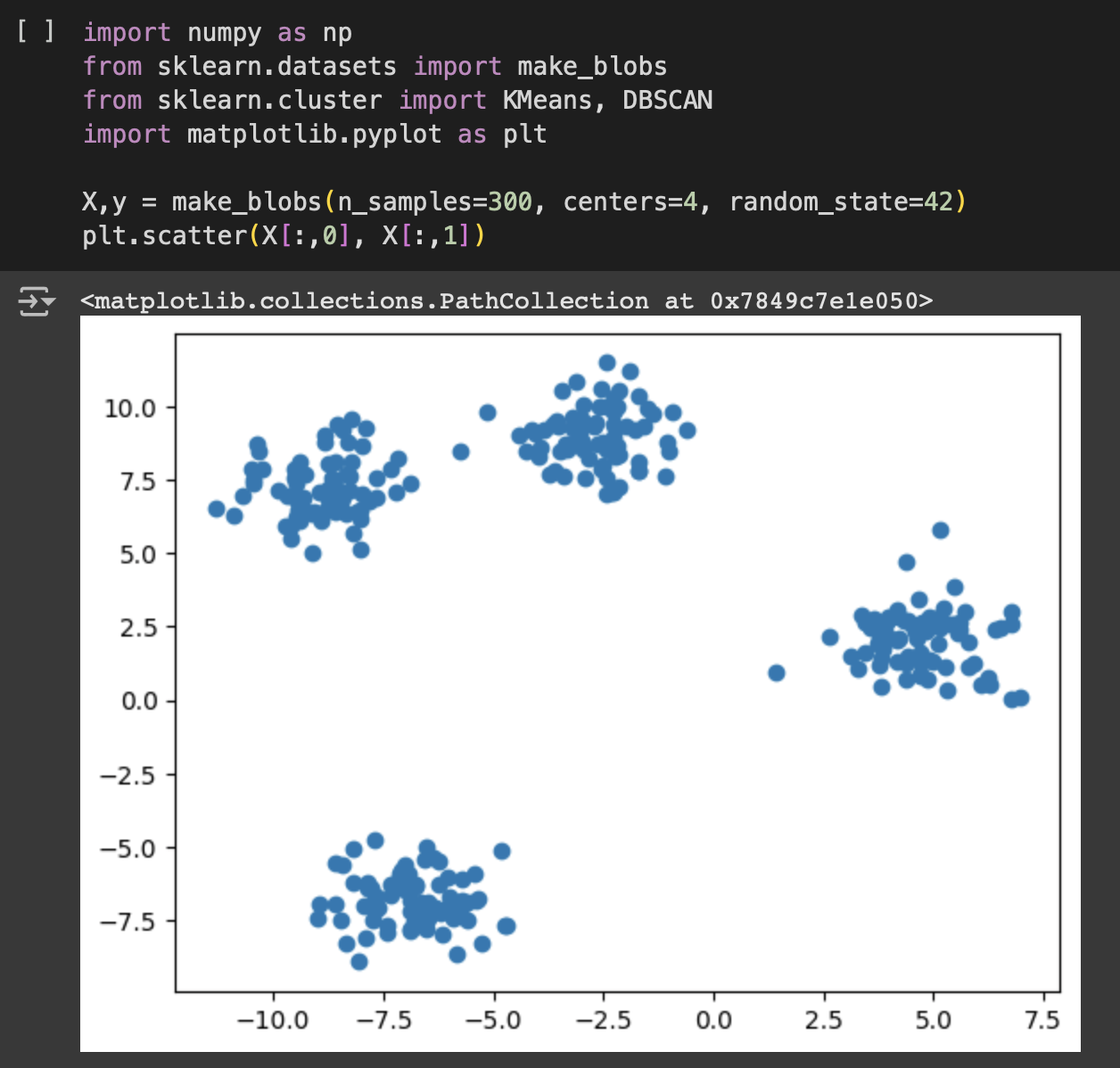


Рисунок 1 – Создание выборки.

С помощью метола локтя определено наилучшее значение для n\_clusters. По графику видно, что это значение 4.

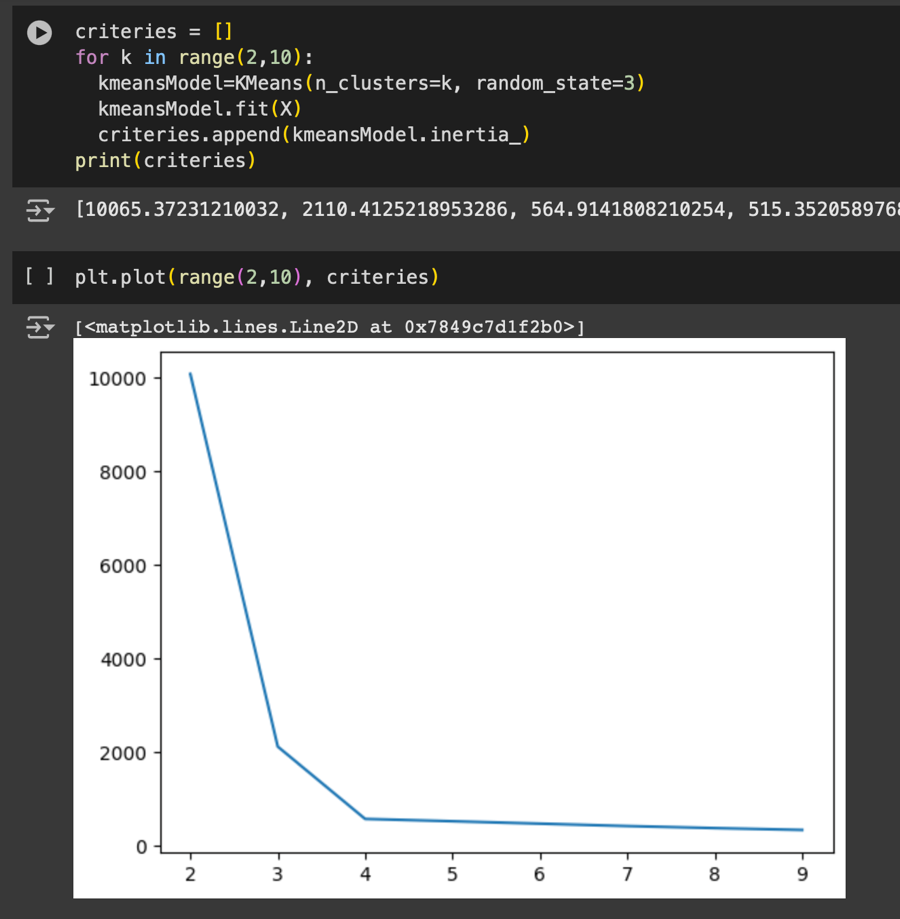


Рисунок 2 – Метод локтя.

Применен метод К-средних с n\_clusters = 4. Код и результат приведены на рисунке 3.

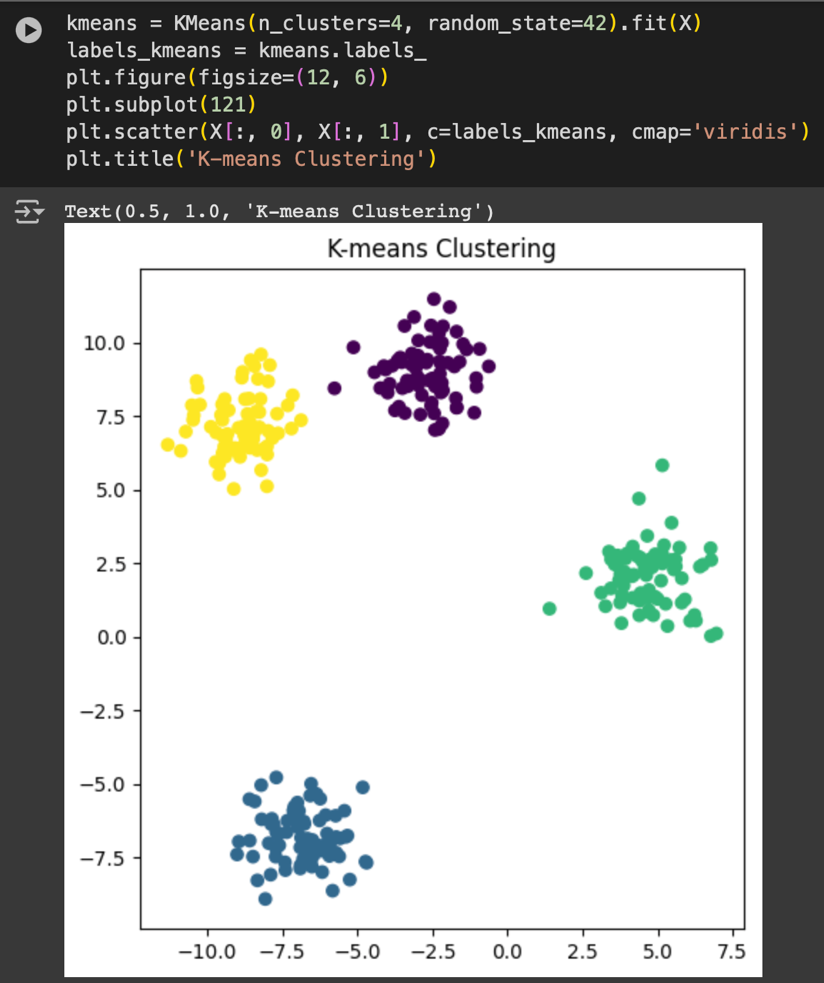


Рисунок 3 - метод К-средних.

Далее применен метод DBSCAN. Параметры подбирались вручную, лучший результат и код приведены на рисунках 4, 5.

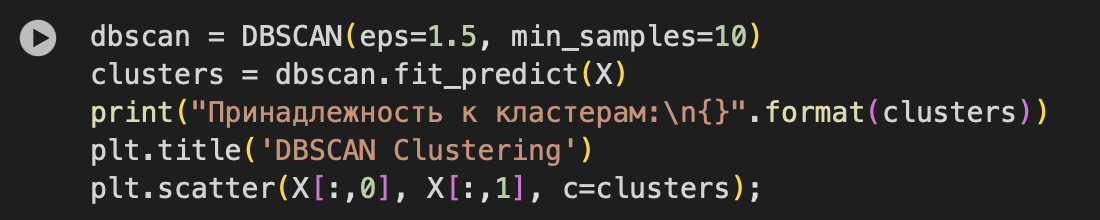


Рисунок 4 – Код DBSCAN.

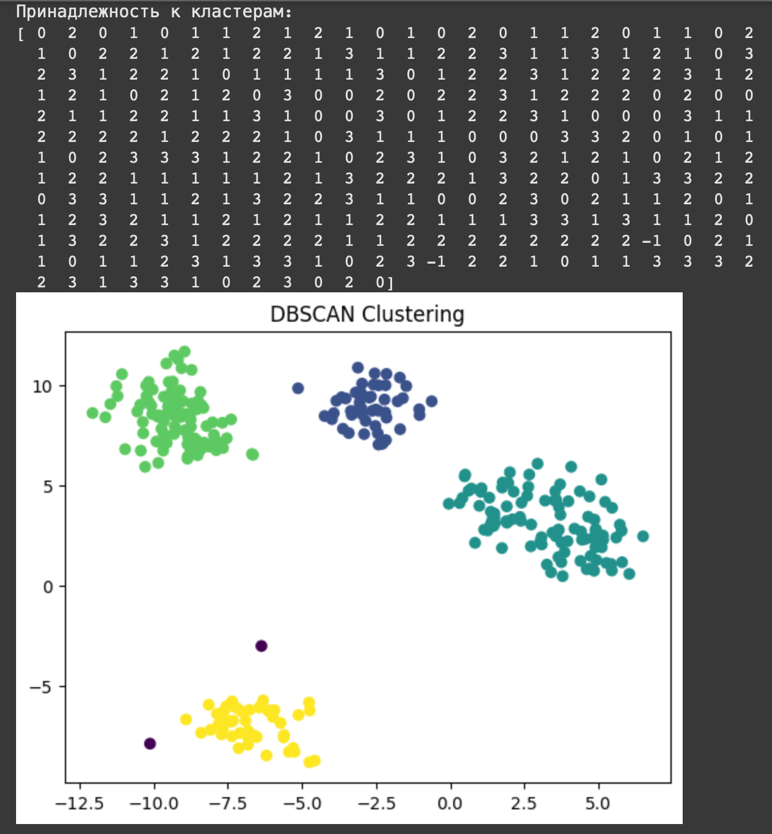


Рисунок 5 – Результат DBSCAN.

Загружены данные из файла 'Mall\_Customers.csv'. Для изучения выбраны Age Spending Score.

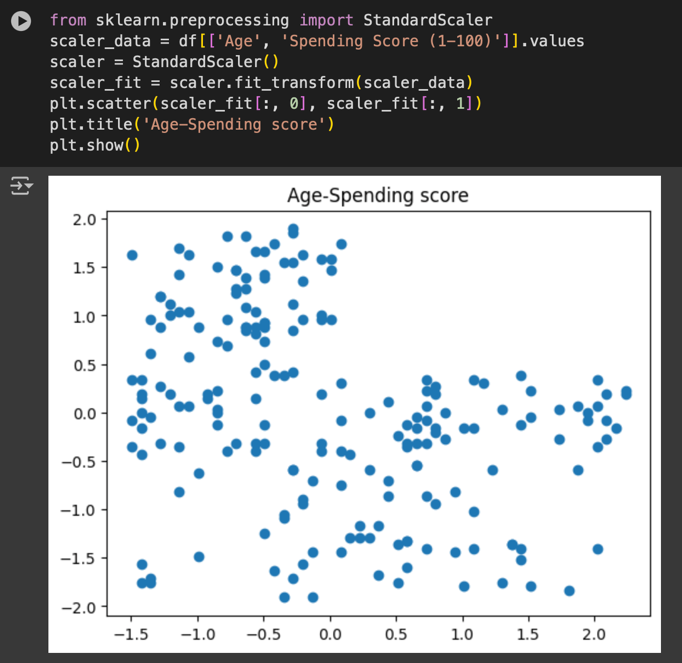


Рисунок 6 – age-spending score.

Далле аналогично с самостаятельно созданной выборкой. На рисунке 7 – метод локтя. На рисунке 8 – результат метода К-средних.

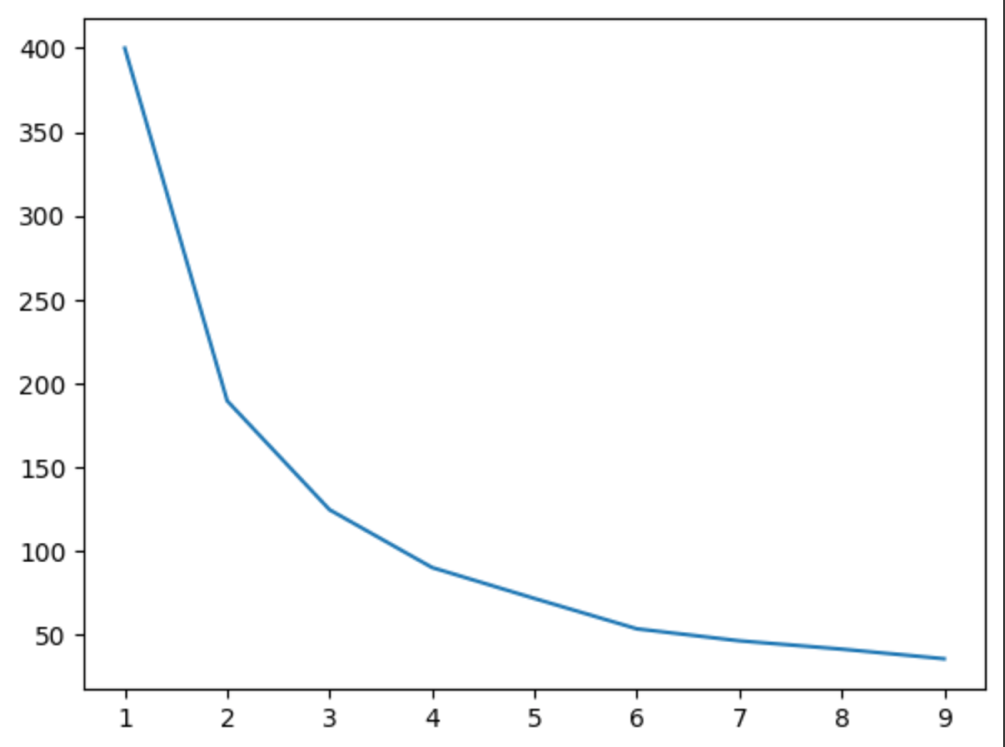


Рисунок 7 – Метод локтя.

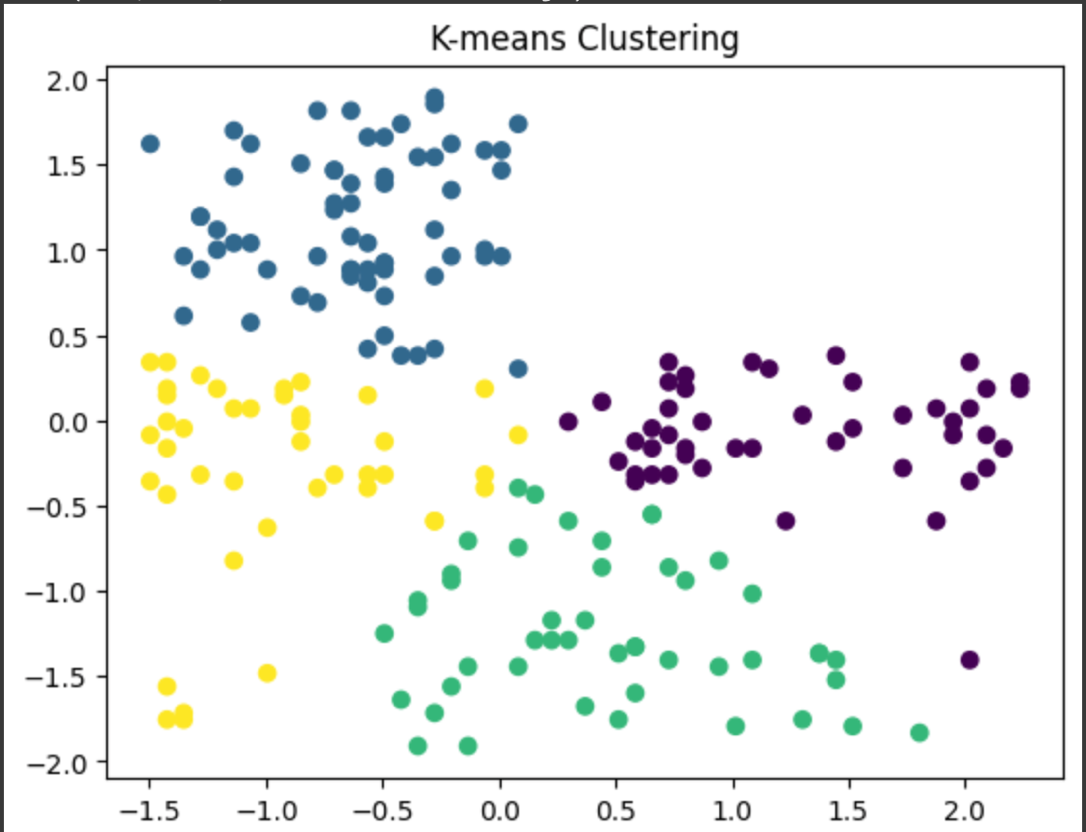


Рисунок 8 – метод К-средних.

На рисунке 9 оптимальный вариант метода DBSCAN (параметры: eps =0.3, min\_samples=4).

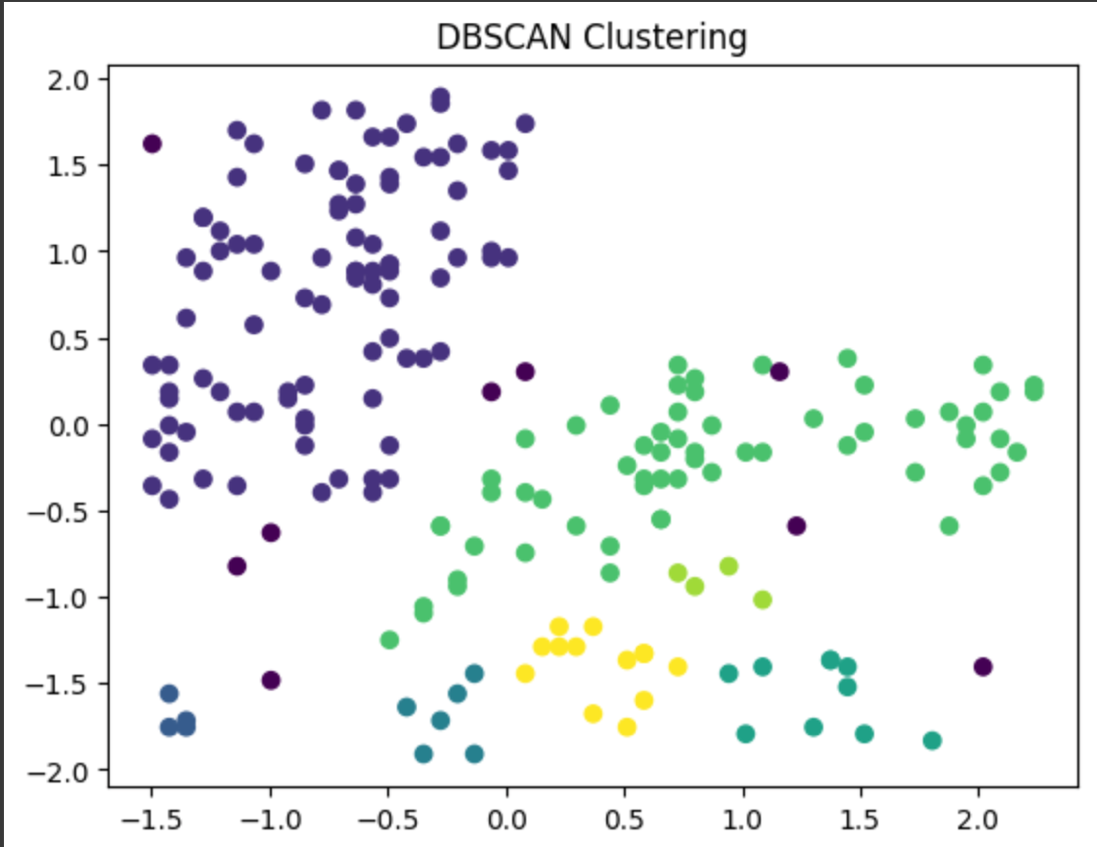


Рисунок 9 - DBSCAN