Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт информационные системы и анализ данных

Центр программной инженерии

**ОТЧЁТ**

к лабораторной работе №5 по дисциплине:

|  |
| --- |
| «Методы анализа данных» |
| Факторный анализ и главные компоненты |
| наименование темы |

Выполнилстудент ИСТб-21-1 Д.И. Морозов

номер группы подпись И. О. Фамилия

дата

Проверил Доцент Е.А. Осипова

Должность подпись И. О. Фамилия

дата

Иркутск – 2023 г.

# 1 Постановка задачи

А. Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта (табл. 1.1).

Б. Загрузить набор данных, отобрать признаки, измеренные в количественной шкале и подготовить данные для анализа. Построить диаграммы рассеяния. С использованием Python выполнить снижение размерности и факторный анализ данных. При этом необходимо:

* выполнить стандартизацию данных;
* построить корреляционную матрицу и график «тепловая карта», определить коллинеарные признаки;
* определить, какую часть общей дисперсии описывают главные компоненты, при пороговом значении 80 % общей дисперсии определить число главных компонент;
* найти две главные компоненты, построить диаграммы рассеяния для новой системы координат; сравнить диаграммы рассеяния с исходными диаграммами;
* выполнить факторный анализ без вращения, вывести матрицу нагрузок и матрицу общностей; выделить факторы и отнести к ним признаки, дать интерпретацию факторам;
* выполнить факторный анализ с вращением для выбранного числа факторов, вывести матрицу нагрузок, матрицу общностей; сравнить результат факторного анализа с результатом, полученным без вращения факторов.

В. Выполнить анализ полученных на каждом этапе результатов   
и оформить отчет по лабораторной работе.

# 2 Описание исходных данных

В качестве набора данных была выбрана статистика по весу людей. В статистику включена информация о весе, росте, о параметрах тела и возрасте.

2.1 Анализ шкал и определение переменных

Таблица 1 – Типы шкал

|  |  |  |  |
| --- | --- | --- | --- |
| **Исходное имя** | **Перевод** | **Тип шкалы** | **Тип данных** |
| Density | Плотность | Количественная | Дробные числа |
| BodyFat | телесный жир | Количественная | Дробные числа |
| Age | Возраст | Количественная | Дробные числа |
| Weight | Масса | Количественная | Дробные числа |
| Height | Высота | Количественная | Дробные числа |
| Neck | Шея | Количественная | Дробные числа |
| Chest | Грудь | Количественная | Дробные числа |
| Abdomen | Брюшная полость | Количественная | Дробные числа |
| Hip | Бедро | Количественная | Дробные числа |
| Thigh | Бедро | Количественная | Дробные числа |
| Knee | Колено | Количественная | Дробные числа |
| Ankle | Лодыжка | Количественная | Дробные числа |
| Biceps | Бицепс | Количественная | Дробные числа |
| Forearm | Предплечье | Количественная | Дробные числа |
| Wrist | Запястье | Количественная | Дробные числа |

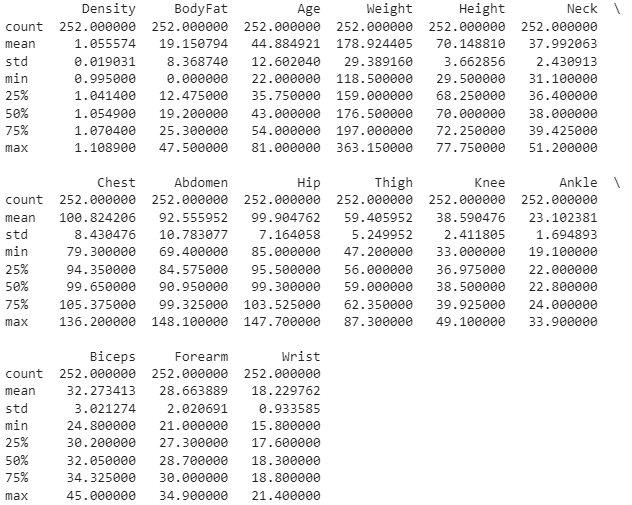
Выведем описательную статистику для числовых столбцов.   


Рисунок 1 – Описательная статистика.

# 1 Построение диаграммы рассеяния и стандартизация данных.

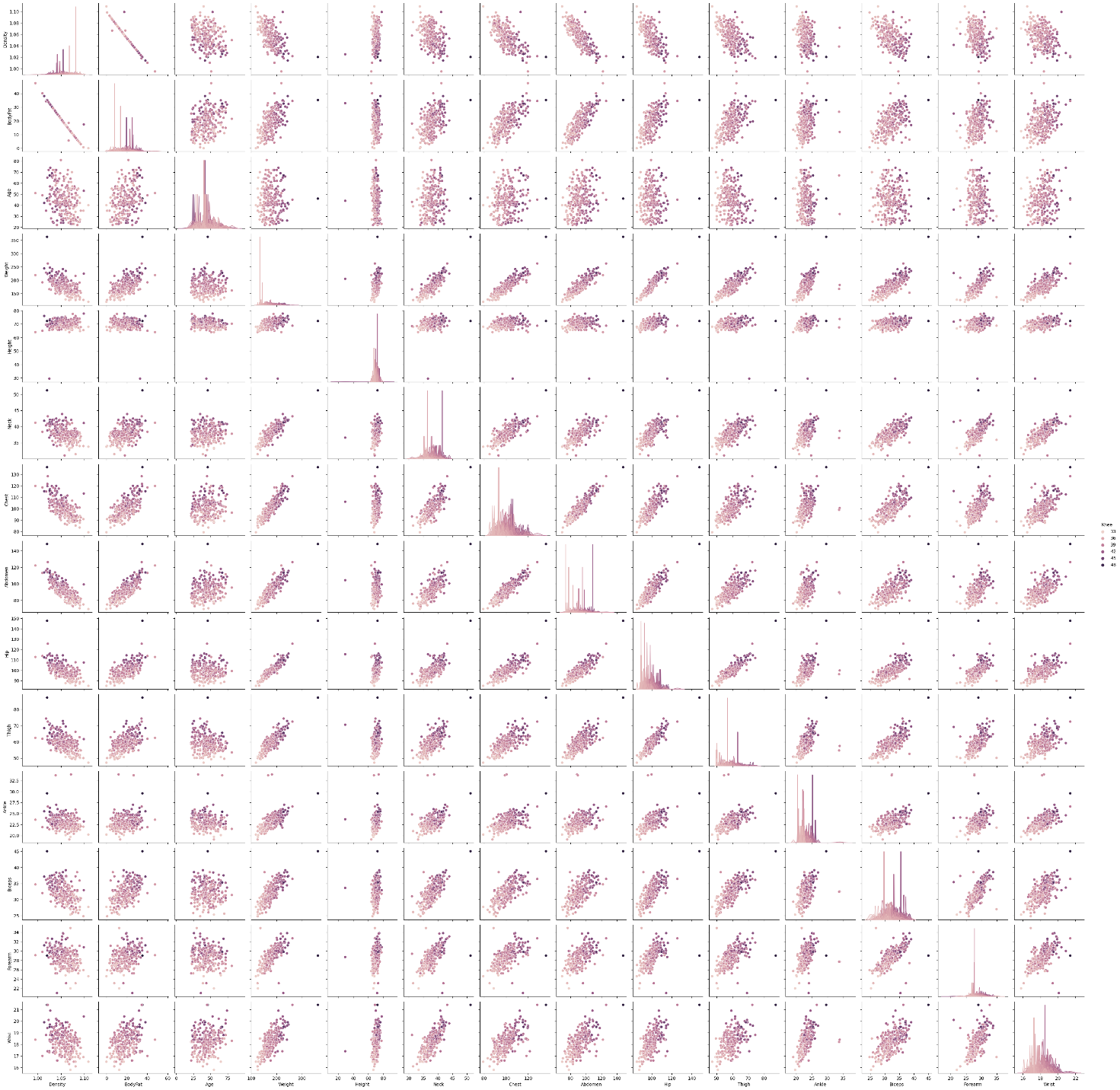


Рисунок 2 – Диаграмма рассеяния.

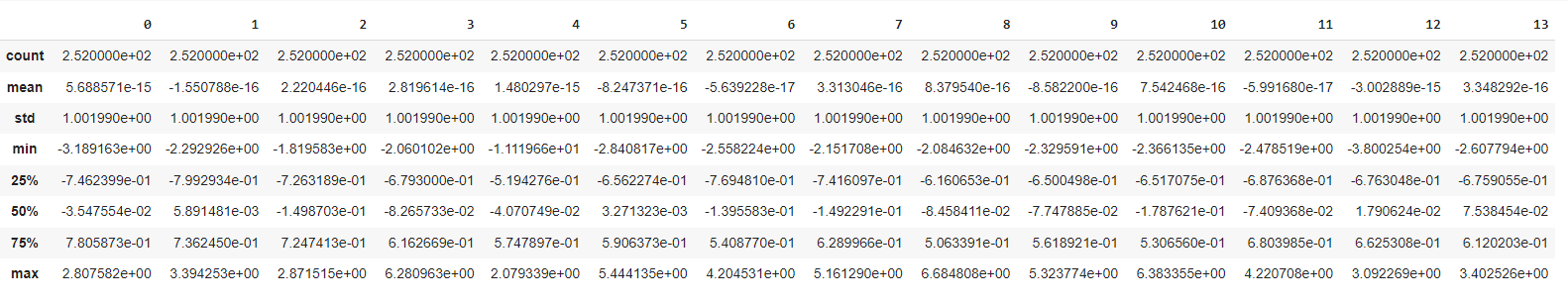


Рисунок 2 – Показатели после стандартизации данных.

# 2 Построение корреляционной матрицы и тепловой карты.

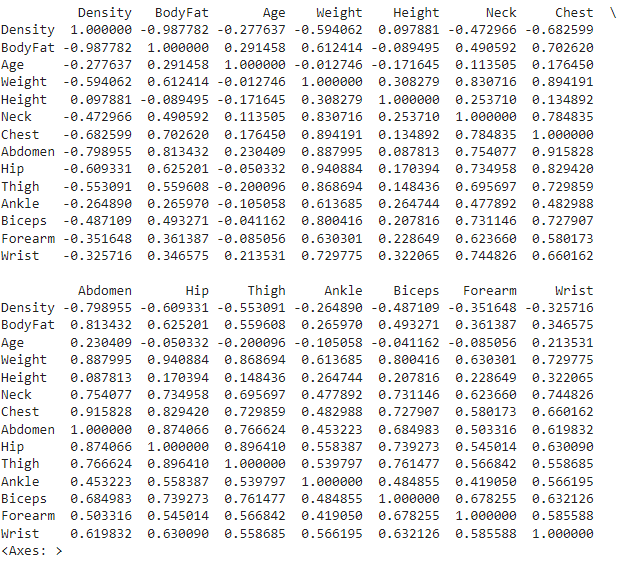


Рисунок 3 – Корреляционная матрица.

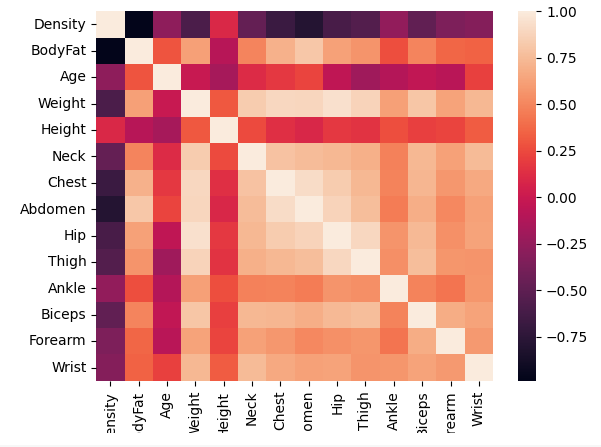


Рисунок 4 – Тепловая карта.

# 3 Определение общей дисперсии.

Определим, какую часть общей дисперсии описывают главные компоненты

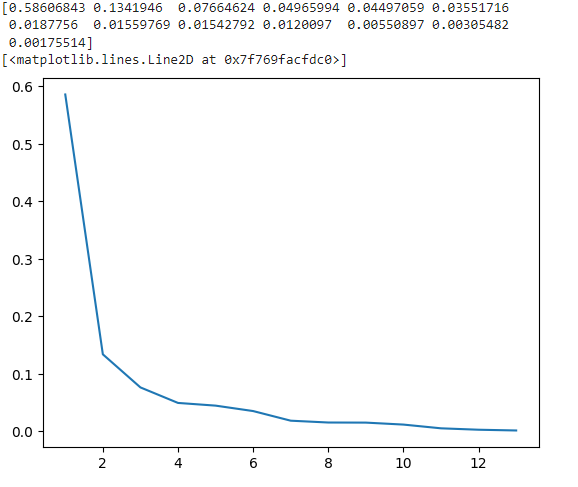


Рисунок 5 – Матрица и график общей дисперсии.

# 4 Построение диаграммы рассеяния для новой системы координат

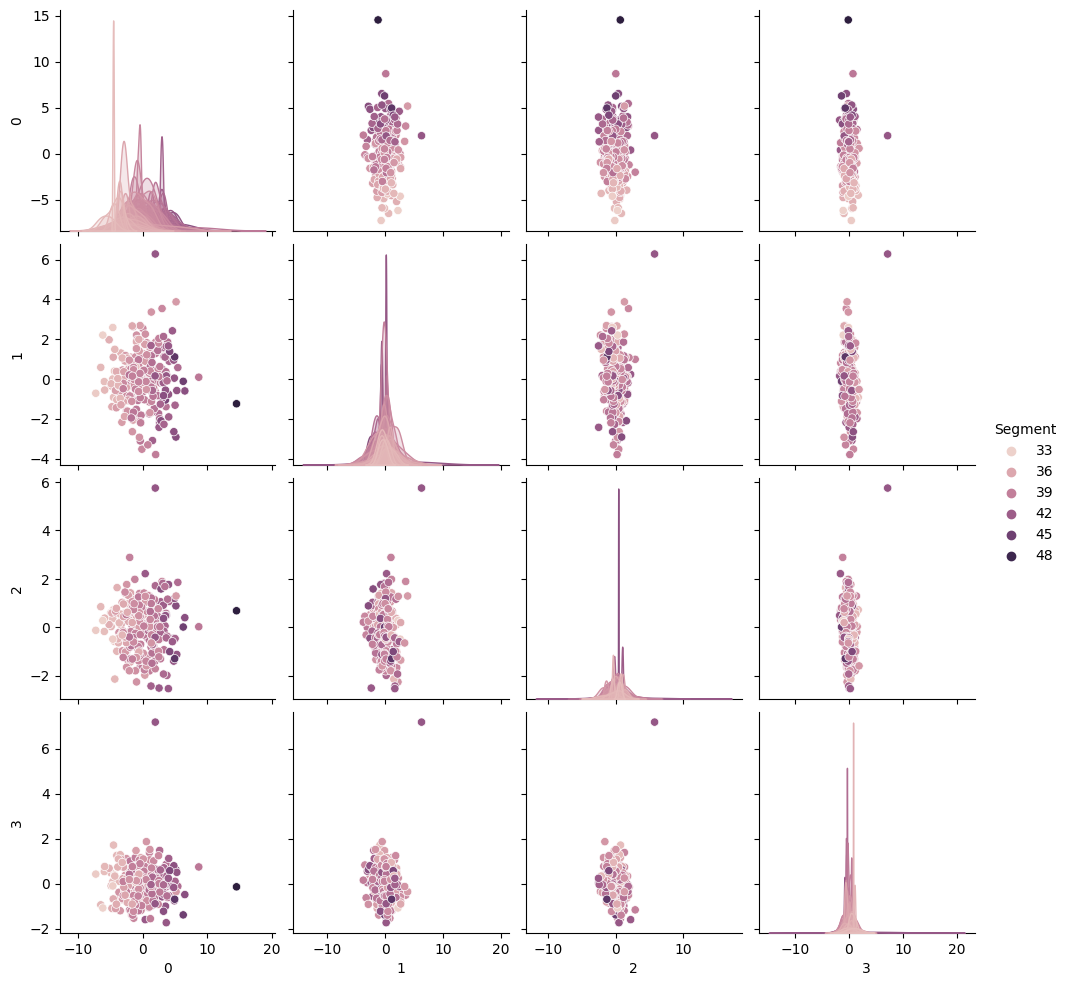
****

Рисунок 6 – Диаграммы рассеяния.

# 5 Факторный анализ

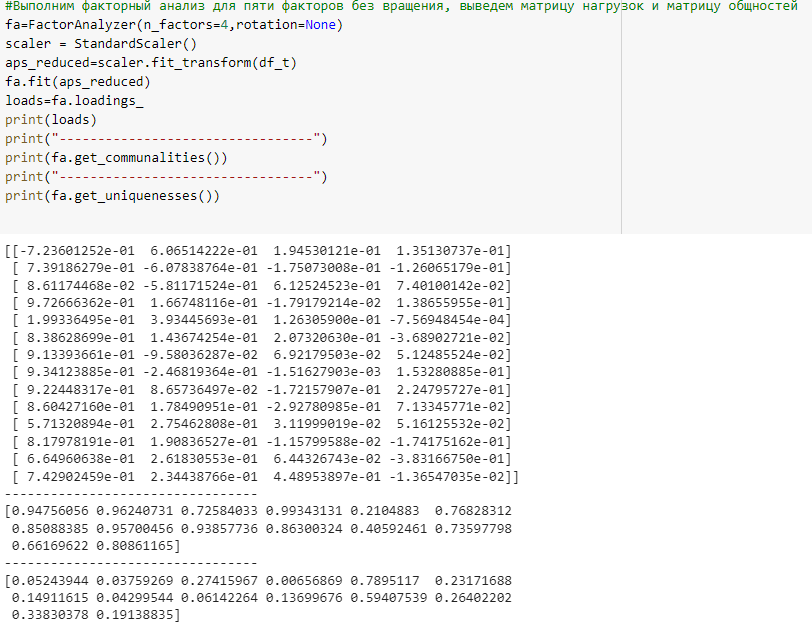


Рисунок 7 – факторный анализ для пяти факторов без вращения.

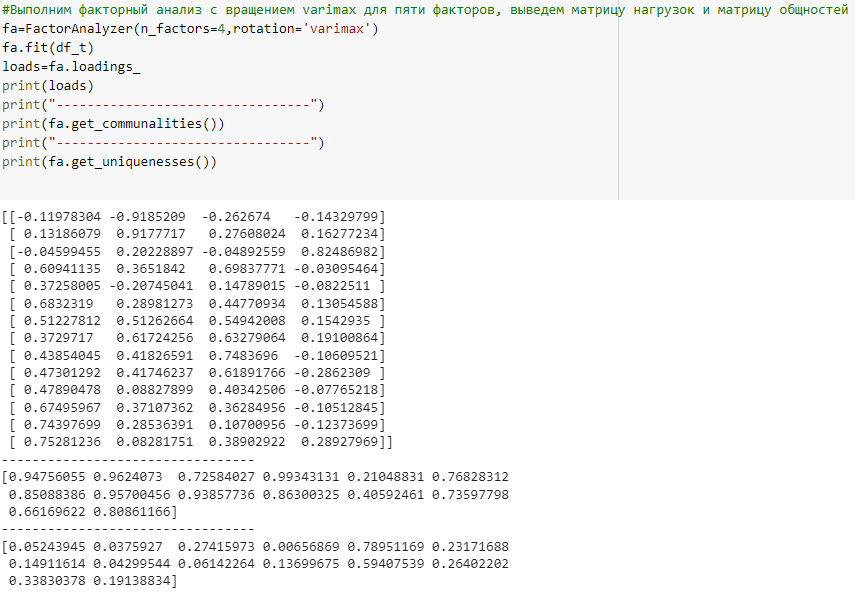


Рисунок 8 – факторный анализ для пяти факторов без вращения.

# 6 Обоснование выбора главных компонент(основной идеи).

Основной идеей метода главных компонент является сокращение размерности исходных данных с минимальной потерей информации. Один из способов определения оптимального числа главных компонент - это построение графика кумулятивной доли объясненной дисперсии от числа главных компонент. Оптимальное число главных компонент - это число, при котором изменение кумулятивной доли объясненной дисперсии становится незначительным.

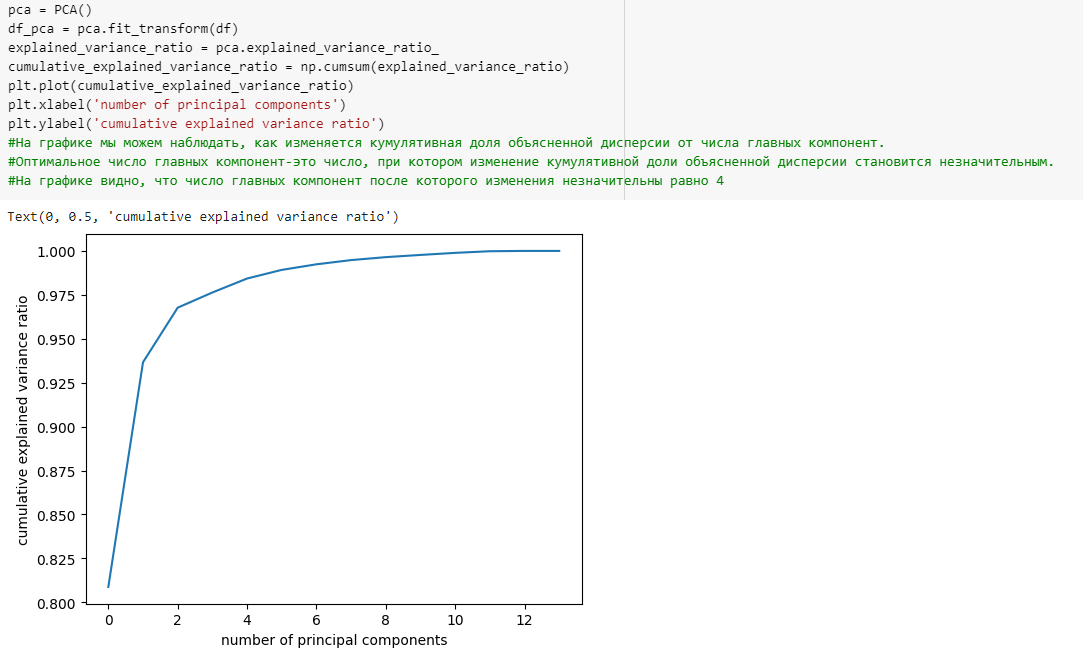


Рисунок 9 – Код и график обоснование выбора числа компонент.

# 7 Решение задачи кластеризации.

Алгоритм решения задачи кластеризации.

1. Для решения задачи кластеризации необходимо выполнить следующие шаги:
2. Подготовить данные: очистить данные от выбросов, пропущенных значений, нормализовать данные.
3. Выбрать подходящий алгоритм кластеризации: иерархический, k-средних, DBSCAN и т.д.
4. Определить количество кластеров: можно использовать методы локтя, силуэта или графика WSS (within-cluster sum of squares).
5. Применить алгоритм кластеризации к данным.
6. Оценить качество кластеризации: с помощью метрик, таких как силуэт или adjusted Rand index.
7. Проанализировать результаты: визуализировать кластеры, проанализировать характеристики объектов в каждом кластере.

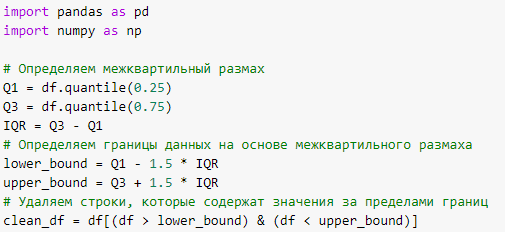


Рисунок 10 – Работа с датасетом.

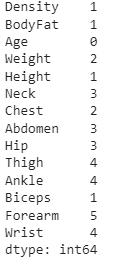


Рисунок 11 – Поиск нулевых значений.

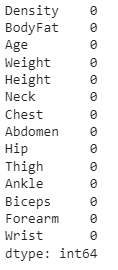


Рисунок 12 – Удаление нулевых значений.

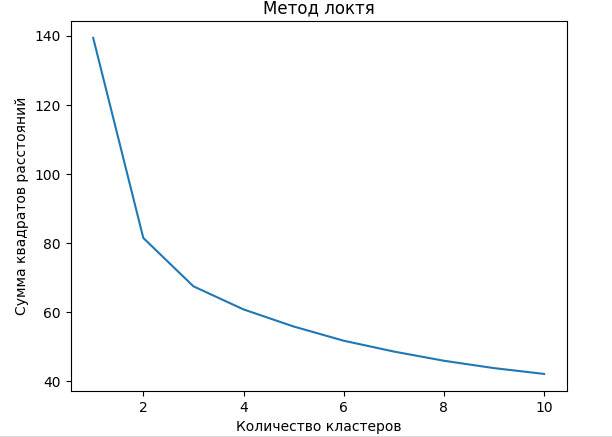


Рисунок 13 – Определяем количество кластеров.

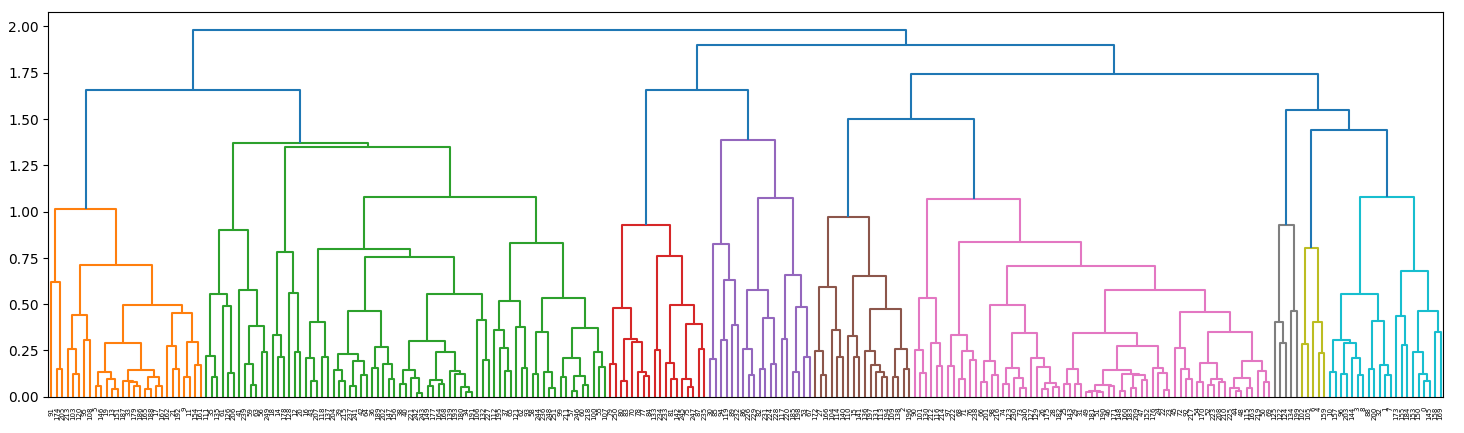


Рисунок 14 – Построение дендрограммы.

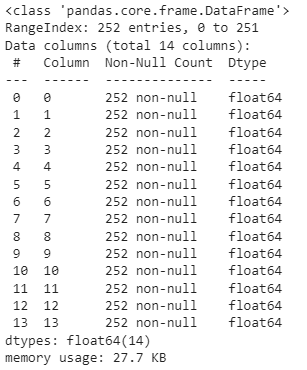


Рисунок 15 – Нормализуем данные.

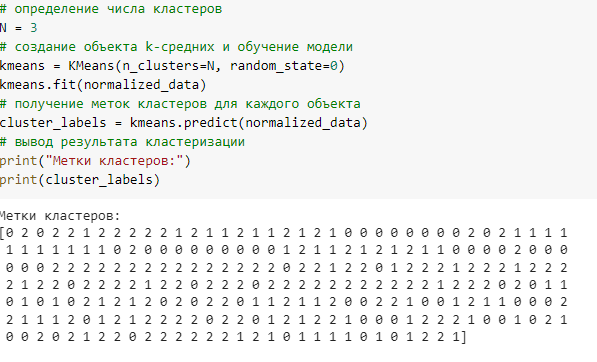


Рисунок 16 – Метки кластеров.

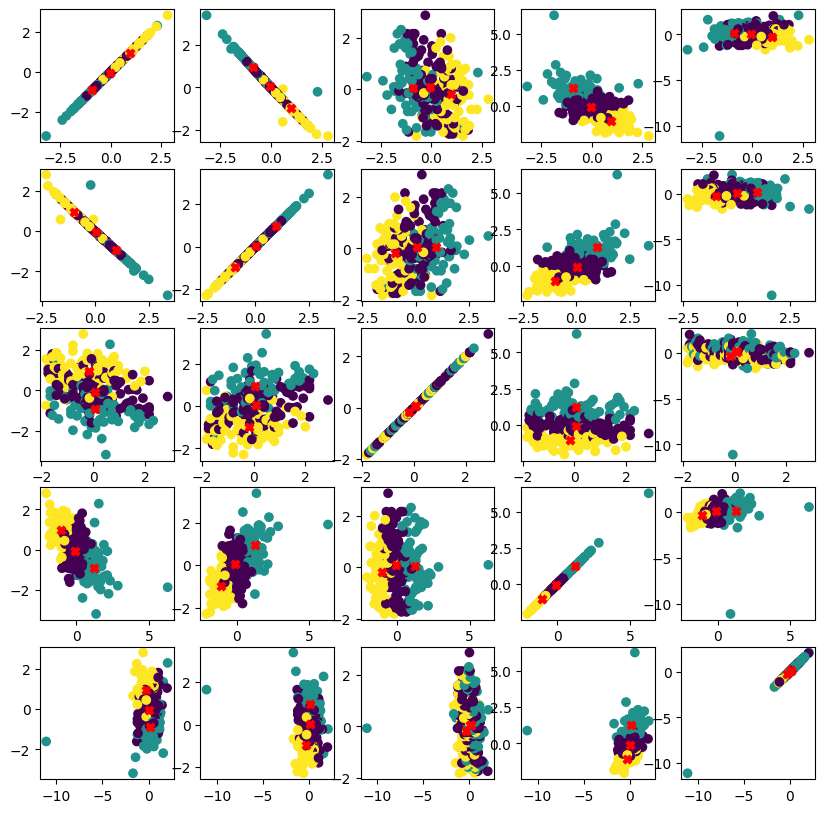


Рисунок 17 – Кластеризация данных методом к-средних.

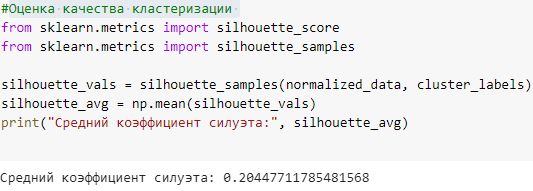


Рисунок 18 – Оценка качества кластеризации.

# 8 Создание нового датасета и решение кластеризации.

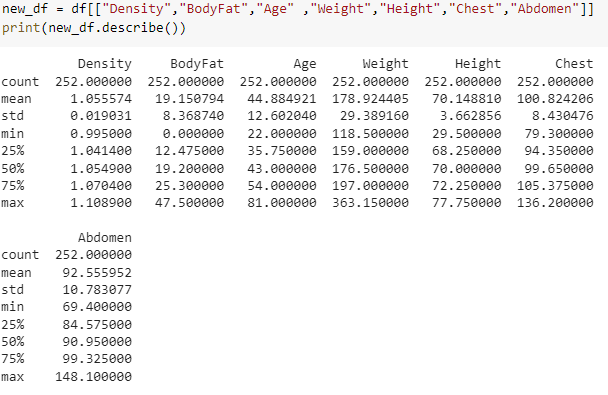


Рисунок 19 – Новый датасет.

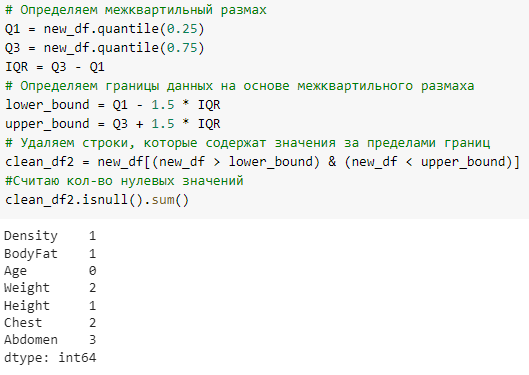


Рисунок 20 – Определение размаха и подсчёт нулевых значений.

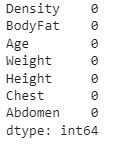


Рисунок 21 – Удаление нулевых значений.

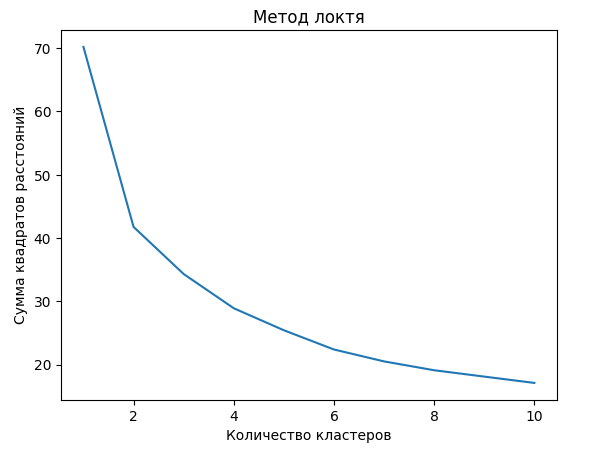


Рисунок 22 – Определяем количество кластеров.

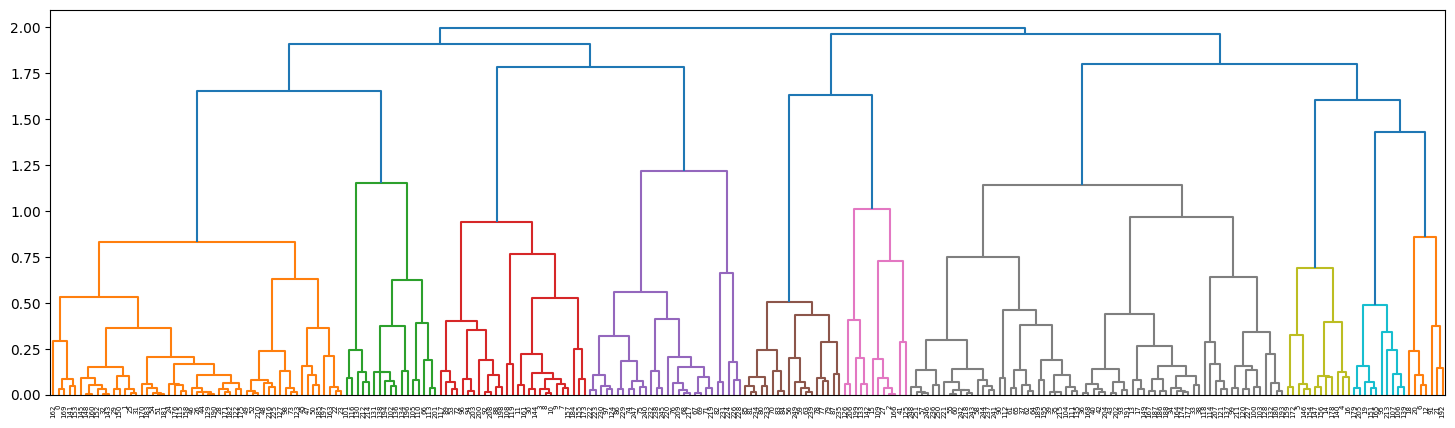


Рисунок 23 – Код для оценки качества кластеризации.

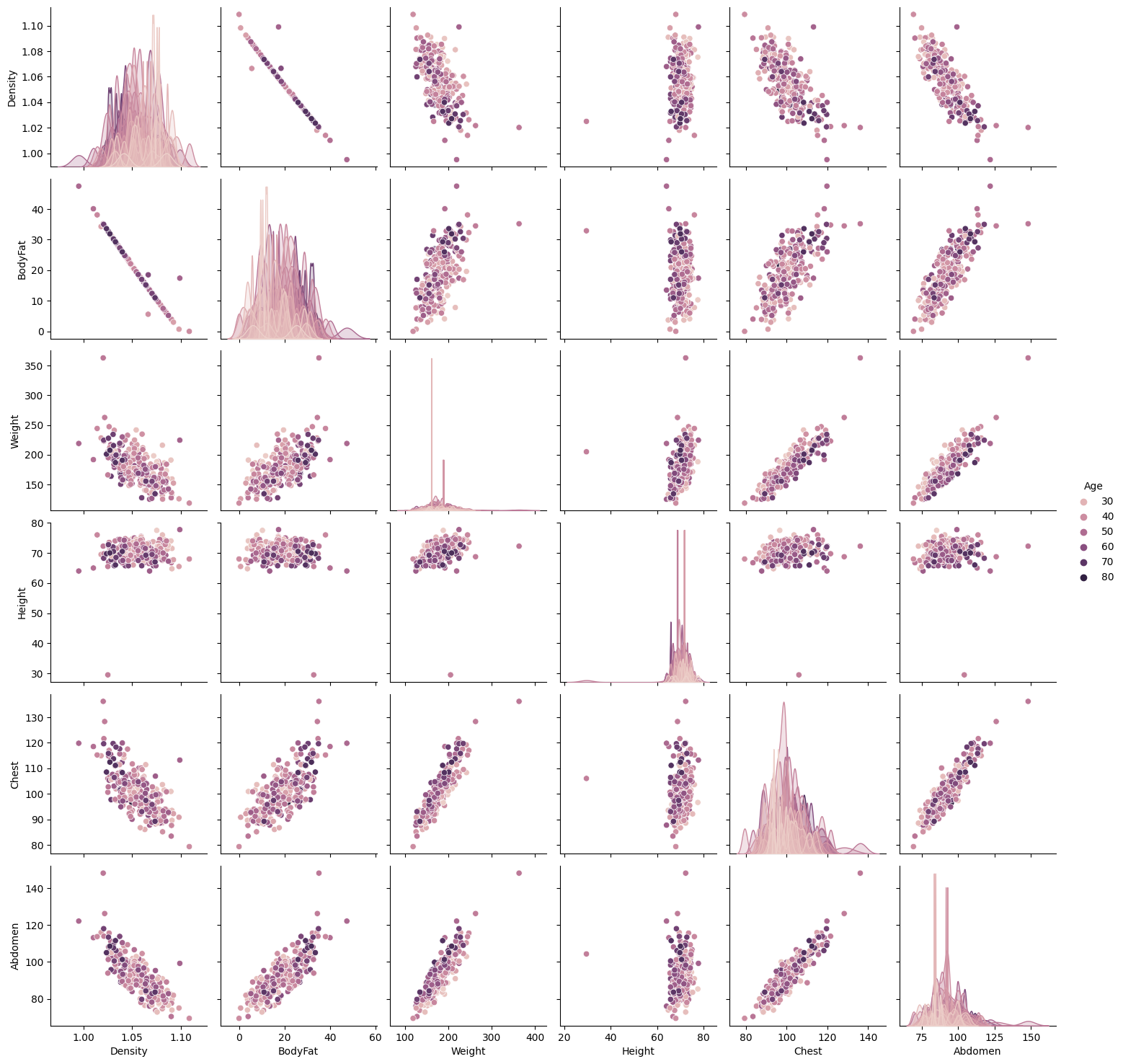


Рисунок 24 – Диаграммы рассеяния.

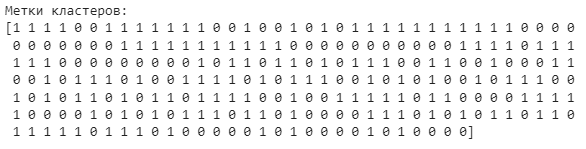


Рисунок 25 – Получаем метки кластеров.

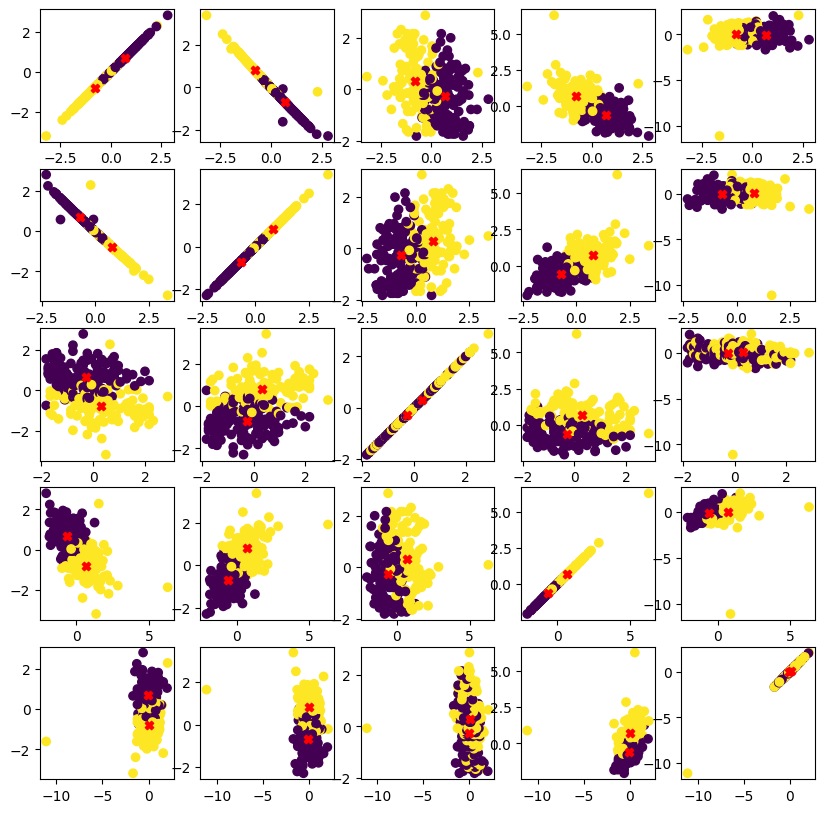


Рисунок 26 – Кластеризацию методом k-средних.



Рисунок 27 – Коэффициент силуэта.

# 9 Проверка качества кластеризации методом главных компонент.

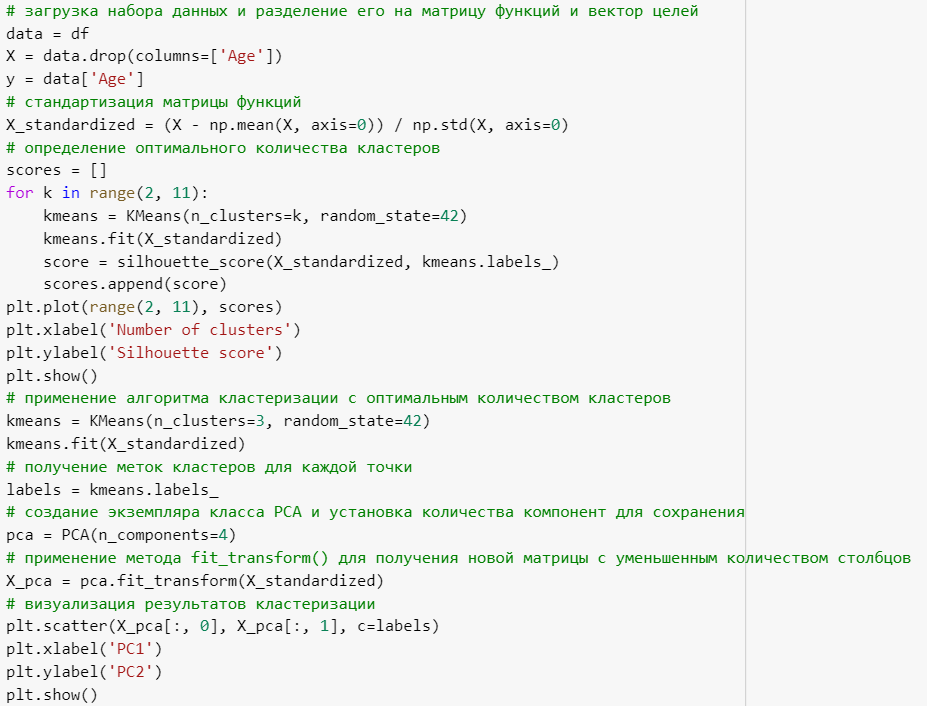


Рисунок 28 – Код для проверки.

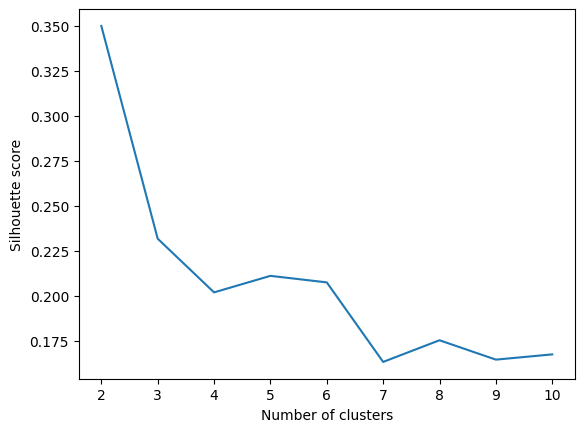
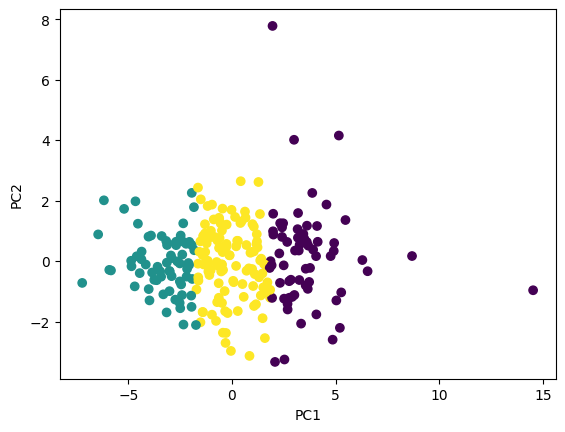
 

Рисунок 29 – Результат вывода.

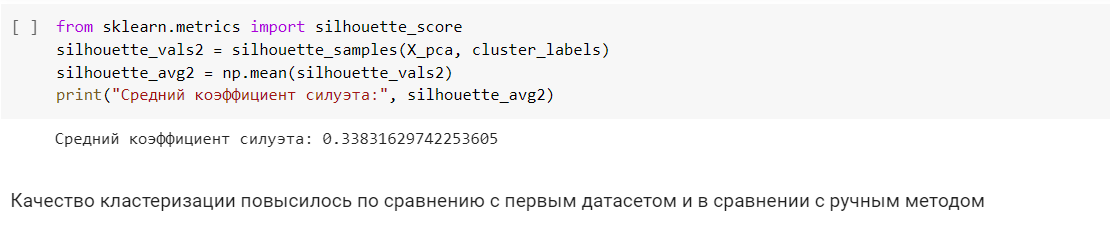


Рисунок 30 – Проверка качества кластеризации.

10 Результаты работы

При изучении задач классификации и кластеризации, были разработаны навыки построения моделей, которые позволяют узнавать закономерности и группировать данные в зависимости от их признаков. Для задачи кластеризации были использованы методы, которые дают возможность сгруппировать данные в несколько кластеров, основываясь на похожих значениях признаков. Для решения задачи классификации были использованы обучающие данные, созданные путем использования готовых данных, где каждый кластер описывает свой отдельный класс. Были разработаны алгоритмы классификации, которые предсказывают классы для новых данных, используя обученную модель.

Одним из главных компонентов исследования была возможность создания нового датасета с помощью генерации синтетических данных, чтобы ускорить и облегчить процесс обучения модели. Также было проанализировано, как классификационная и кластеризационная модели могут быть использованы в различных областях для выявления закономерностей и группировки данных.

В целом, проведенное исследование обладает значительным значением, так как может найти практическое применение в многих областях, включая медицину, маркетинг, банковское дело и другие.

# Список использованных источников

1 Петров А.В., Бучнев О.С. Лабораторный практикум — Иркутск: ФГБОУ ВО ИРНИТУ, 2022. — 114 с.