Санкт-Петербургский Национальный Исследовательский Университет

Информационных Технологий, Механики и Оптики

Факультет Инфокоммуникационных Технологий

**Современные инструменты анализа данных**

**Лабораторная работа 2**

**Уменьшение размерности данных и сравнение методов.**

Выполнила

Улитина М. С.

Проверила

Добренко Н.В.

Санкт-Петербург, 2024

**Цель работы.** Освоить практические навыки работы с методами уменьшения размерности данных, такими как PCA и t-SNE, а также интерпретации их результатов.

**Задачи:**

1. Сбор и предобработка данных: a) Загрузить данные о погоде за период не менее 6 месяцев с ресурса [rp5.ru](https://rp5.ru/) (город на выбор). b) Провести предварительную обработку данных.
2. Реализация и применение PCA: a) Реализовать алгоритм PCA (можно использовать [sklearn.decomposition.PCA](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)). b) Вычислить объясненную дисперсию для каждой компоненты. c) Вычислить процент объясненной дисперсии для первых n компонент. d) Построить проекцию данных на новое пространство;
3. Реализация и применение t-SNE: a) Реализовать алгоритм t-SNE (можно использовать [sklearn.manifold.TSNE](https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html)). b) Применить t-SNE с разными значениями перплексии. c) Визуализировать результаты;
4. Сравнительный анализ: a) Сравнить результаты PCA и t-SNE на исследуемых данных. b) Проанализировать преимущества и недостатки каждого метода в контексте ваших данных.

**Ход работы:**

На рисунке 1 загрузка данных о погоде за 6 месяцев в Санкт-Петербурге с сайта rp5.ru.

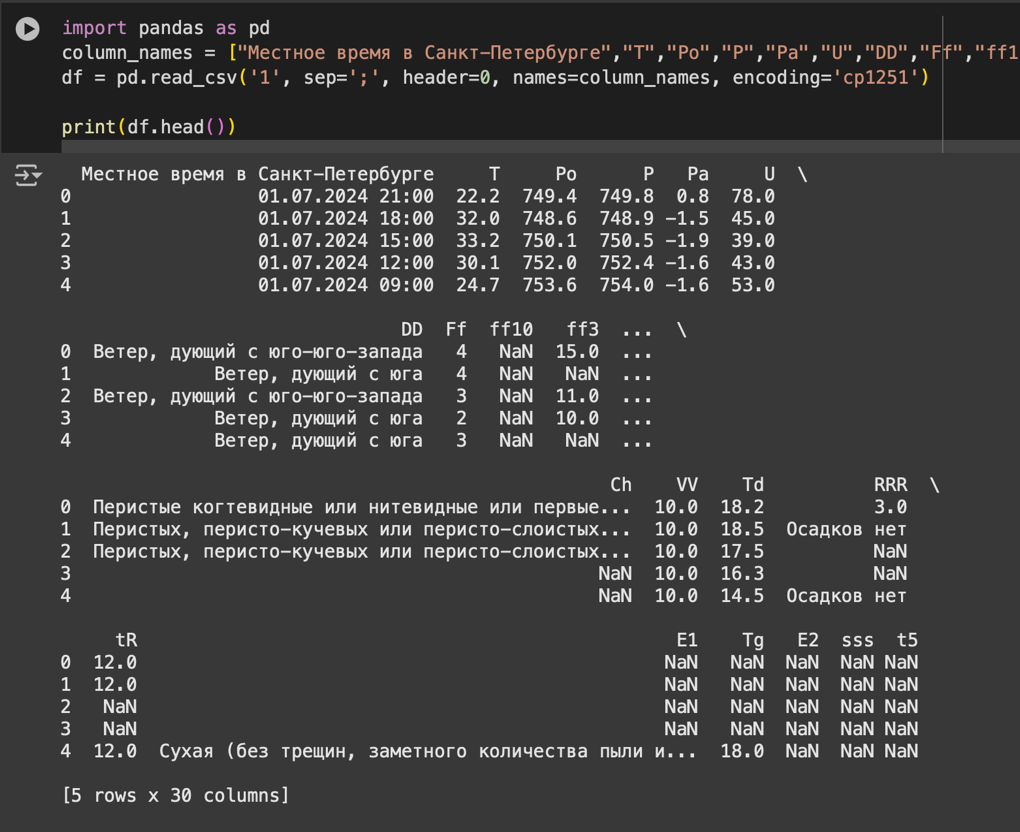


Рисунок 1 – Загрузка данных.

На рисунке 2 преобразованный датафрейм, выбрала данные: температура, давление, вложность и ветер.

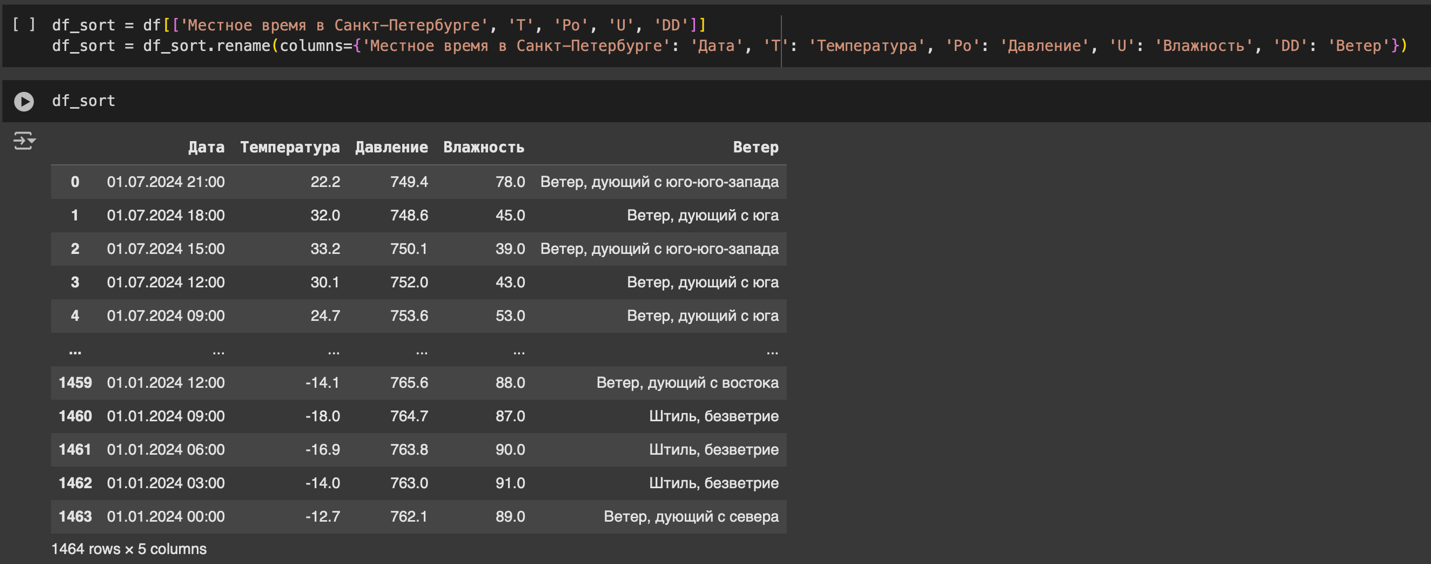


Рисунок 2 – Выбор данных.

Заменены нечисловые данные на числовые, выделен один отчет в сутки, удалены переменные с нулевой дисперсией, оставлены 10-12 параметров погоды (рисунок 3).

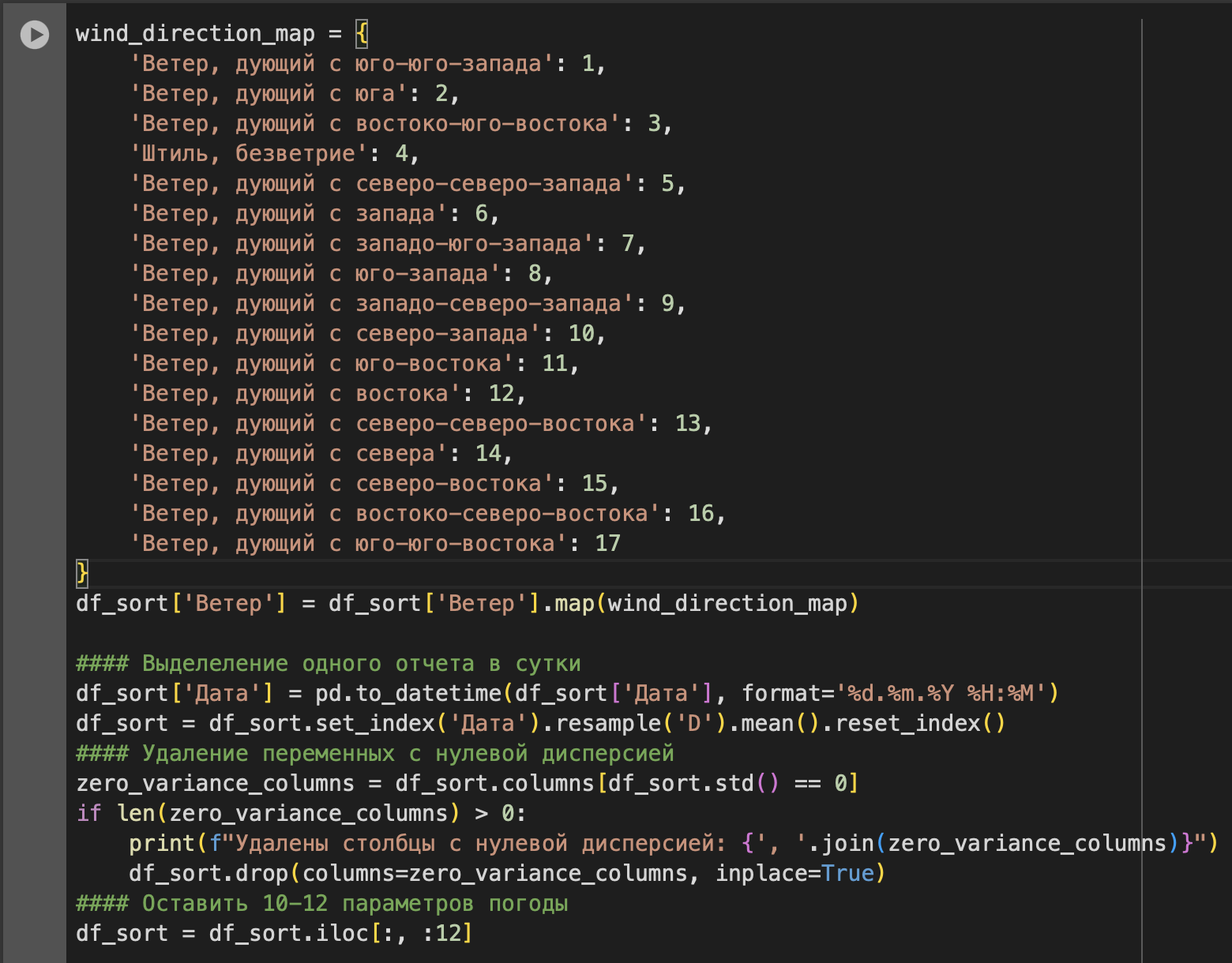


Рисунок 3 – Обработка данных.

На рисунке 4 созданная корреляционная матрица параметров погоды.

Интерпретация корреляций:

Дата и Температура имеют сильную корреляцию (0.90)

Дата и Влажность имеют сильную корреляцию (-0.72)

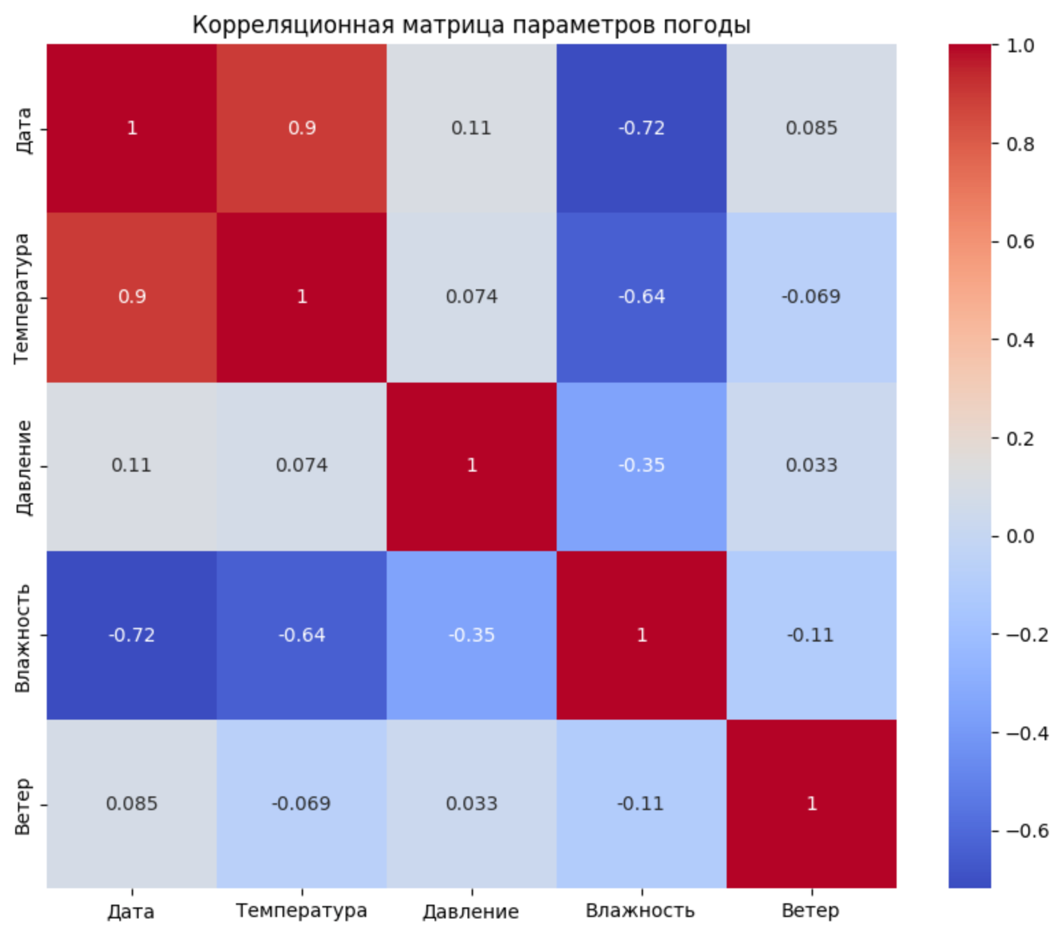


Рисунок 4 – Корреляционная матрица.

На рисунке 5 график каменистой осыпи.

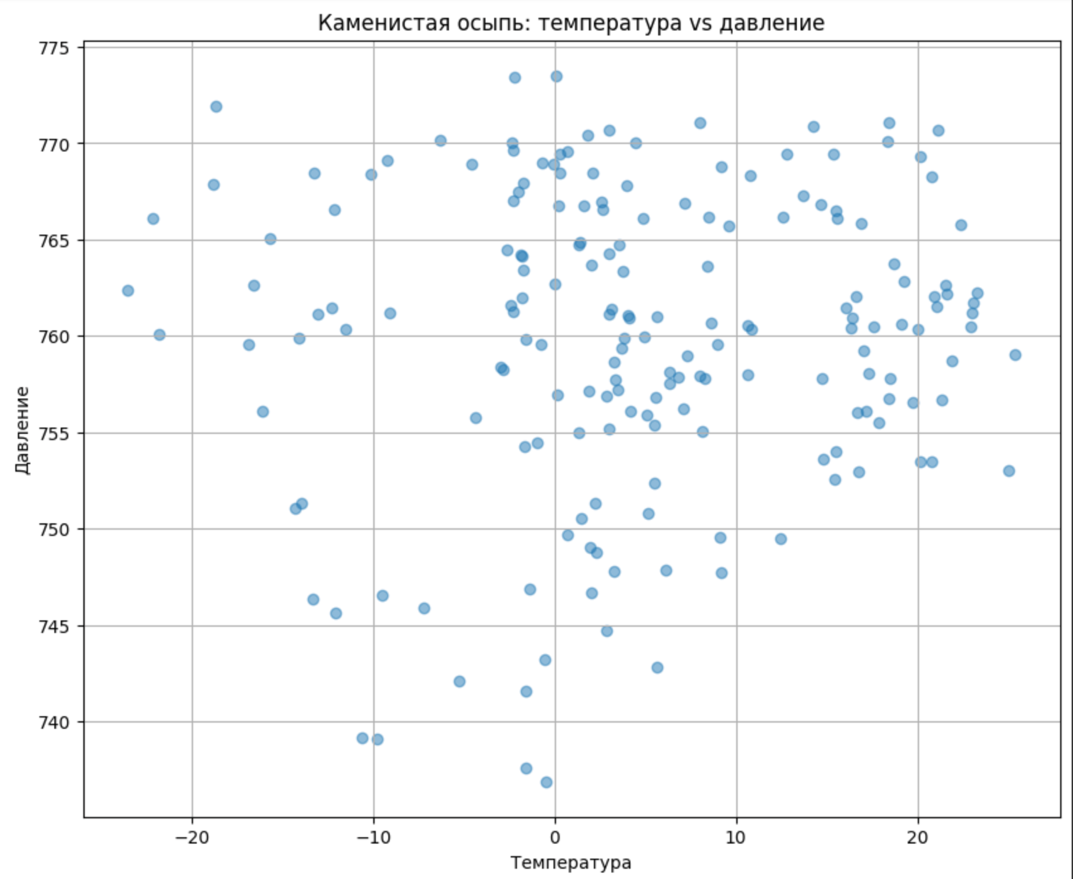


Рисунок 5 - График каменистой осыпи.

Реализован алгоритм PCA, вычислена объясненную дисперсию для каждой компоненты, вычислен процент объясненной дисперсии для первых n компонент (рисунок 6).

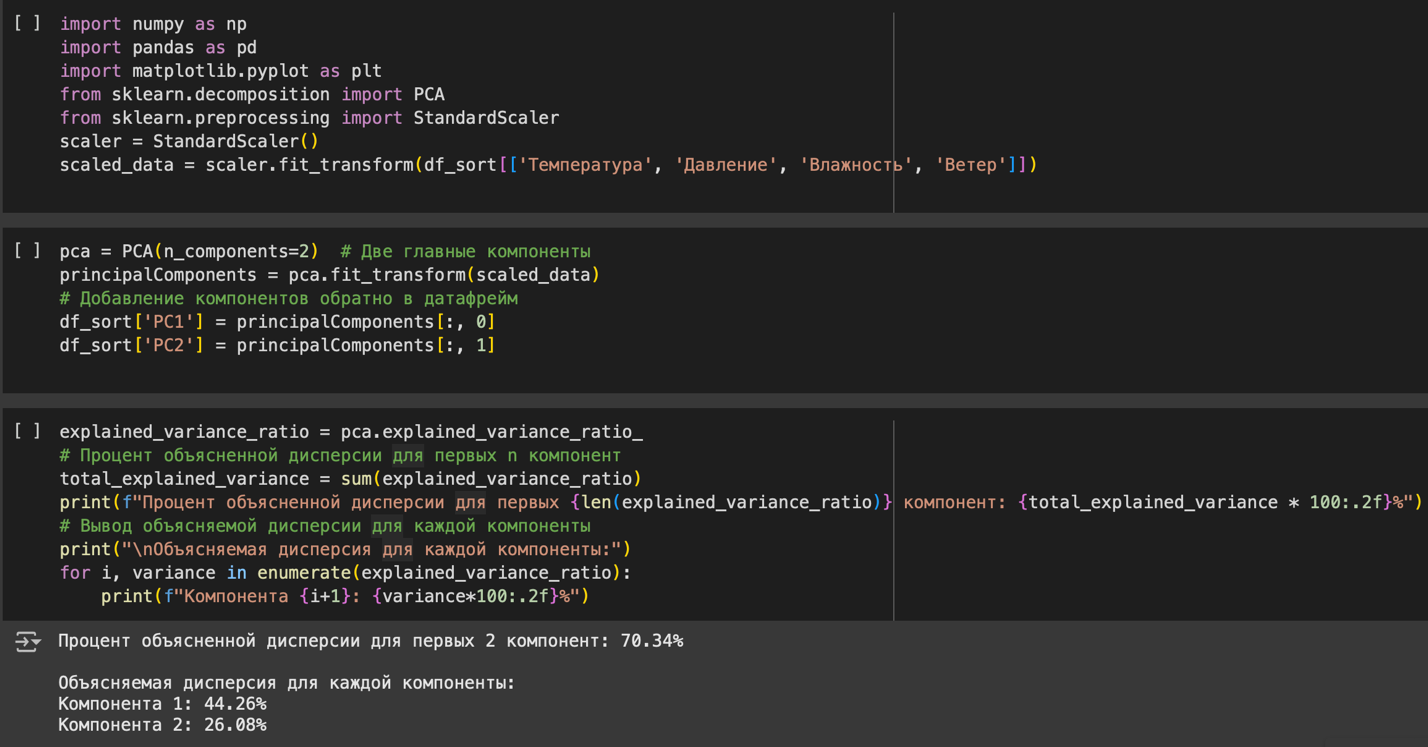


Рисунок 6 – Алгоритм PCA.

Построена проекция данных на новое пространство на рисунке 7.



Рисунок 7 – Проекция данных.

Реализован алгоритм t-SNE, применен t-SNE с разными значениями перплексии, визуализированы результаты. Код на рисунке 8. Визуализированные результаты 9-10.

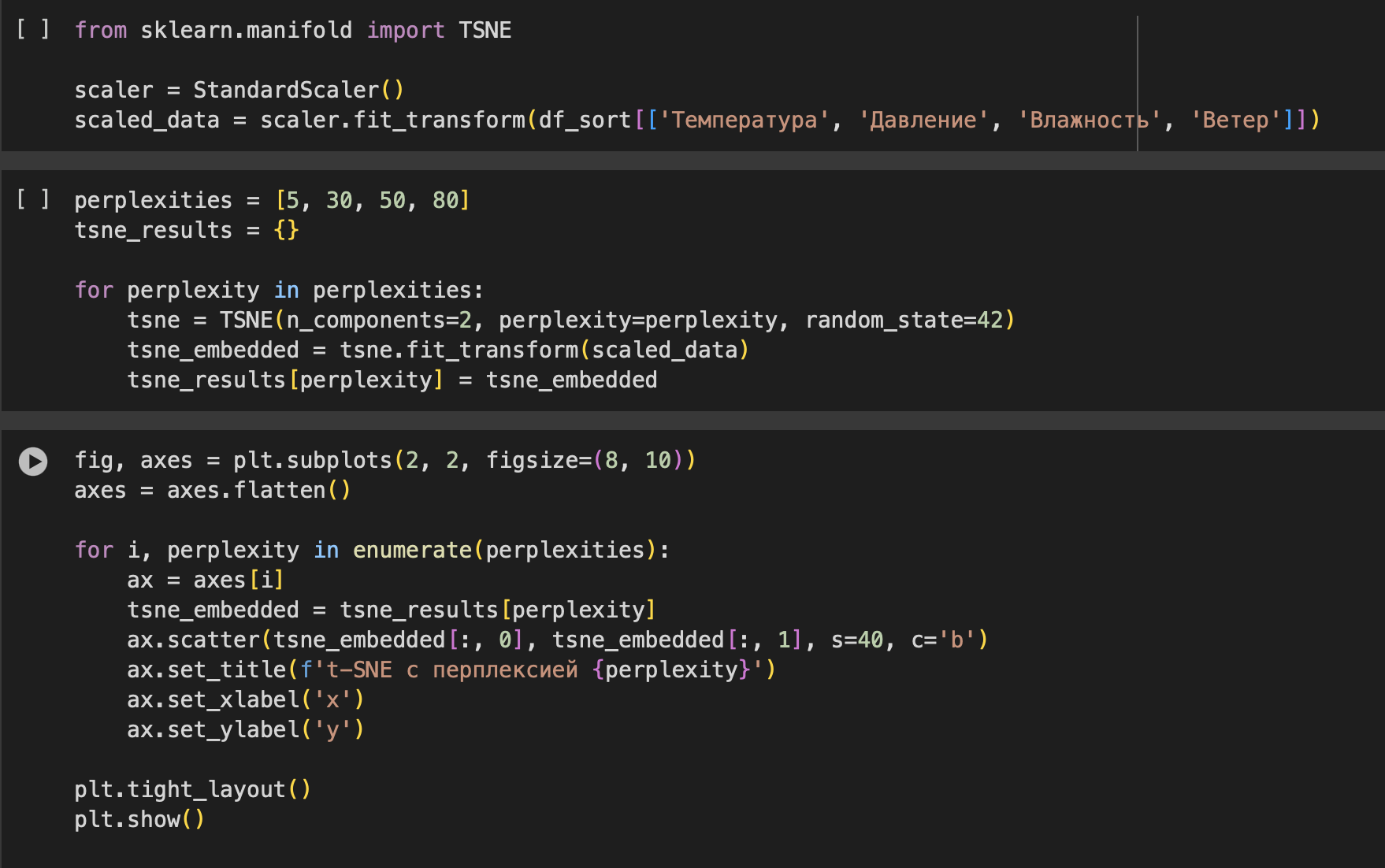


Рисунок 8 – Код.

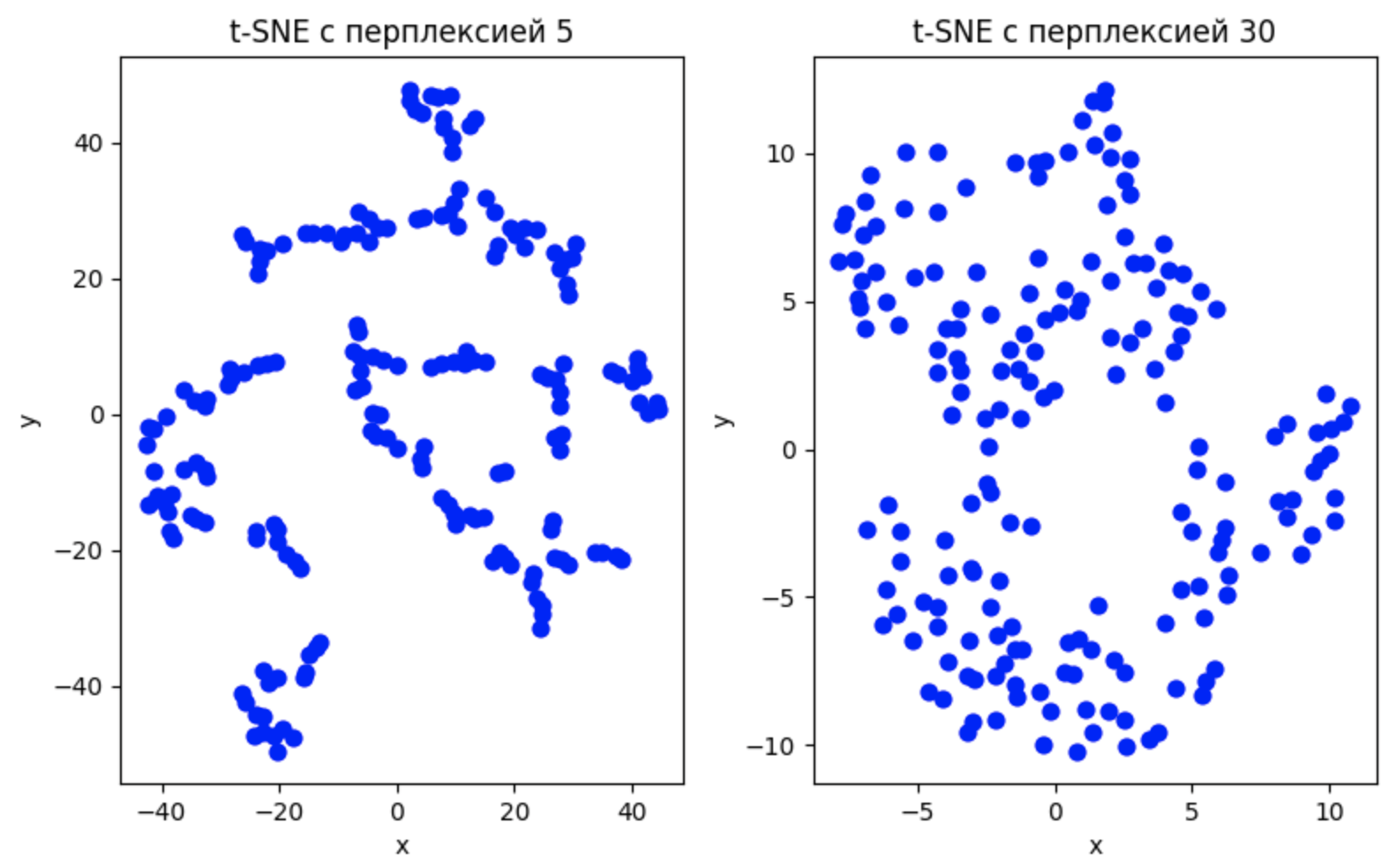


Рисунок 9 – t-SNE 5, 30.

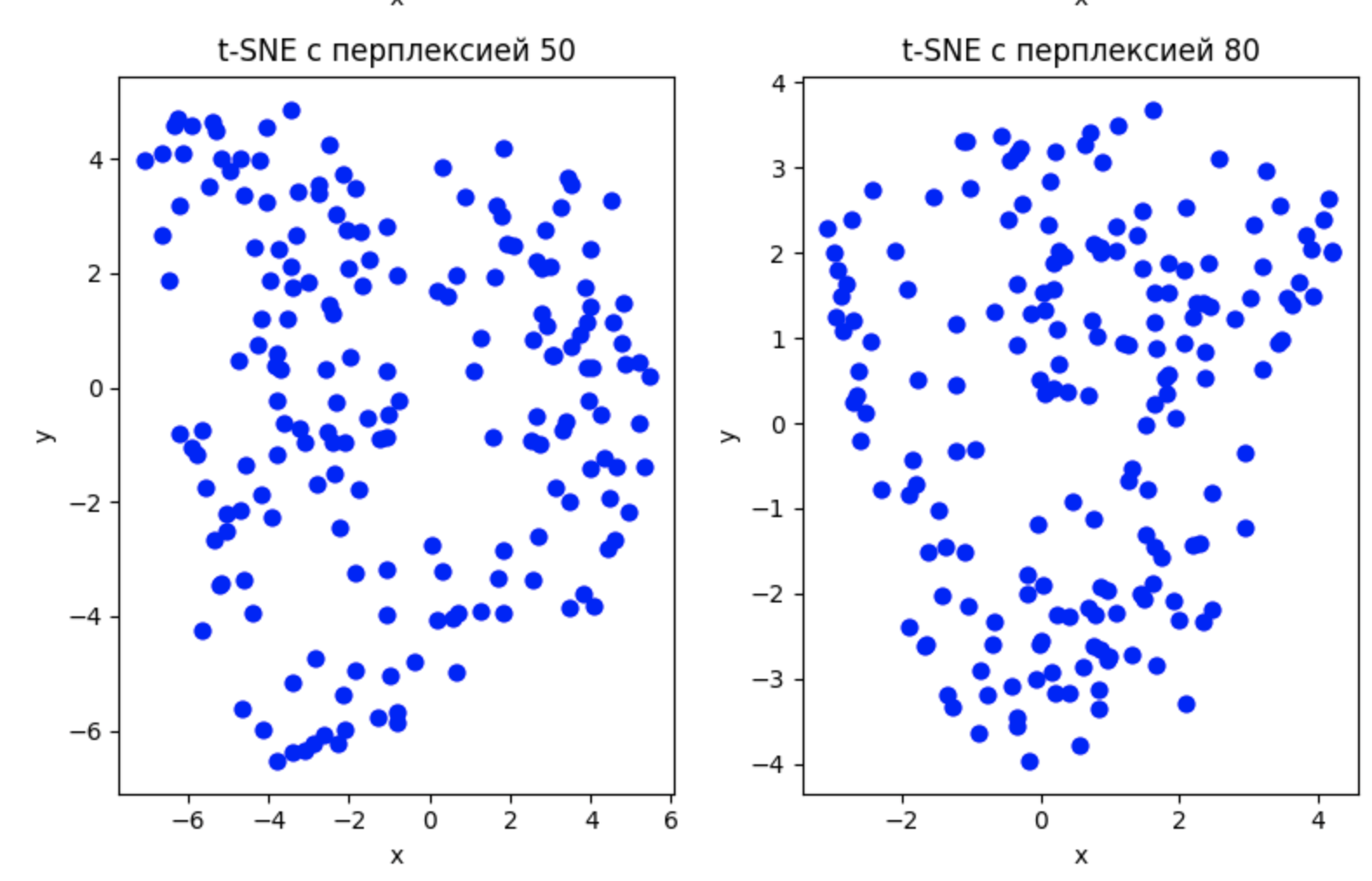


Рисунок 10 – t-SNE 50, 80.

Проанализировано влияние значения перплексии:

- Низкая перплексия: образуются кластеры, состояще из похожих объектов, но общая картина распределения данных менее понятна.

- Высокая перплексия: лучше видна общая структура данных, но меньше заметны некоторые мелкие детали и различия внутри отдельных групп.

- Средняя перплексия: оптимальный вариант, отображающий и общую картину и отдельные детали (в случае построенных графиков оптимальным является значение 30).

Сравнение результатов PCA и t-SNE на исследуемых данных. Код и визуализация на рисунказ 11-13.

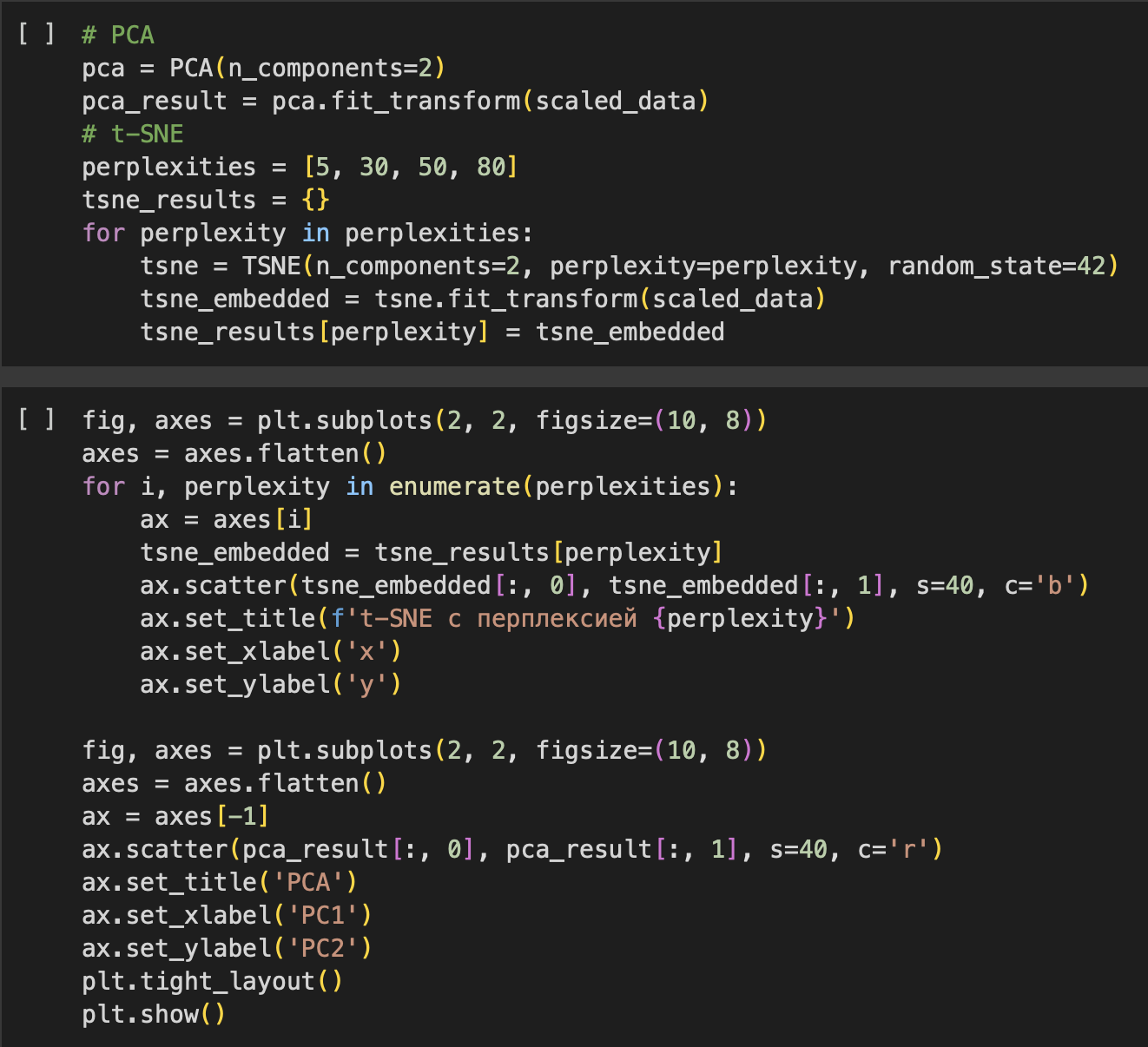


Рисунок 11 – Код.

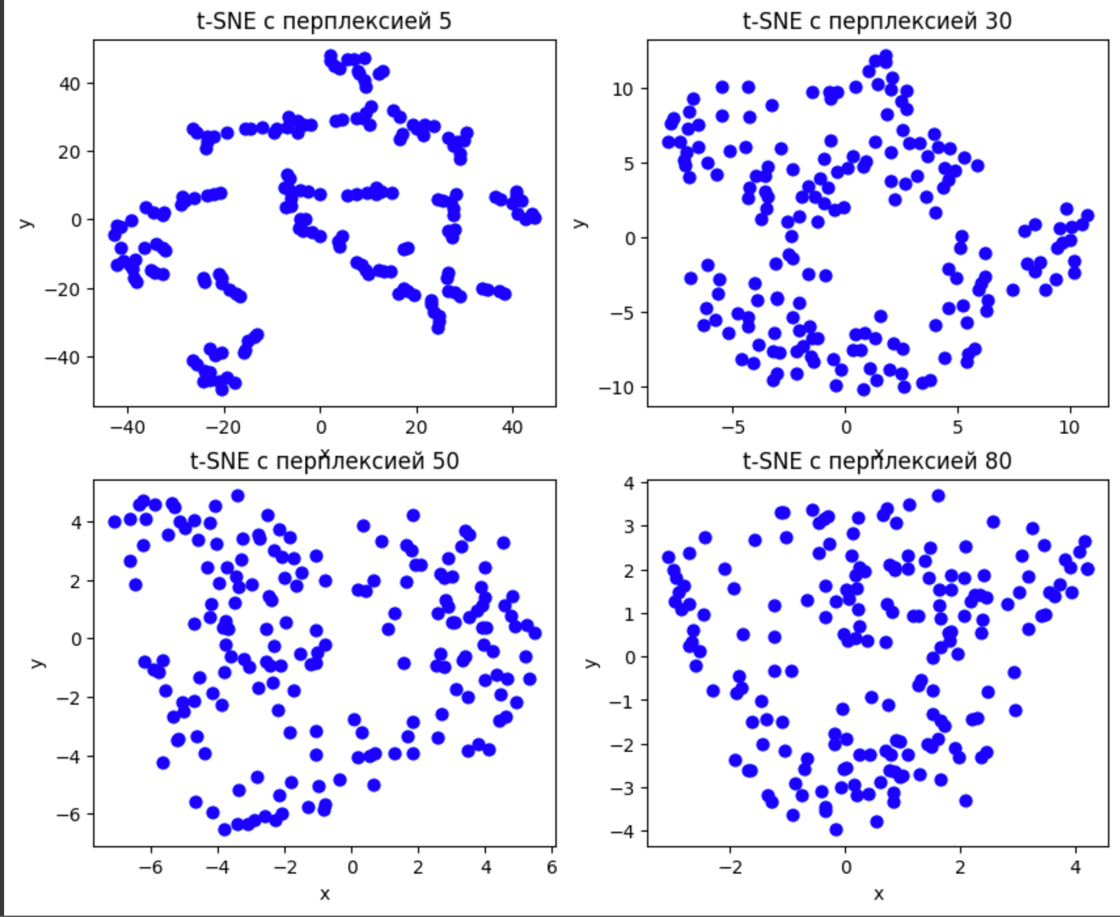


Рисунок 12 – t-SNE.

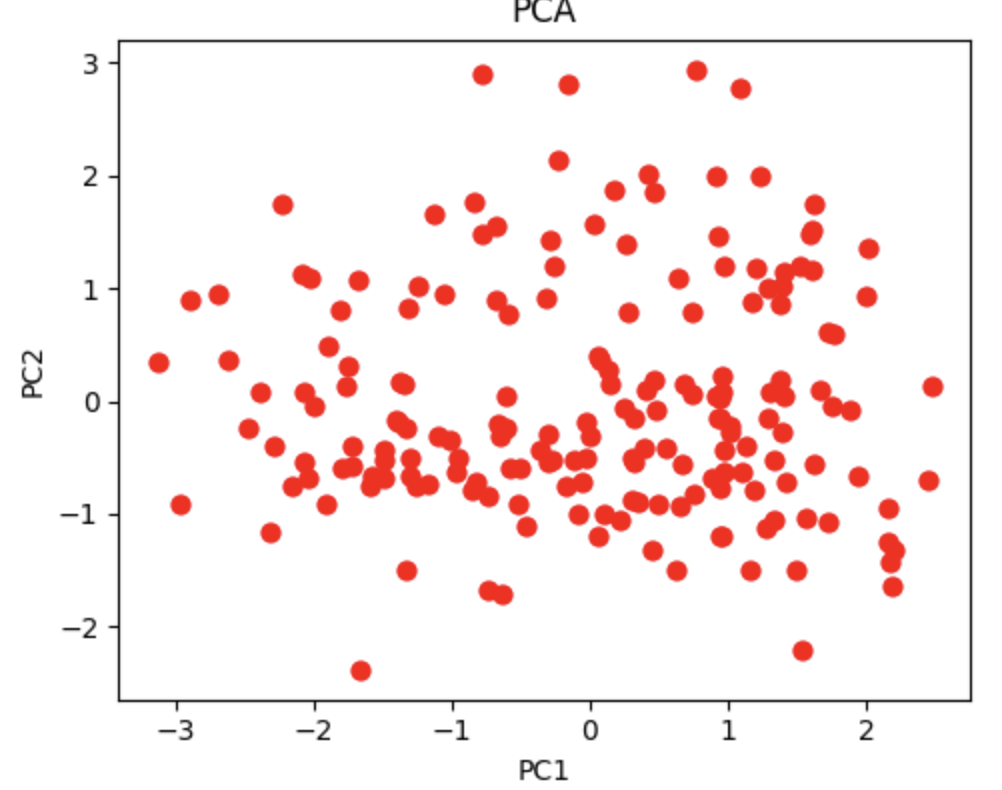


Рисунок 13 – PCA.

Сравнение:

1. Преимущества и недостатки:

- PCA: преимуществами является простота реализации, скорость работы и мало необходимых настроек, но есть и недостатки, такие как, линейсность и проблемы в визуализации объектов большой размерности

-t-SNE: преимуществами являются удобная визуализация и возможность работать с нелинейностью, но сложнее настраивается и дольше работает.

2. Когда использовать:

- PCA: анализ главных компонент, быстрая оценка.

-t-SNE: исследование нелинейных и сложных данных.

**Вывод:**

Выполнены все задачи лабораторной работы, освоены практические навыки работы с методами уменьшения размерности данных, такими как PCA и t-SNE, а также интерпретации их результатов.