|  |  |
| --- | --- |
| *voenmeh* | МИНОБРНАУКИ РОССИИ  федеральное государственное бюджетное образовательное учреждение  высшего образования  **«Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова»**  **(БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова»)** |
| БГТУ.СМК-Ф-4.2-К5-02 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Факультет |  | О |  | “Естественнонаучный” |
|  |  | шифр |  | наименование |
| Кафедра |  | О7 |  | “Информационные системы и программная инженерия” |
|  |  | шифр |  | наименование |
| Дисциплина |  | “Системы искусственного интеллекта” | | |

|  |
| --- |
| Практическая работа №3 |
| Обучение без учителя |
| Вариант 4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил студент группы | | | |  |  |
|  | | | | | |
| Фамилия И.О. | | | | | |
| **РУКОВОДИТЕЛЬ** | | | | | |
|  | |  |  | | |
| Фамилия И.О. Подпись | | | | | |
| Оценка |  | | | |  |
| «\_\_\_\_\_» |  | | | | 202 г. |

САНКТ-ПЕТЕРБУРГ

202 г.

**СОДЕРЖАНИЕ**

[Общая постановка задачи 3](#_Toc218991513)

[Часть 1 5](#_Toc218991514)

[Часть 2 7](#_Toc218991515)

[Часть 3 10](#_Toc218991516)

[Часть 4 13](#_Toc218991517)

[Вывод 19](#_Toc218991518)

# **Общая постановка задачи**

Для оформления работы использовать приложенный файл PR3.ipynb.

**Часть 1**

1. Загрузить приложенный датасет;
2. Подготовить его для исследования – убрать объекты с неполными данными;
3. Преобразовать текстовые значения всех полей в числовые;
4. Вывести информацию о датасете.

**Часть 2**

1. Используя любой метод поиска аномалий найти выбросы в датасете;
2. Отобрать результат на графиках как минимум для трех признаков. Желательно выбрать те признаки, на которых наиболее заметны выбросы;
3. Сделать вывод по успешности поиска аномалий выбранным методом;
4. Сохранить датасет без аномалий в отдельную переменную и создать его стандартизированную версию.

**Часть 3**

1. Используя любой метод понижения размерности, снизить размерность полученного датасета до минимально возможной (желательно до двух признаков);
2. Вывести полученный «сжатый» датасет;
3. Для получения дополнительного балла добавить ещё один метод понижения размерности;
4. Сформулировать вывод по эффективности понижения размерности.

**Часть 4**

1. Используя метод кластеризации, указанный в варианте, разбить датасет, полученный после выполнения второй части работы;
2. Используя полученные метки, вывести графики тех признаков, которые наиболее явно показывают принцип разбиения. Признаки должны быть взяты из нестандартизированной версии датасета из второй части задания;
3. Используя тот же метод кластеризации, разбить на кластеры «сжатый» датасет, полученный после выполнения третьей части работы;
4. Используя полученные метки, вывести как график «сжатого» датасета, так и графики тех же признаков оригинального датасета, что и в пункте два;
5. Сравнить результаты;
6. Для получения дополнительного балла добавить ещё один метод кластеризации;
7. Сформулировать вывод по эффективности кластеризации.

В отчёте должны быть приведены задание, вариант, скриншоты из файла с полученными результатами, сформулированные выводы. К отчёту должен быть приложен файл ipynb с результатами.

В варианте указаны датасет и метод кластеризации.

Вариант 4.

– датасет 4 Country-data.csv;

– на этапе подготовки необходимо убрать поле country;

– метод – Агломератиный.

# **Часть 1**

На рисунке 1 представлено подключение библиотек для работы с датасетом 4 Country-data.csv.

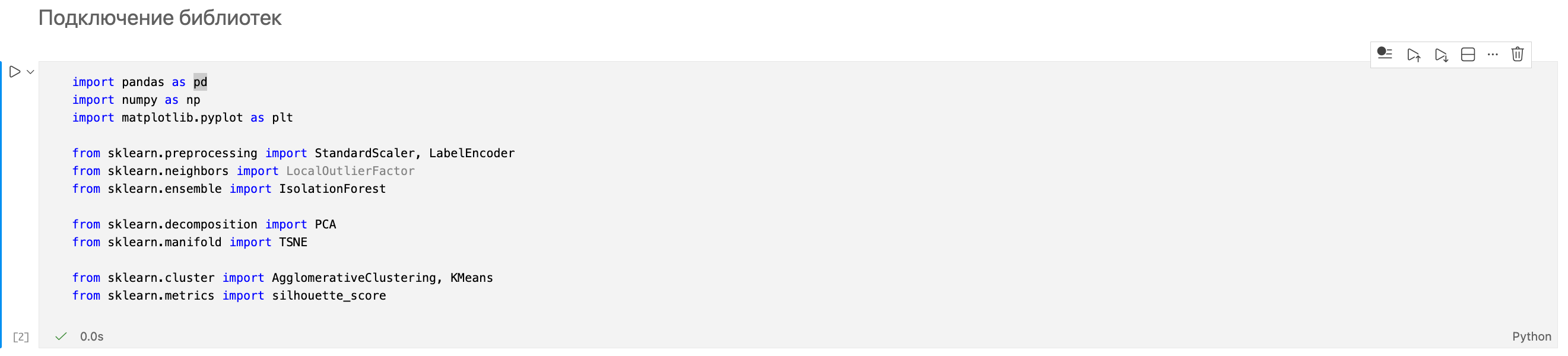


Рисунок 1 – Подключение библиотек

Загрузка датасета и вывод его содержимого представлены на рисунке 2.

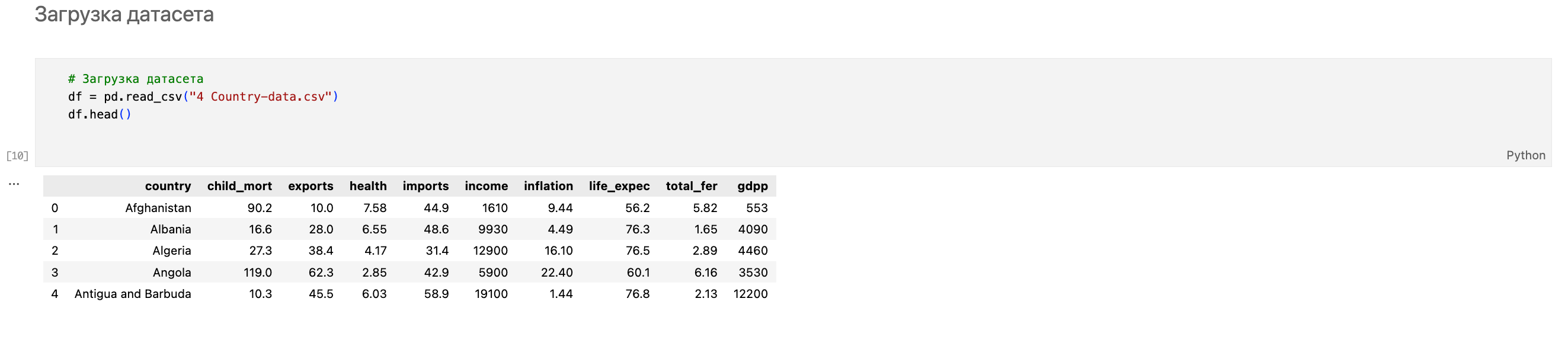


Рисунок 2 – Загрузка датасета

Вывод информации о датасете, представлен на рисунках 3-4.

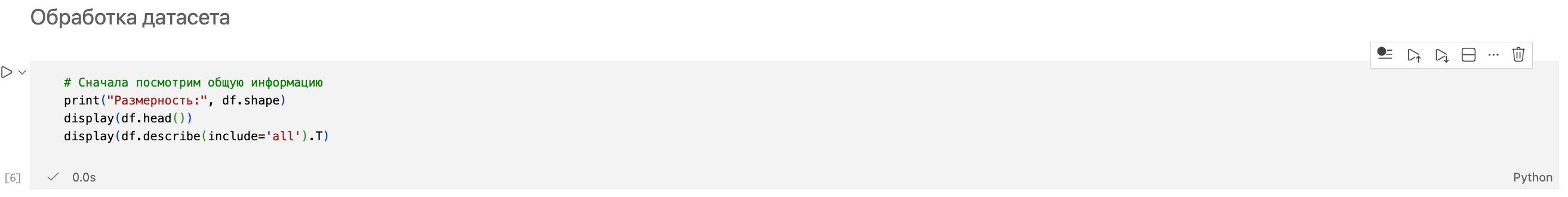


Рисунок 3 – Обработка датасета



Рисунок 4 – Вывод информации о датасете

На данном этапе была выполнена предобработка данных, включающая удаление неинформативного признака country, обработку пропущенных значений и кодирование категориальных признаков. В результате был сформирован очищенный числовой датасет, пригодный для дальнейшего анализа и применения алгоритмов машинного обучения, представлено на рисунке 5.

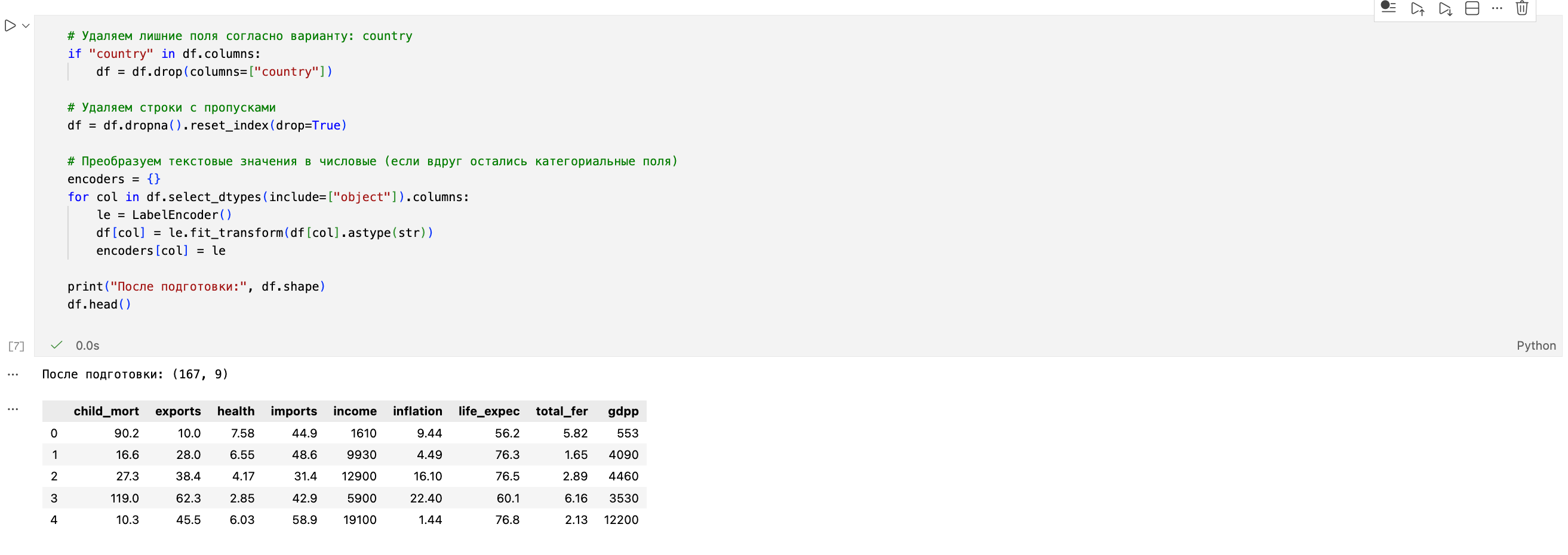


Рисунок 5 – Подготовка данных к анализу

Разведочный анализ данных, включающий изучение структуры датасета и визуализацию распределений отдельных числовых признаков. Анализ позволил оценить диапазоны значений и форму распределений показателей, а также подтвердить необходимость масштабирования данных перед применением методов кластерного анализа, представлено на рисунке 6.

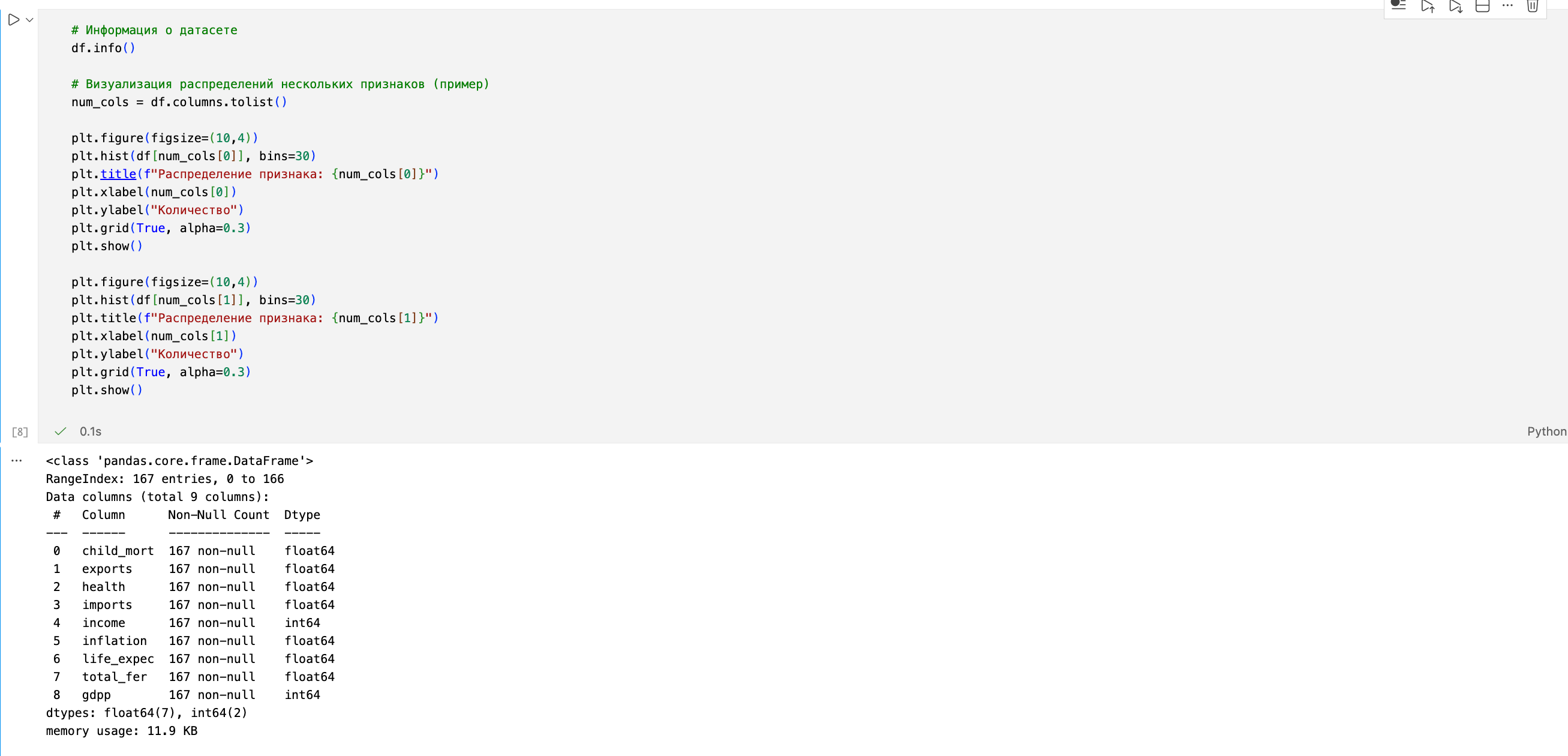


Рисунок 6 – Анализ структуры и распределений признаков

Для анализа распределений числовых признаков были построены гистограммы показателей child\_mort и exports. Анализ показал, что оба признака имеют выраженную правостороннюю асимметрию и содержат выбросы. Большинство стран характеризуются низкими значениями детской смертности и умеренными значениями экспорта, однако присутствуют отдельные страны с экстремальными показателями. Полученные результаты подтверждают необходимость масштабирования данных перед выполнением кластерного анализа, представлено на рисунке 7.

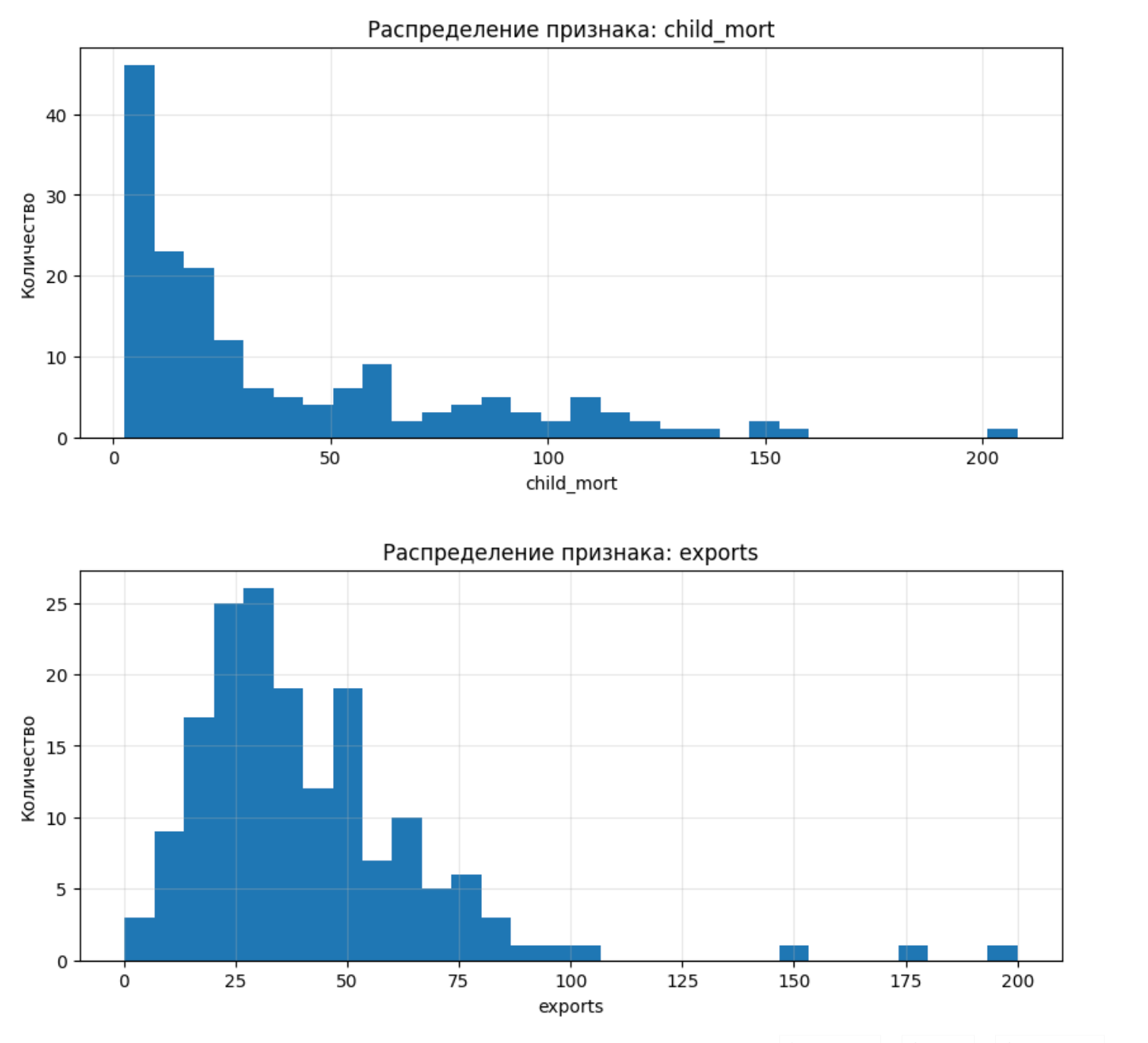


Рисунок 7 – Гистограммы распределения числовых признаков

# **Часть 2**

На данном этапе был выполнен поиск аномалий в данных с использованием алгоритма Isolation Forest. Перед обучением модели все признаки были стандартизированы, что обеспечило корректную работу алгоритма в многомерном пространстве. В качестве параметра contamination было выбрано значение 0.05, что соответствует предположению о наличии около 5% выбросов в данных. В результате каждому наблюдению была присвоена метка, указывающая на его нормальность или аномальность, что позволило количественно оценить число выявленных выбросов и использовать эту информацию на последующих этапах анализа, представлено на рисунках 8-9.

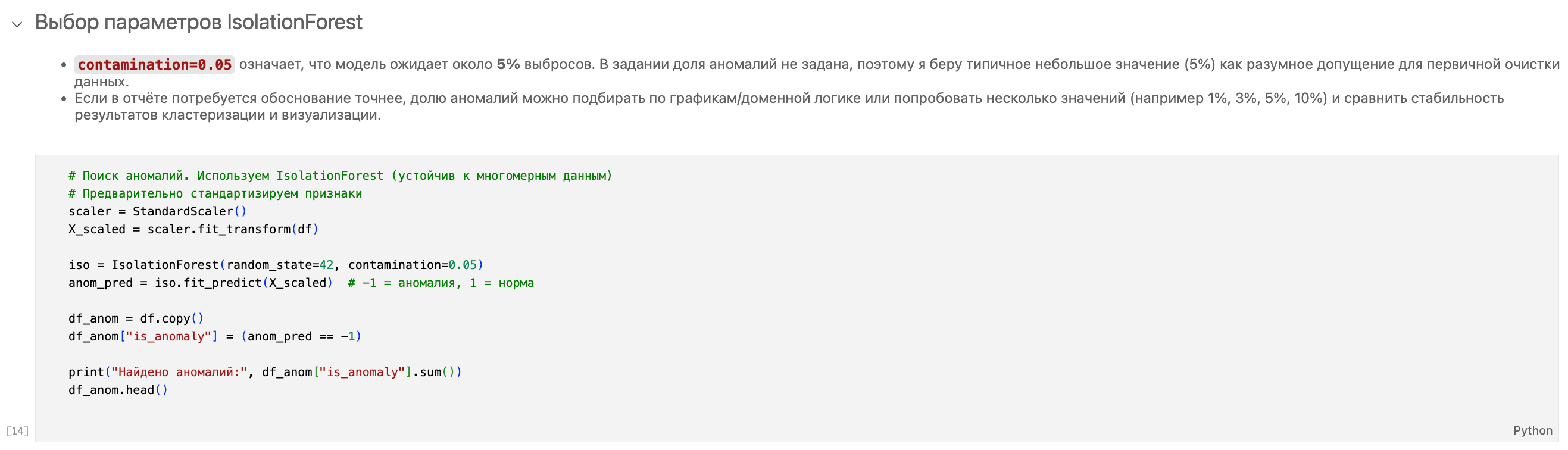


Рисунок 8 – Поиск аномалий

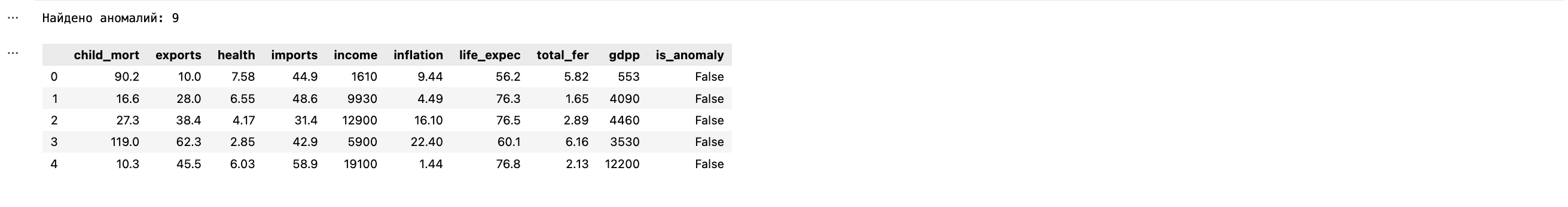


Рисунок 9 –Аномалии

Для визуального анализа результатов поиска аномалий были выбраны три числовых признака с наибольшей дисперсией. Такой выбор обусловлен тем, что признаки с высокой вариативностью обладают наибольшим разбросом значений, вследствие чего аномальные наблюдения на них наиболее заметны визуально. Для каждого выбранного признака были построены диаграммы рассеяния, на которых нормальные объекты и аномалии выделены разными цветами. Полученные графики позволяют наглядно подтвердить наличие выбросов и проанализировать характер отклонений по отдельным социально-экономическим показателям, представлено на рисунках 10-13.



Рисунок 10 – Признаки аномалий



Рисунок 11 – Income аномалии



Рисунок 12 – Gdpp аномалии

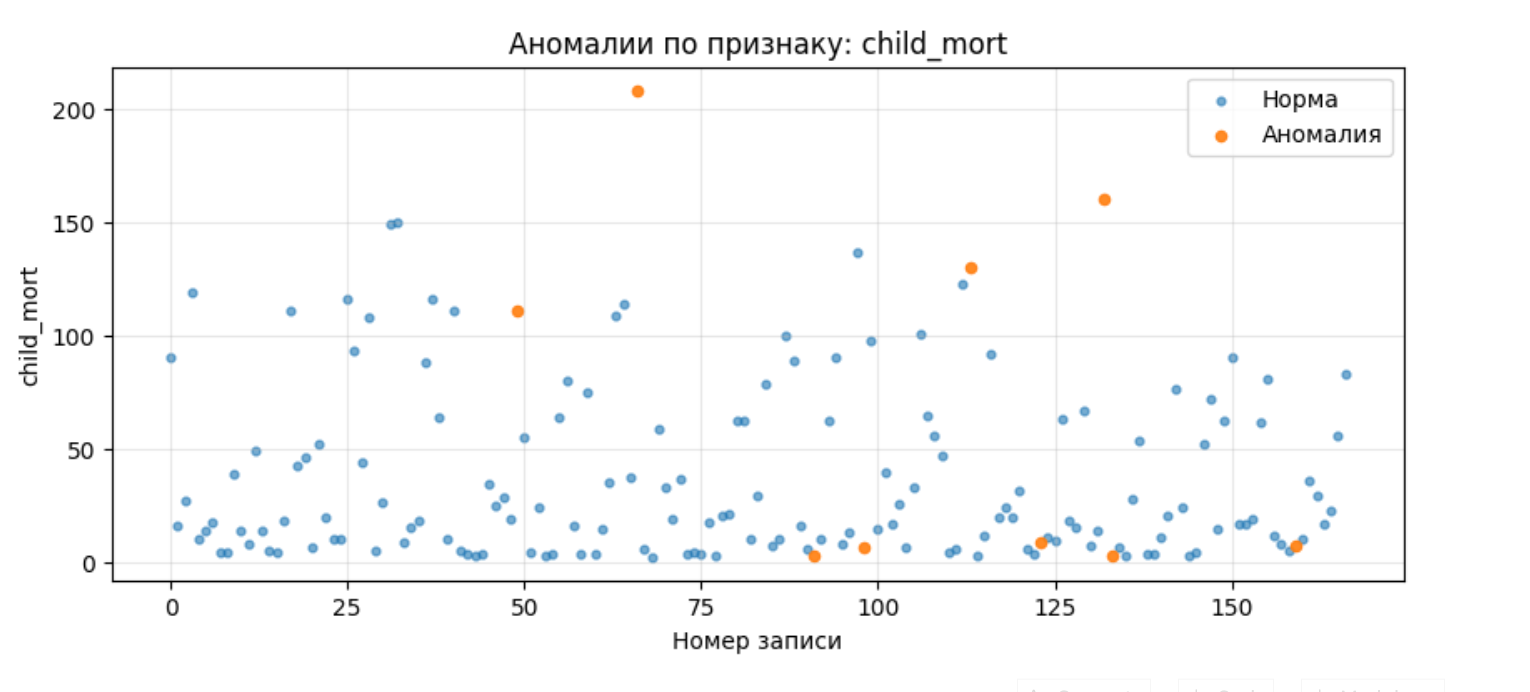


Рисунок 13 – Child\_mort аномалии

На основе результатов алгоритма Isolation Forest был сформирован очищенный датасет путём удаления наблюдений, идентифицированных как аномальные. После удаления выбросов данные были стандартизированы с использованием метода StandardScaler, что обеспечило приведение признаков к единому масштабу. В результате был получен финальный набор данных без аномалий, пригодный для дальнейшего анализа и применения методов машинного обучения. Размерность датасета сократилась с 167 до 158 наблюдений, что соответствует заданной доле аномалий, представлено на рисунке 14.



Рисунок 14 – Удалении аномалий

# **Часть 3**

На данном этапе был выполнен переход к пространству меньшей размерности с использованием метода главных компонент (PCA). Исходные стандартизированные данные были преобразованы в двумерное пространство, представленное двумя главными компонентами (PC1 и PC2), которые сохраняют наибольшую долю дисперсии исходных признаков. Полученное представление используется для последующей визуализации структуры данных и применения методов кластеризации, представлено на рисунке 15.

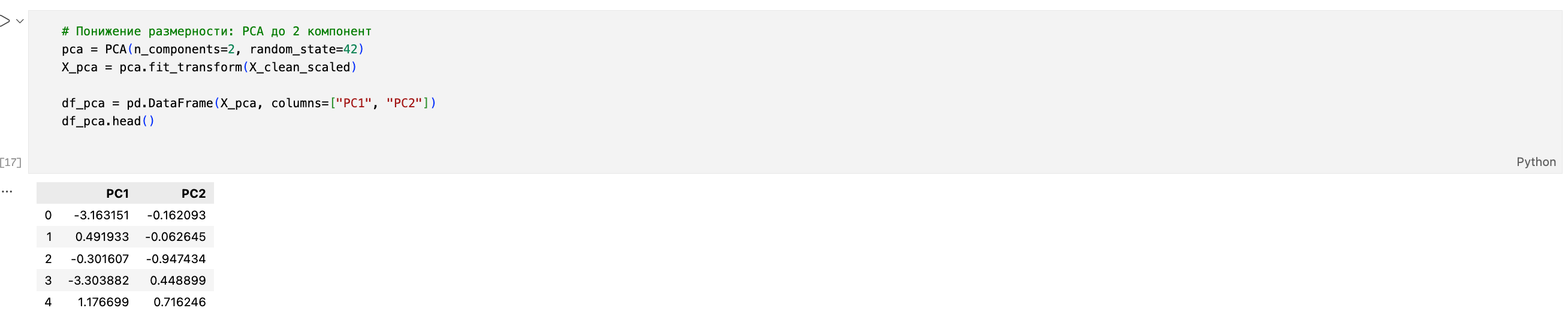


Рисунок 15 – Понижение размерности

оценена эффективность понижения размерности методом главных компонент. Для двух выбранных компонент вычислена доля объяснённой дисперсии: первая главная компонента объясняет около 44,6% общей вариативности данных, вторая — около 17,0%. В совокупности две компоненты сохраняют примерно 63,6% информации исходного датасета, что свидетельствует о корректности выбранного числа компонент. Также была построена визуализация данных в пространстве главных компонент PC1 и PC2, позволяющая наглядно проанализировать структуру данных и использовать полученное представление для последующего кластерного анализа, представлено на рисунках 16-17.

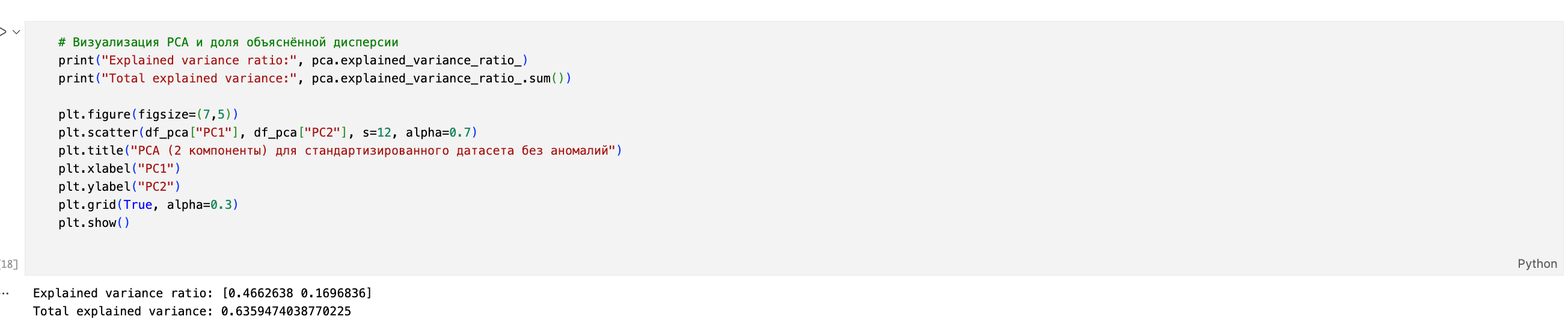


Рисунок 16 – Уменьшение признаков до двух



Рисунок 17 – PCA-представление очищенного и стандартизированного датасета

На рисунках 18-19 представлен дополнительный метод понижения размерности t-SNE



Рисунок 18 – t-SNE метод

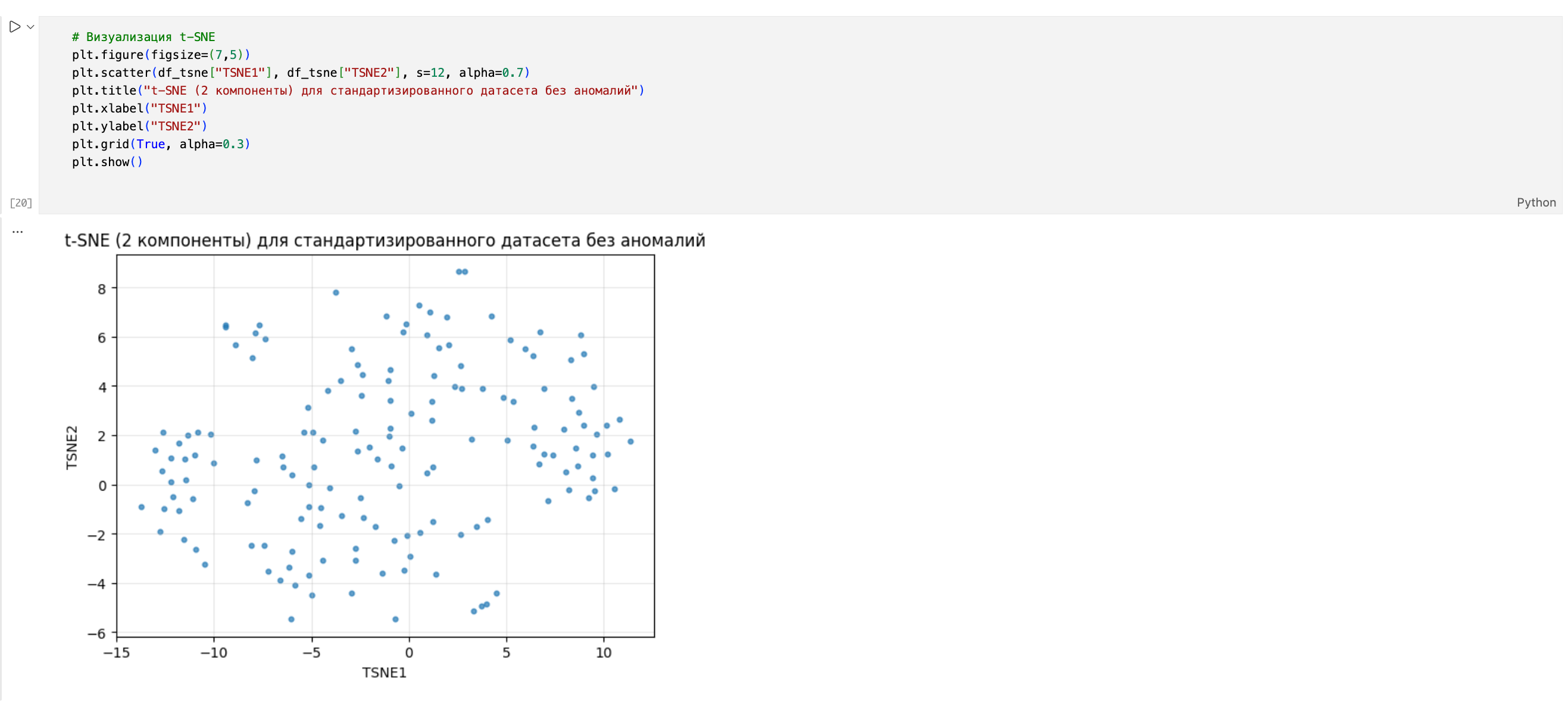


Рисунок 19 – Очищенный и стандартизированного датасета

**Вывод**

В ходе работы было выполнено понижение размерности датасета методом главных компонент (PCA) до двух компонент на предварительно стандартизированных и очищенных от аномалий данных. Анализ доли объяснённой дисперсии показал, что первые две главные компоненты суммарно объясняют около 63,6% общей вариативности данных, что свидетельствует о сохранении значительной части информации при существенном сокращении размерности признакового пространства. t‑SNE даёт более «раздвинутую» визуализацию локальной структуры и может лучше показывать возможные группы, но хуже интерпретируется количественно. Для кластеризации далее используется PCA‑представление и сравнение с кластеризацией исходных признаков.

# **Часть 4**

На данном этапе была выполнена агломеративная кластеризация стандартизированного датасета без аномалий. Оптимальное число кластеров было определено на основе метрики silhouette score путём перебора значений от 2 до 6. Максимальное значение метрики было получено при k = 2, что позволило выделить два наиболее устойчивых кластера в данных, представлено на рисунке 20.

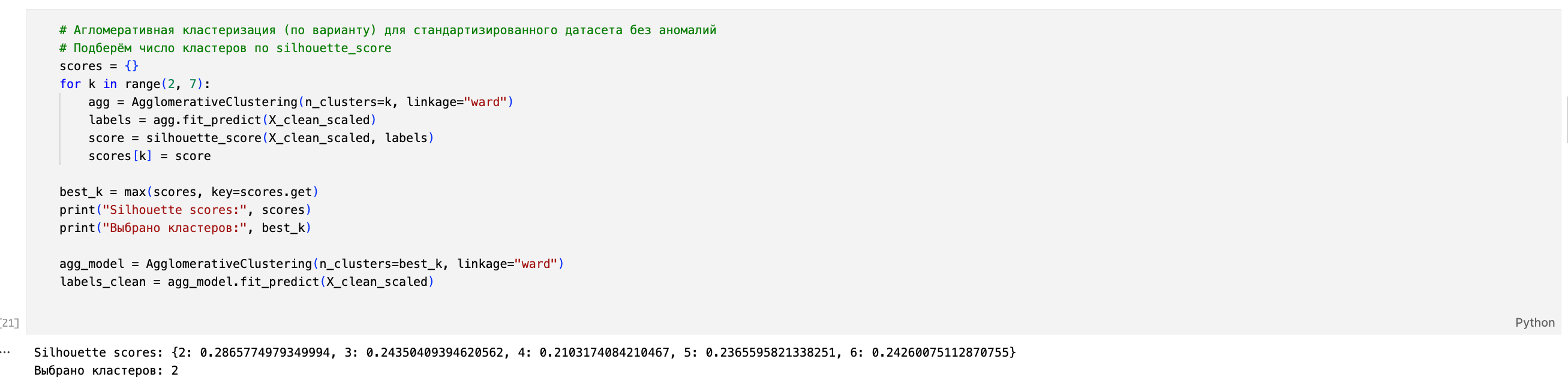


Рисунок 20 – Агломеративная кластеризация

На графике представлена зависимость метрики качества кластеризации silhouette score от количества кластеров k для алгоритма агломеративной кластеризации. Анализ графика позволяет визуально оценить, при каком числе кластеров достигается наилучшее качество разбиения данных, представлено на рисунке 21.

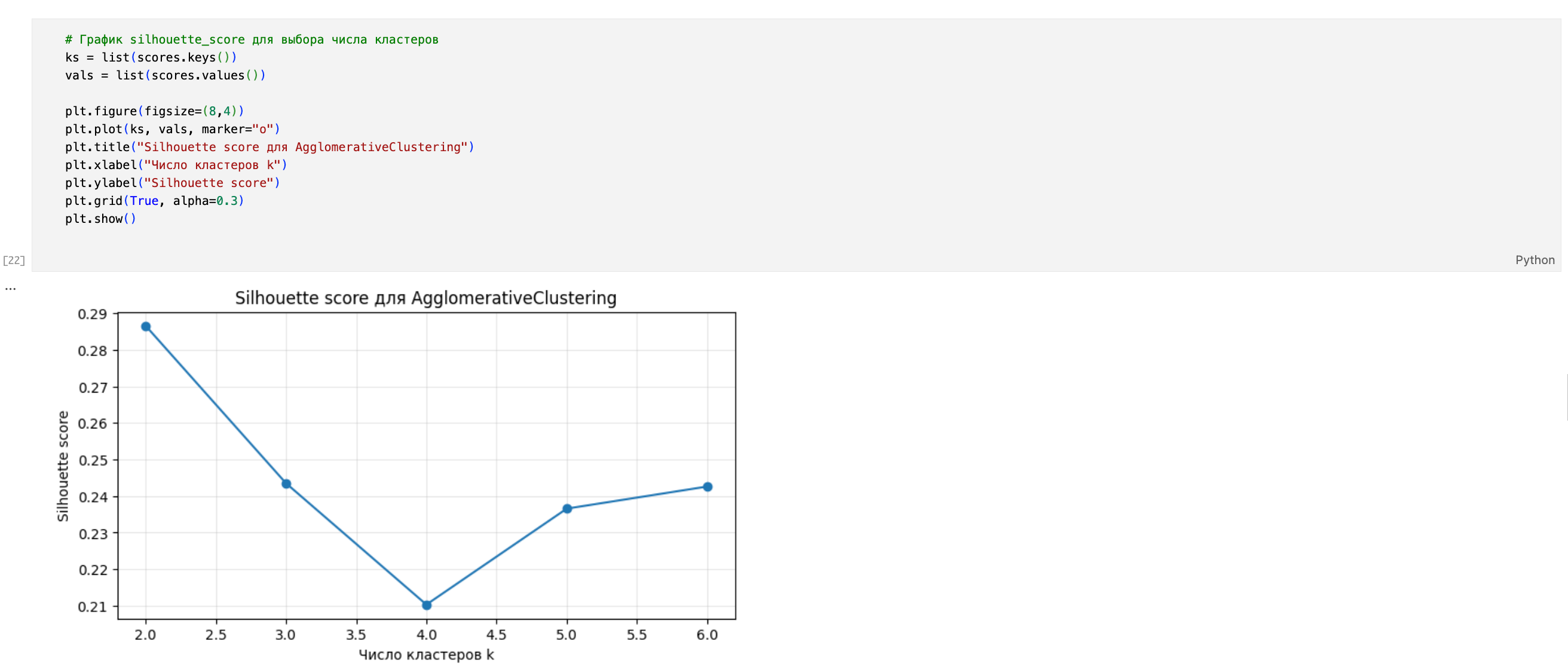


Рисунок 21 – Оценка качества кластеризация

визуализация результатов агломеративной кластеризации для не стандартизированного датасета. Для анализа выбраны три числовых признака с наибольшей вариативностью. Для каждого признака строится диаграмма рассеяния, на которой объекты окрашены в соответствии с принадлежностью к кластерам, что позволяет оценить характер разделения данных и интерпретируемость полученных кластеров, представлено на рисунках 22-25.



Рисунок 22 – Визуализации результатов кластеризации



Рисунок 23 – Income



Рисунок 24 – Gdpp



Рисунок 25 – Child\_mort

Агломеративная кластеризация стандартизированного датасета после понижения размерности методом главных компонент (PCA). Кластеризация проводится в двумерном пространстве первых двух главных компонент, после чего результаты проецируются обратно на исходные числовые признаки для сравнения структуры кластеров. Такой подход позволяет оценить, насколько понижение размерности сохраняет информацию, необходимую для корректной кластеризации и интерпретации данных, представлено на рисунках 26-30.

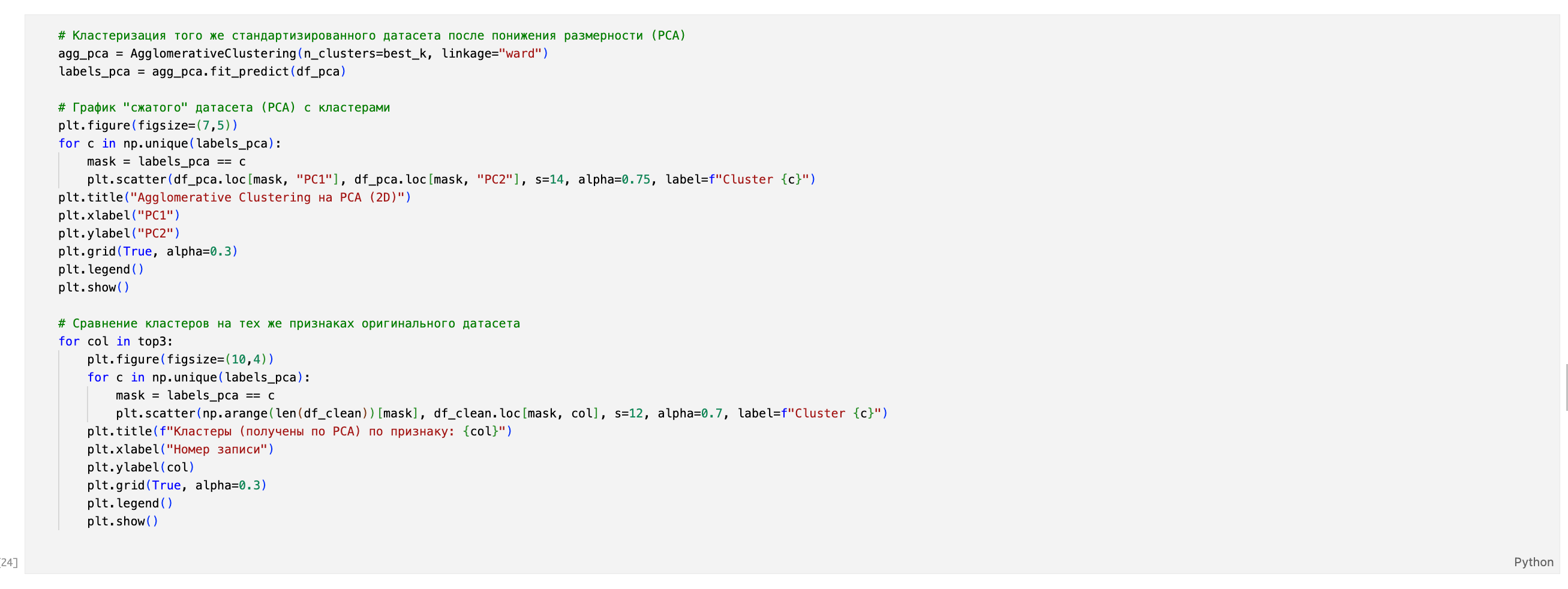


Рисунок 26 – Проекция кластеров, полученных после PCA

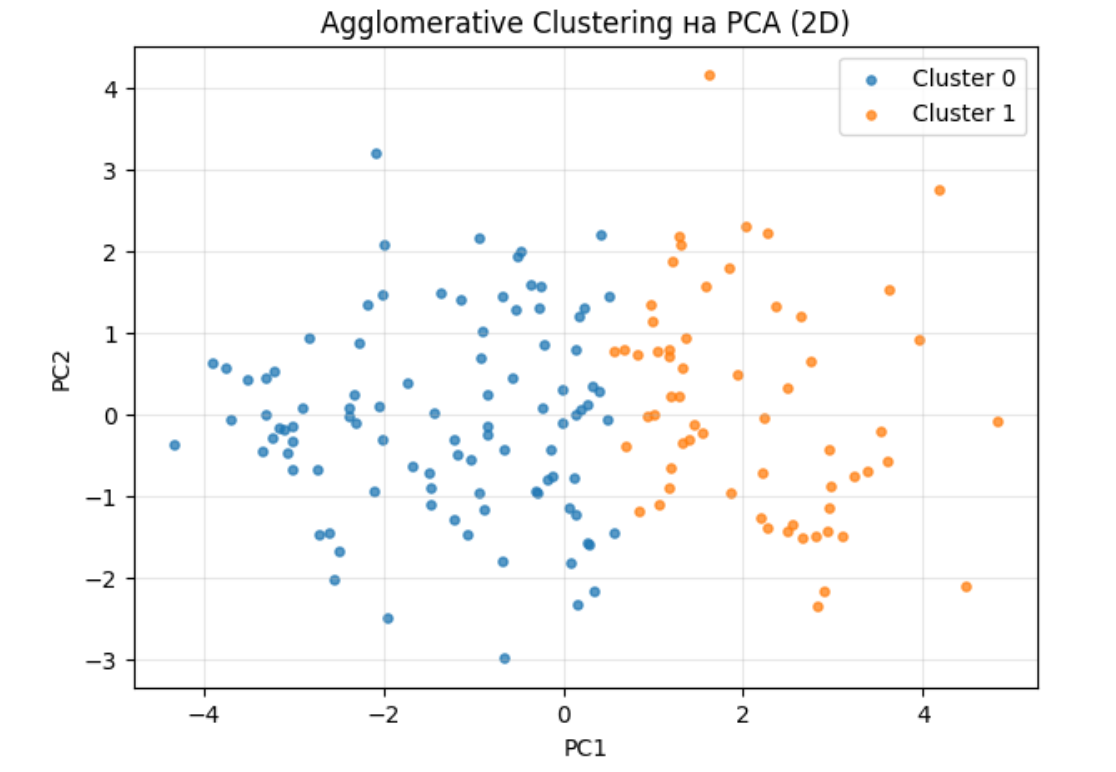


Рисунок 27 – График кластеризации в двумерном пространстве

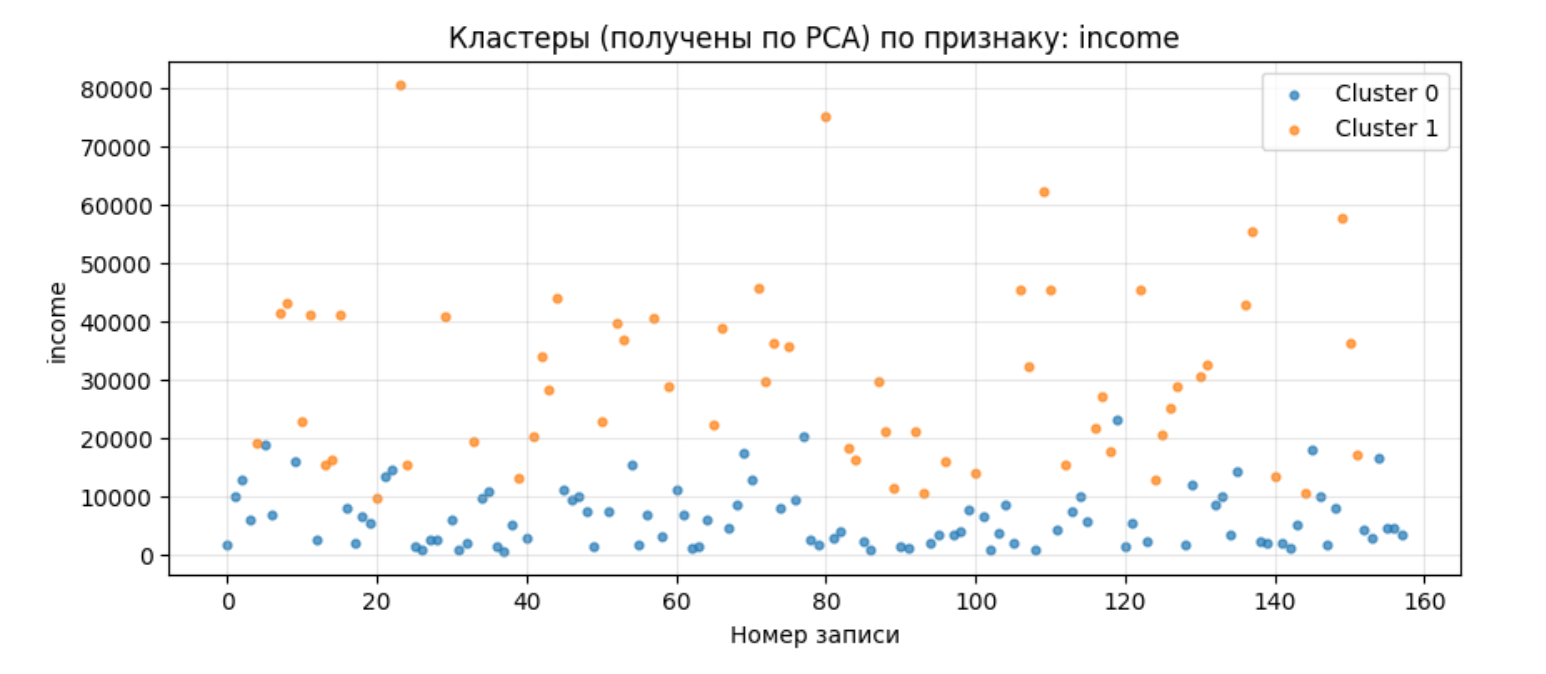


Рисунок 28 – График распределения значений income



Рисунок 29 – График распределения значений gdpp

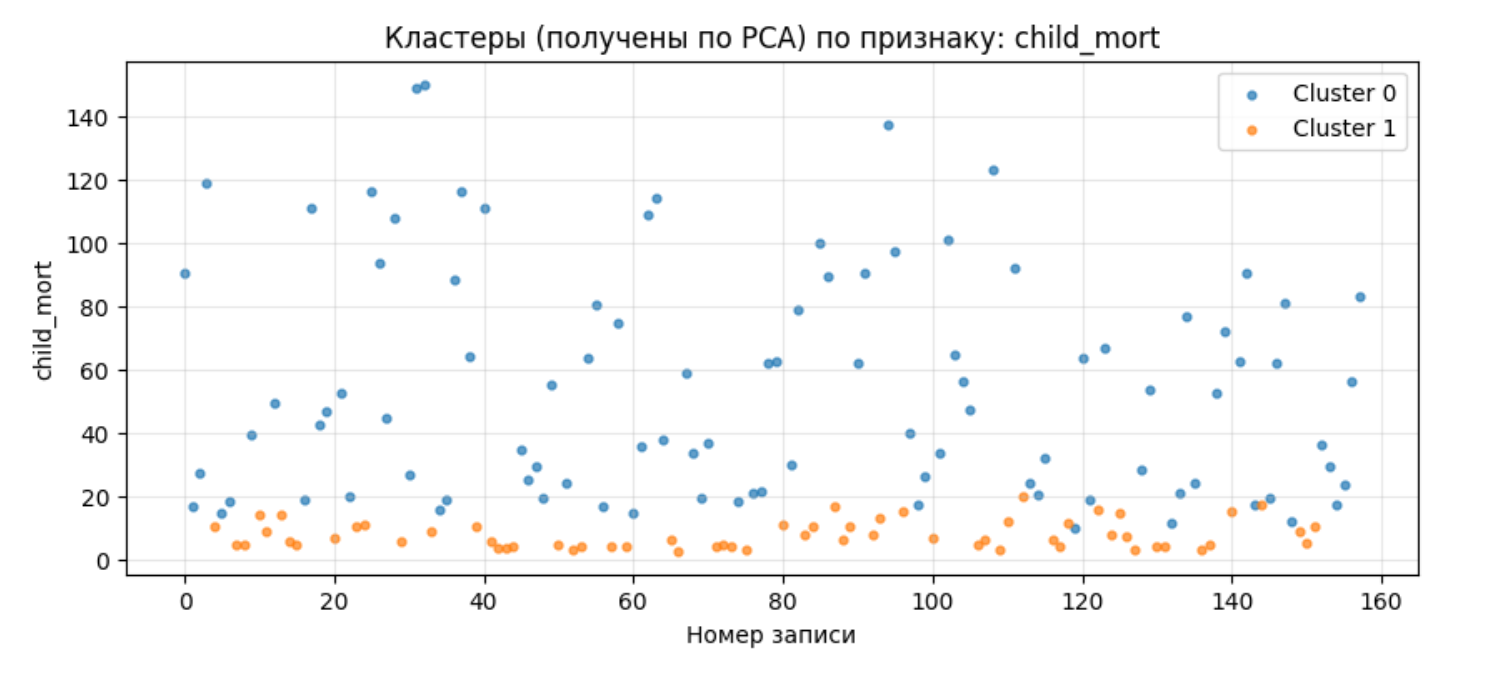


Рисунок 30 – График распределения значений child\_mort

Сравнение результатов агломеративной кластеризации, полученных на исходном стандартизированном датасете и на данных после понижения размерности методом PCA. Для количественной оценки была рассчитана доля совпадения кластерных меток, которая составила около 0.80, что свидетельствует о высокой степени сохранения структуры данных после применения PCA. Дополнительно была построена таблица сопряжённости, показывающая соответствие кластеров до и после понижения размерности. Анализ таблицы подтверждает, что большинство объектов сохранили кластерную принадлежность, а изменения затронули преимущественно пограничные наблюдения, представлено на рисунке 31.



Рисунок 31 – График распределения значений child\_mort

**Вывод**

В рамках работы была выполнена агломеративная кластеризация стандартизированного датасета без аномалий с использованием метода Ward. Оптимальное количество кластеров было выбрано на основе метрики **silhouette score**, максимальное значение которой достигалось при **k = 2**, что указывает на наилучшее соотношение внутрикластерной компактности и межкластерной разделимости для данного набора данных.

Полученное значение silhouette score находится на умеренном уровне, что свидетельствует о наличии различимой, но не строго выраженной кластерной структуры. Это означает, что объекты внутри кластеров в целом схожи между собой, однако между кластерами сохраняется частичное перекрытие, что характерно для реальных социально-экономических данных.

Визуальный анализ кластеров в пространстве главных компонент (PCA) подтвердил корректность разбиения: кластеры формируют устойчивые группы с заметным разделением вдоль первой главной компоненты. Дополнительное рассмотрение кластеров в исходном пространстве признаков (например, по показателю **income**) показало, что выделенные группы отличаются по уровню ключевых социально-экономических характеристик, что повышает интерпретируемость результатов.