ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ |
| КЛАСТЕРИЗАЦИЯ ДАННЫХ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4017 |  |  |  | Т. А. Михайлова |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2022

**Цель работы:** изучить алгоритмы и методы кластерного анализа на практике.

**Вариант №9.**

Записи о болезнях сердца.

1. возраст: возраст пациента (лет) - age

2. анемия: снижение количества эритроцитов или гемоглобина (логическое значение) - anemia

3. высокое кровяное давление: если у пациента гипертония (логическое значение) - hypertension

4. креатининфосфокиназа (КФК): уровень фермента КФК в крови (мкг/л) - creatinine\_phosphokinase

5. диабет: если у пациента диабет (логическое значение) - diabetes

6. фракция выброса: процент крови, покидающей сердце при каждом сокращении (в процентах) - ejection\_fraction

7. тромбоциты: тромбоциты в крови (килотромбоциты/ мл) - platelets

8. пол: женщина или мужчина (бинарный) - sex

9. креатинин сыворотки: уровень креатинина сыворотки в крови (мг/дл) - serum\_creatinine

10. натрий сыворотки: уровень натрия сыворотки в крови (мэкв/л) - serum\_sodium

11. курение: если пациент курит или нет (логическое) – smoking

12.время: период наблюдения (дни) – time

13. событие смерти: если пациент умер в течение периода наблюдения (логическое значение) – death\_event

**Ход работы:**

1. Был подключен файл по ссылке с репозитория гитхаба. С помощью метода .head() были выведены первые 5 записей таблицы.

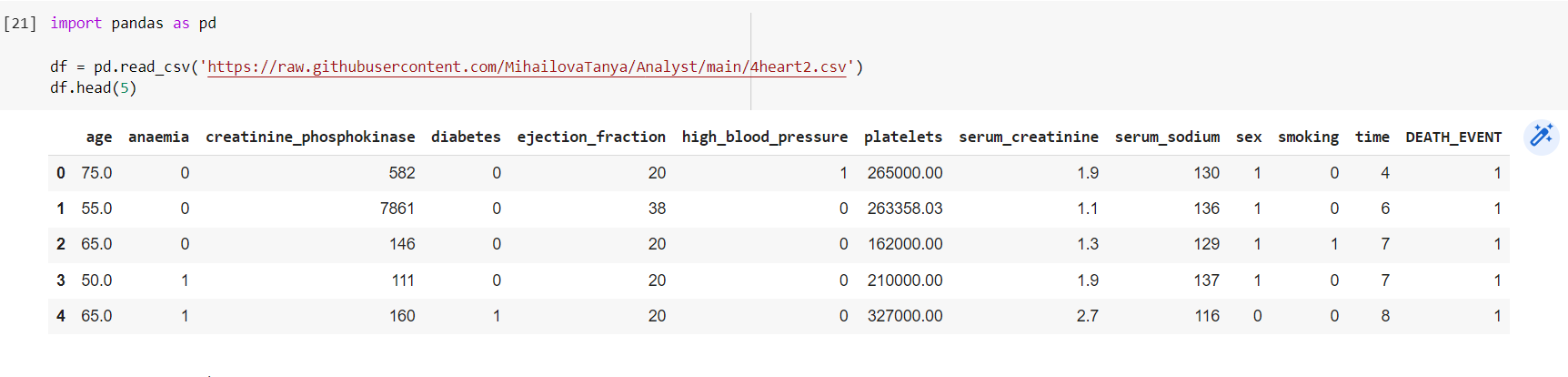


Рисунок 1 – Чтение файла

1. С помощью метод .info() были выведены типы данных файла. После вызова стало известна информация о типах данных, названиях самих столбцов и отсутствии пропусков.

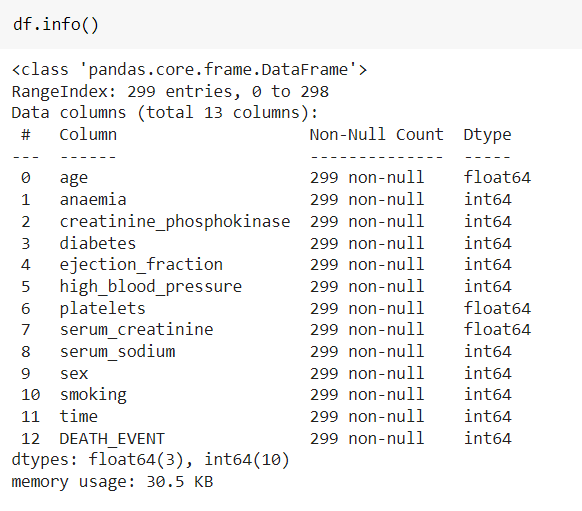


Рисунок 2 – Сведения о датасете

1. С помощью метода .columns были выведены названия столбцов. Некоторые из названий столбцов введены неверно или плохо отражают информацию, поэтому с помощью метода .rename() были переименованы, а именно:

anaemia -> anemia

high\_blood\_pressure -> hypertension

DEATH\_EVENT -> death\_event

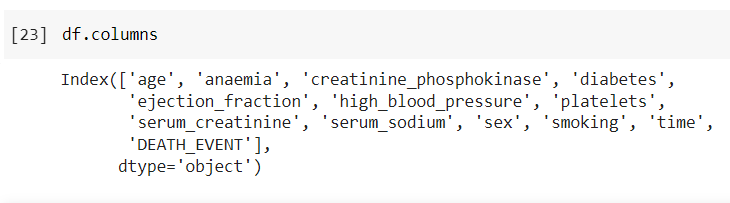


Рисунок 3 – Названия колонок до переименования

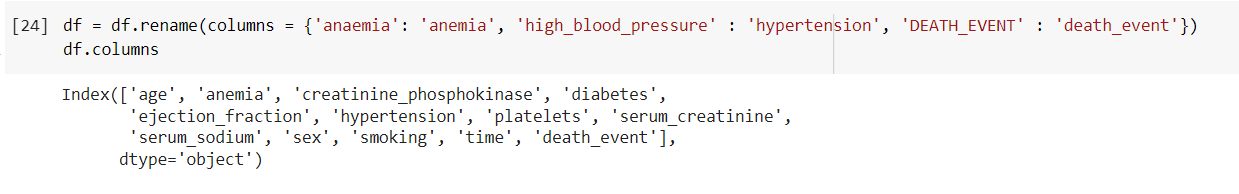


Рисунок 4 – Названия столбцов после переименования

1. С помощью метода .isna() находим количество пустых ячеек в датасете. По результатам этого метода не было выявлено пропусков.

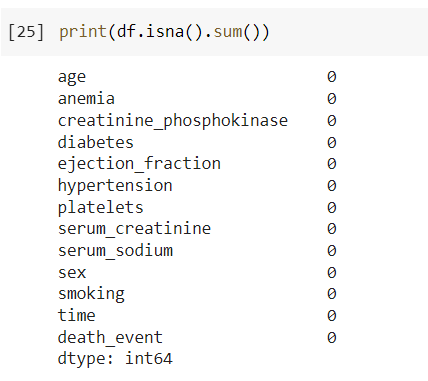


Рисунок 5 – Проверка на наличие пропусков

С помощью метода .duplicated() находим повторяющиеся записи в таблице. По результатам данного метода не было выявлено явных дубликатов в датасете.

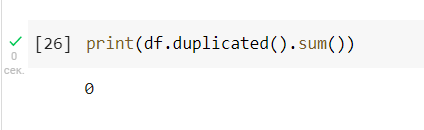


Рисунок 6 – Проверка на наличие явных дубликатов

С помощью метода .unique() проверяю датасет на неявные дубликаты. По результатам этого метода не было выявлено неявных дубликатов. Однако можно заменить значение 263358.03 на 263358 в столбце platelets (тромбоциты), а также в столбце age (возраст) значение 60.677 было заменено на 60. Изменение значения было сделано с помощью метода .replace().

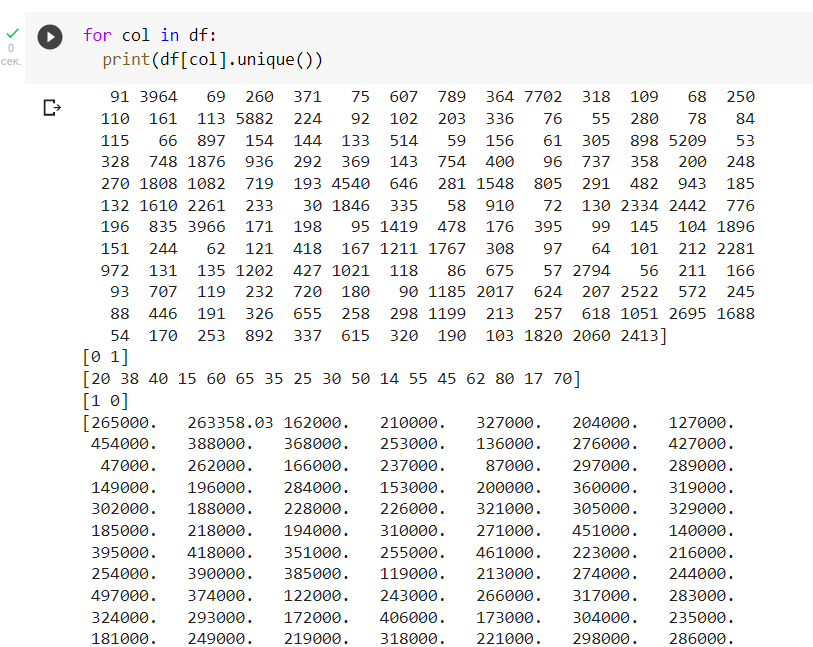


Рисунок 7 – Проверка на наличие неявных дубликатов

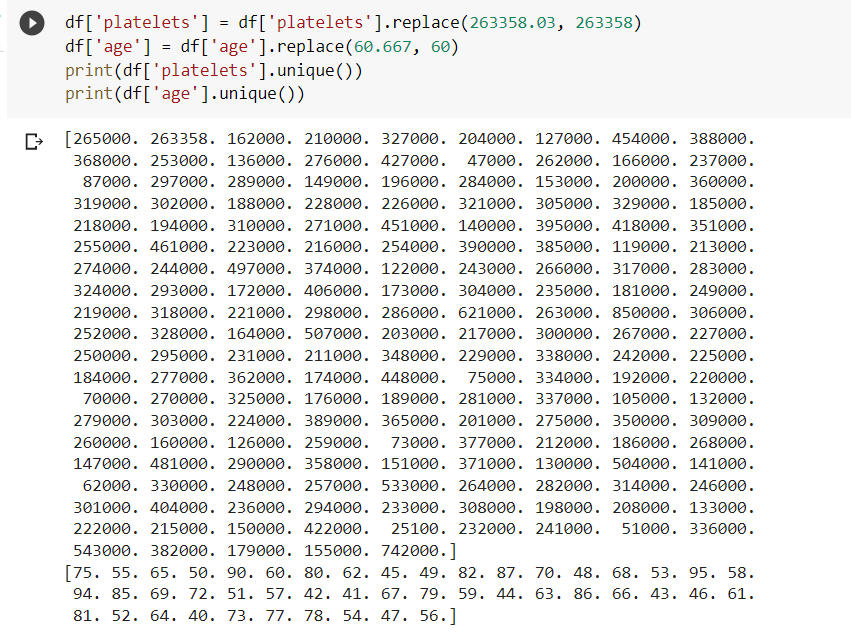


Рисунок 8 – Результат изменения значений

Теперь для столбцов age (возраст) и platelets (тромбоциты) можно заменить тип данных с float64 на int64.

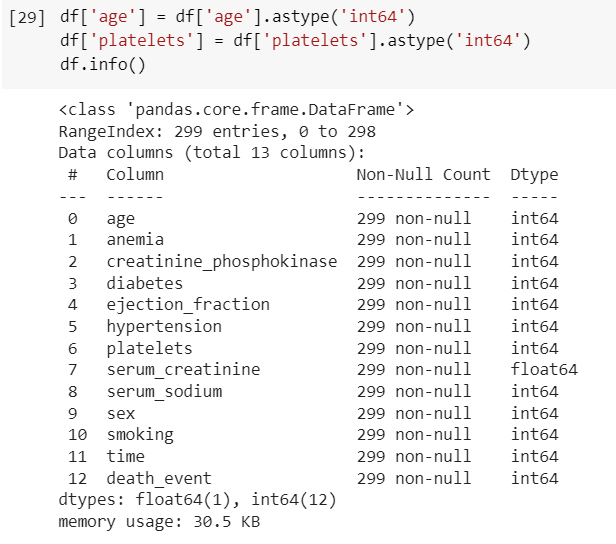


Рисунок 9 – Результат изменения типа данных

1. С помощью метода .describe() были описаны данные в датасете.

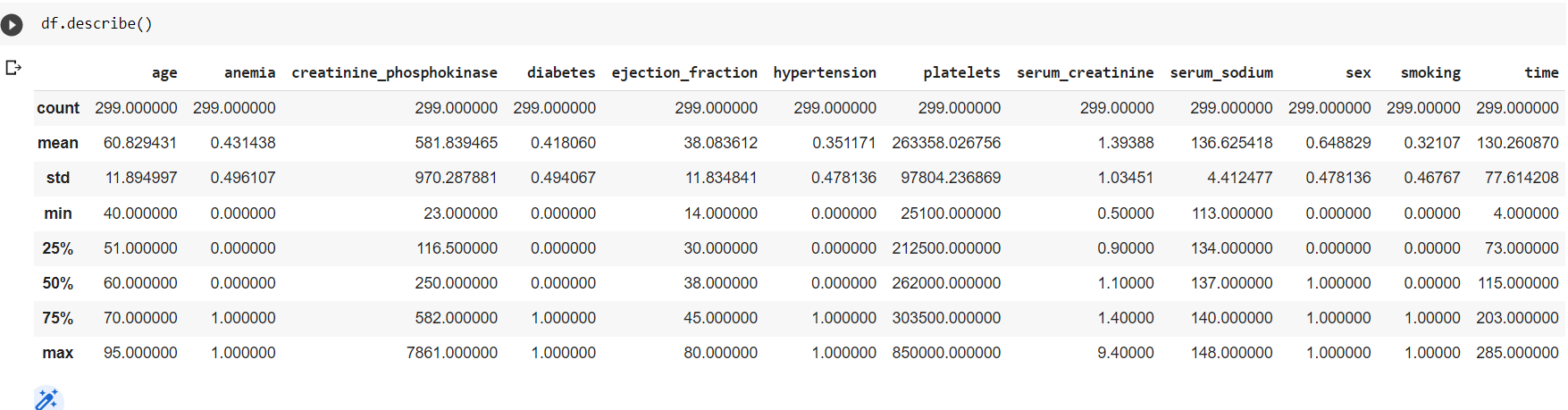


Рисунок 10 – Описательная статистика датасета

* Строка count показывает 299 строк в каждом столбце.
* Cтрока mean подсчитывает среднее значение по столбцу.
* Cтрока std определяет среднеквадратичное отклонение.
* Cтрока min отображает минимальное значение в столбце.
* В строках 25%, 50%, 75% показываются средние значения в каждом квартиле.
* В строке max отображается максимальное значение в столбце.

1. Были удалены столбцы с информацией, не имеющих отношения к анализам по заболеваниям сердца.

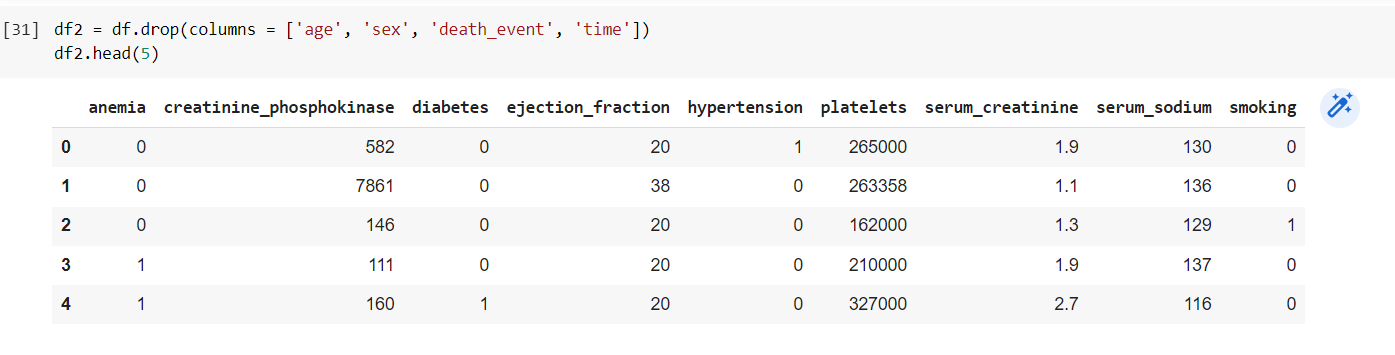


Рисунок 11 – Таблица без данных, не имеющих отношения к исследованию

1. Выполняю стандартизацию данных с помощью модуля sklearn библиотеки StandardScaler.

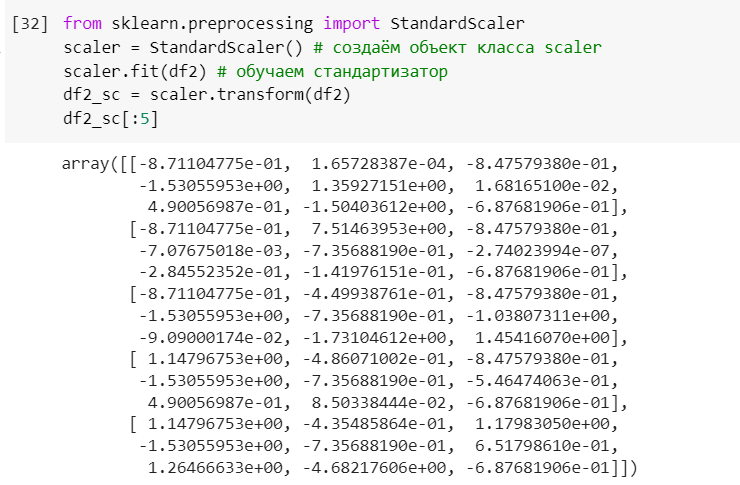


Рисунок 12 – Стандартизация данных

1. Кластеризация агломеративным иерархическим методом.

Передаю стандартизированную таблицу в качестве параметра функции linkage(). Был использован параметр ward для показательной диаграммы.

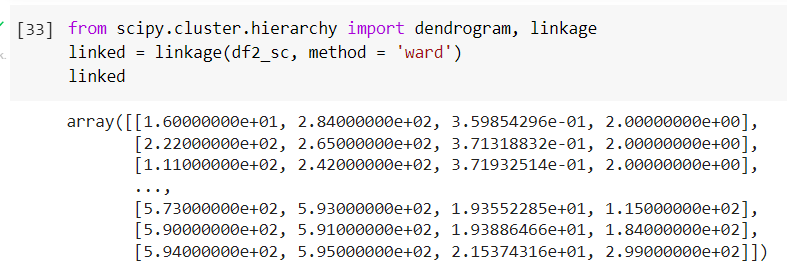


Рисунок 13 – Передача стандартизированной таблицы

Визуализирую таблицу связок linked как дендрограмму. Было использовано усечение через параметр truncate\_mode с режимом level, отображаемым 6 уровней.

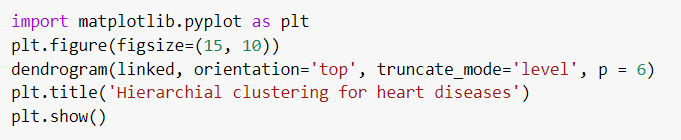


Рисунок 14 – Визуализация таблицы

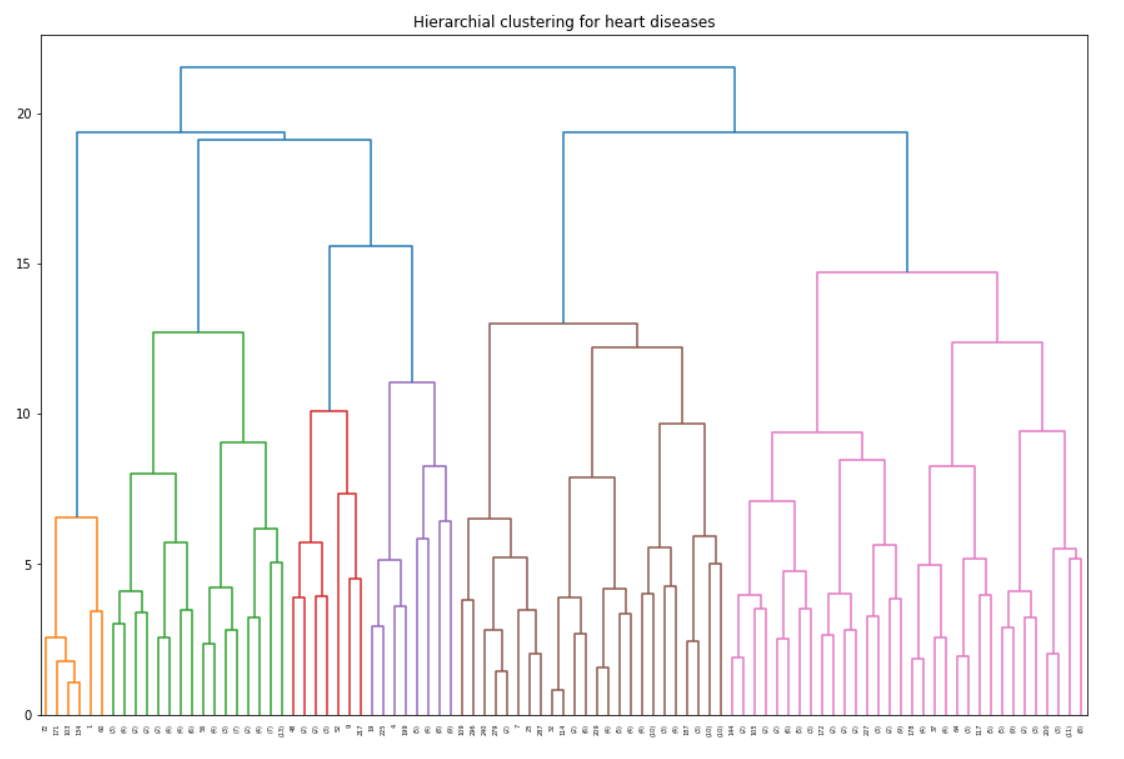


Рисунок 15 – Дендограмма

Глядя на дендограмму, можно сделать вывод, что оптимальное число кластеров равно 6. С помощью данного метода можно увидеть количество элементов в каждом кластере.

Список кластеров по возрастанию элементов в них:

* Оранжевый
* Красный
* Фиолетовый
* Зеленый
* Коричневый
* Розовый

Список элементов, которые имеют наибольшее влияние на выделение кластера:

1. Оранжевый: 72, 171, 109, 134, 1, 60
2. Зеленый: 126
3. Красный: 42, 52, 9, 217
4. Фиолетовый: 19, 225, 4, 199
5. Коричневый: 109, 298, 249, 279, 7, 29, 297, 32, 114, 208, 157
6. Розовый: 144, 106, 172, 227, 175, 17, 64, 117

Индексы элементов соответствуют номеру строк в датасете.

1. Кластеризация методом К-средних.

Был выполнен алгоритм k-means. Количество кластеров равно 6, random\_state = 0 для сохранения и воспроизводимости результата. В переменной labels сохраняются индексы предложенных алгоритмом групп и формируется вектор кластеров. Распределяю данные по кластерам с помощью их индексов, найденных в алгоритме.

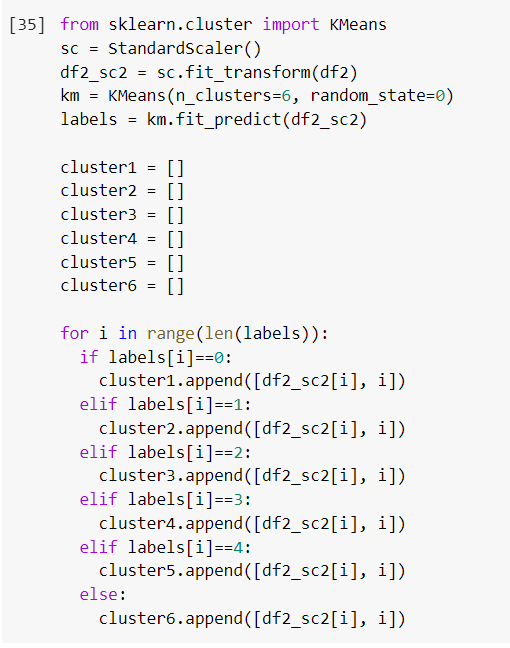


Рисунок 16 – Распределение данных по их кластерам

Были найдены центры кластеров.

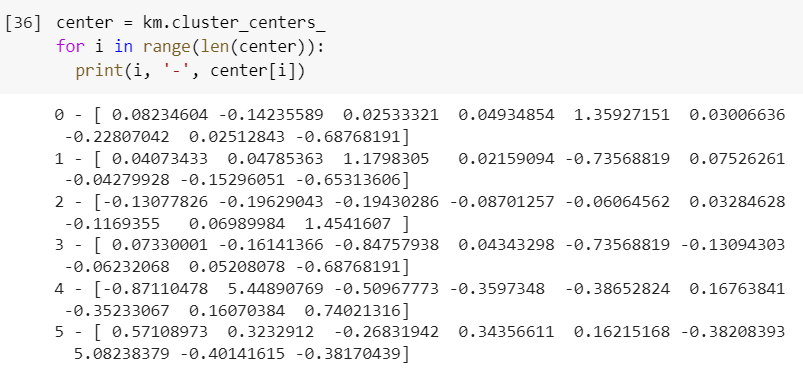


Рисунок 17 – Нахождение центра кластеров

Рассчитала евклидово расстояние между кластерами.

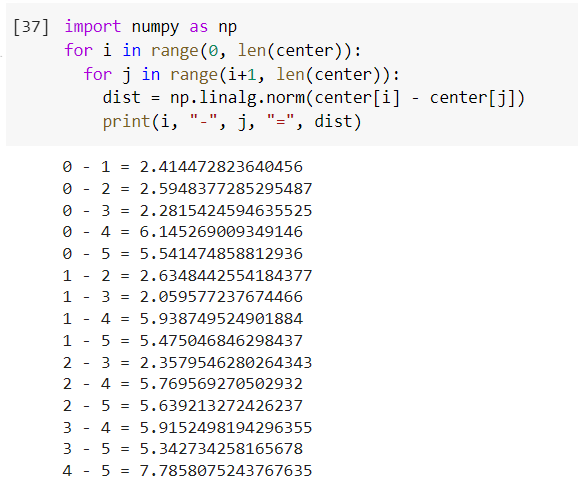


Рисунок 18 – Нахождение евклидового расстояния

Были определены признаки, оказавшие наибольшее влияние на выделение кластеров с помощью значений евклидова расстояния и центрами кластера.

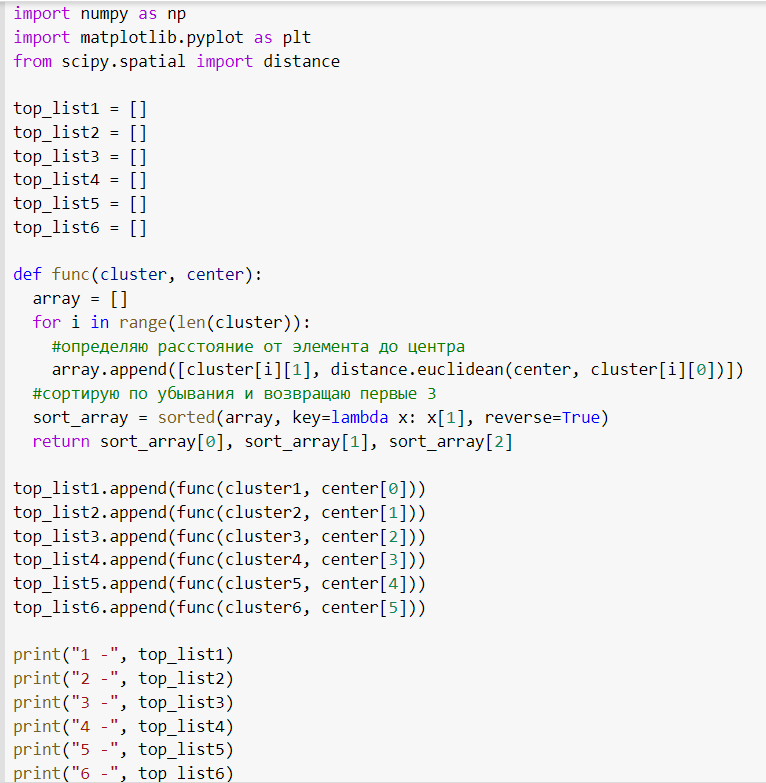


Рисунок 19 – Нахождение признаков, влияющих на выделение кластеров

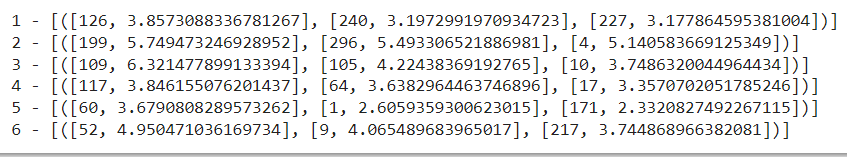


Рисунок 20 – Признаки

Список индексов элементов, которые имеют наибольшее влияние на выделение кластера:

1. 126, 240, 227
2. 199, 296, 4
3. 109, 105, 10
4. 117, 64, 17
5. 60, 1, 171
6. 52, 9, 217

В этом списке элементов, имеющих наибольшее влияние на выделение кластера, можно заметить элементы, которые были отмечены ещё в агломеративном иерархическом методе (126, 4, 199, 109, 117, 64, 17, 60, 1, 52, 9, 217).

1. Рассчитываю оптимальное количество кластеров с помощью метрики силуэта. Итоговый результат может принимать значение от -1 до 1. Значение данной метрики получилось ближе к нулю. Это может говорить о том, что кластеры накладываются друг на друга.

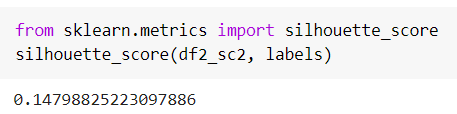


Рисунок 21 – Метрика силуэта для 6 кластеров

Так как оптимальное количество кластеров не было найдено с помощью метода к-средних, то найдем это значение с помощью метрики силуэта. Исходя из результатов видно, что оптимальное количество кластеров равно 18.

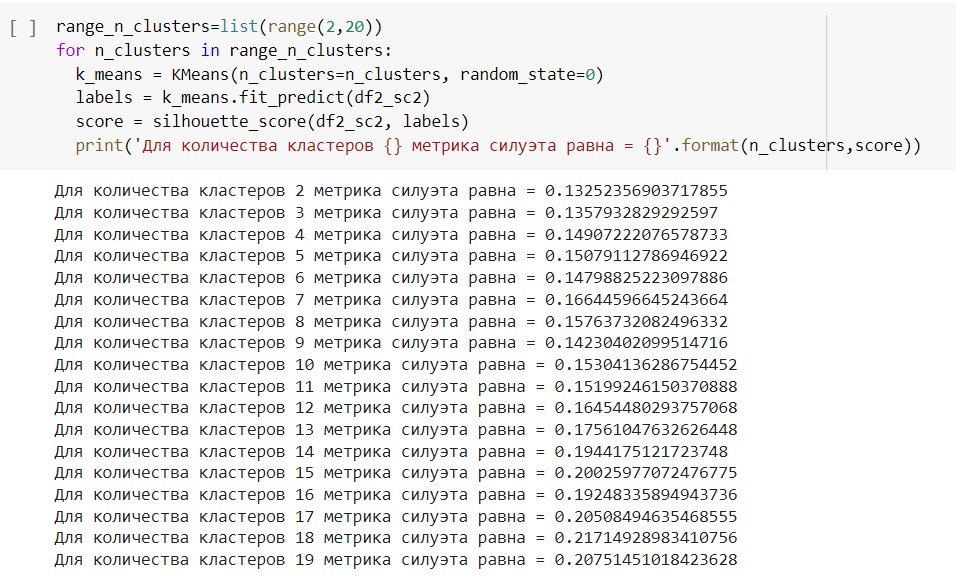


Рисунок 22 – Метрика силуэта для 5 кластеров

**Ссылка на notebook:**

https://colab.research.google.com/drive/174IJ-4-n51XcbQakWmFB-tgr\_-M-YjOf?authuser=1#scrollTo=Oenhb\_ykJf5K

**Вывод:**

Была выполнена предобработка датасета перед проведением исследования: описаны столбцы, тип данных в них, изменены названия колонок на точные и понятные наименования. Также были проверены данные на наличие пропусков, явных и неявных дубликатов, изменены типы данных в необходимых колонках. Пропусков и дубликатов не было обнаружено.

Перед стандартизацией данных столбцы были проанализированы и удалены лишние, не имеющие отношения к исследованию.

С помощью иерархического агломеративного метода было выделено 6 кластеров. А используя усечения было выяснено сколько элементов принадлежит каждому кластеру. Также были выделены признаки, оказавшие наибольшее влияние на выделение кластеров.

Была выполнена кластеризация методом К-средних с заданным количеством кластеров, равных 6. Были высчитаны центры кластеров и евклидово расстояние. Также были определены признаки, оказавшие наибольшее влияние на выделение кластеров.

С помощью метрики силуэта было рассчитано оптимальное количество кластеров. Результатом метрики было оптимально выделить 5 кластеров.