Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение

высшего образования

«Пермский государственный национальный исследовательский

университет»

**Цифровая кафедра**

Отчёт по работе

**«Анализ факторов, влияющих на расходы покупателей в магазине»**

**Выполнили:**

Студенты 4-го курсов

физико-математического института факультета

Т. М. Овчинников

М. Л. Мазязин

**Проверил:**

к.ф.-м.н. А. В. Ратт

г. Пермь 2024

**Содержание**

Введение

1. Подготовка данных к анализу

2. Предварительный анализ данных

3. Кластерный анализ

Вывод

**Введение**

Расходы покупателей являются важным аспектом функционирования рынка и экономики в целом. Они влияют на потребительский спрос, распределение доходов и уровень благосостояния населения. Изучение расходов покупателей позволяет понять закономерности поведения потребителей, определить их предпочтения и потребности, а также разработать стратегии для удовлетворения этих потребностей и стимулирования спроса на определённые товары и услуги.

В современном мире, где рынок становится всё более конкурентным, компаниям необходимо учитывать расходы своих потенциальных клиентов, чтобы успешно конкурировать и привлекать внимание потребителей. Изучение расходов покупателей также помогает выявить возможные тенденции и изменения в потребностях и предпочтениях аудитории, что позволяет компаниям адаптировать свои продукты и услуги в соответствии с этими изменениями.

**Цель работы:** проанализировать данные о клиентах магазина, выполнить разбиение данных на кластеры и оценить качество кластеризации.

**Техническое задание:** требуется проанализировать данные о клиентах магазина (файл Customers.csv), выявить зависимости между факторными переменными, разбить данные на кластеры. Дать интерпретацию полученным результатам. Сделать выводы.

**1. Подготовка данных к анализу**

Была выполнена загрузка данных в датафрейм:

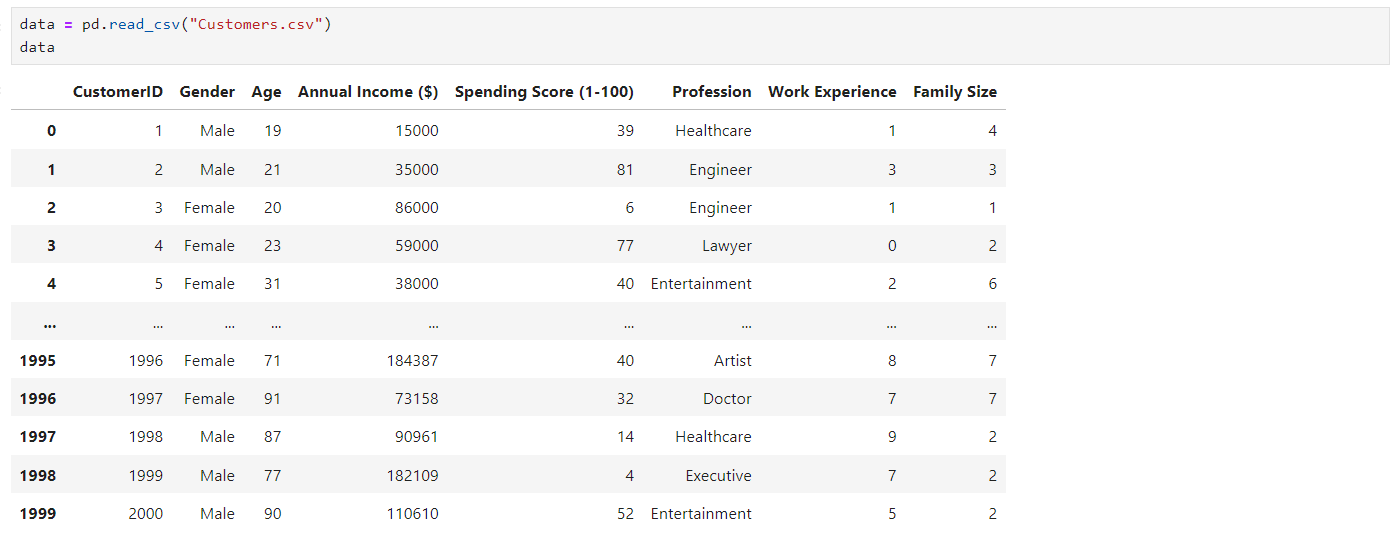


Рисунок 1. Данные датафрейма

Далее выполнена проверка, что все количественные столбцы имеют числовой тип, с последующим преобразованием типа столбца к числовому. А также проверка на пропуски, строки с пропусками были удалены из датафрейма:

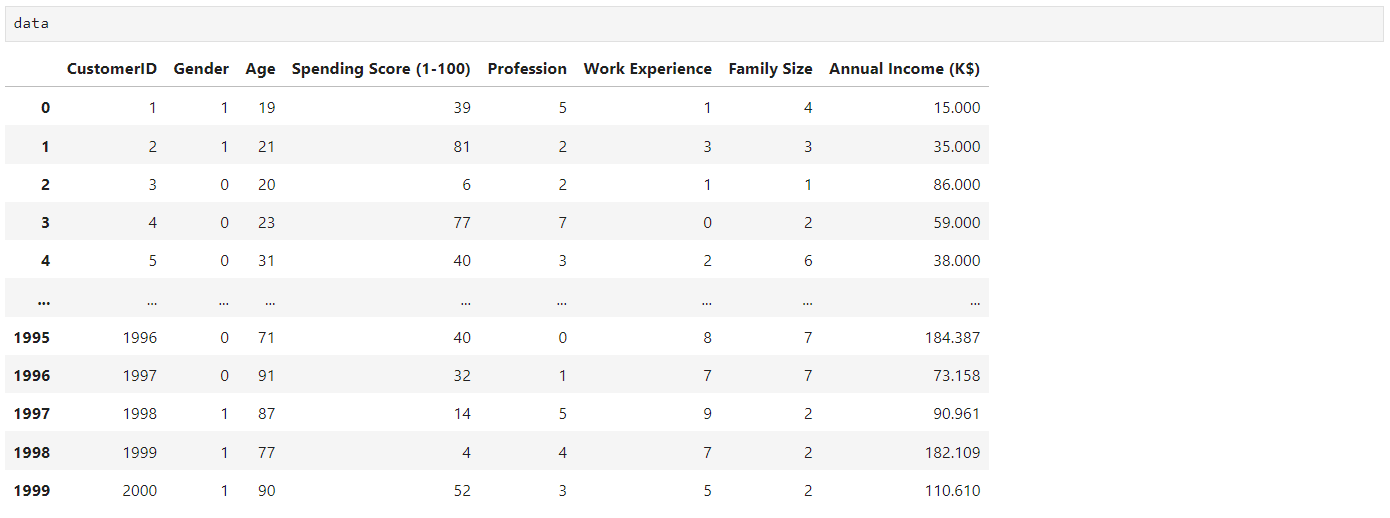
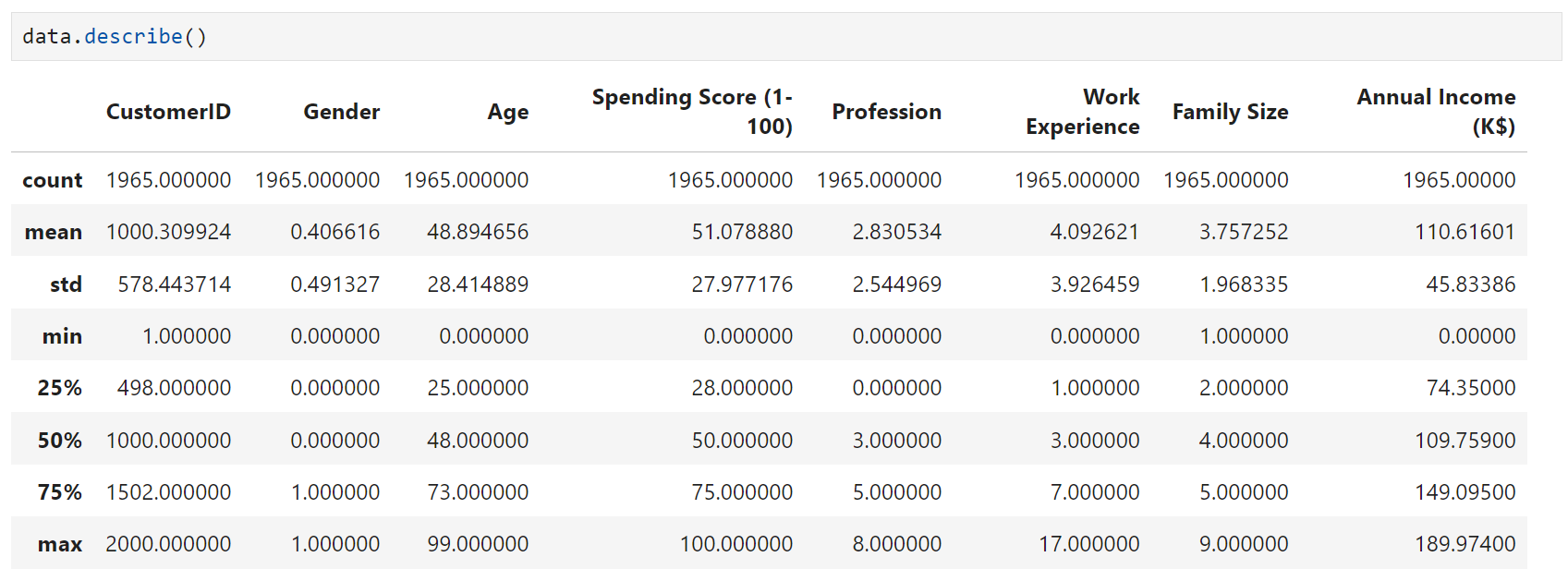


Рисунок 2. Данные датафрейма после подготовки

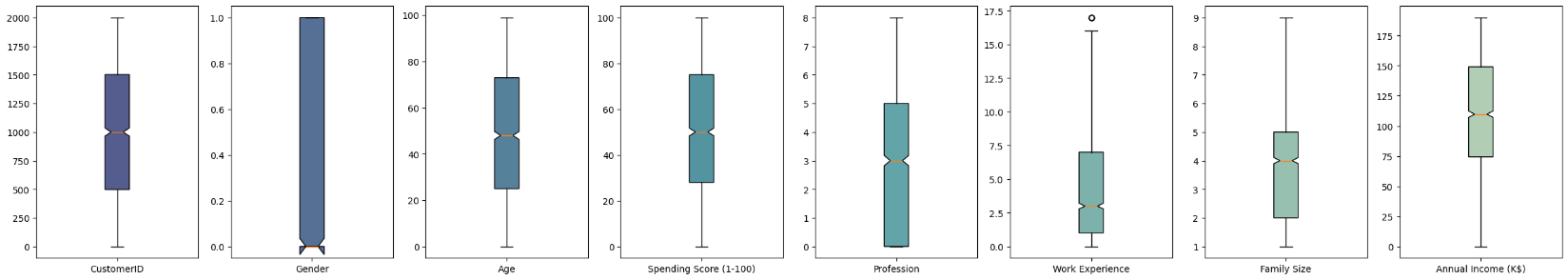
**2. Предварительный анализ данных**

В предварительном анализе данные было произведено вычисление описательных статистик по колонкам (среднее, моду, медиану, стандартное отклонение, квартили):



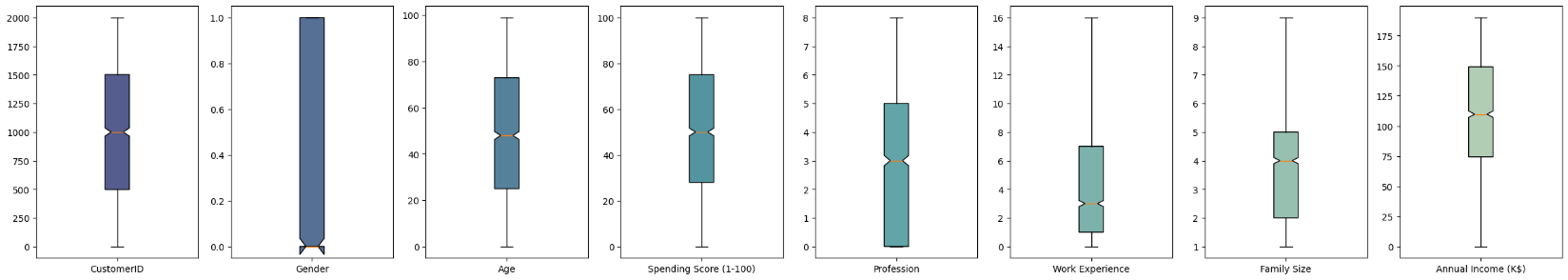
**Рисунок 3.** Описательные статистики

Далее проверка числовых колонок на наличие выбросов, для этого можно использовать диаграмму «ящик с усами» (boxplot), у которого был выявлен выброс на «Work experience»:



**Рисунок 4.** Ящики с усами

От выброса можно избавиться заменой на 99-ый процентиль:



**Рисунок 5**. Ящики с усами без выбросов

После произведена проверка колонок на нормальность распределения с помощью:

* Гистограммы рассеяния
* статического теста

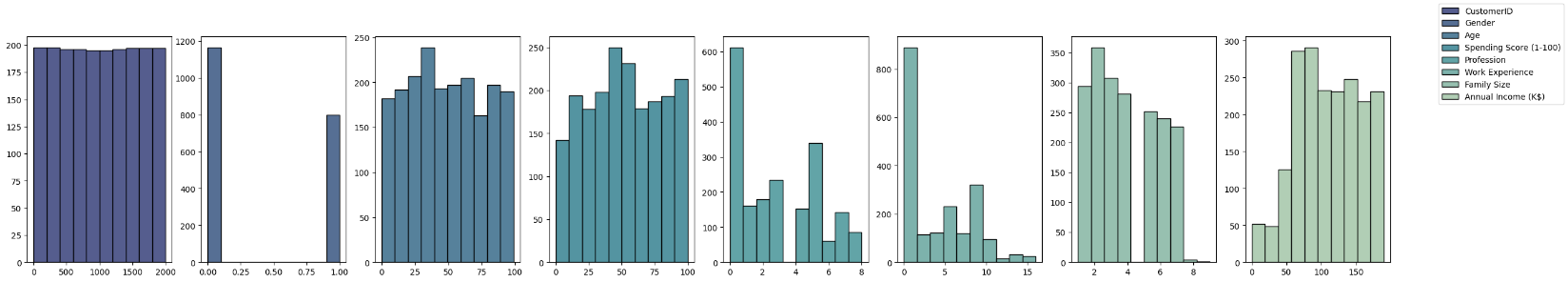


Рисунок 6. Гистограммы

Но ни на одной гистограмме не наблюдается нормального распределения, а для статистического тест была следующая нулевой гипотезой: **«Распределение столбцов является нормальным»:**



Рисунок 7. Статистический тест

Из теста видно, что p-критерий для каждого из столбцов стремится к нулю, а значит: ***«Нулевая гипотеза не выполняется, столбцы не имеют нормального распределения».***

Была построена корреляционная матрица с последующим отбором признаков для кластеризации – такие, которые как можно меньше зависят друг от друга.

Рисунок 8. Корреляционная матрица с признаками для кластеризации

***Результатом*** **предварительного анализа данных** должна быть выдвинутая гипотеза о том, что данные можно разделить на некоторое количество кластеров (предположить, каково это количество), используя для группировки отобранные признаки и произведём нормализацию выбранных для кластеризации признаки:

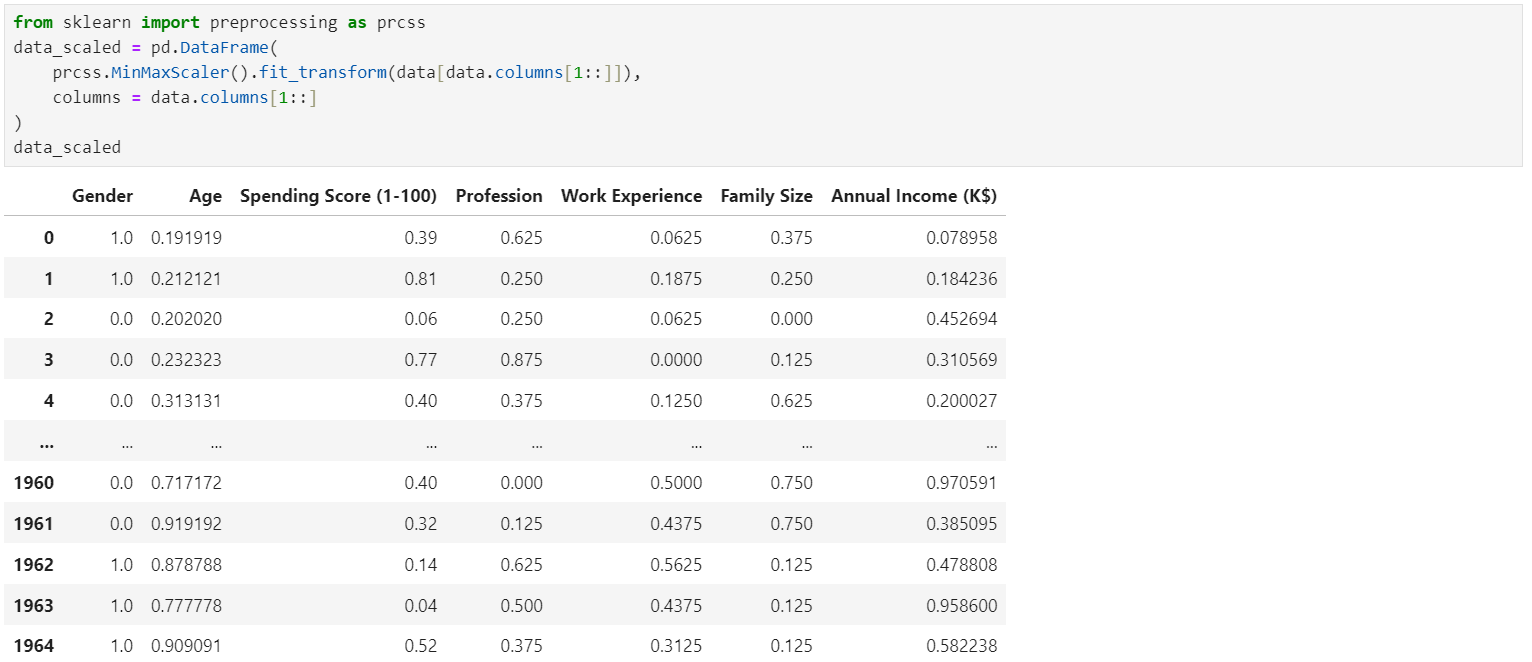


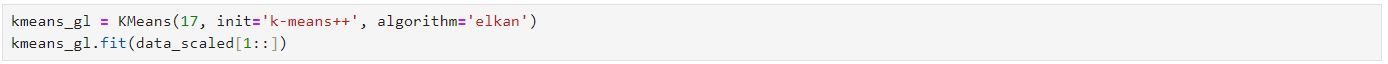
Рисунок 9. Нормализация для кластеризации признаки

Из полученных данных можно сделать вывод, что данные можно разделить на ***2, 9 и 17 кластеров***, используя признаки: Gender, Age, Spending Score (1-100), Profession, Work Experience, Family Size, Annual Income (K$). 2 по числу полов, 9 по числу профессий и 17 по опыту работы.

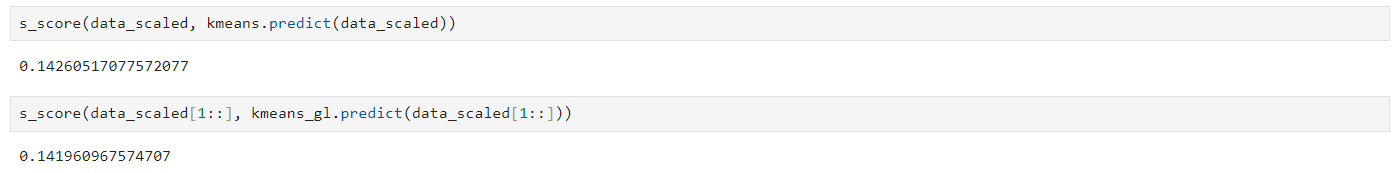
**3. Кластерный анализ**

Осуществлена кластеризация данных методом кластеризации K-means с предположенным ранее числом кластеров (17 по опыту работы).



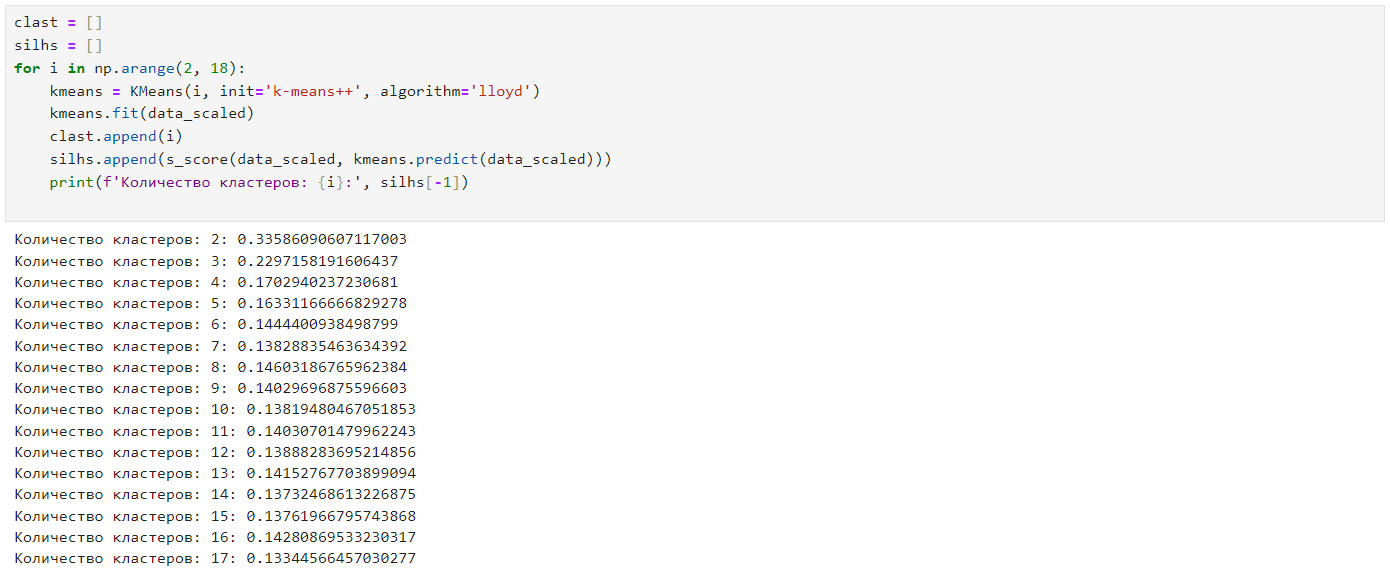
Рисунок 10. Метод К-means

Вычислим для него силуэт:

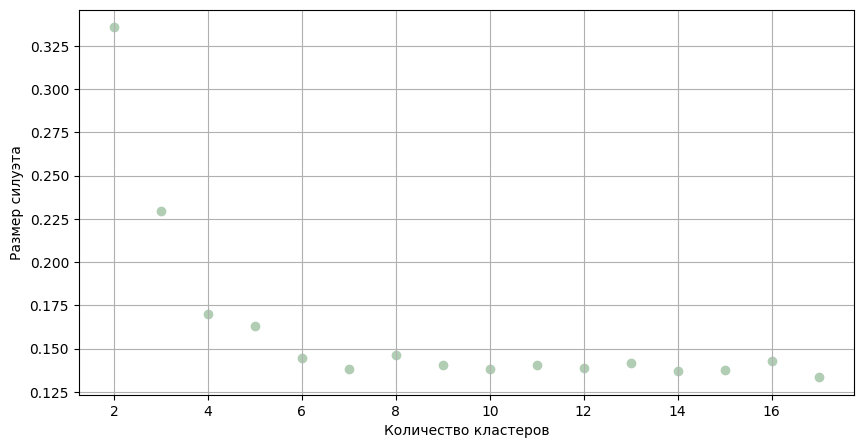


*Рисунок 11. Метод К-means*

Найдем количество кластеров, при котором значение силуэта максимально:

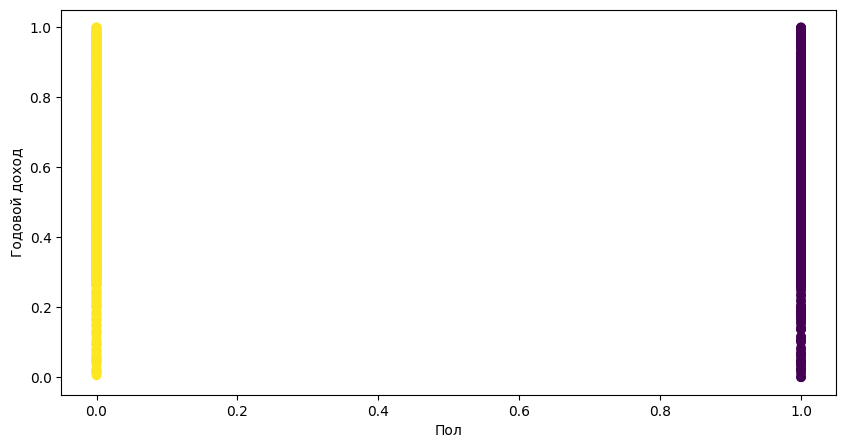


*Рисунок 12. Значение силуэта от количества кластеров*

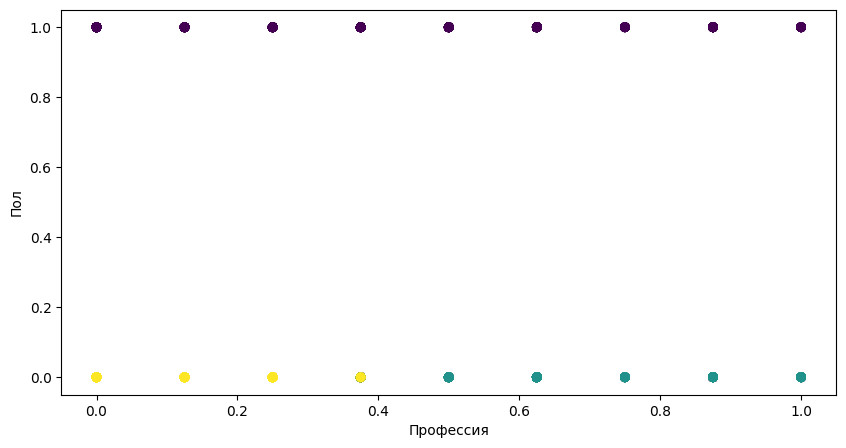


*Рисунок 13. Зависимость силуэта от количества кластеров*

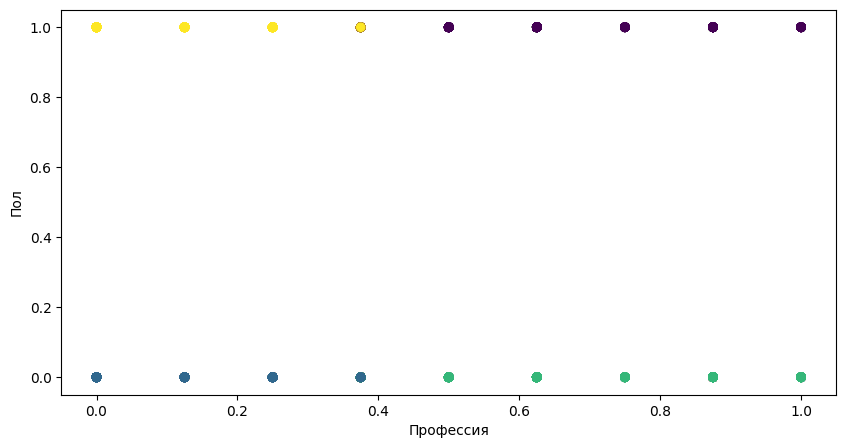
Рассмотрим эти кластеры:



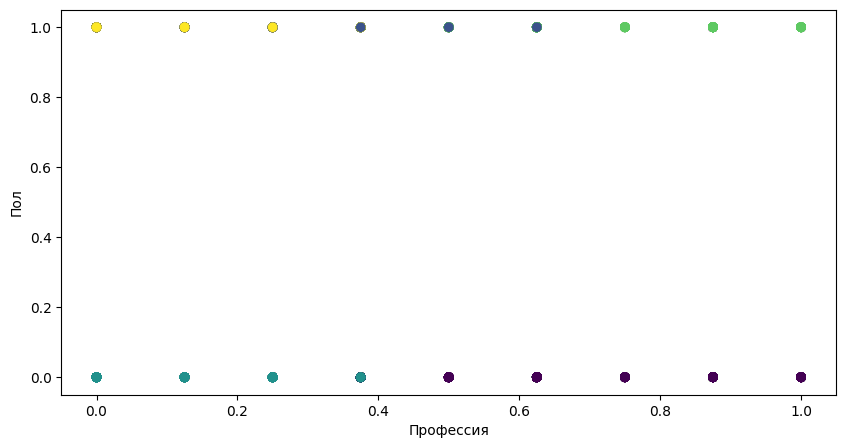
*Рисунок 14. Количество кластеров 2*



*Рисунок 15. Количество кластеров 3*

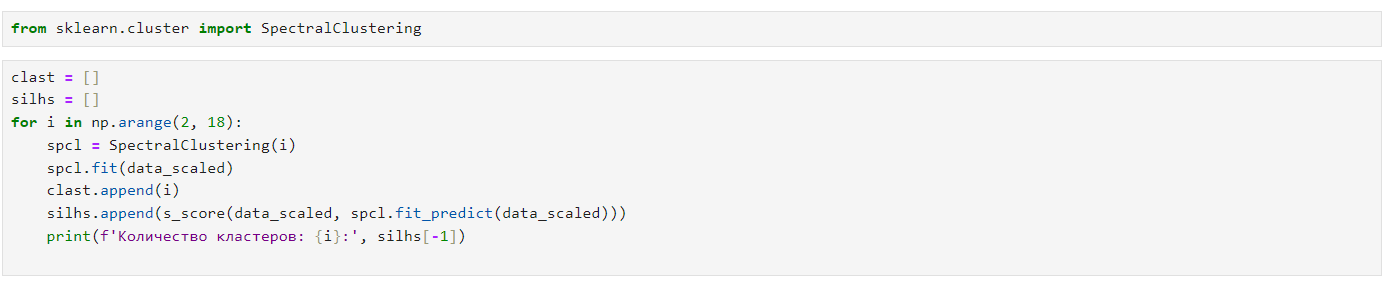


*Рисунок 16. Количество кластеров 4*

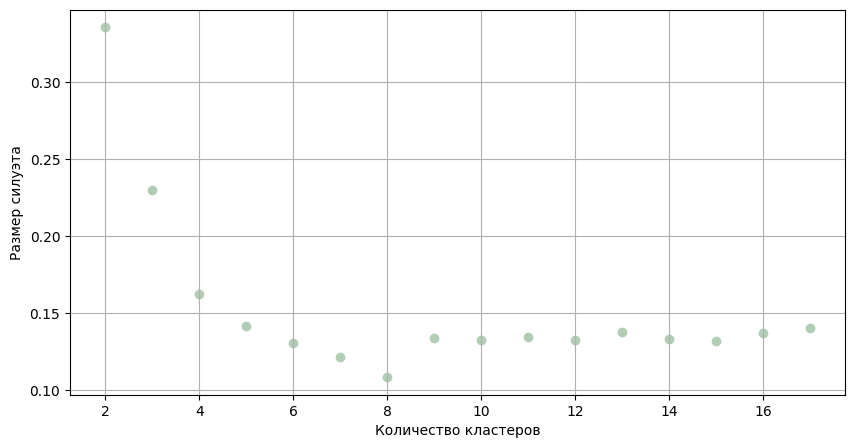


*Рисунок 17. Количество кластеров 5*

Также осуществлена кластеризация данных методом кластеризации SpectralClustering:

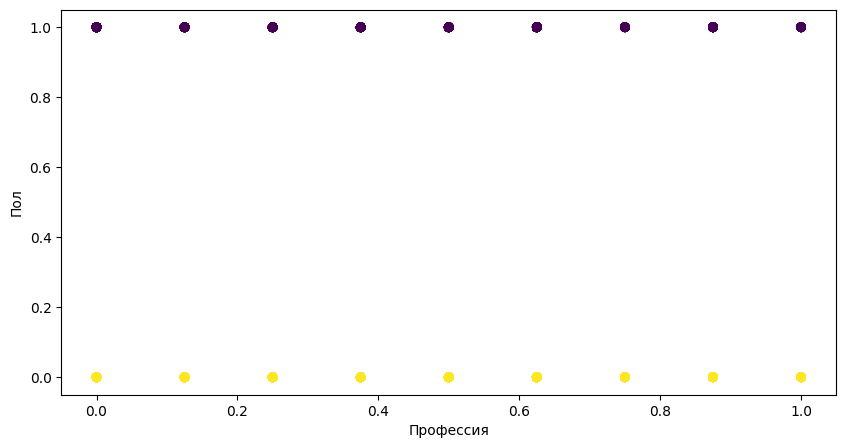


*Рисунок 18. Метод кластеризации SpectralClustering и значение силуэта от количества кластеров*

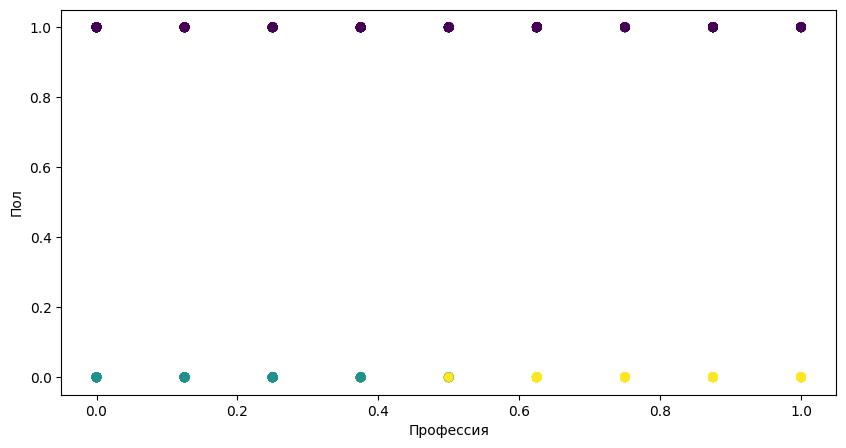


*Рисунок 19. Зависимость силуэте от количества кластеров*

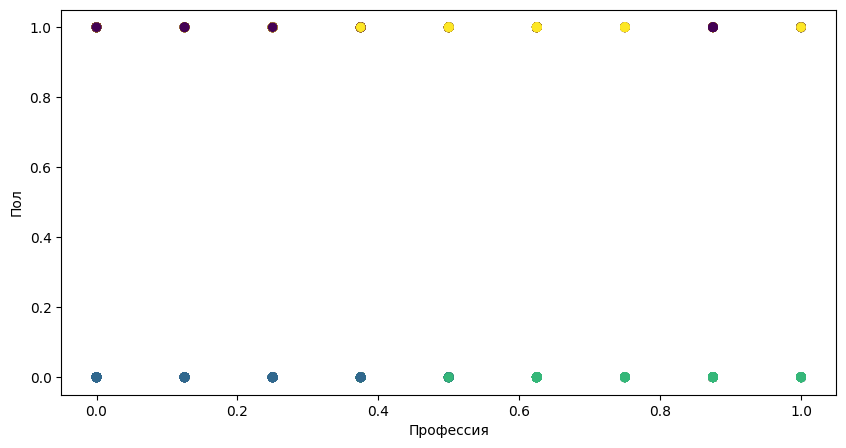
Рассмотрим эти кластеры:



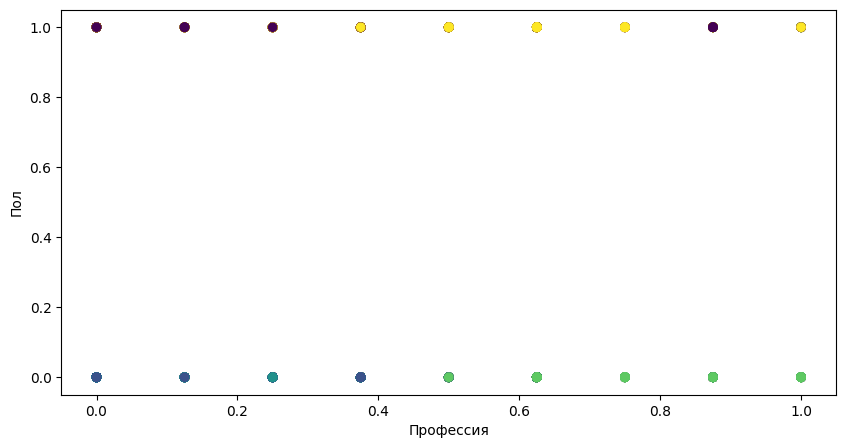
*Рисунок 20. Количество кластеров 2*



*Рисунок 21. Количество кластеров 3*

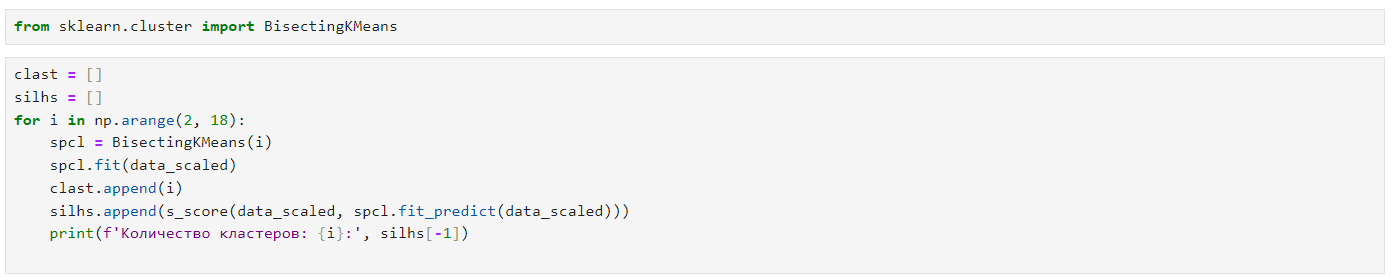


*Рисунок 22. Количество кластеров 4*

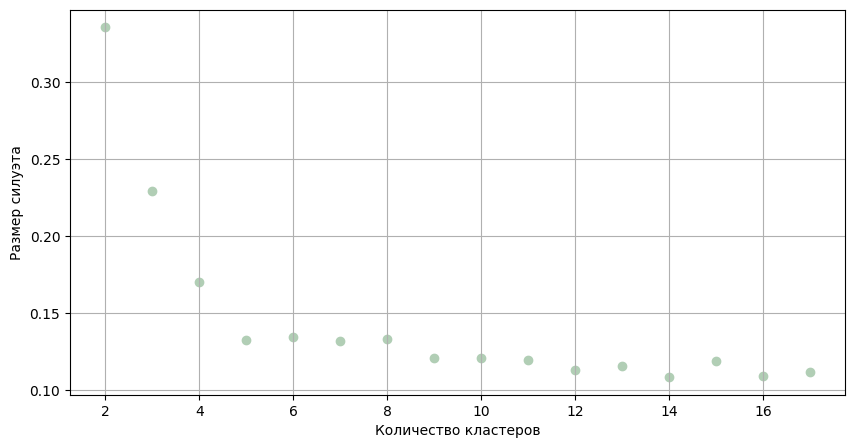


*Рисунок 23. Количество кластеров 5*

Осуществлена кластеризация данных методом кластеризации BisectingKMeans:

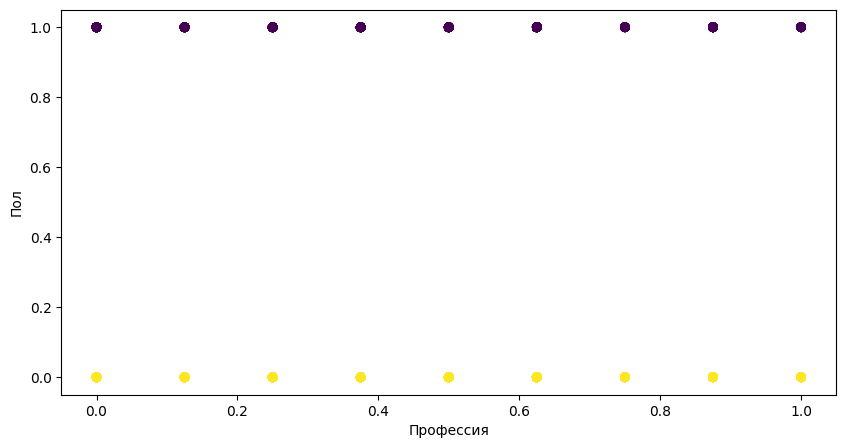


*Рисунок 24. Метод кластеризации BisectingKMeans и значение силуэта от количества кластеров*

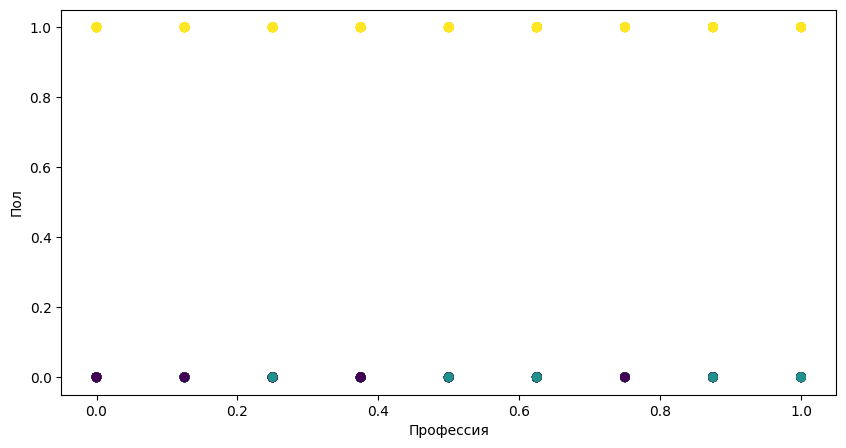


*Рисунок 25. Зависимость силуэта от количества кластеров*

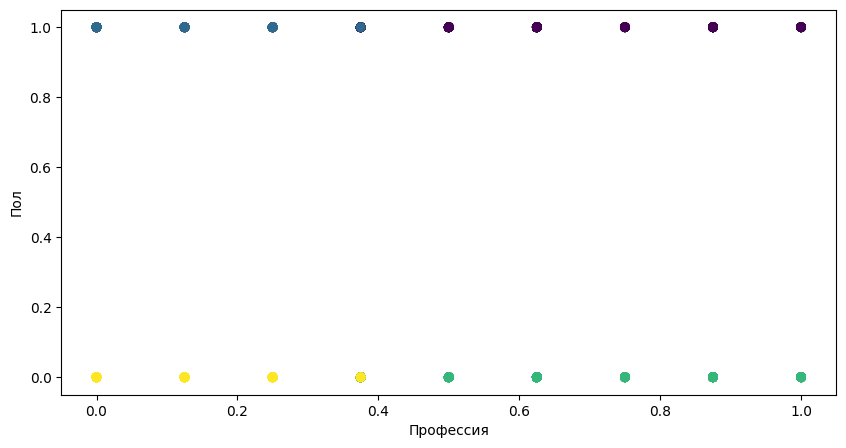
Рассмотрим эти кластеры:



*Рисунок 26. Количество кластеров 2*



*Рисунок 27. Количество кластеров 3*



*Рисунок 28. Количество кластеров 4*

**Вывод**

Были проанализировать данные о клиентах магазина:

1. Выявлены зависимости по следующим факторным переменным:

* **Gender** (Пол),
* **Age** (Возраст),
* **Spending Score (1-100)** (Оценка расходов (1-100)),
* **Profession** (Профессия),
* **Work Experience** (Опыт работы),
* **Family Size** (Размер семьи),
* **Annual Income (K$)** (Годовой доход (K$)).

1. Данные были разбиты на кластеры (рис. 14–17, 20-23, 26-28).
2. Оценено качество кластеризации для методов **K-means**, **SpectralClustering**, **BisectingKMeans** (рис. 13, 19, 25), и для всех методов максимальным размером силуэт (равен **0.33586090607117003**) был при количестве кластеров равным **2**.