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

**Цель**: 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.

**Порядок выполнения проекта**

**Шаг 1.** Выполнить подготовку данных к анализу.

1.1. Загрузить данные в датафрейм.

1.2. Убедиться, что все количественные столбцы имеют числовой тип. Если это не так, выполнить преобразование типа столбца к числовому.

1.3. Проверить данные на пропуски. Если в столбцах или строках очень много пропусков (порог выбрать самостоятельно), такие признаки и объекты для анализа малоинформативны, их следует удалить. Оставшиеся пропуски необходимо заполнить любым подходящим методом (средним, модой, медианой, линейной интерполяцией).

***Результатом*** ***первого шага*** должен быть готовый к анализу набор данных в виде датафрейма.

**Шаг 2.** Выполнить предварительный анализ данных.

2.1. Вычислить описательные статистики по колонкам (среднее, моду, медиану, стандартное отклонение, квартили).

2.2. Проверить числовые колонки на наличие выбросов, для этого можно использовать диаграмму «ящик с усами» (boxplot). Если выбросов мало, следует их сгладить любым методом (заменой на 1-ый или 99-ый процентиль, среднее, моду, медиану, линейной интерполяцией). Если выбросов много, и они не сконцентрированы в узком диапазоне значений, а рассеяны по широкому диапазону, тогда можно ничего с ними не делать.

2.3. Проверить колонки на нормальность распределения двумя способами:

1) построить гистограмму и сделать предположение о том, являются ли данные нормально распределенными;

2) выполнить статистический тест на нормальность (уровень значимости выбрать самостоятельно) и убедиться, что выдвинутое ранее предположение о нормальности верно (ошибочно).

2.4. Выполнить корреляционный анализ данных с помощью матрицы корреляции. В случае нормальности всех колонок использовать корреляцию Пирсона, в противном случае – ранговые коэффициенты. Матрицу корреляции отобразить с помощью диаграммы «тепловая карта» (heatmap).

2.5. Используя результаты корреляционного анализа, отобрать признаки для кластеризации – такие, которые как можно меньше зависят друг от друга (отбросить наиболее взаимно скоррелированные признаки).

2.6. При наличии категориальных признаков, которые могут влиять на группировку объектов, закодировать их значения числами и тоже использовать для анализа.

2.7. Выполнить нормализацию отобранных столбцов путем приведения их значений к диапазону от 0 до 1.

***Результатом*** ***второго шага*** должна быть выдвинутая гипотеза о том, что данные можно разделить на некоторое количество кластеров (предположить, каково это количество), используя для группировки отобранные признаки.

**Шаг 3.** Выполнить кластерный анализ.

3.1. Осуществить кластеризацию данных, используя отобранные ранее столбцы, применив какой-либо алгоритм кластеризации, требующий указать число кластеров (например, k-means). В качестве числа кластеров взять предполагаемое ранее.

3.2. Вычислить показатель качества кластеризации силуэт.

3.3. Перебрать несколько значений числа кластеров, для каждого выполнить кластеризацию и вычислить силуэт. Найти такое число кластеров, при котором величина силуэта максимальна.

3.4. Используя два любых числовых факторных признака в качестве координат, построить диаграмму рассеяния, при этом объекты, отнесенные к разным кластерам, раскрасить разными цветами.

***Результатом третьего шага*** должны быть выводы о том, насколько хорошо удалось кластеризовать данные (подтвердилась ли выдвинутая на втором шаге гипотеза), какие факторные переменные оказывают наибольшее влияние на группировку объектов, каково объяснение этого, как можно описать полученное разбиение данных на кластеры, в чем может быть отличие объектов из разных кластеров, каким характеристикам объекта из жизни могут соответствовать полученные кластеры, как можно применить на практике полученные результаты.

**Шаг 4.** Представить отчет по проекту в виде документа Word. Отчет должен включать подробное описание всех выполненных действий с анализом, предположениями о зависимостях в данных и т.д. Обязательно приведение кода на языке Python и результатов его выполнения. Также к отчету должна быть приложена ссылка на блокнот с расчетами в Google Colab (или блокнот Jupyter Notebook, или файл с исходными кодами – если использована иная среда разработки на Python).

**Шаг 5.** Подготовить презентацию по полученным результатам (в формате PowerPoint). Презентация должна содержать краткое описание хода выполнения проекта (описание данных, их подготовка, предварительный анализ, выдвинутые гипотезы, кластерный анализ) и выводы, сделанные по итогам анализа (подтвердились ли гипотезы?). Количество слайдов – не более 20.

**Файлы:**

Customers.csv

**Данные:**

CustomerID - Пользовательский ИД

Gender - Пол

Age - Возраст

Annual Income ($)- Годовой доход

Spending Score (1-100) - Spending Score (оценка, присваиваемая магазином на основе поведения покупателей и характера расходов)

Profession - Профессия

Work Experience - Опыт работы - лет

Family Size - Семейный размер