Задание

Импортируйте датасет retail\_data\_transactions.csv

1. Проведите базовый анализ данных: какие данные, какие столбцы и что они значат.
2. Action-based аналитика. Проведите статистический тест (например, t-тест) для оценки какого-либо параметра (количество покупок и траты).
3. Decision-based аналитика. Примените один из методов машинного обучения (например, кластеризацию) для сегментации клиентов или прогнозирования продаж.
4. Опишите, какие проблемы и трудности возникли в процессе работы. Поделитесь выводами и замечаниями по анализу данных.
5. Подготовьте небольшой отчёт в формате PPTX, PDF или любом другом, который демонстрирует ваши находки и результаты анализа.

# Решение

1. Исходные данные загрузили на Google-диск.

from google.colab import drive

drive.mount('/content/drive')

1. Загруженные данные представлены 125000 сущностями (записями). Всего в наборе данных 3 столбца: customer\_id (идентификационный номер покупателя от 0 до 124999), trans\_date (дата транзакции в формате DD-Mmm-YY) и tran\_amount (сумма транзакции). В столбцах отсутствуют пропуски.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 125000 entries, 0 to 124999

Data columns (total 3 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 customer\_id 125000 non-null object

1 trans\_date 125000 non-null object

2 tran\_amount 125000 non-null int64

dtypes: int64(1), object(2)

memory usage: 2.9+ MB.

1. Группировка по датам показала наличие 1401 различных дат.

Name: trans\_date, Length: 1401, dtype: int64.

При группировке по покупателям выявлено всего 6889 групп. Т.е. некоторые покупатели делали несколько покупок.

Name: customer\_id, Length: 6889, dtype: int64.

При этом средние расходы покупателей колеблются в пределах 26.7-87.4, а количество транзакций на одного покупателя от 4-39.

1. По результатам -теста на зависимость trans\_date и tran\_amount p-значение составило 0.44149481628064347, что больше 0,05, следовательно, данные столбцы независимы.

По результатам -теста на зависимость customer\_id и tran\_amount p-значение составило 0.0, что менее 0,05, следовательно, customer\_id и tran\_amount зависимы.

1. Визуальный анализ данных показал, что распределение сумм похоже на нормальное (рисунок 1).

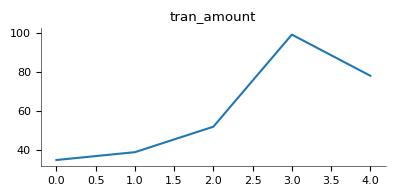


Рисунок 1 – Распределение сумм.

В остальном визуализация в частности 2d не была полезна для анализа данных.

1. t-test показал, что p-value=0.0<0.05, следовательно, количество покупок не связано со средними затратами покупателя.
2. Визуализация средних затрат и количества покупок позволила предположить сегментацию покупателей на две группы (рисунок 2).

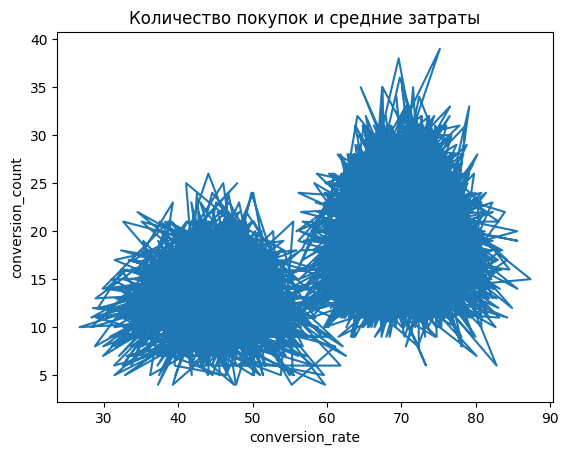


Рисунок 2 – Средние затраты покупателей и количество покупок.

1. Для кластеризации данных применим метод k-mean в предположении, что имеется два кластера. Визуализация результатов модели кластеризации подтвердила предположение о двух кластерах (рисунок 3).

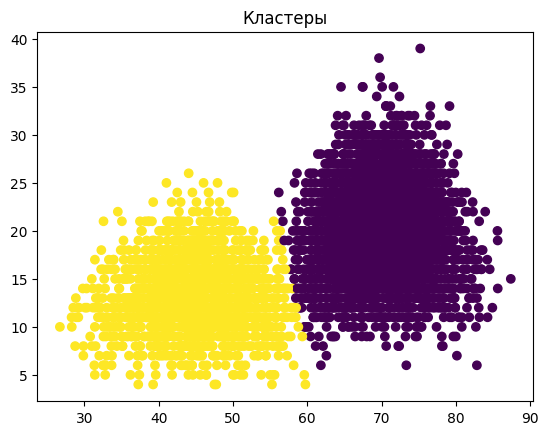


Рисунок 3 – Кластеризация датасета.

На основе полученной модели можно сделать выводы о том, что покупатели совершающие меньшее количество покупок в среднем затрачивают меньше средств. Такое разбиение можно объяснить возможной гомогенностью товаров со слабой дифференциацией цен.

**Выводы:** вPython имеется большое количество библиотек и моделей для анализа данных. Самое сложное выбрать подходящую модель машинного обучения и определить все параметры модели. В зависимости от модели параметрами могут выступать разные типы данных, поэтому требуется преобразование данных в ходе работы. Несмотря на сложность изучения, методы машинного обучения Python дают исследователям мощный инструмент для моделирования и исследования данных.