Процесс анализа данных.

Шаг 1. Откройте таблицу и изучите общую информацию о данных
Библиотеки: pandas, requests, bs4, sqlalchemy

1. при необходимости напишите парсер для сбора данных со сторонних ресурсов;
2. сделайте необходимые SQL запросы к базе данных;

Шаг 2. Предобработка данных
Библиотеки: nltk.stem, pymystem3, collections (эти три библиотеки нужны для лемматизации), datetime

1. определите и заполните пропущенные значения:
   - опишите, какие пропущенные значения вы обнаружили;
   - приведите возможные причины появления пропусков в данных;
   - объясните, по какому принципу заполнены пропуски;
2. замените вещественный тип данных на целочисленный:
   - поясните, как выбирали метод для изменения типа данных;
3. удалите дубликаты:
   - поясните, как выбирали метод для поиска и удаления дубликатов в данных;
   - текстовые данные приведите к нижнему регистру
   - приведите возможные причины появления дубликатов;
4. выделите леммы в значениях столбца:
   - опишите, как вы проводили лемматизацию;
5. категоризируйте данные:
   - перечислите, какие «словари» вы выделили для этого набора данных, и объясните, почему.

Шаг 3. Добавьте в таблицу необходимые для расчетов данные.

1. дополнительные метрики, необходимые для анализа:
   - средний чек;
   - кумулятивная сумма;
   - кумулятивный средний чек;
   - соотношения разных метрик;
2. дополнительные параметры даты и времени:
   - минуты, часы, дни недели, дни месяца, недели, месяца, года, в том числе год_недели, год_месяца;
   - любые из вышеперечисленных параметров для первого события у каждого пользователя;
3. посчитать время жизни в необходимых параметрах даты и времени;

Шаг 4. Проведите исследовательский анализ
Библиотеки: matplotlib.pyplot

1. изучите необходимые параметры, постройте гистограммы;
2. посчитайте среднее, медиану, процентили, дисперсию, стандартное отклонение, постройте ящик с усами;
3. уберите редкие и выбивающиеся значения;
4. изучите зависимость разных метрик друг от друга;

Шаг 5. Изучите воронку событий

1. посмотрите, какие события есть в логах, как часто они встречаются, отсортируйте события по частоте;
2. посчитайте, сколько пользователей совершали каждое из этих событий;
3. по воронке событий посчитайте, какая доля пользователей проходит на следующий шаг воронки (от числа пользователей на предыдущем);
4. проанализируйте, где и сколько теряем пользователей;

Шаг 6. Проведите статистический анализ
Библиотеки: math, numpy, scipy

1. определите, какие нулевые гипотезы вы будете проверять;
2. поясните, как вы формулировали нулевую и альтернативную гипотезы;
3. задайте пороговое значение alpha;
4. поясните, какой критерий использовали для проверки гипотез и почему;

Шаг 7. Постройте отчёты
Библиотеки: matplotlib, seaborn, plotly

1. отобразите на графиках, как разные метрики отличаются по разным параметрам;
2. используйте разные виды графиков:
   - столбчатые диаграмы;
   - круговые диаграмы;
   - графики;

Шаг 8. Проведите приоритизацию гипотез

1. примените фреймворк ICE для приоритизации гипотез;
2. примените фреймворк RICE для приоритизации гипотез;
3. укажите, как изменилась приоритизация гипотез при применении RICE вместо ICE, объясните, почему так произошло.

Шаг 9. Проведите анализ A/A, A/B-теста

1. постройте графики по кумулятивным метрикам;
2. постройте графики относительного изменения кумулятивных метрик;
3. постройте график кумулятивной конверсии по группам;
4. постройте график относительного изменения кумулятивной конверсии;
5. выявите «всплески» данных;
6. посчитайте статистическую значимость различий в конверсии между группами по «сырым» данным;
7. посчитайте статистическую значимость различий в конверсии между группами по «очищенным» данным;
8. примите решение по результатам теста и объясните его;

Шаг 10. Постройте модель прогнозирования
Библиотеки: sklearn

1. разбейте данные на обучающую и валидационную выборку;
2. обучите модель на train-выборке разными способами;

3. оцените метрики accuracy, precision и recall для выбранных моделей на валидационной выборке, сравните по ним модели;

Шаг 11. Сделайте кластеризацию пользователей
Библиотеки: scipy, joblib

1. стандартизируйте данные;
2. постройте матрицу расстояний на стандартизованной матрице признаков и нарисуйте дендрограмму.;
3. обучите модель кластеризации на основании алгоритма K-Means и спрогнозируйте кластеры клиентов;

Шаг 12. Напишите общий вывод

Шаг 13. Подготовьте презентацию исследования

Шаг 14. Подготовьте дашборд
Библиотеки: dash, dash_core_components, dash_html_components,

1. составьте техническое задание;
2. спроектируйте структуру агрегирующих таблиц;
3. создайте агрегирующие таблицы;
4. создайте пайплайн;
5. создайте дашборд;