27/12/2022

Проект №5
“Такси Нью-Йорка”

автор: Еропова Т.
тел. 8 900 000 001, email: sample@email.ru

Цели исследования и источники данных:

Основные **цели** исследования:

- рассчитать процент поездок по каждому дню в зависимости от числа пассажиров (в разрез 1, 2, 3, 4 и более человек)
- проанализировать зависимость размера чаевых от числа пассажиров и дальности поездки

Источниками для исследования послужили “сырые” данные о использовании такси.

Для проведения работы использован ЯП Python с применением библиотек pandas, numpy, matplotlib и seaborn. В качестве среды разработки Jupyter-notebook.

План реализации

1. Загрузка первичных “сырых” данных
2. Проверка качества “сырых” данных, обработка согласно бизнес-требованиям
3. Построение итоговой таблицы с полями по бизнес-требованиям
4. Выгрузка итоговых данных в parquet-файл
5. Проведение анализа зависимости показателя “чаевые” с числом пассажиров и дальности поездки

Описание итогового parquet-файла

'date_trip' - дата начала поездки (посадки пассажира), **тип данных** дата,
'percentage_1p' - доля поездок с 1 пассажиром по каждому дню, **тип данных** float,
'percentage_2p'- доля поездок с 2 пассажирами по каждому дню, **тип данных** float,
'percentage_3p'- доля поездок с 3 пассажирами по каждому дню, **тип данных** float,
'percentage_4p_plus'- доля поездок с 4 и более пассажирами по каждому дню, **тип данных** float,
'percentage_zero'- доля поездок без пассажиров по каждому дню, **тип данных** float,
'min_percentage_1p' -минимальное значение стоимости поездки с 1 пассажиром по каждому дню, **тип данных** float,
'min_percentage_2p' - минимальное значение стоимости поездки с 2 пассажирами по каждому дню, **тип данных** float,
'min_percentage_3p' - минимальное значение стоимости поездки с 3 пассажирами по каждому дню, **тип данных** float,
'min_percentage_4p_plus' - минимальное значение стоимости поездки с 4 и более пассажирами по каждому дню, **тип данных** float,
'min_percentage_zero' -минимальное значение стоимости поездки без пассажиров по каждому дню, **тип данных** float,
'max_percentage_1p' - максимальное значение стоимости поездки с 1 пассажиром по каждому дню, **тип данных** float,
'max_percentage_2p' - максимальное значение стоимости поездки с 2 пассажирами по каждому дню, **тип данных** float,
'max_percentage_3p' -максимальное значение стоимости поездки с 3 пассажирами по каждому дню, **тип данных** float,
'max_percentage_4p_plus' -максимальное значение стоимости поездки с 4 и более пассажирами по каждому дню, **тип данных** float,
'max_percentage_zero' - максимальное значение стоимости поездки без пассажиров по каждому дню, **тип данных** float,

Выводы

1. В ходе работы проведена загрузка и обработка “сырых” данных (6405008 строк, 22 столбца)
2. Итоговый файл содержит 31 строку и 16 столбцов.
3. В ходе обработки и подготовке последующей выгрузки учтены все бизнес задачи по содержанию и формату данных.
4. В аналитической части проведен исследования по выявлению зависимости размера чаевых от количества пассажиров и дальности поездки.
5. Было выяснено, что максимальный размер чаевых зафиксирован:
    - в поездках 1-2 пассажира,
    - на расстояние 197-230 км/миль.