Проект № 5

Такси

Выполнил:
Макутин Евгений

Проект № 5 - Такси

Описание:

В проекте проанализированы данные поездок такси, получен **"parquet"** в виде результата работы

Название и общее описание проекта

Итоговый проект №5 для курса Data Engineer на тему: Клиенты и счета (Такси)

Цель проекта: на основе данных поездок Taxi г. Нью-Йорк построить таблицу-отчет (далее "parquet") со следующей информацией для каждого дня:

- процент поездок по количеству человек в машине (5 групп пассажиров)
- Самая дорогая поездка для каждой группы пассажиров
- Самая дешевая поездка для каждой группы пассажиров

Доп. задача: Провести аналитику и построить график на тему "как пройденное расстояние и количество пассажиров влияет на размер чаевых"

План реализации

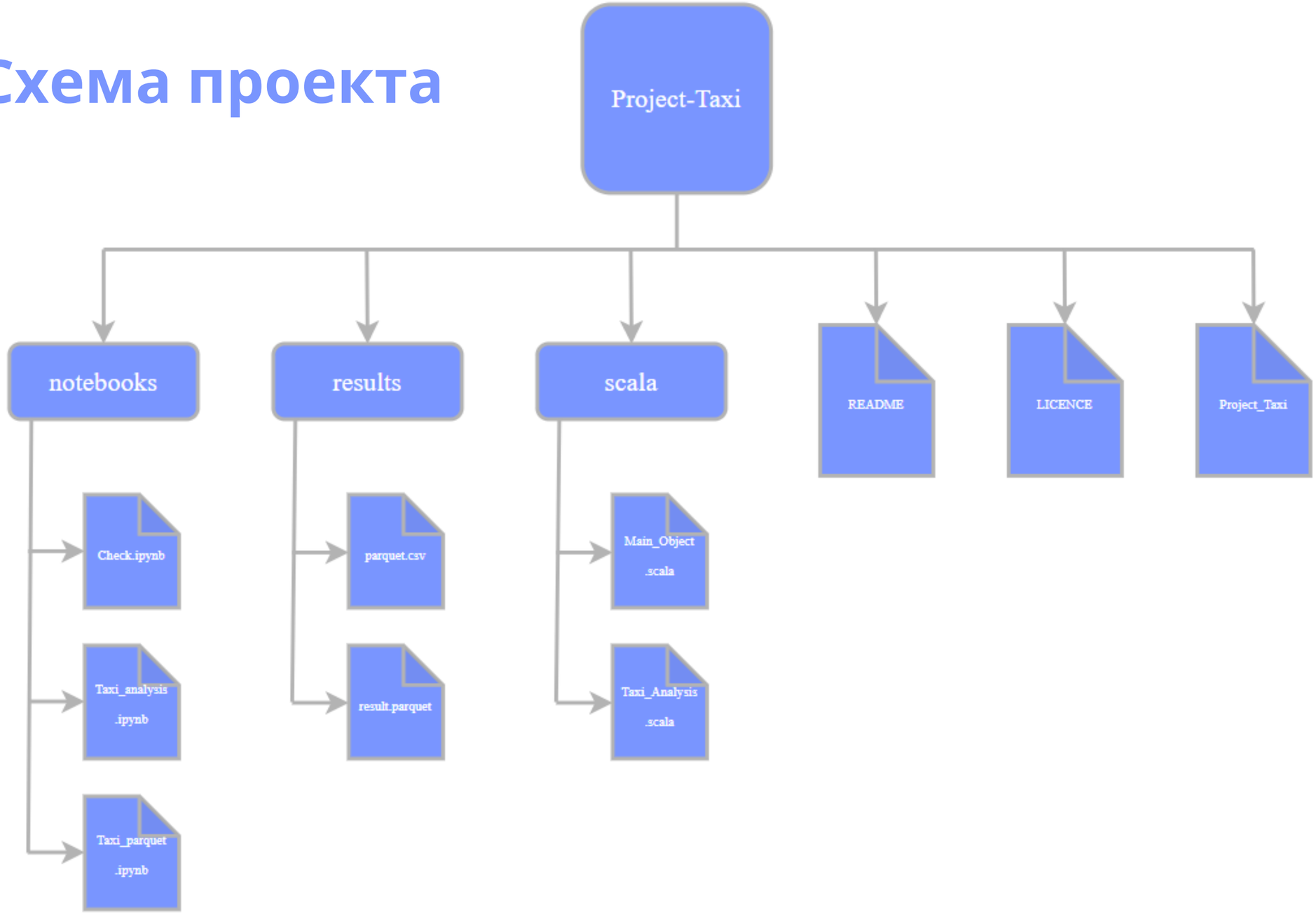
- Скачал данные csv формата
- Написал код на Scala в IntelliJ IDEA, создал два объекта, один решал основную задачу: parquet, второй решал доп. задачу: Аналитика
- Красиво оформил в формате ipynb (Ноутбуки: осн. задача + доп. задача + проверка результата)
- Сохранил parquet в удобном формате
- Оформил описание проекта в README.md
- Залил проект на github

Используемые технологии

- **Git** - для удобства разработки и хранения проекта
- **Scala** v. 2.11.8 (библиотека Vegas v 0.3.11 для анализа данных) - так как на данном проекте разрешено писать только на Scala (Условие организаторов)
- **Spark** v. 2.4.0 - мощный инструмент для обработки больших данных (версия выбрана исходя из совместимости с версией Scala)
- **Docker** - на нем открыл образ JupyterLab + Spark со всеми предустановками, изначально использовал для доп. задачи, чтобы красиво отображались графики
- **Canva** для выполнения презентации

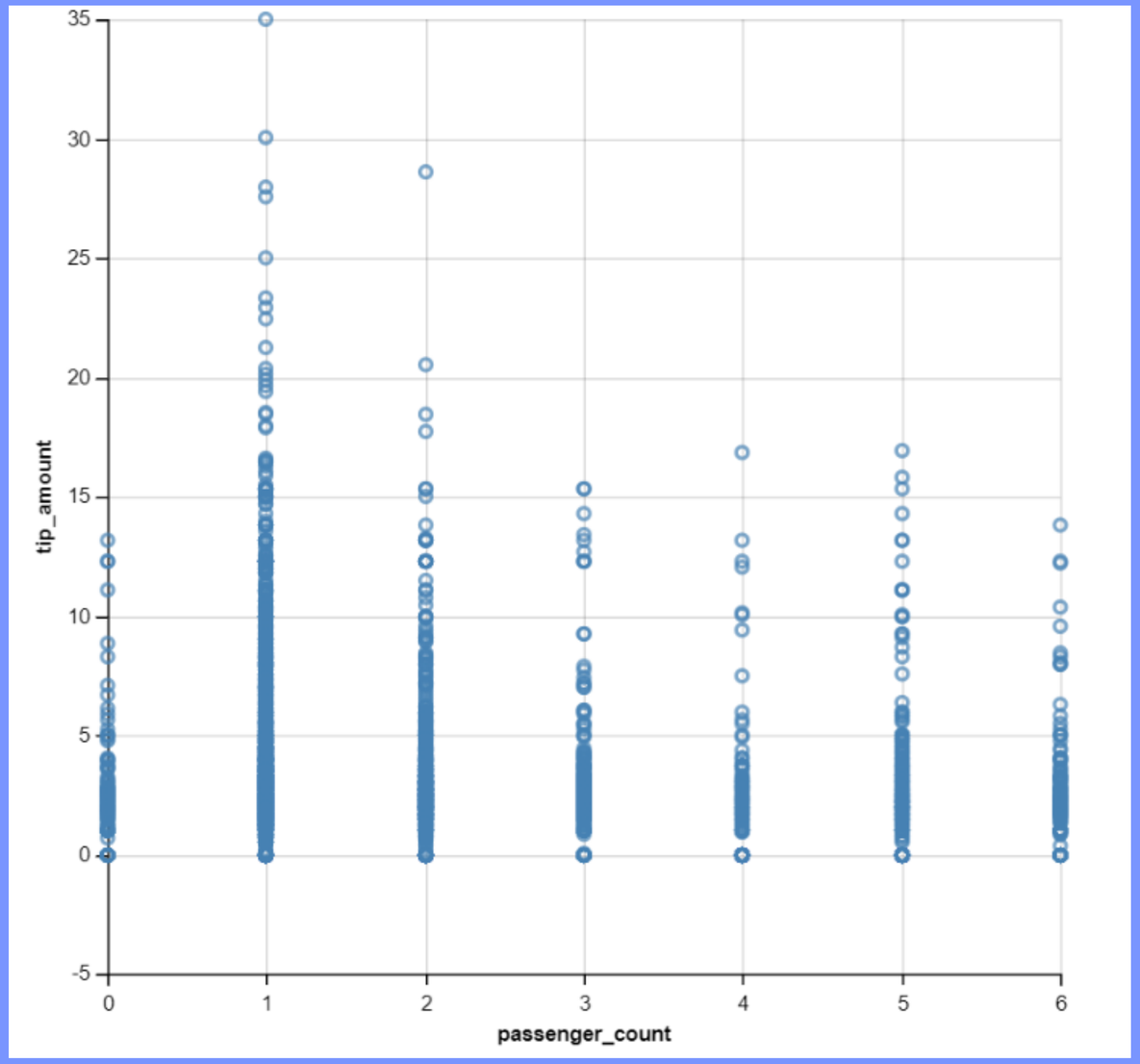
- **Draw.io -** для составления схемы проекта

Схема проекта

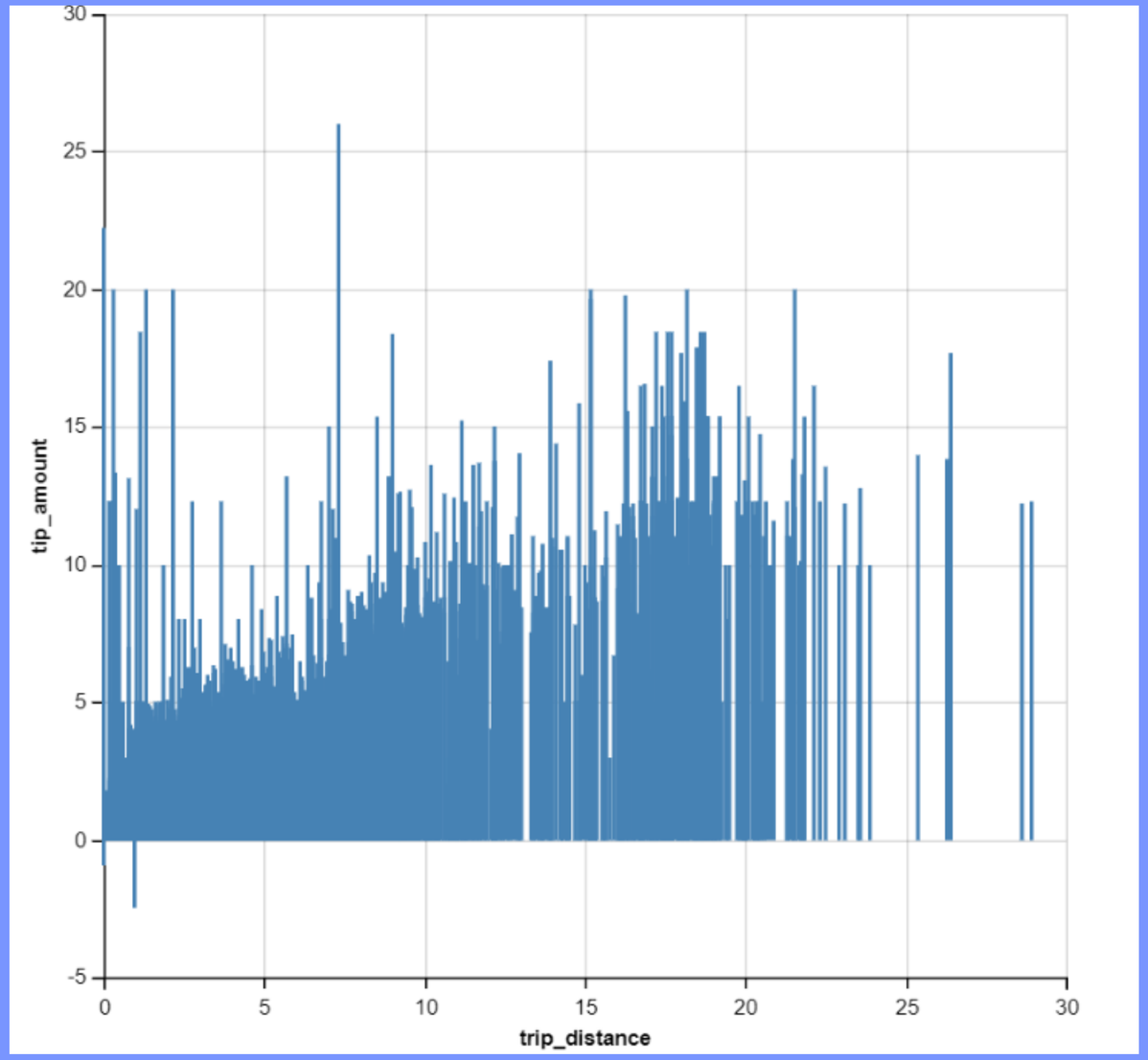


Графики

Зависимость размера чаевых от кол-ва пассажиров



Зависимость размера чаевых от расстояния поездки



Выводы по графикам

Из графика зависимости размера чаевых от кол-ва пассажиров следует, что:

- Самые большие чаевые давала первая группа пассажиров (1 человек)
- Чаще всего чаевые давала первая группа пассажиров, реже всего 3 группа и 4 группа (именно где 4 пассажира)
- В среднем можно предположить, что первая и вторая группа пассажиров самые благоприятные в плане чаевых

Из графика зависимости размера чаевых от пройденного расстояния следует, что:

- Не учитывая выброс (чаевые ~ 255 при расстоянии 0), то наибольшие чаевые получали при расстоянии 16-19 миль
- Минимальные чаевые при расстоянии 2-4 мили
- В итоге можно предположить, что лучше брать заказ на расстояние 7-10 миль или 15-20 миль, если нужны хорошие чаевые

Результаты разработки

- Parquet, который содержит результаты по всем группам пассажиров
- Графики зависимости размера чаевых от кол-ва пассажиров и пройденного расстояния, а также вывод по графикам
- Схема проекта, выполненная в **Draw.io**
- Интересный проект в github

Выводы

- Узнал много полезных вещей по работе со Scala и Spark
- Познакомился, как работать с настоящими большими данными (6400000 записей)
- Закрепил навыки по работе с Docker
- Научился запускать scala kernel для работы со Scala в Jupyter Notebook