

Проект «Служба такси»

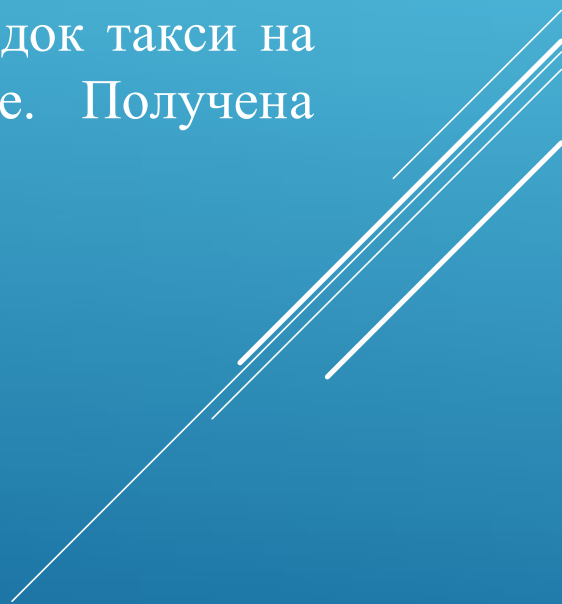
Выполнил:
Цыкунов Д. С.

2023

ВВЕДЕНИЕ

Клиенты и счета (Такси)

В представленном проекте проведён анализ данных поездок такси на основе таблицы, состоящей из поездок такси в г. Нью-Йорке. Получена таблица-отчёт как результат проекта.

Several thin, white, parallel diagonal lines are positioned in the bottom right corner of the slide, extending from the right edge towards the center.

Название и общее описание проекта

Итоговый проект №5 / курс Data Engineer / Клиенты и счета (Такси)

Цель проекта: на основе данных поездок такси в городе Нью-Йорк построить таблицу-отчёт (далее – parquet) с информацией каждого дня, в которую входят:

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

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

План реализации

Реализован проект на следующих этапах:

1. Скачать данные csv формата
2. Написать код на Scala в IntelliJ и создать два объекта для основной и дополнительной задачи
3. Оформление ноутбука в формате ipynb
4. Сохранение parquet в csv и parquet форматах для удобства
5. Создать описание проекта в README.md
6. Создать репозиторий и отправить 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 - развёртка образа JupyterLab и Spark со всеми предустановками
- PowerPoint – для выполнения презентации
- Draw.io - для составления схемы проекта

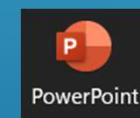
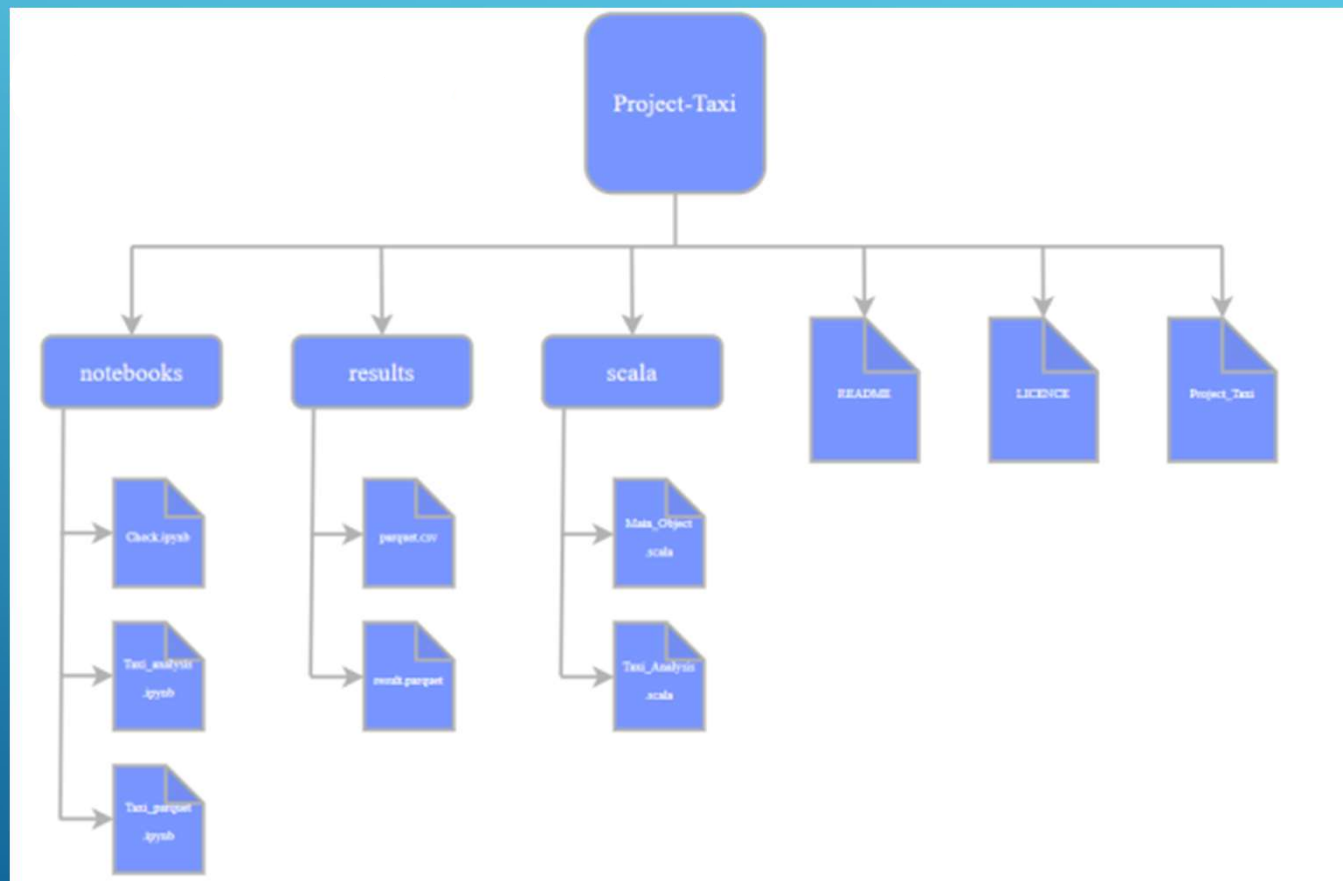
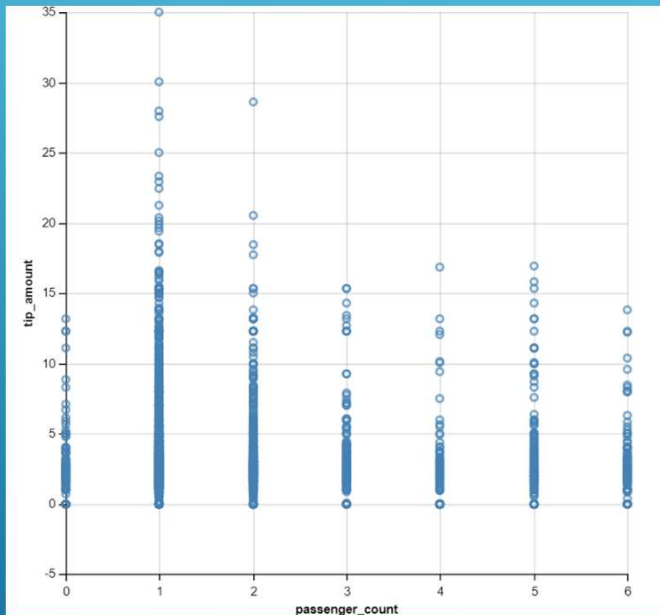


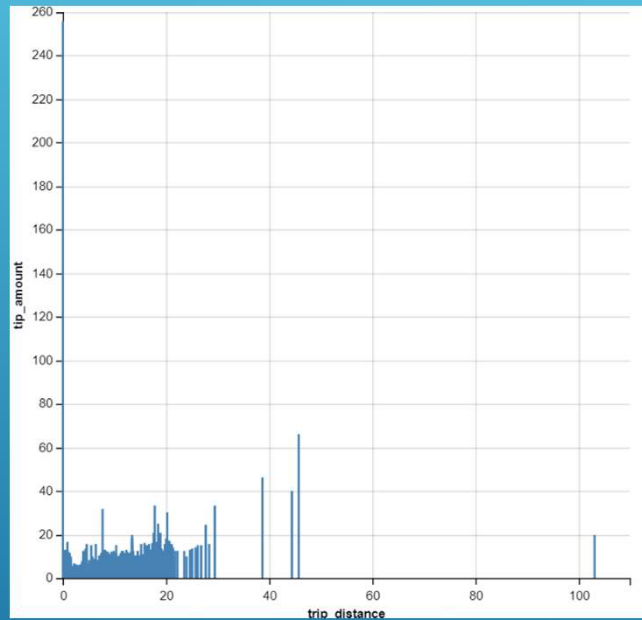
Схема проекта



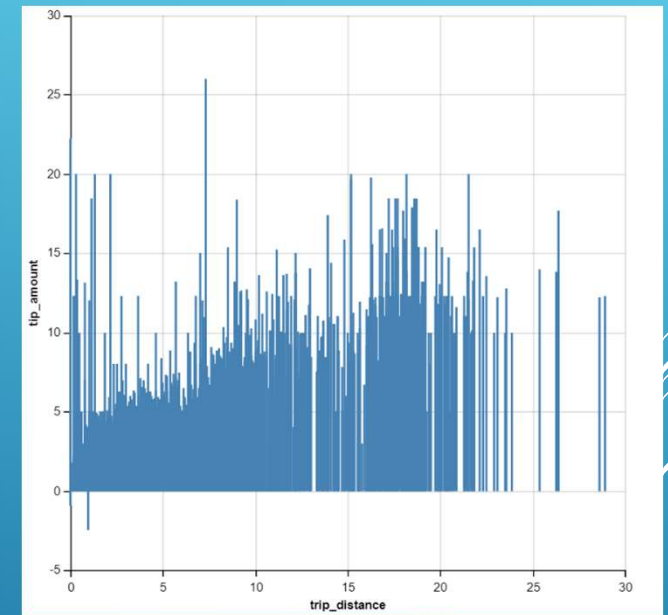
Графики



1



2



3

1 – зависимость чаевых от количества пассажиров; 2 – зависимость чаевых от расстояния поездки (all values); 3 – зависимость чаевых от расстояния поездки (emission-free)

Выводы по полученным 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 (4 пассажира)

На основании результатов можно предположить, что первая и вторая группа пассажиров лидируют по показателям

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

- Не учитывая выброс (чаевые ~ 255 при расстоянии 0) наибольшие чаевые получали при расстоянии 16-19 миль
- Минимальные чаевые клиенты давали при поездках на расстояние 2-4 мили

В итоге можно предположить, что заказы с усреднёнными расстояниями 7-10 миль и 15-20 миль проносят лучшую прибыль по чаевым.

Результаты разработки

В результате работы мы получили:

- Parquet , содержащий результаты по всем группам пассажиров
- Графики зависимости размера чаевых от кол-ва пассажиров и пройденного расстояния, на основании которых были сделаны выводы
- Схему проекта, выполненную в Draw.io
- Готовый проект на GitHub

Выводы

В ходе проекта были лучше освоены и закреплены навыки по работе с большими данными в основных средствах и платформах, в числе которых: Docker, Scala, Spark, Jupiter Notebook, GitHub. Также были выполнены цели, поставленные для проекта: построена таблица-отчёт (parquet) с информацией каждого дня, в которую входят: процент поездок по количеству человек в машине (5 групп пассажиров), самая дорогая поездка для каждой группы пассажиров и самая дешёвая поездка для каждой группы пассажиров.