**Проект № 5**

**Служба такси.**

Все операции должны считаться локально.

**Данные** - [Служба такси](https://disk.yandex.ru/d/DKeoopbGH1Ttuw" \o "https://disk.yandex.ru/d/DKeoopbGH1Ttuw)

Есть таблица, состоящая из поездок такси в Нью-Йорке.

|  |  |
| --- | --- |
| Поле | Описание |
| VendorId | ИД компании |
| Trep\_pickup\_datetime | Время и дата, когда пассажир сел в такси |
| Trep\_dropoff\_datetime | Время и дата, когда пассажир вышел из такси |
| Passanger\_count | Количество пассажиров |
| Trip\_distance | Пройденное расстояние |
| Ratecodeid | Код скорости |
| Store\_and\_fwd\_flag | Флаг, отвечающий за сохранение записи поездки перед ее отправкой поставщику |
| PulocationId | Широта, где была начата поездка |
| Dolocationid | Долгота, где была начата поездка |
| Payment\_type | Тип оплаты |
| Fare\_amount | Стоимость поездки |
| Mta\_tax | Комиссия автопарка |
| Tip\_amount | Чаевые |
| Tools\_amount | Оплата за платные дороги |
| Improvement\_surchange | Доплата за страховку |
| Total\_amount | Полная стоимость поездки |
| Congestion\_surchange | Дополнительный сбор |

Необходимо, используя таблицу поездок для каждого дня, рассчитать процент поездок по количеству человек в машине (без пассажиров, 1, 2, 3, 4 и более пассажиров). По итогу должна получиться таблица (parquet) с колонками date, percentage\_zero, percentage\_1p, percentage\_2p, percentage\_3p, percentage\_4p\_plus. Технологический стек — sql, scala (что-то одно).   
Также добавить столбцы к предыдущим результатам с самой дорогой и самой дешевой поездкой для каждой группы.

**Дополнительно:** также провести аналитику и построить график на тему «Как пройденное расстояние и количество пассажиров влияет на чаевые» в любом удобном инструменте.

Результат выполнения задания необходимо выложить в github/gitlab и указать ссылку на Ваш репозиторий (не забудьте: репозиторий должен быть публичным).