YELLOW TAXI TRIP DATA 2017 EDA

ПАВЛОВА АРИНА АЛЕКСАНДРОВНА, МФТИАД22

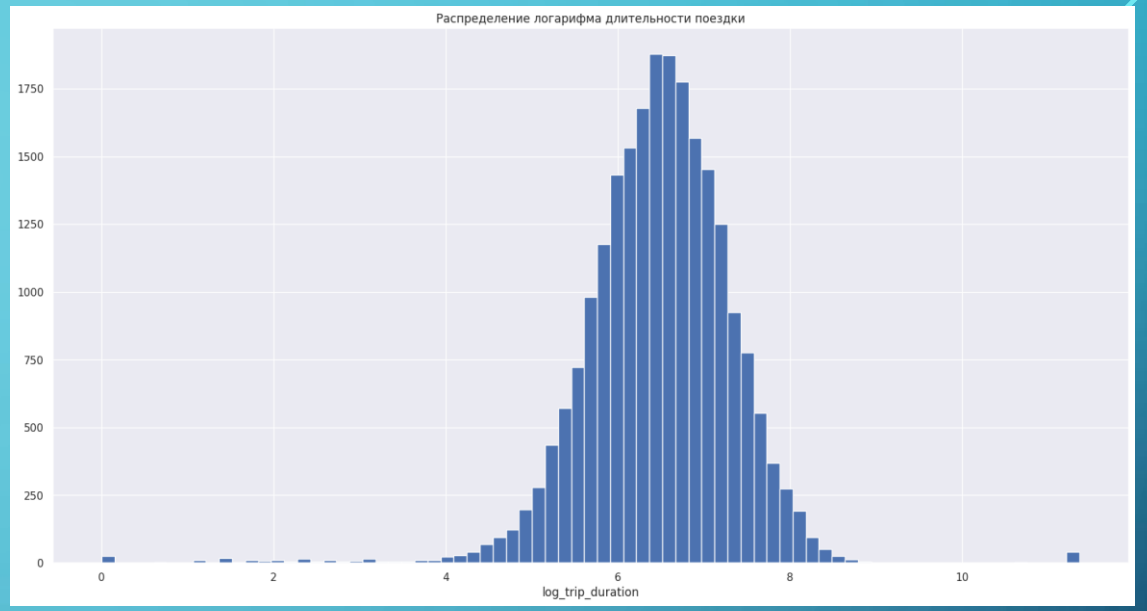
ДАННЫЕ ПО ПОЕЗДКАМ НА ТАКСИ ЗА 2017 ГОД

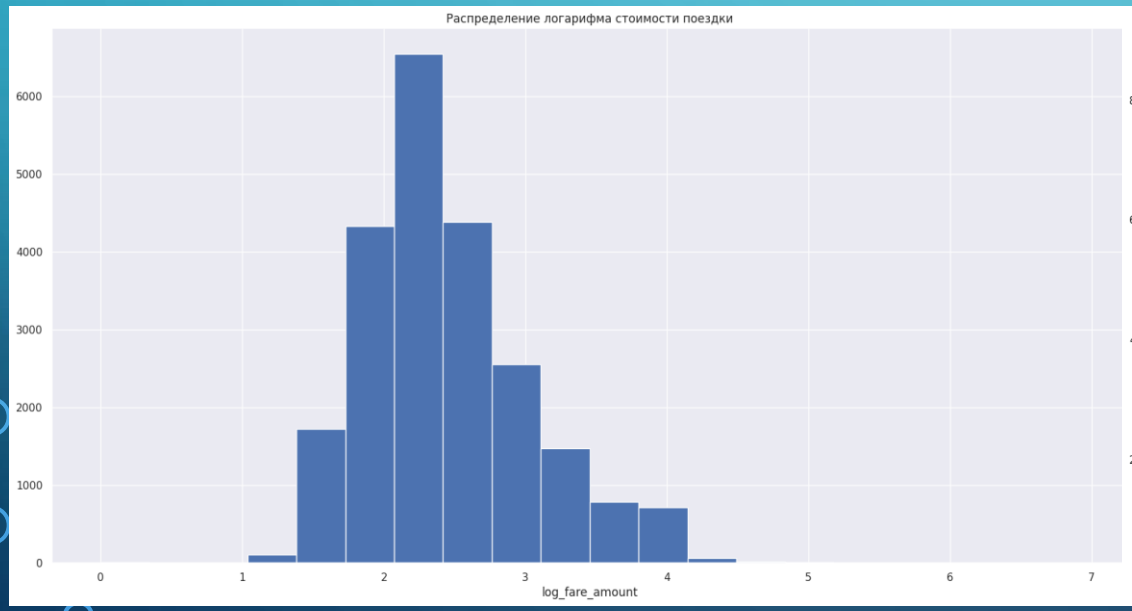
- https://www.kaggle.com/datasets/helddata/yellow-taxi-trip-data-2017

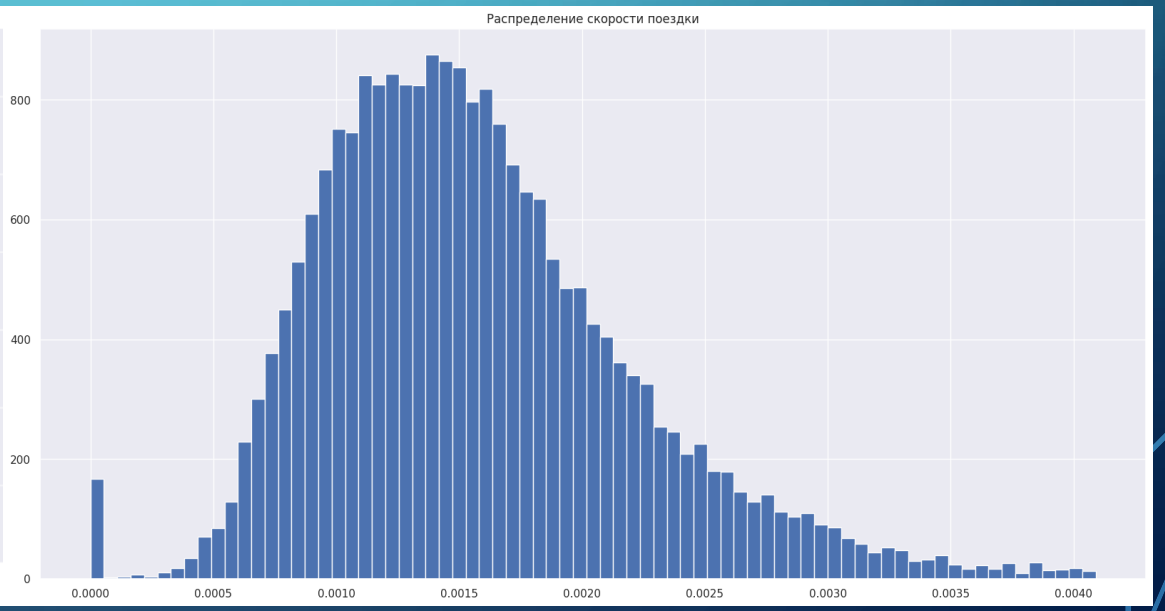
Описание столбцов:
•**ID** - Идентификационный номер поездки.
•**VendorID** - Код провайдера, предоставившего запись.
1) VeriFone Inc.
2) Creative Mobile Technologies, LLC;
•**tpep_pickup_datetime** - Дата и время, когда счетчик был включен.
•**tpep_dropoff_datetime** - Дата и время, когда счетчик был выключен.
•**Passenger_count** - Количество пассажиров в транспортном средстве. Это значение, введенное водителем.
•**Trip_distance** - Пройденное расстояние в милях, показанное таксометром.
•**PULocationID** - Зона такси TLC, в которой был включен таксометр.
•**DOLocationID** - Зона такси TLC, в которой был выключен таксометр.
•**RateCodeID** - Окончательный код тарифа, действующий в конце поездки.
•**Store_and_fwd_flag** - Этот флаг указывает, хранилась ли запись о поездке в памяти транспортного средства перед отправкой поставщику
•**Payment_type** - Цифровой код, обозначающий, как пассажир оплатил поездку.
•**Fare_amount** - Стоимость проезда по времени и расстоянию, рассчитанная с помощью счетчика.
•**Extra** - Прочие дополнительные услуги и доплаты. В настоящее время сюда входят только сборы в размере 0,50 и 1 доллара США в час пик и за ночь.
•**MTA_tax** - Налог в размере 0,50 доллара США в год, который автоматически взимается в зависимости от используемой учетной ставки.
•**Improvement_surcharge** - Надбавка за улучшение в размере 0,30 доллара США начислялась за поездки при снятии флага. Надбавка за улучшение начала взиматься в 2015 году.
•**Tip_amount** - Сумма чаевых – Это поле автоматически заполняется для чаевых по кредитной карте. Чаевые наличными не включены.
•**Tolls_amount** - Общая сумма всех сборов, уплаченных за поездку.
•**Total_amount** - Общая сумма, взимаемая с пассажиров. Не включает чаевые наличными.

РАСПРЕДЕЛЕНИЯ ДЛИТЕЛЬНОСТИ, СТОИМОСТИ И СКОРОСТИ ПОЕЗДОК

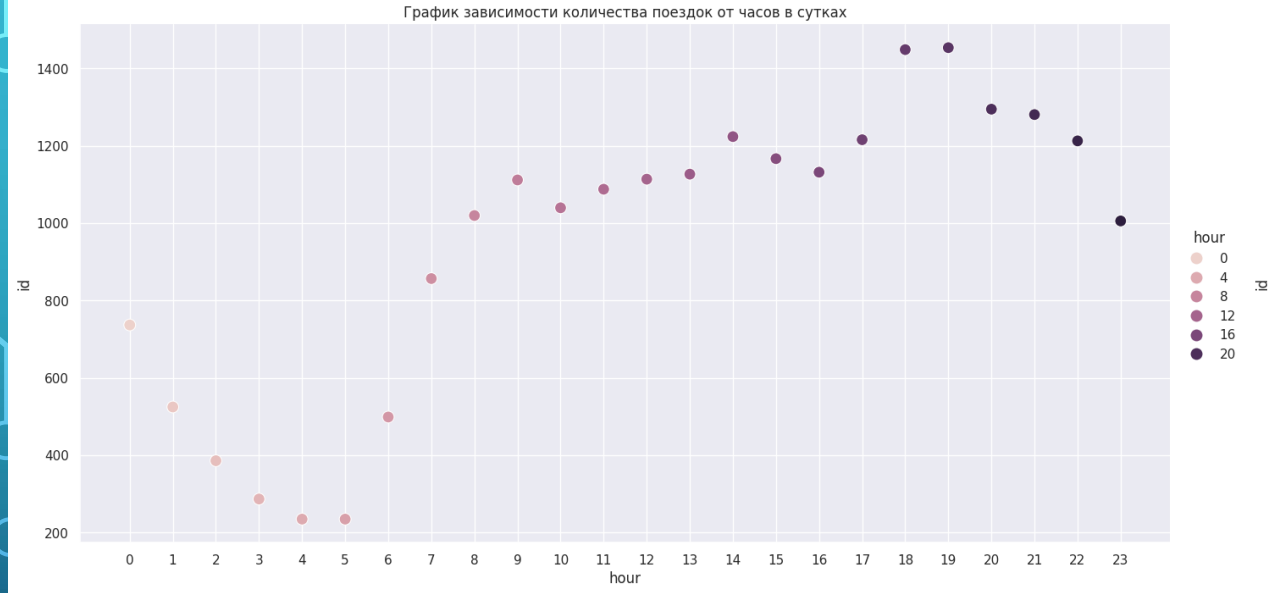
МОЖЕМ ГОВОРИТЬ О НОРМАЛЬНОСТИ

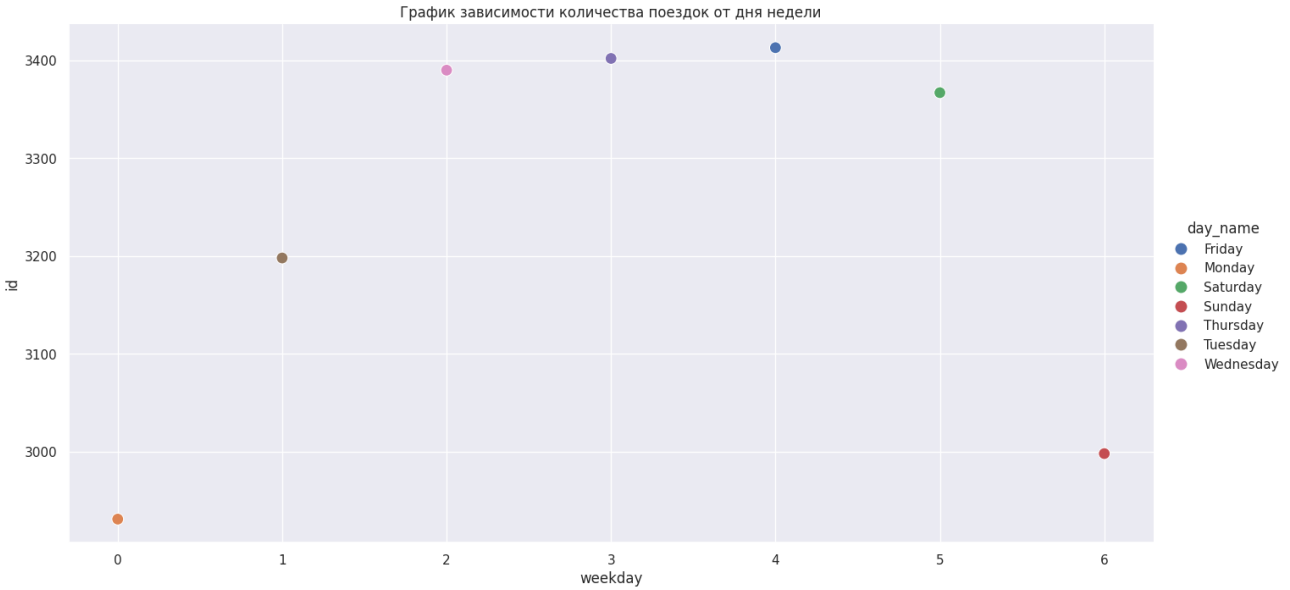






ЗАВИСИМОСТЬ КОЛИЧЕСТВА ПОЕЗДОК ОТ ВРЕМЕНИ

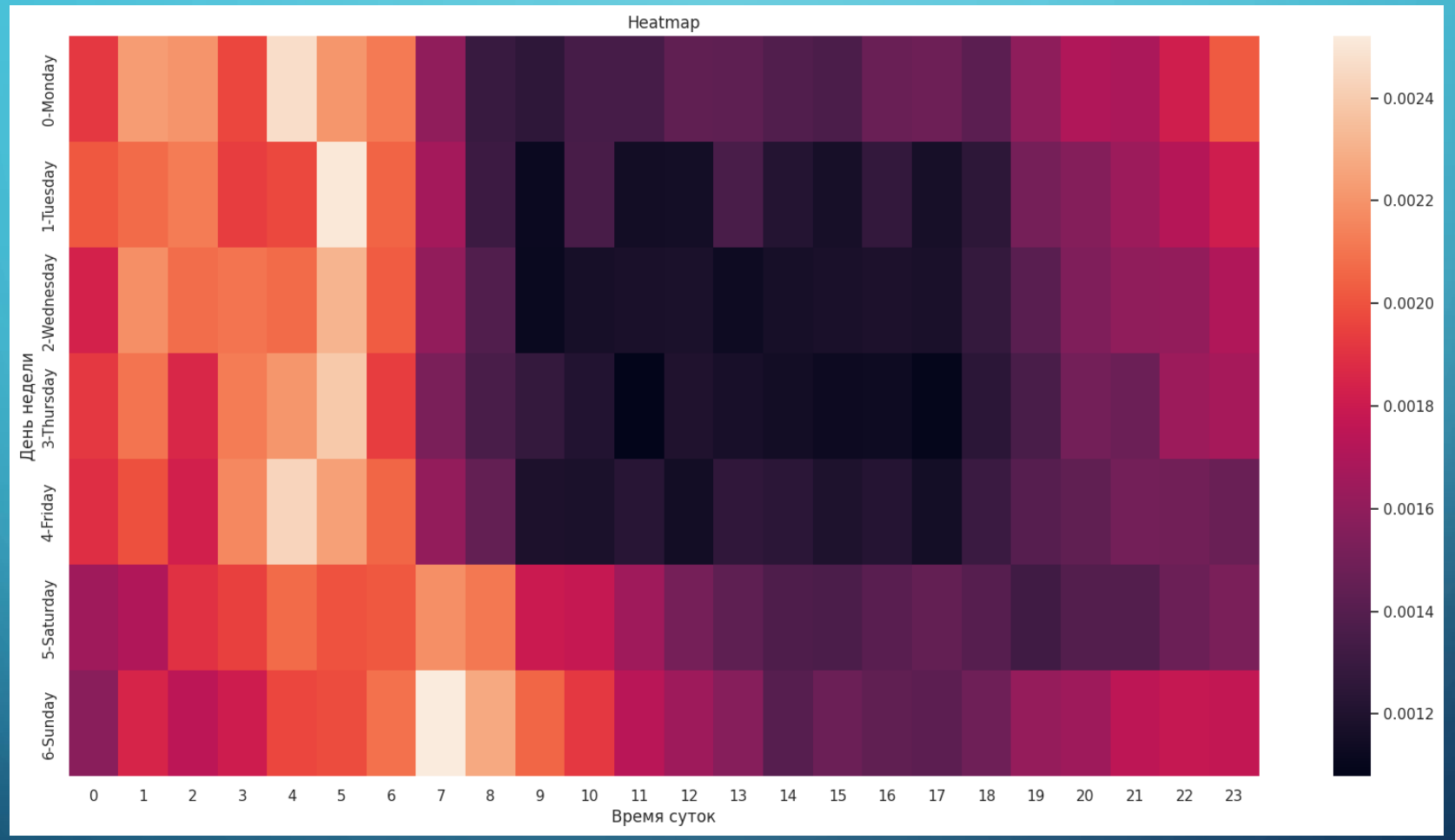




По графику можно сделать вывод о том, что больше всего поездок выпадает на 18-19 часов и меньше всего на 4-5 часов. При этом основная доля поездок приходится на рабочее время.

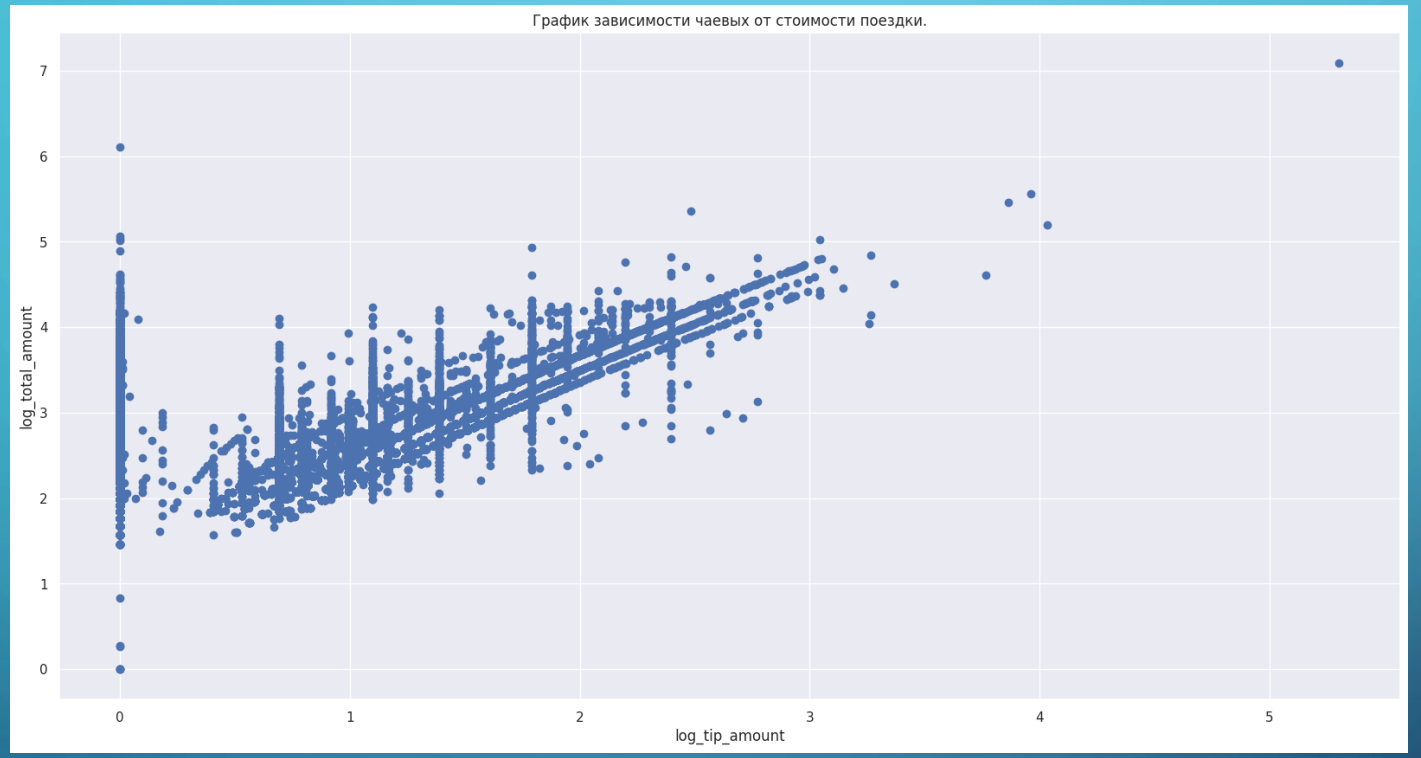
По графику можно сделать вывод о том, что больше всего поездок выпадает на пятницу и меньше всего на понедельник. При этом основная доля поездок приходится на будние дни.

ГРАФИК ЗАГРУЖЕННОСТИ (СКОРОСТИ ПОЕЗДОК)



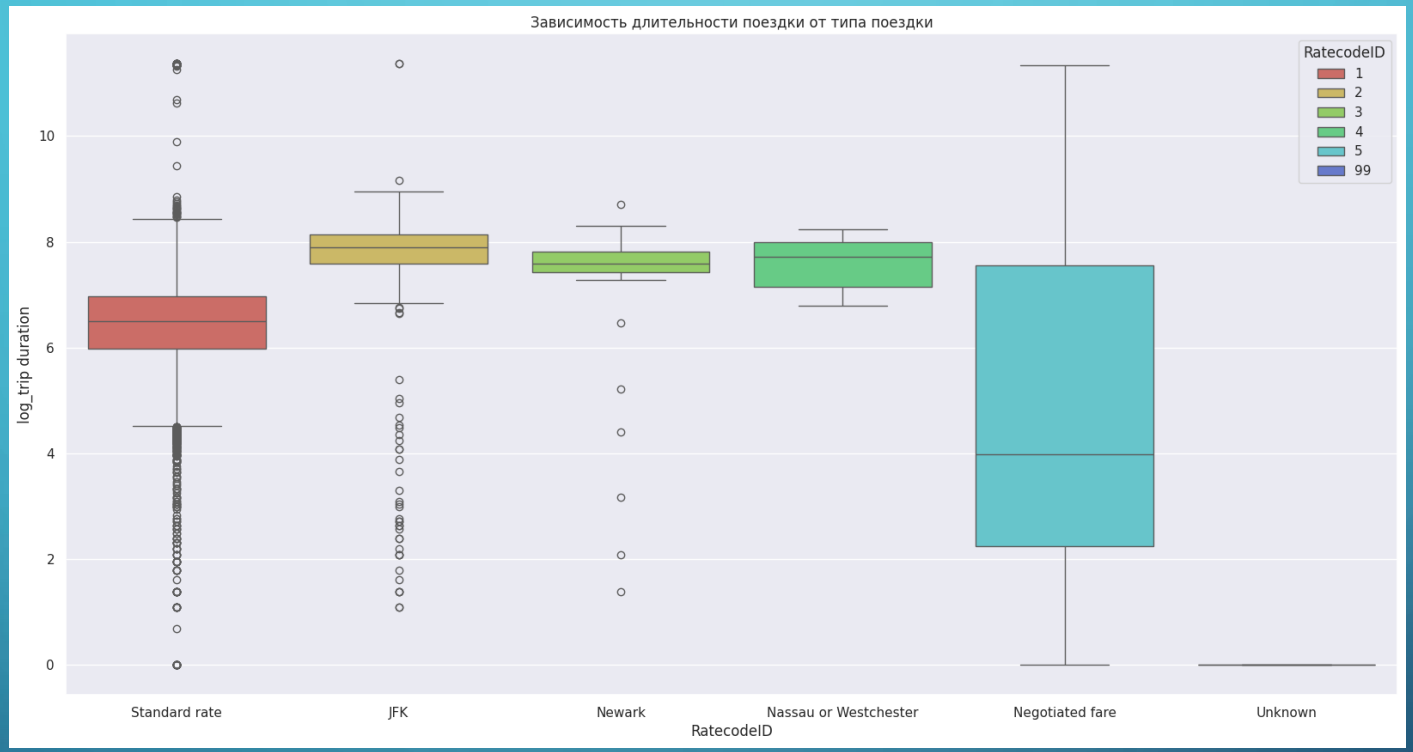
Исходя из графика, мы можем сказать, что скорость максимальна ранним утром (4-5 утра по будням и 6 часов в воскресенье) и минимальна в рабочее время со вторника по пятницу.

ВОПРОС О ЛИНЕЙНОСТИ ЗАВИСИМОСТИ ЧАЕВЫХ ОТ СТОИМОСТИ ПОЕЗДКИ



Так как разброс значений небольшой, мы взяли логарифмы. По графику отслеживается линейная зависимость логарифма чаевых от логарифма стоимости. Столбики точек говорят о склонности людей округлять, так как чаевые учтены только по карте, а не наличными.

BOXPLOT ЗАВИСИМОСТИ ДЛИТЕЛЬНОСТИ ПОЕЗДКИ ОТ ТИПА ПОЕЗДКИ

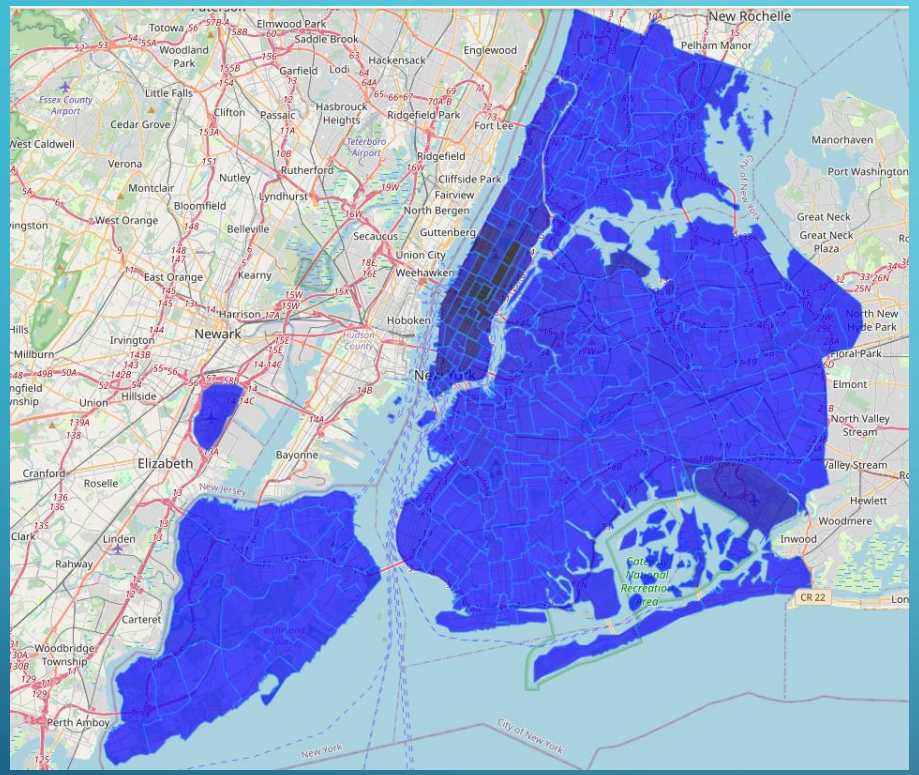


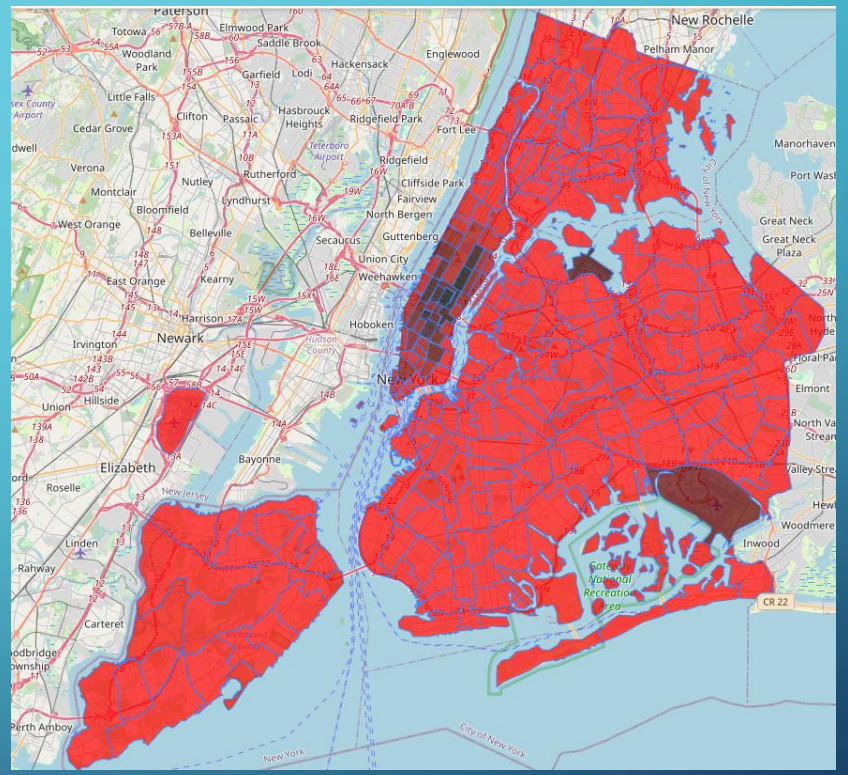
Мы видим, что самые длительные поездки до аэропортов JFK и Newark. Также замечаем новое категориальное значение 99, которое не было указано в описании признака. Скорее всего, тип поездки был не указан. Также смотрим на межквартильный размах у согласованного тарифа длительность поездки по согласованному тарифу очень сильно разнится.

ДАННЫЕ О ЛОКАЦИЯХ

- Так как такси очень сильно завязано на локациях, я решила использовать дополнительные данные о локациях, в которых совершались поездки, для визуализации.
- Данные взяты: https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc

НАРИСУЕМ КАРТУ ДЛЯ ВИЗУАЛИЗАЦИИ, КУДА И ОТКУДА ЧАЩЕ ВСЕГО ЕЗДЯТ В ТАКСИ

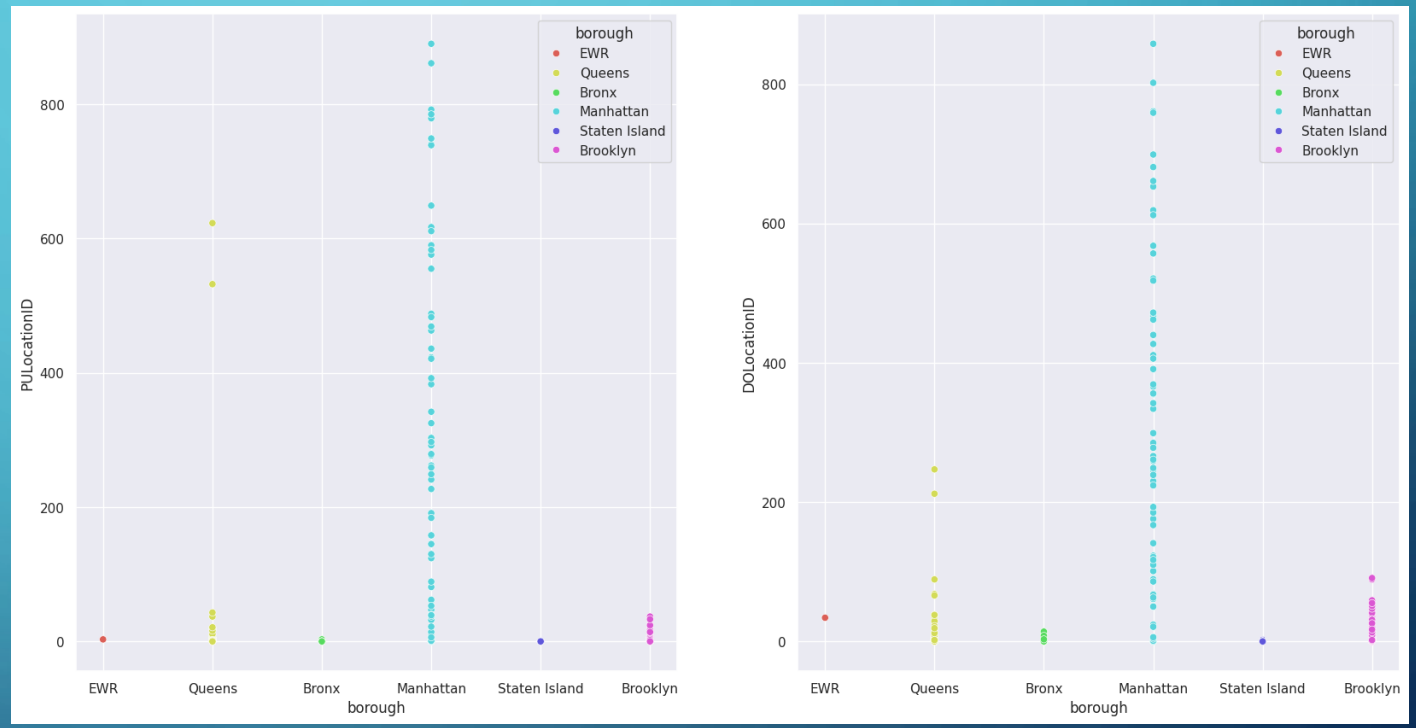




В основном самые частые направления и отправления принадлежат Манхэттену и двум аэропортам. Мы видим это по интенсивности цветов полигонов на карте. Для карты отправлений (красной) аэропорты выделяются еще ярче. Из аэропортов часто вызывают такси до домов/отелей.

САМЫЕ ЧАСТЫЕ РАЙОНЫ

borough	DOLocationID	PULocationID
Manhattan	20008	20602
Queens	1088	1399
Brooklyn	1068	317
Bronx	133	18
EWR	34	3
Staten Island	4	1



САМЫЕ ЧАСТЫЕ ЗОНЫ НАПРАВЛЕНИЯ/ОТПРАВЛЕНИЯ

LocationID	zone	borough	PULocationID
237	Upper East Side South	Manhattan	890
161	Midtown Center	Manhattan	861
186	Penn Station/Madison Sq West	Manhattan	792
236	Upper East Side North	Manhattan	785
162	Midtown East	Manhattan	779
...	...	...	...
157	Maspeth	Queens	0
156	Mariners Harbor	Staten Island	0
155	Marine Park/Mill Basin	Brooklyn	0
154	Marine Park/Floyd Bennett Field	Brooklyn	0
182	Parkchester	Bronx	0

Здесь опять видим Манхэттэн в топе.

LocationID	zone	borough	DOLocationID
161	Midtown Center	Manhattan	858
236	Upper East Side North	Manhattan	802
230	Times Sq/Theatre District	Manhattan	761
237	Upper East Side South	Manhattan	759
170	Murray Hill	Manhattan	699
...	...	...	...
20	Belmont	Bronx	0
108	Gravesend	Brooklyn	0
203	Rosedale	Queens	0
109	Great Kills	Staten Island	0
206	Saint George/New Brighton	Staten Island	0

Ситуация с Манхэттеном сохраняется. Если присмотреться, можно увидеть, что люди часто уезжают из центра и Times Square, видимо домой.