NYC Green Taxi Price Analysis

Team 18: Anton, Anatoly, Sofia, Anastasiia

Table of contents

01

Objectives of the project

Our aim

The project objective is to develop a machine learning model based on the NYC Green taxi dataset to predict taxi fares based on several factors: date, time, number of passengers, etc.

02

Data Description

83,000,000

Records of taxi rides

2014-2024

Period covered by the data

20

Features

Features

DateTime

- lpep_pickup_datetime - time of pickup
- lpep_dropoff_datetime - time of drop off

Categorical

- VendorID - LPEP provider
- RatecodeID - rate type
- Payment_type
- trip_type - street-hail or dispatch
- PULocationID, DOLocationID - start and end locations

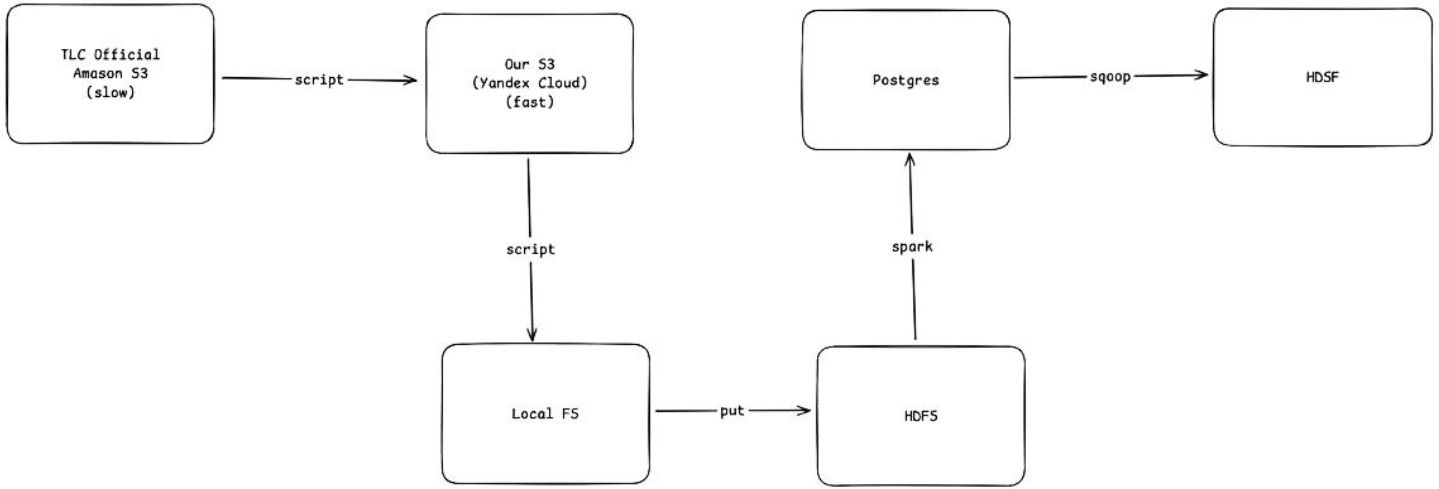
Numerical

- passanger_count
- trip_distance
- fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, congestion_surcharge, cbd_congestion_fee
- total_amount

03

Data Preparation

Pipeline


TLC Official
Amason S3
(slow)
script
Our S3
(Yandex Cloud)
(fast)
Postgres
sqoop
HDSF
script
spark
Local FS
put
HDFS

Hive

Partitioning

We partitioned the Hive dataset by year and month, creating separate table for each. This way it is easier to find correlations during certain period of time.

Bucketing

We did not use bucketing in our project since there were no suitable columns: we have already partitioned by year and month, and other columns could not be used to generate balanced data splits.

04

Data Analysis

Queries

Missing values

Find out how many missing values are in each rows

Invalid rows

Find out number of rows with weird values

Price Correlations

Calculate correlation between price and duration/distance/passenger count

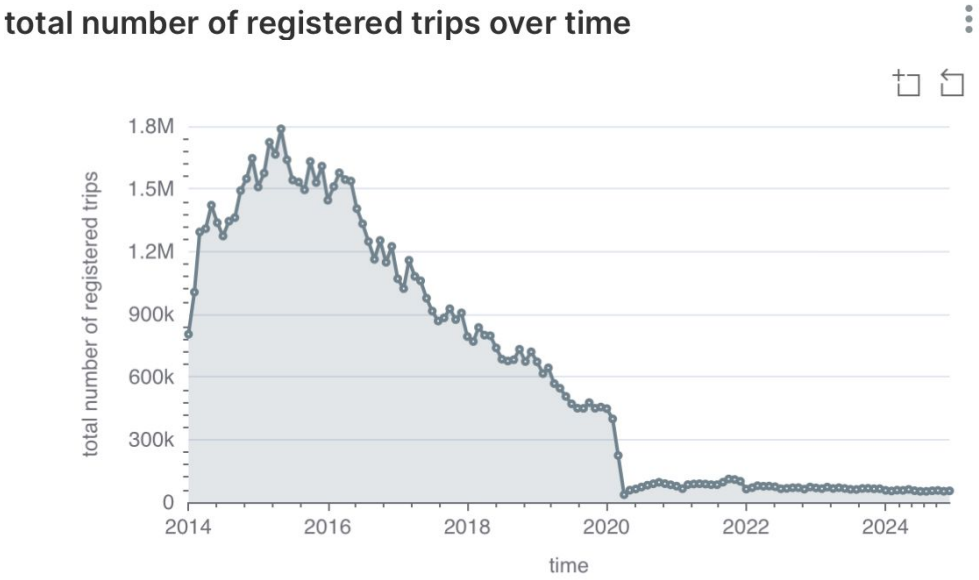
Duration

Create table with pre-calculated trip duration

Hour

Create table with drop-off and pick-up hours and price of the trip

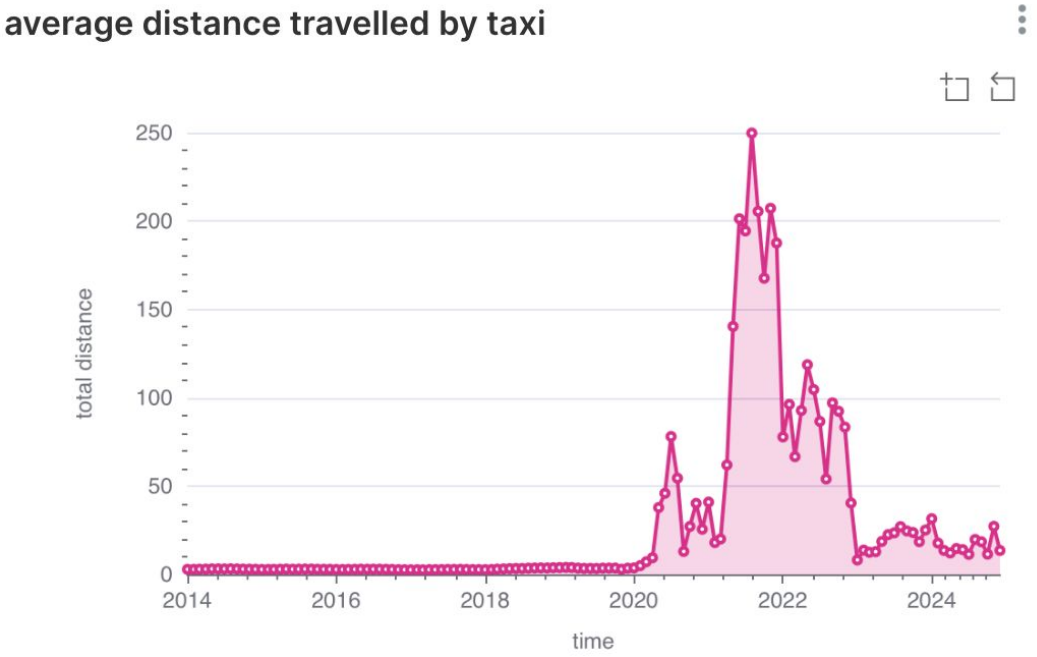
Findings



Big Drop in Taxi Rides

NYC taxis saw a gradual decline in the number of registered trips due to the popularity of Uber. Then a complete drop in COVID times (2020).

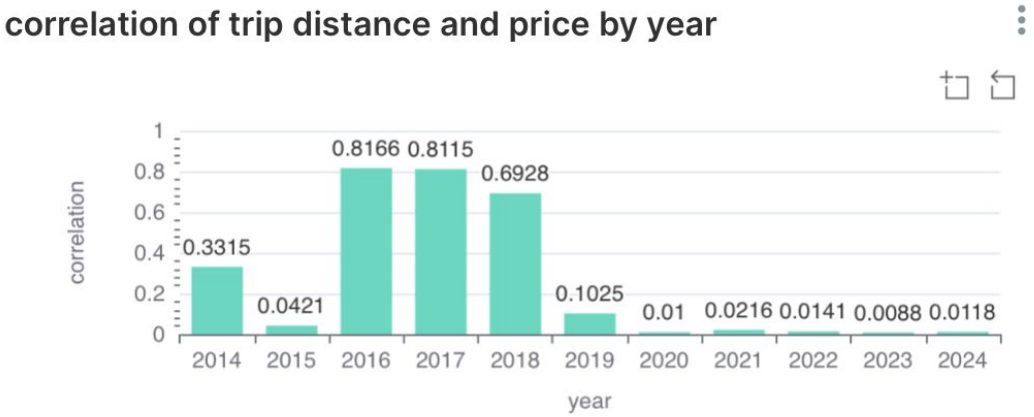
Findings



Shift towards Long Distance

In 2021 especially, the few taxi trips people took were incredibly long. This was a big, temporary change in how people used taxis. This most probably happened due to COVID, remote jobs and a housing crisis in NYC, forcing people to move further from the city.

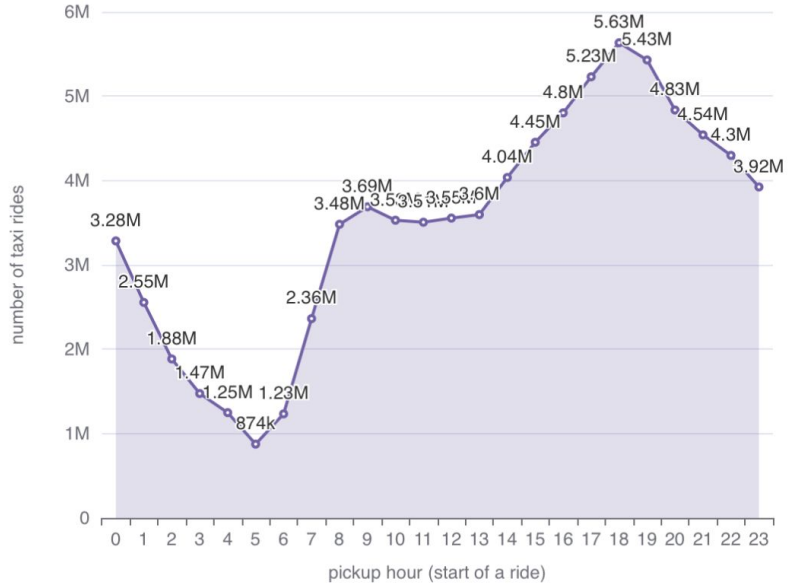
Findings



How Prices Changed: Distance Matters Less

Before 2020, longer trips cost significantly more. Now, trip distance doesn't affect the price, which helped make those long trips possible.

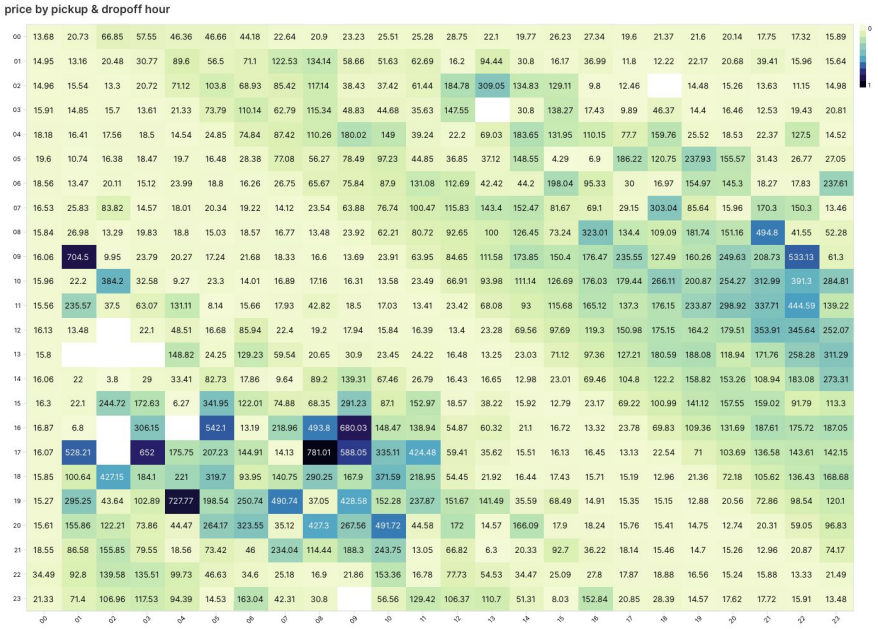
Findings



Busiest Times: Morning & Evening Rush Hours

Most people take taxis during morning and evening travel times. The evening rush hour (around 5-7 PM) is the busiest for taxi rides, making them more expensive.

Findings

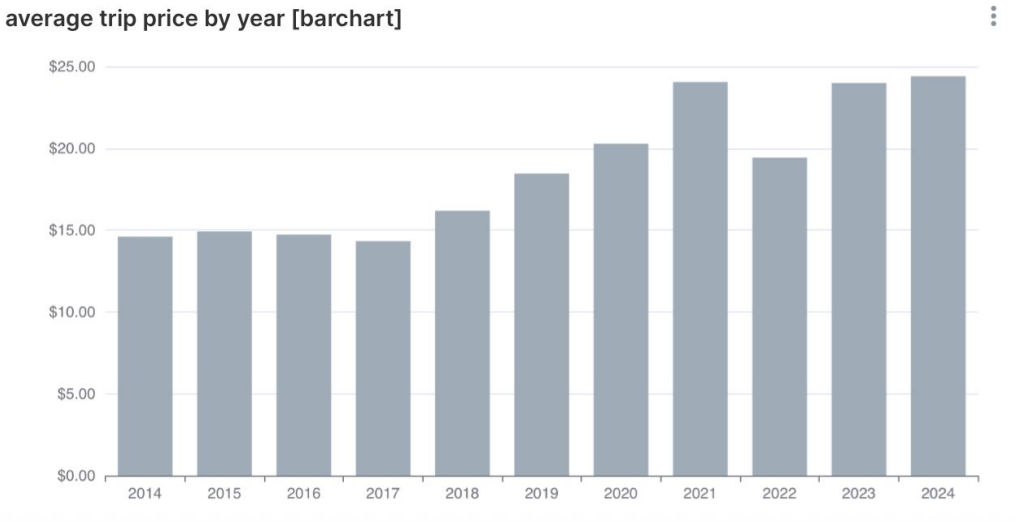


Peak Times and Long Trips Pay More

This heatmap shows how taxi prices change based on when a trip starts and ends. Trips during morning and evening rush hour or those lasting a very long time (like overnight) cost the most, helping maximize earnings.

(Y axis - pickup hour, X axis - drop off)

Findings



Steep Increase in Prices

After 2020, fare prices have increased faster than the US inflation rate at the time. This could be due to several reasons:
- Losing rentability with lower prices due to the decreased market share
- COVID restrictions
- Longer rides costing more

05

ML Modeling

Data Preparation for modeling

- Since some time in 2020 the formula for the price have possibly changed (as wel as the average distance and demand), we decided to train the model only on the data from 2021 and after
- We have dropped the features which do not influence the fare_amount (according to the official NYC taxi documentation and our analysis).
- One-hot encoded the categorical features
- Encoded datetime features as year, month, weekday, hour, minute, and second
- Scaled all numerical values with the standard scaler to prevent bias from the model.

Models

Linear Regression

Hyperparameters:

- regParam: [0.01, 0.1, 1]
- elasticNetParam: [0.0, 0.5, 1.0]

Random Forest Regression

Hyperparameters:

- numTrees: [25, 50, 75]
- maxDepth: [3, 4, 5]

Model comparison

model	rmse	r2
LinearRegression	14.752092861088691	0.22984644283559685

Best train parameters from cross-validation:

- regParam: 0.01
- elasticNetParam: 0

model	rmse	r2	numtree	maxdepth
RandomForestRegressor	11.60975685222611	0.5230022844384274	75	5

As we can see, the best performing model turned out to be Random Forest Regression with maximum depth equal to 5 and number of trees equal to 75.

This means that the trees capture correlation better, and rely on both depth, and the number of trees.

06

Data Presentation

Apache Superset Dashboard

Overview

Description of datasets and steps performed during the project

Data Insights

Useful data findings gained during EDA

Modeling

Results of the Modeling part

Conclusion

We have analyzed a Green Taxi trips dataset and build an ML Model using subset of this dataset.

We have found multiple helpful business insights during EDA stage that can be used by the company to improve their business. However, certain parameters did not have a correlation to fare amount on their own.

The modeling part confirmed our suspicions: the formula for the price must be more complex than we initially thought, so most probably better features need to be created from existing data to get better results.

07

Bonus Slides (legacy EDA)

Findings

null count in columns [table view]

column_name	null_count	null_percent
ehail_fee	83481636	0.9999634423979641
congestion_surcharge	74423160	0.8914588026010231
improvement_surcharge	14191088	0.16998432095715563
trip_type	3667901	0.04393501476582149
ratecodeid	1885146	0.02258073959622392
store_and_fwd_flag	1885146	0.02258073959622392
payment_type	1885146	0.02258073959622392
passenger_count	1885146	0.02258073959622392

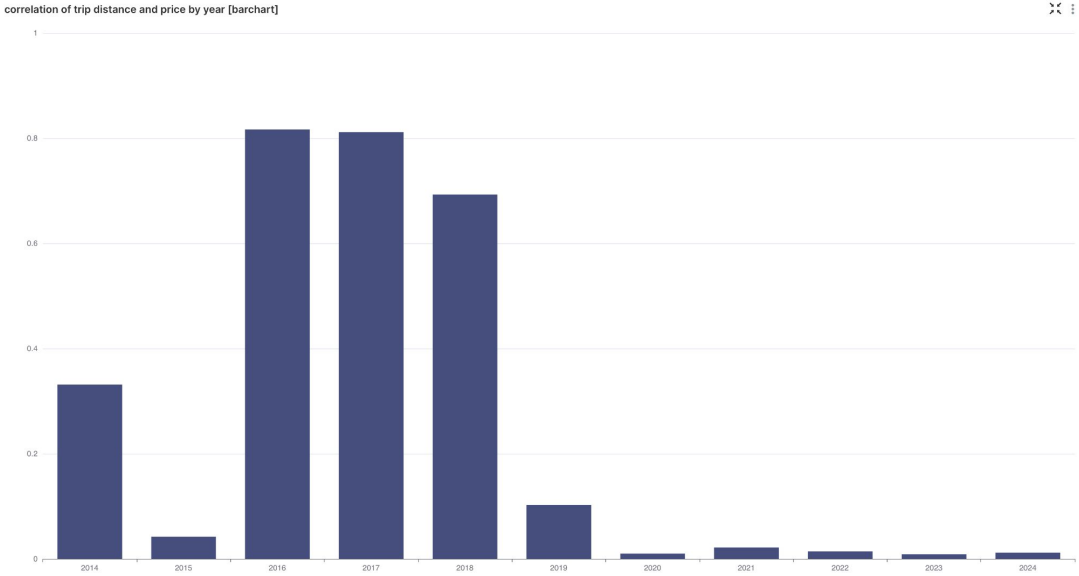
Null values percentage

Most of the important columns have an insignificant amount of null values.

14.4%

Rows are invalid

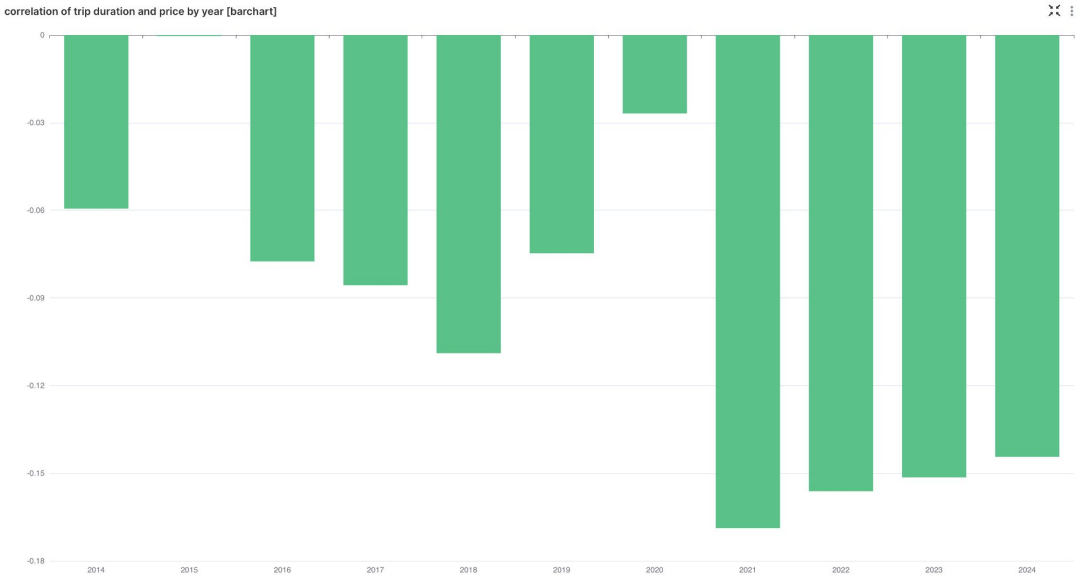
Findings



Correlation of distance and price

There is no significant correlation between trip distance and price after 2019. That might signal a change in formula.

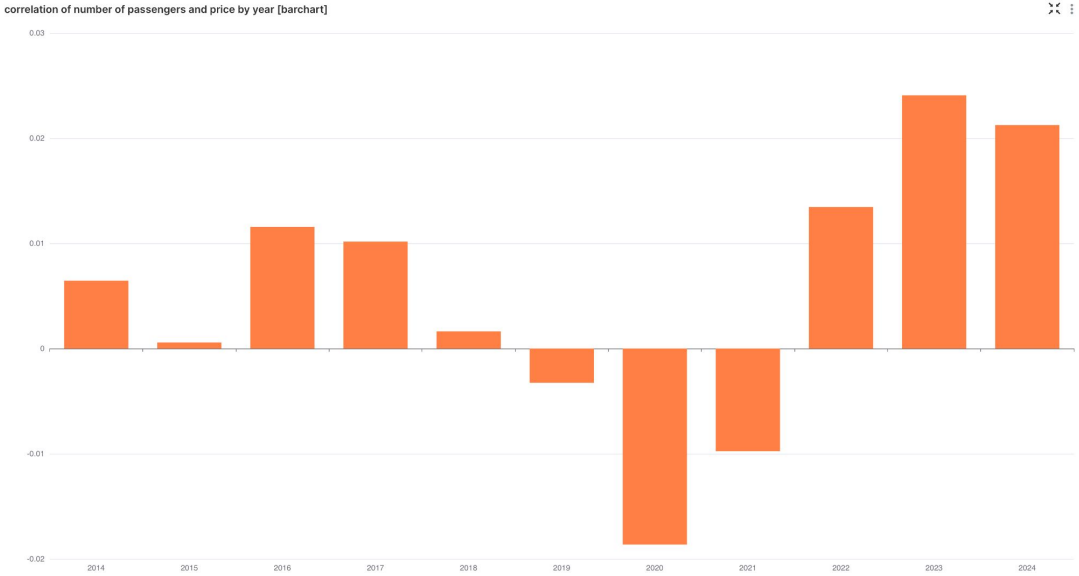
Findings



Correlation of duration and price

There is no significant correlation between trip duration and price in all observed years.

Findings



Correlation of # passengers and price

There is no significant correlation with the number of passengers either.

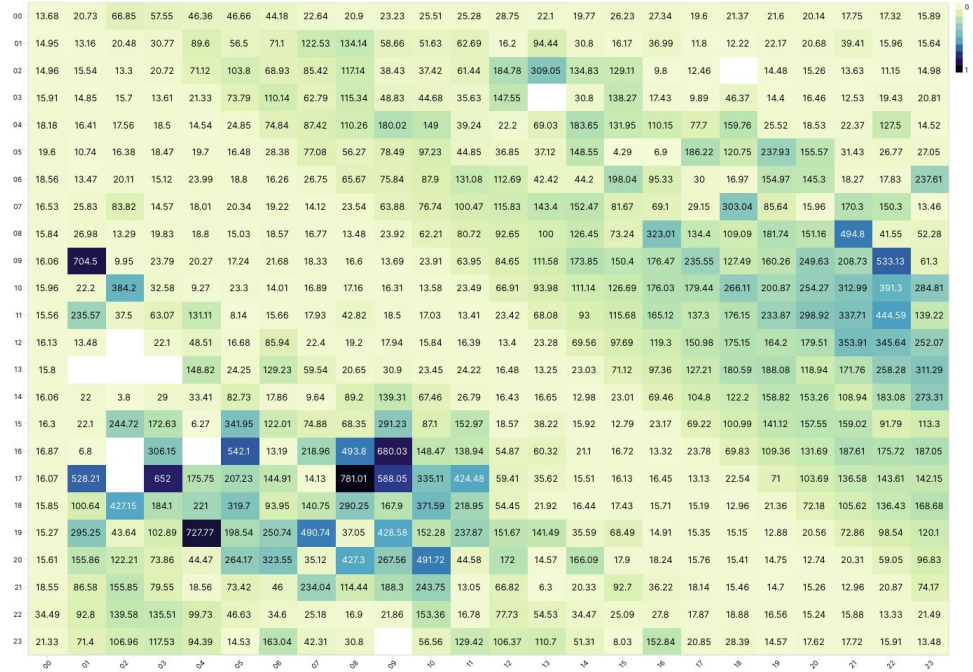
Findings

Average trip duration by time

There is no significant change in trip duration over time, meaning that the company didn't significantly change the area covered by green taxi.

Findings

price by pickup & dropoff hour



	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
00	13.68	20.73	66.85	57.55	46.36	46.66	44.18	22.64	20.9	23.23	25.51	25.28	28.75	22.1	19.77	26.23	27.34	19.6	21.37	21.6	20.14	17.75	17.32	15.89
01	14.95	13.16	20.48	30.77	89.6	56.5	71.1	122.53	134.14	58.66	51.63	62.69	16.2	94.44	30.8	16.17	36.99	11.8	12.22	22.17	20.68	39.41	15.96	15.64
02	14.96	15.54	13.3	20.72	71.12	103.8	68.93	85.42	117.14	38.43	37.42	61.44	184.78	309.05	134.83	129.11	9.8	12.46		14.48	15.26	13.63	11.15	14.98
03	15.91	14.85	15.7	13.61	21.33	73.79	110.14	62.79	115.34	48.83	44.68	35.63	147.55		30.8	138.27	17.43	9.89	46.37	14.4	16.46	12.53	19.43	20.81
04	18.18	16.41	17.56	18.5	14.54	24.85	74.84	87.42	110.26	180.02	149	39.24	22.2	69.03	183.65	131.95	110.15	77.7	159.76	25.52	18.53	22.37	127.5	14.52
05	19.6	10.74	16.38	18.47	19.7	16.48	28.38	77.08	56.27	78.49	97.23	44.85	36.85	37.12	148.55	4.29	6.9	186.22	120.75	237.93	155.57	31.43	26.77	27.05
06	18.56	13.47	20.11	15.12	23.99	18.8	16.26	26.75	65.67	75.84	87.9	131.08	112.69	42.42	44.2	198.04	95.33	30	16.97	154.97	145.3	18.27	17.83	237.61
07	16.53	25.83	83.82	14.57	18.01	20.34	19.22	14.12	23.54	63.88	76.74	100.47	115.83	143.4	152.47	81.67	69.1	29.15	303.04	85.64	15.96	170.3	150.3	13.46
08	15.84	26.98	13.29	19.83	18.8	15.03	18.57	16.77	13.48	23.92	62.21	80.72	92.65	100	126.45	73.24	323.01	134.4	109.09	181.74	151.16	494.8	41.55	52.28
09	16.06	704.5	9.95	23.79	20.27	17.24	21.68	18.33	16.6	13.69	23.91	63.95	84.65	111.58	173.85	150.4	176.47	235.55	127.49	160.26	249.63	208.73	533.13	61.3
10	15.96	22.2	384.2	32.58	9.27	23.3	14.01	16.89	17.16	16.31	13.58	23.49	66.91	93.98	111.14	126.69	176.03	179.44	266.11	200.87	254.27	312.99	391.3	284.81
11	15.56	235.57	37.5	63.07	131.11	8.14	15.66	17.93	42.82	18.5	17.03	13.41	23.42	68.08	93	115.68	165.12	137.3	176.15	233.87	298.92	337.71	444.59	139.22
12	16.13	13.48		22.1	48.51	16.68	85.94	22.4	19.2	17.94	15.84	16.39	13.4	23.28	69.56	97.69	119.3	150.98	175.15	164.2	179.51	353.91	345.64	252.07
13	15.8				148.82	24.25	129.23	59.54	20.65	30.9	23.45	24.22	16.48	13.25	23.03	71.12	97.36	127.21	180.59	188.08	118.94	171.76	258.28	311.29
14	16.06	22	3.8	29	33.41	82.73	17.86	9.64	89.2	139.31	67.46	26.79	16.43	16.65	12.98	23.01	69.46	104.8	122.2	158.82	153.26	108.94	183.08	273.31
15	16.3	22.1	244.72	172.63	6.27	341.95	122.01	74.88	68.35	291.23	87.1	152.97	18.57	38.22	15.92	12.79	23.17	69.22	100.99	141.12	157.55	159.02	91.79	113.3
16	16.87	6.8		306.15		542.1	13.19	218.96	493.8	680.03	148.47	138.94	54.87	60.32	21.1	16.72	13.32	23.78	69.83	109.36	131.69	187.61	175.72	187.05
17	16.07	528.21		652	175.75	207.23	144.91	14.13	781.01	588.05	335.11	424.48	59.41	35.62	15.51	16.13	16.45	13.13	22.54	71	103.69	136.58	143.61	142.15
18	15.85	100.64	427.15	184.1	221	319.7	93.95	140.75	290.25	167.9	371.59	218.95	54.45	21.92	16.44	17.43	15.71	15.19	12.96	21.36	72.18	105.62	136.43	168.68
19	15.27	295.25	43.64	102.89	727.77	198.54	250.74	490.74	37.05	428.58	152.28	237.87	151.67	141.49	35.59	68.49	14.91	15.35	15.15	12.88	20.56	72.86	98.54	120.1
20	15.61	155.86	122.21	73.86	44.47	264.17	323.55	35.12	427.3	267.56	491.72	44.58	172	14.57	166.09	17.9	18.24	15.76	15.41	14.75	12.74	20.31	59.05	96.83
21	18.55	86.58	155.85	79.55	18.56	73.42	46	234.04	114.44	188.3	243.75	13.05	66.82	6.3	20.33	92.7	36.22	18.14	15.46	14.7	15.26	12.96	20.87	74.17
22	34.49	92.8	139.58	135.51	99.73	46.63	34.6	25.18	16.9	21.86	153.36	16.78	77.73	54.53	34.47	25.09	27.8	17.87	18.88	16.56	15.24	15.88	13.33	21.49
23	21.33	71.4	106.96	117.53	94.39	14.53	163.04	42.31	30.8		56.56	129.42	106.37	110.7	51.31	8.03	152.84	20.85	28.39	14.57	17.62	17.72	15.91	13.48

Correlation of price with pickup and dropoff hour

1) Trips under 2 hours are the cheapest throughout the day
2) Long trips during daytime (morning to evening) cost significantly more than night trips (from night to morning)

(Y axis - pickup hour, X axis - drop off)