



# NYC Green Taxi Price Analysis

Team 18: Anton, Anatoly, Sofia, Anastasiia

# Table of contents

01

Objectives

02

Data Description

03

Data Preparation

04

Data Analysis

05

ML Modeling

06

Data Presentation



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

- passenger\_count
- trip\_distance
- fare\_amount, extra, mta\_tax, tip\_amount, tolls\_amount, improvement\_surcharge, congestion\_surcharge, cbd\_congestion\_fee
- total\_amount



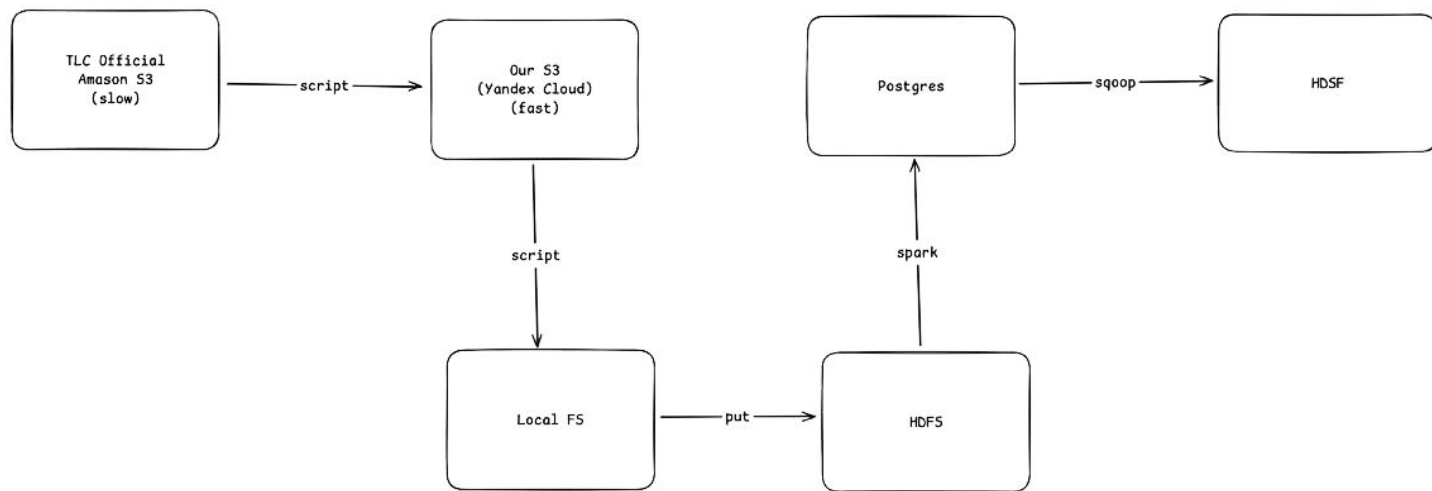


03

# Data Preparation



# Pipeline



# 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

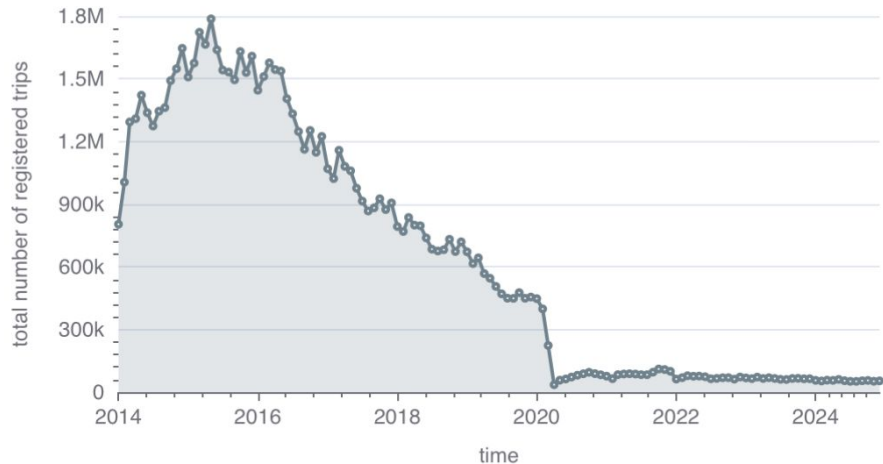


## Hour

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

# Findings

total number of registered trips over time

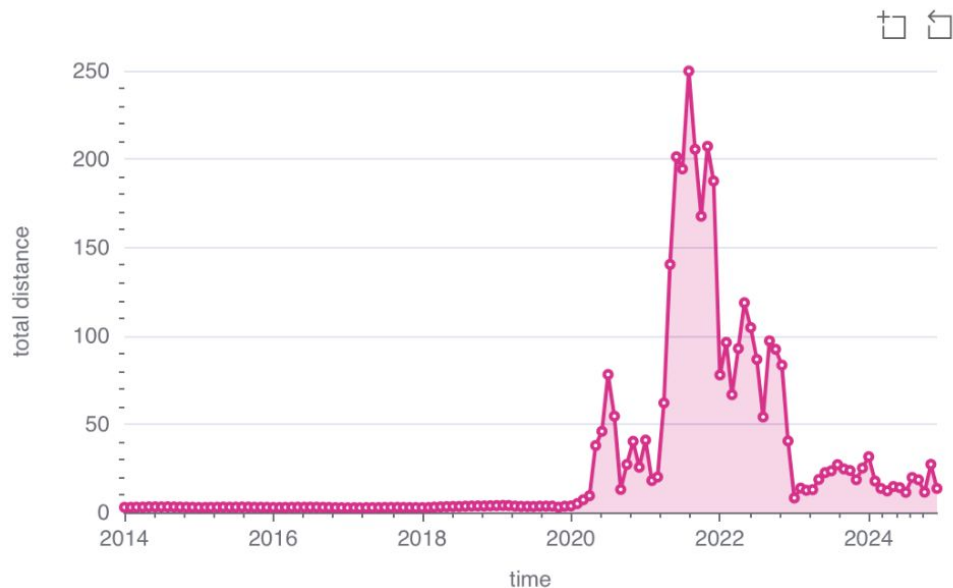


## Big Drop in Taxi Rides

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

average distance travelled by taxi



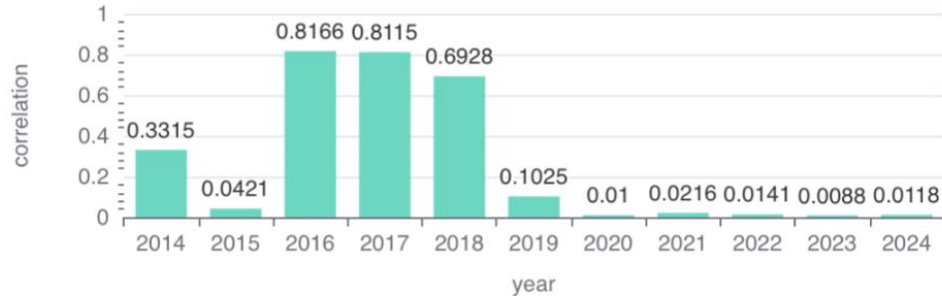
## Shift towards Long Distance

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

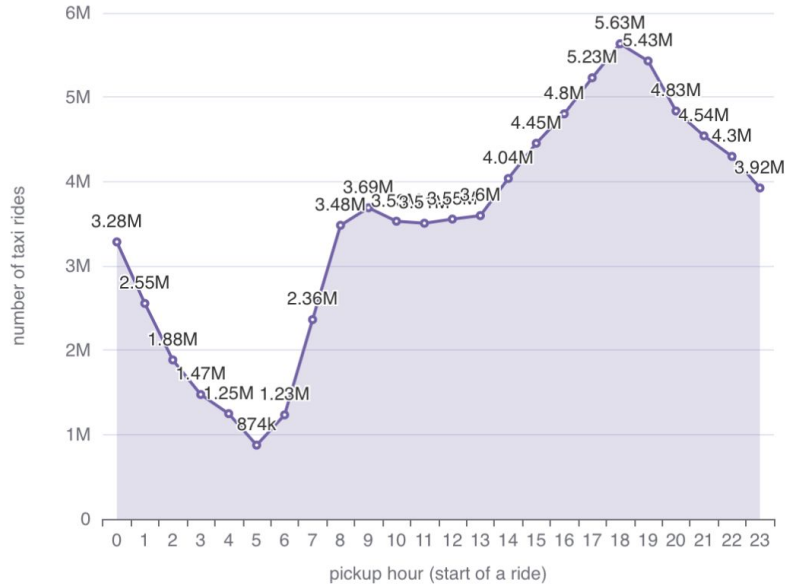
correlation of trip distance and price by year



## 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.

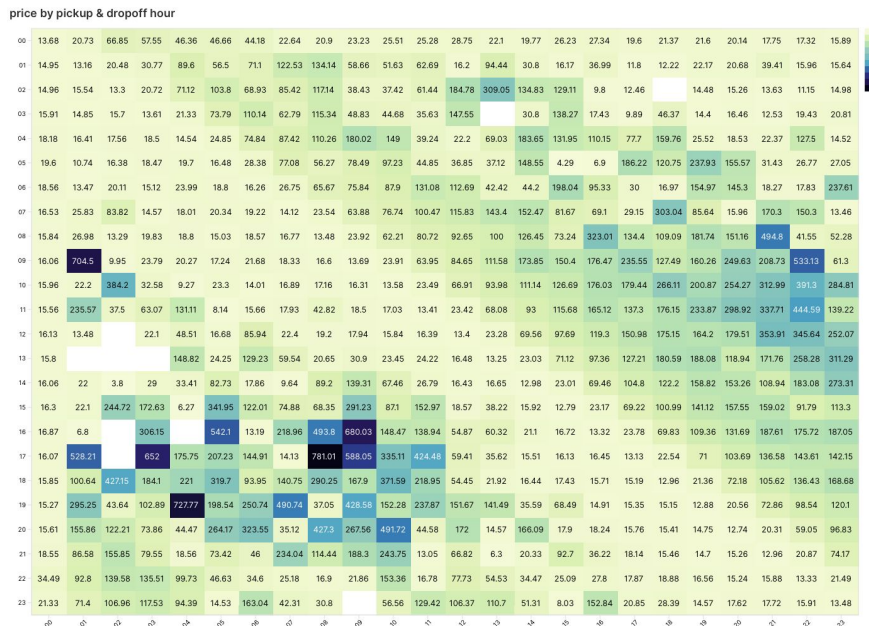
# Findings



## 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



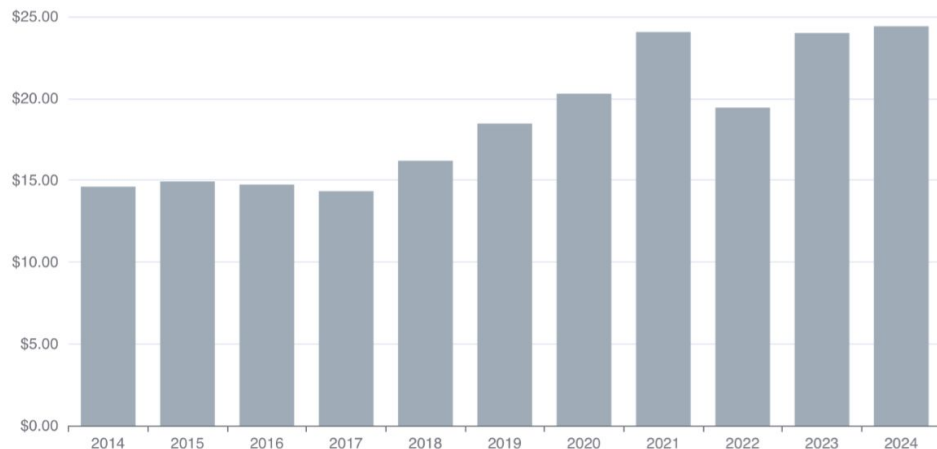
## Peak Times and Long Trips Pay More

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

average trip price by year [barchart]



## 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, 5, 9]

# Model comparison

model	rmse	r2
LinearRegression	14.752092861088691	0.22984644283559685

Figure 18: Linear Regression results, for regParam = 0.01, elasticNetParam = 0.0

name	r2	rmse
RandomForestRegressor(maxDepth=9, numTrees=50)	0.79891905480599	4.220788142870471

Figure 19: Random Forest Regression results

As we can see, the best performing model turned out to be Random Forest Regression with maximum depth equal to 9 and number of trees equal to 50.

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

Nonetheless, the Random forest model was trained on a smaller portion of the data.



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.

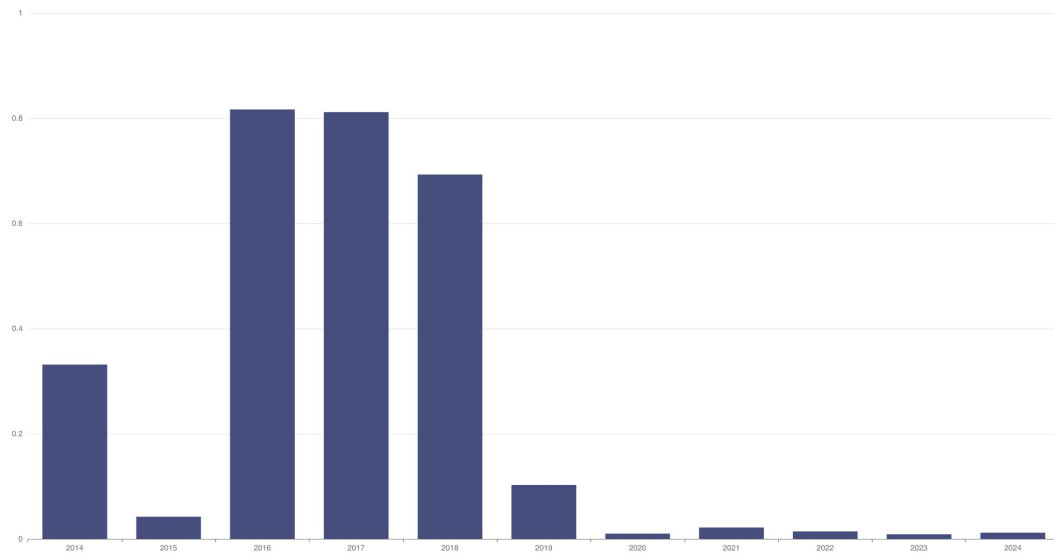


14.4%

Rows are invalid

# Findings

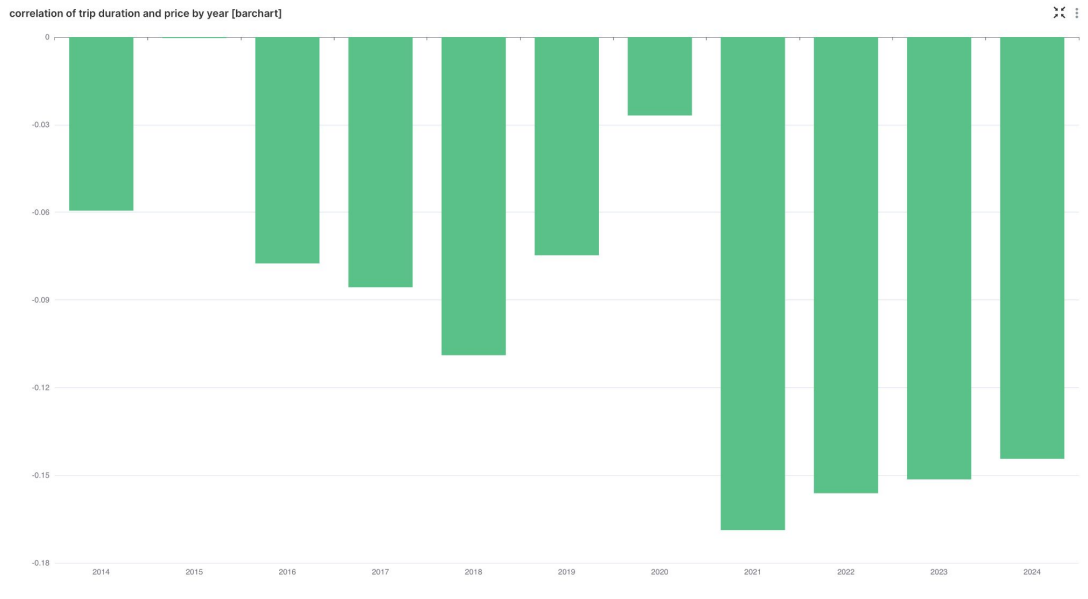
correlation of trip distance and price by year [barchart]



## 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

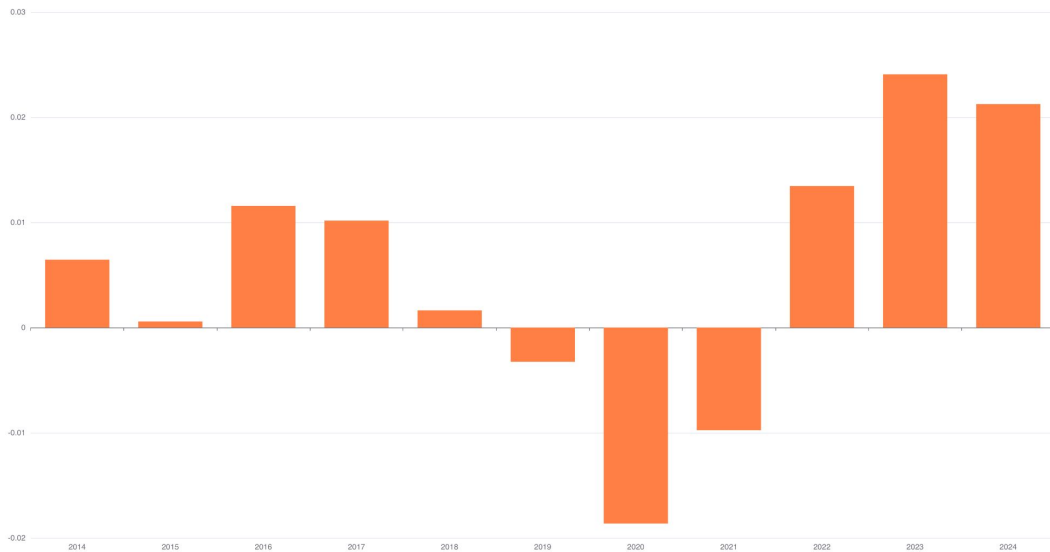


## Correlation of duration and price

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

# Findings

correlation of number of passengers and price by year [barchart]



## Correlation of # passengers and price

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

# Findings

average trip duration by time [barchart]



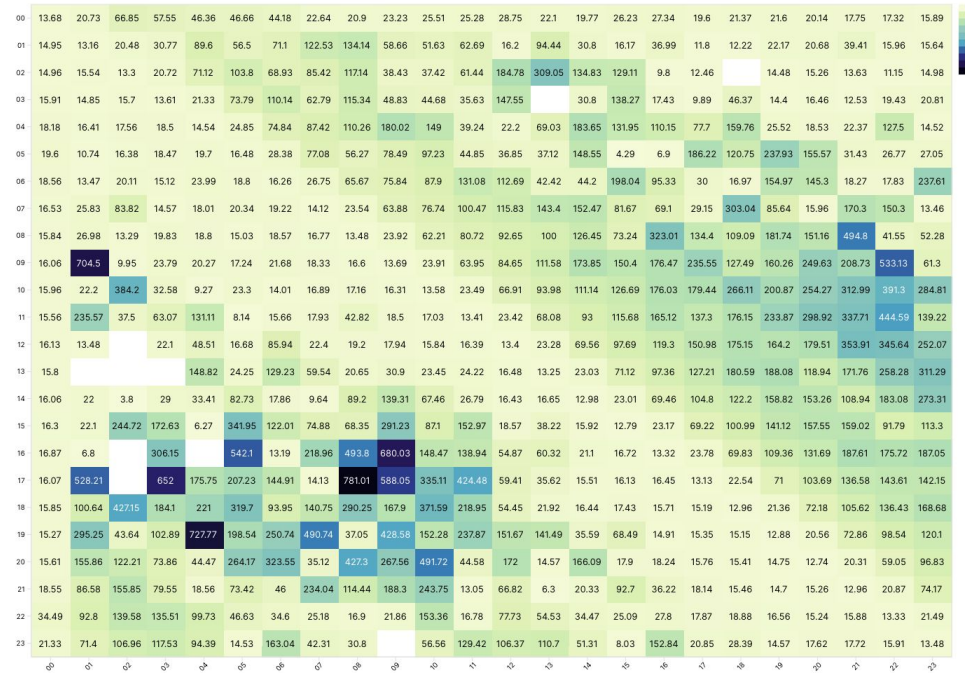
## Average trip duration by time

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



## 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)