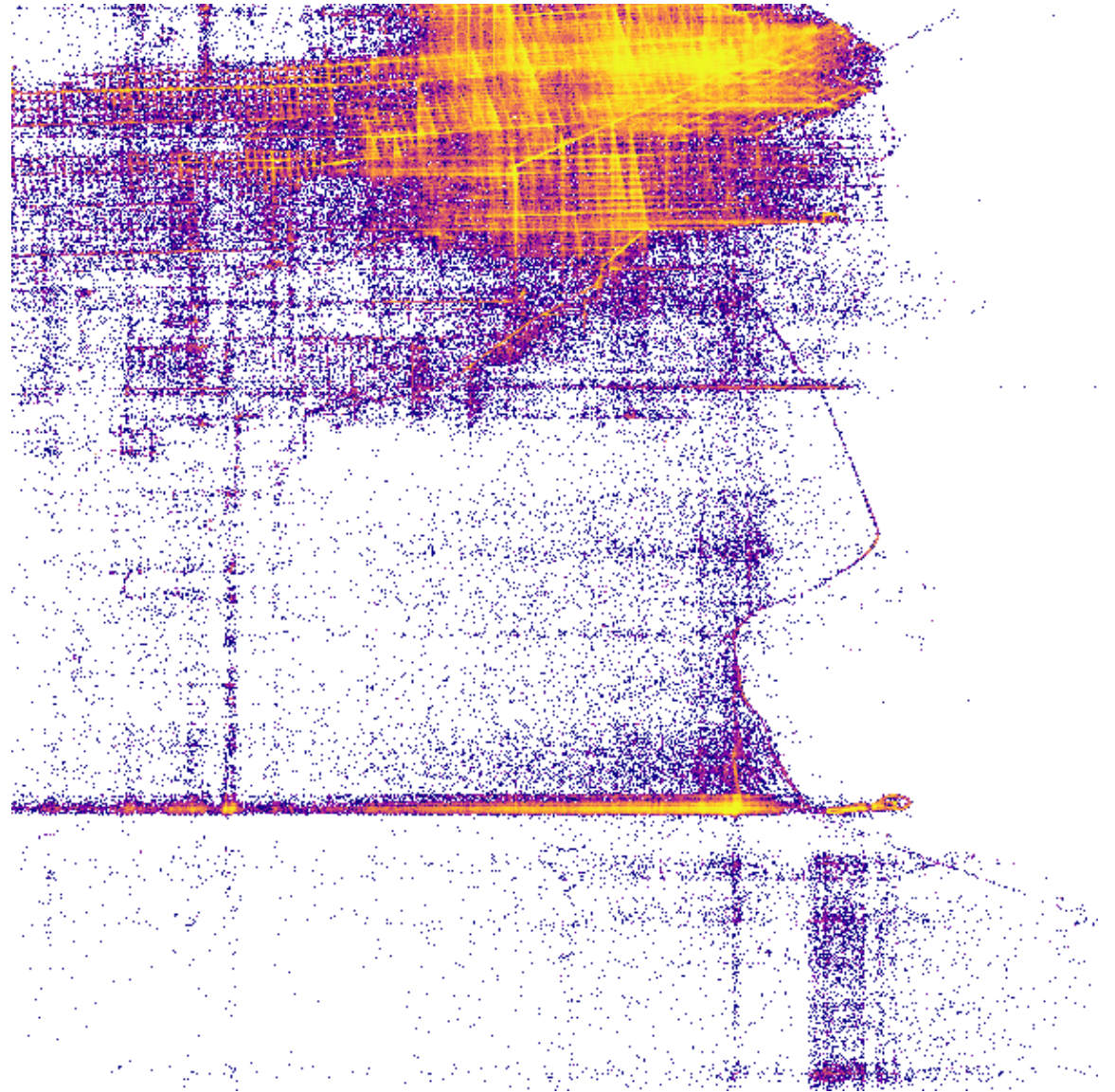# San Francisco Taxi Cabs

Pavel Troshenkov
May 2022

Data and Tasks

- 537 files with taxi mobility traces
- 11 million records with 5 weeks of history
- Only latitude, longitude, occupancy and time are known
- Data provided for a period from 2008-05-17 – 2008-06-10

|	latitude	longitude	occupancy	time	filename
0	37.75153	-122.39447	0.0	1.211034e+09	new_abboip.txt
1	37.75149	-122.39447	0.0	1.211034e+09	new_abboip.txt
2	37.75149	-122.39447	0.0	1.211034e+09	new_abboip.txt
3	37.75149	-122.39446	0.0	1.211034e+09	new_abboip.txt
4	37.75144	-122.39449	0.0	1.211035e+09	new_abboip.txt

There are 3 tasks:

1. To calculate the potential for a yearly reduction in CO_2 emissions caused by the taxi cabs roaming without passengers
2. To build a predictor for taxi drivers, predicting the next place a passenger will hail a cab.
3. Identify clusters of taxi cabs that you find being relevant from the taxi cab company point of view.
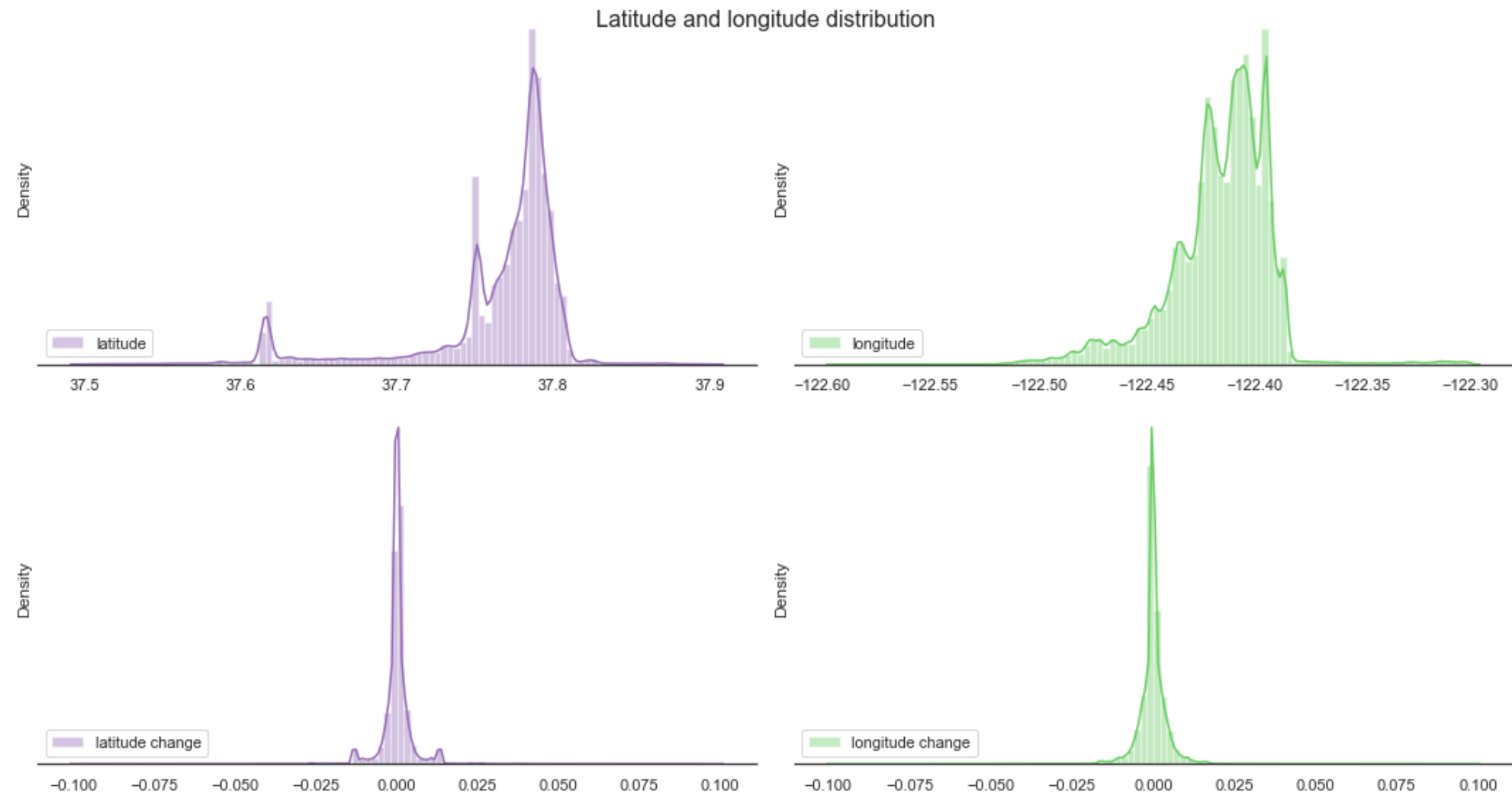
Raw Data Brief Overview

Total number of traces
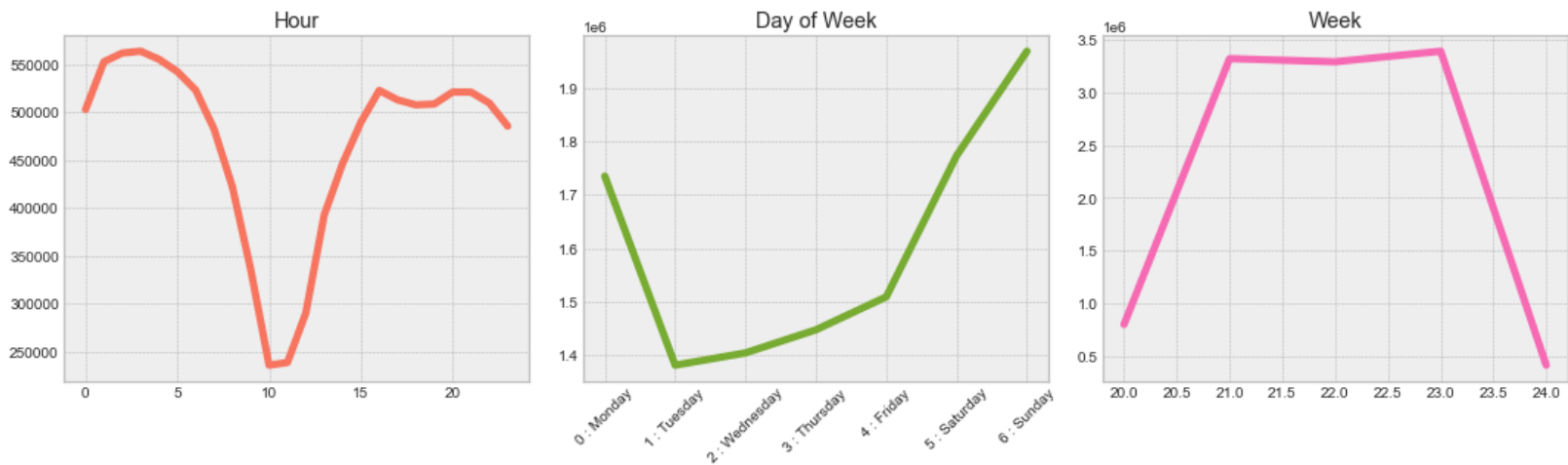
Total traces distribution by occupancy

Raw Data Brief Overview

Latitude and longitude distribution
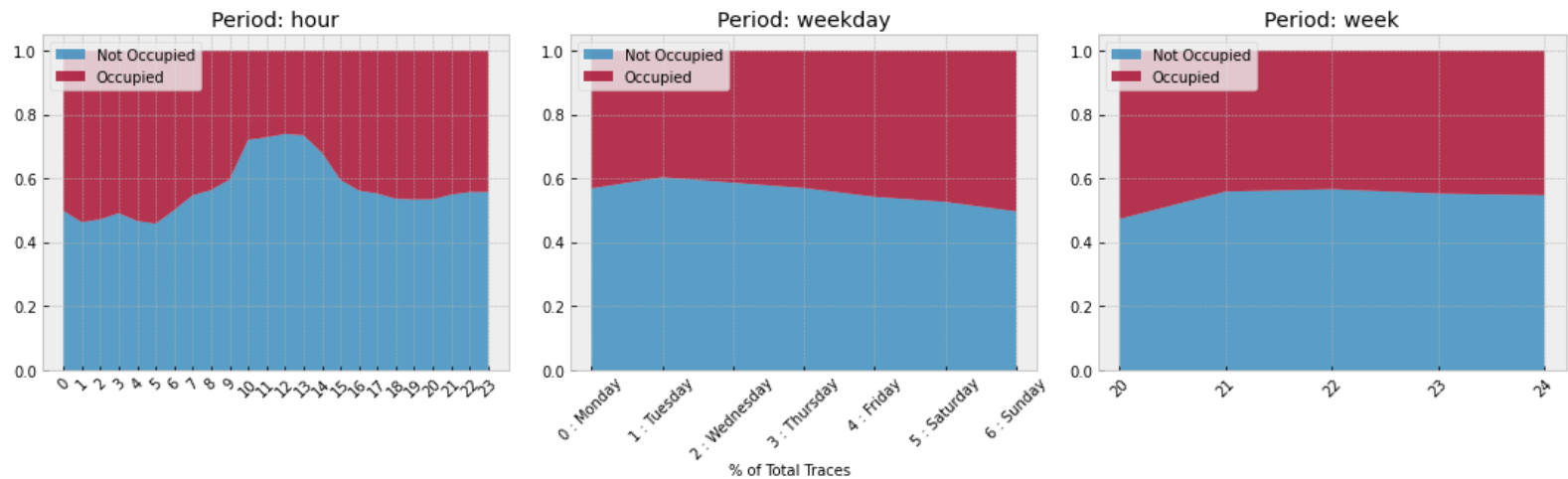
Data Processing

Key assumptions

- trip is defined as a period with same sequential positive occupancy status
- there is always a break between trips thus we are excluding cases when next trips starts immediately after the previous one ends
- outliers ignored
- trip distance was calculated by haversine formula ignoring real landscape

In this way raw data was reduced from 11kk to 900k records

|	filename	trip_number	occupancy	latitude_start	latitude_end	longitude_start	longitude_end	start_time	end_time	trip_distance
0	new_abboip.txt	0	0.0	37.74978	37.75153	-122.39709	-122.39446	1.211034e+09	1.211036e+09	0.347694
1	new_abboip.txt	1	1.0	37.74831	37.75552	-122.41438	-122.39724	1.211036e+09	1.211036e+09	2.246390
2	new_abboip.txt	2	0.0	37.75042	37.76523	-122.42291	-122.41441	1.211036e+09	1.211037e+09	3.578681
3	new_abboip.txt	3	1.0	37.75053	37.75206	-122.43101	-122.42086	1.211037e+09	1.211038e+09	0.996433
4	new_abboip.txt	4	0.0	37.74833	37.77219	-122.43172	-122.41402	1.211038e+09	1.211039e+09	5.123720

EDA

- Most of the trips lay down inside 10 km range and withing 30 minutes
- There are some outliers that can be ignored (~2% of records)
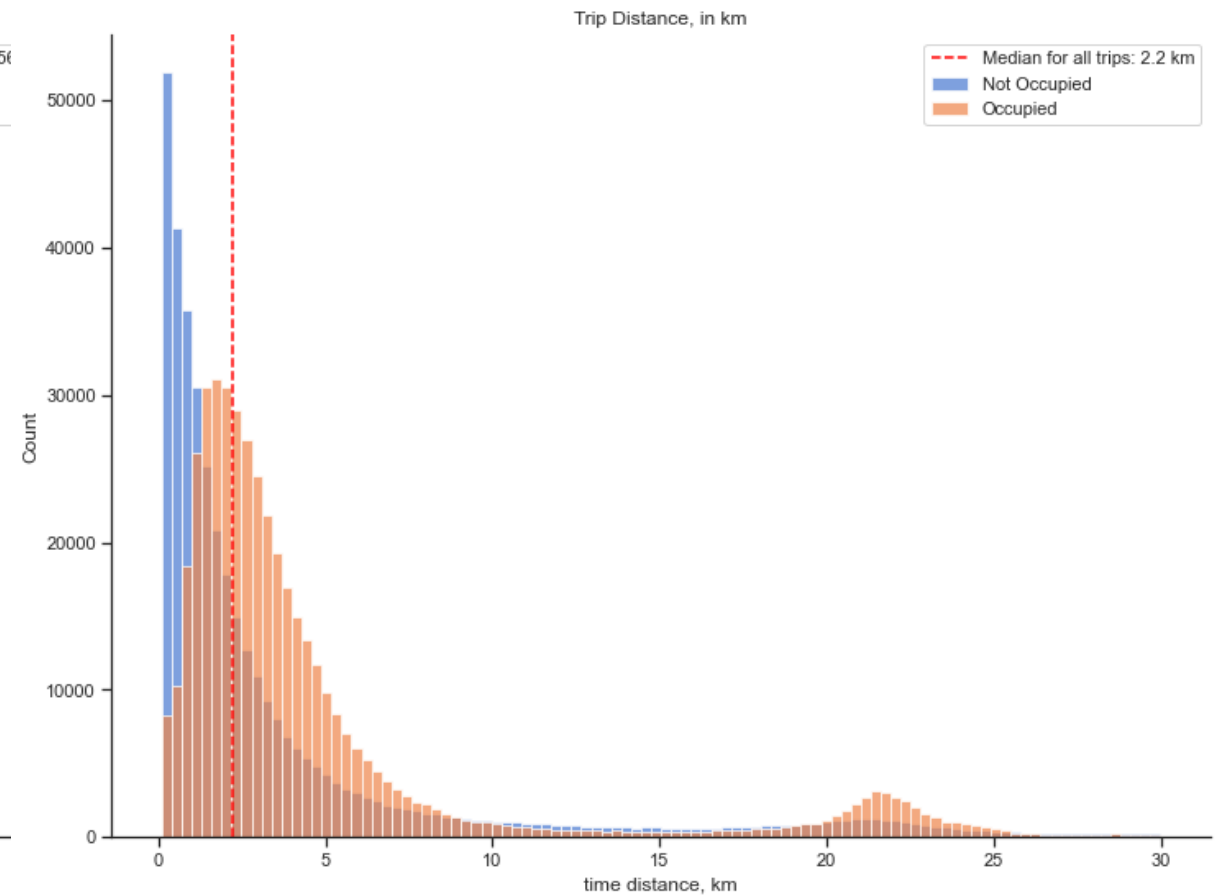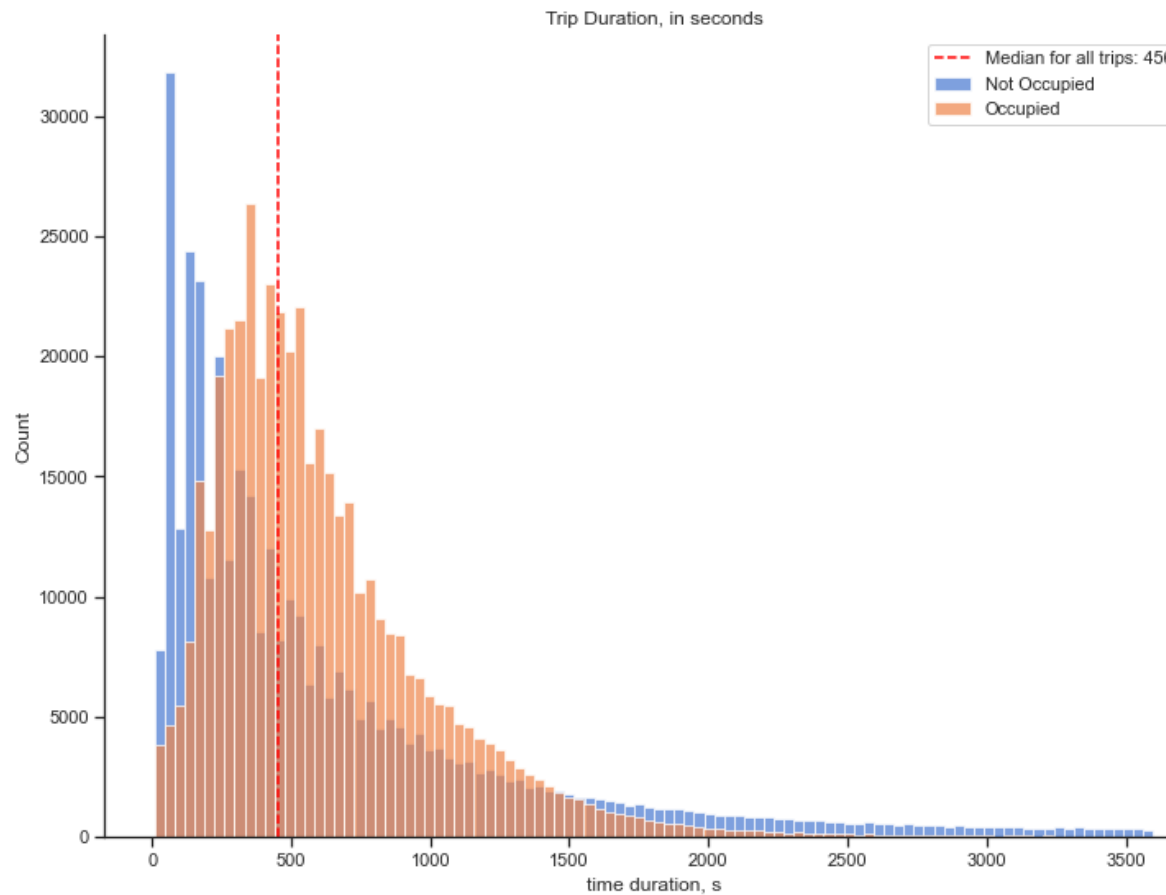
EDA

- Latitude and longitude changes seem to be close to normal distribution
- It makes sense to predict changes of coordinates instead of absolute values
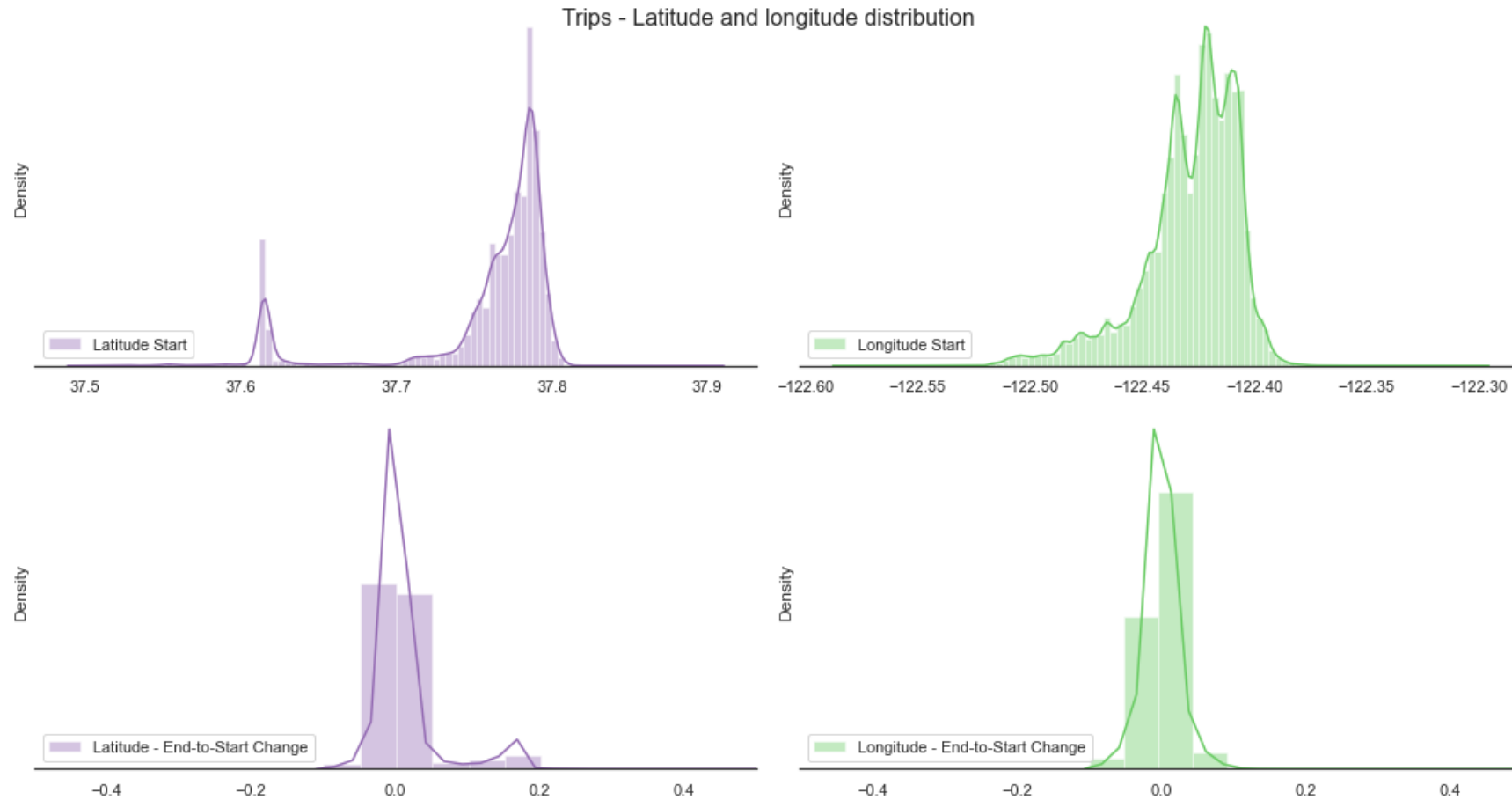
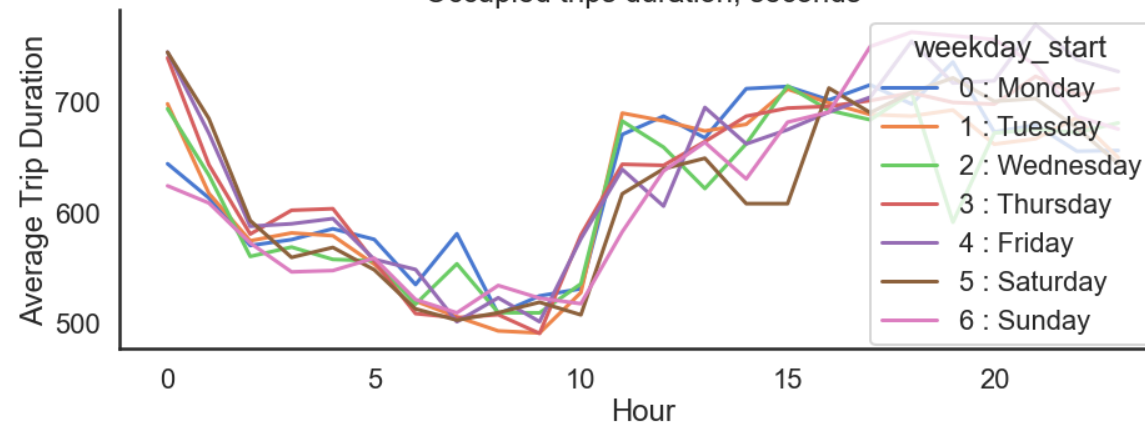Trips - Latitude and longitude distribution

EDA

- Most of the trips happen between 3pm – 5am on weekends
- Trips are longer during busy hours which makes sense (traffic jams, people are going home/work, etc.)

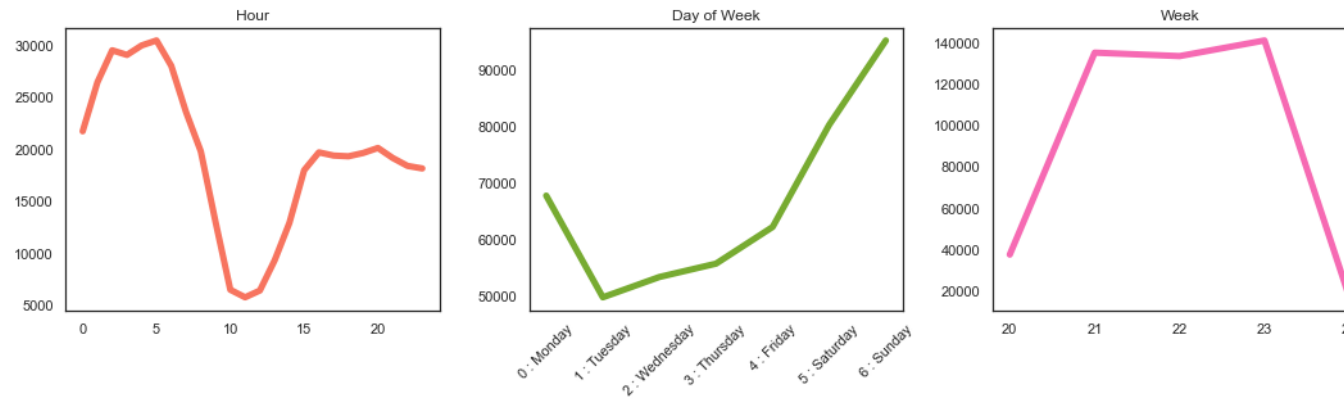Total number of occupied trips
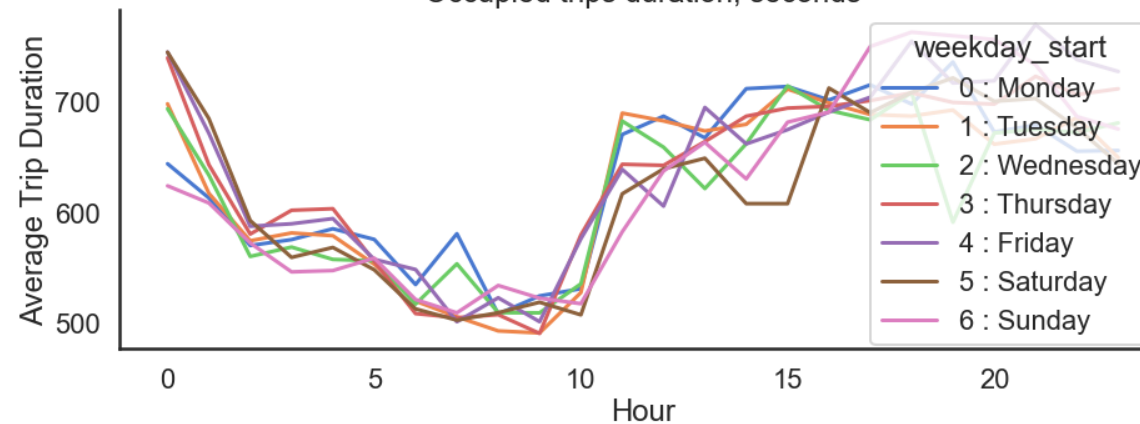

Occupied trips duration, seconds

EDA

- Most of the trips happen between 3pm – 5am on weekends
- Trips are longer during busy hours which makes sense (traffic jams, people are going home/work, etc.)

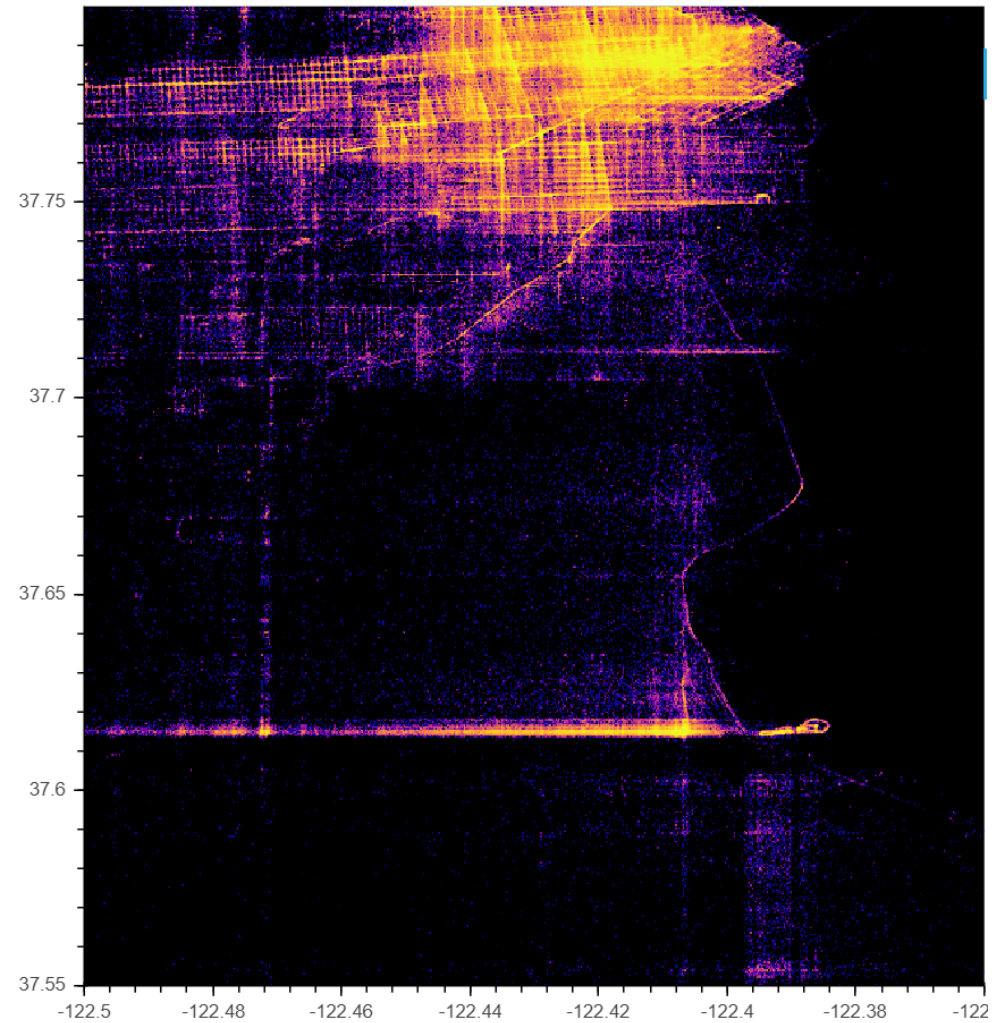Total number of occupied trips
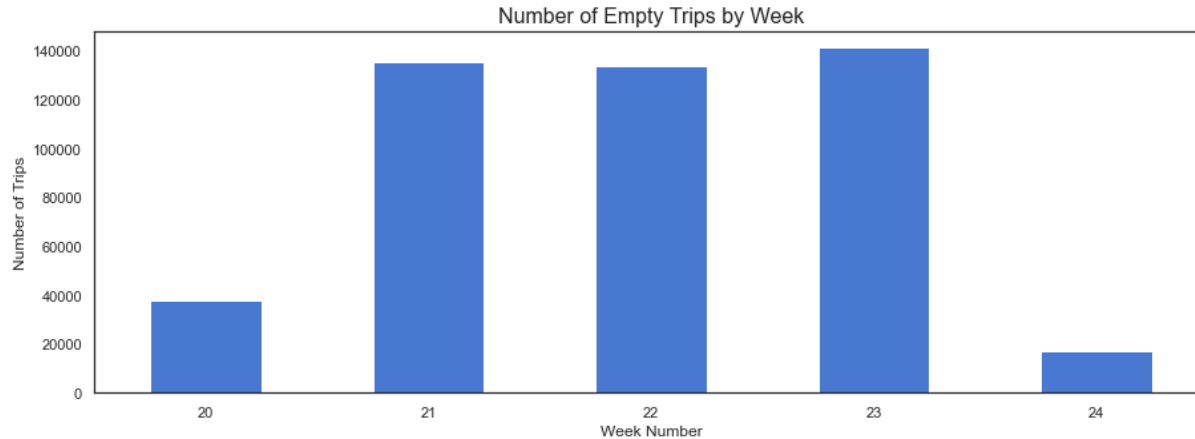

Occupied trips duration, seconds

EDA
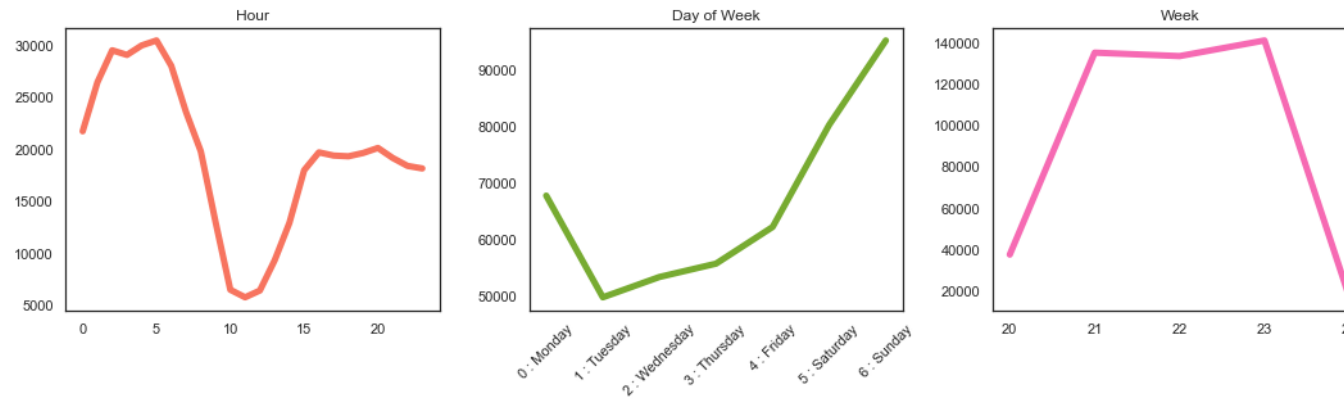
- Most of the trips located around the city center

Task 1. The potential for a yearly reduction in CO2 emissions

- We have 3 complete weeks of data for 500 cars

Number of Empty Trips by Week

- We can calculate total mileage by

Total mileage by Week

- We know that number of gasoline cars drops by 15% per month
- Also, we can consider that number of drivers grows by 1% in the pessimistic scenario[1]
- Another important assumption is how we are considering gaps between trips lasting more that 1 hour

The formula to follow:

$$M_t = M_{t-1} \times GrowthRate \times (1 - ReductionRate)$$

where
M – is a monthly CO2 emission,
Growth Rate – 0% or 1% (depends on scenario),
Reduction Rate – 15% based on this task
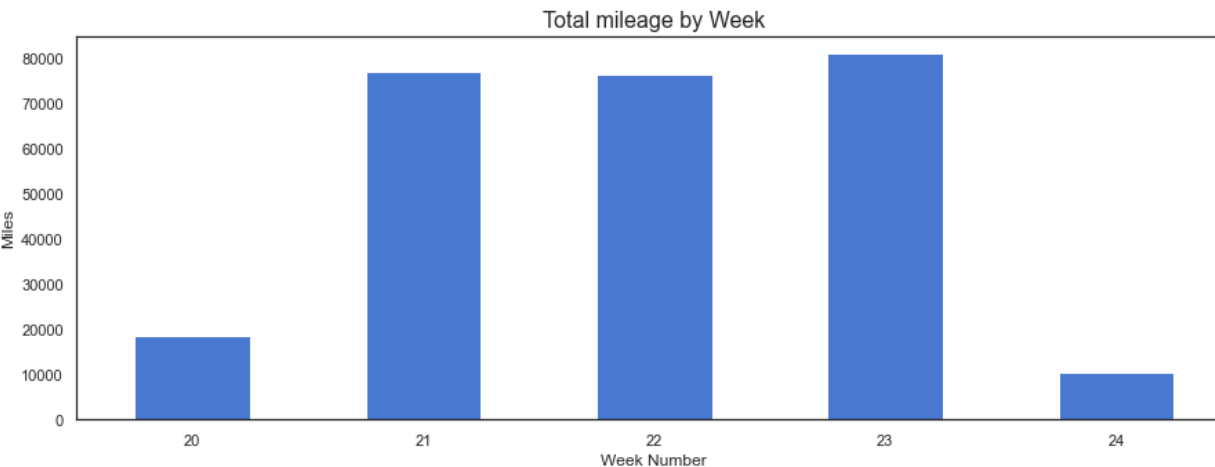
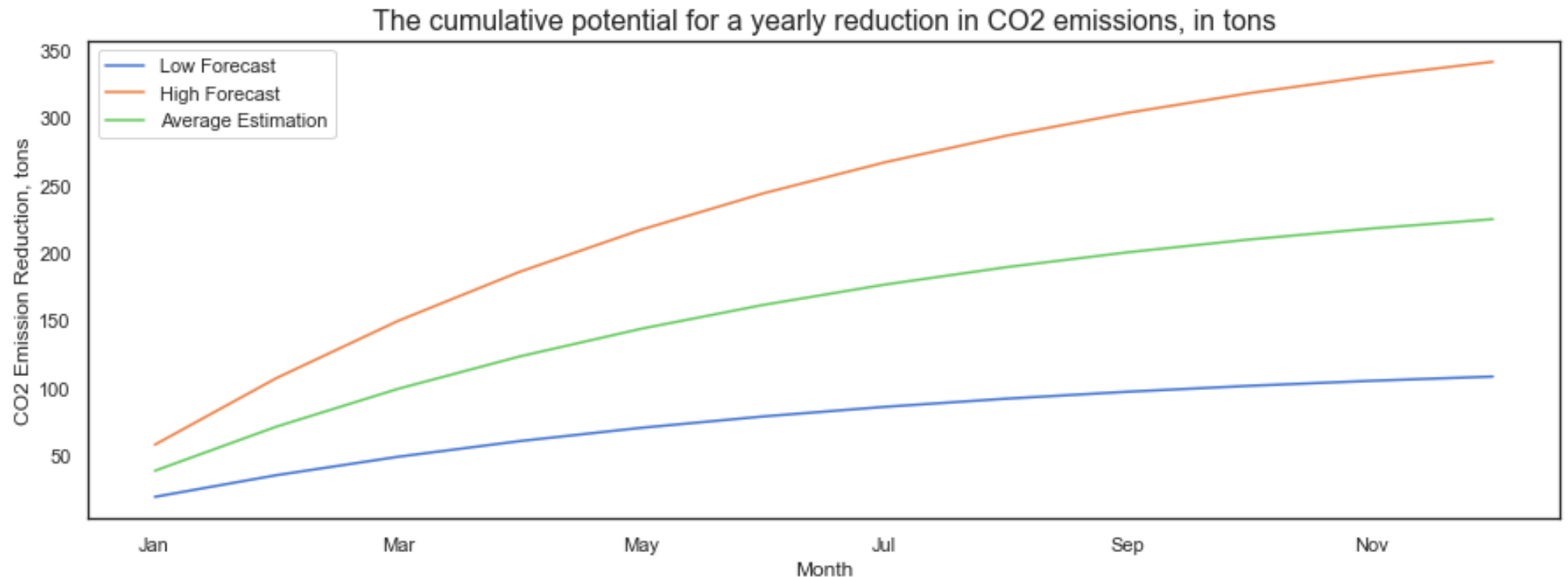# Task 1. The potential for a yearly reduction in CO2 emissions

Example

For complete weeks 21, 22, 23 the average mileage is 77,942 miles with the low emission scenario.
Monthly emission = 4 x 77,942 = 311,769 miles

Given this starting point we can predict CO2 reduction for one month as:

311,769 x (1 - 0.85) x 404 (CO2 grams) = 18.9 tones of CO2

The cumulative potential for a yearly reduction in CO2 emissions, in tons

Task 2. Predict the next place a passenger will hail a cab

Key assumptions

- trip is a sequence of records with the same occupancy status
- driver's shift is a sequence of records separated by more than 5 hours gap
- ML predicts a delta of two sequences with occupancy = 1 status separated by a sequence with occupancy status = 0

What we want to predict

- Longitude change between dropoff and pickup locations
- Latitude change between dropoff and pickup locations

What we want to change

E.g. we can hope that CO_2 emission will be lower if drivers would use ML-based routing

Task 2. Predict the next place a passenger will hail a cab

Key assumptions

- trip is a sequence of records with the same occupancy status
- driver's shift is a sequence of records separated by more than 5 hours gap
- ML predicts a delta of two sequences with occupancy = 1 status separated by a sequence with occupancy status = 0

What we want to predict

- Longitude change between dropoff and pickup locations
- Latitude change between dropoff and pickup locations

What we want to change

E.g. we can hope that CO_2 emission will be lower if drivers would use ML-based routing
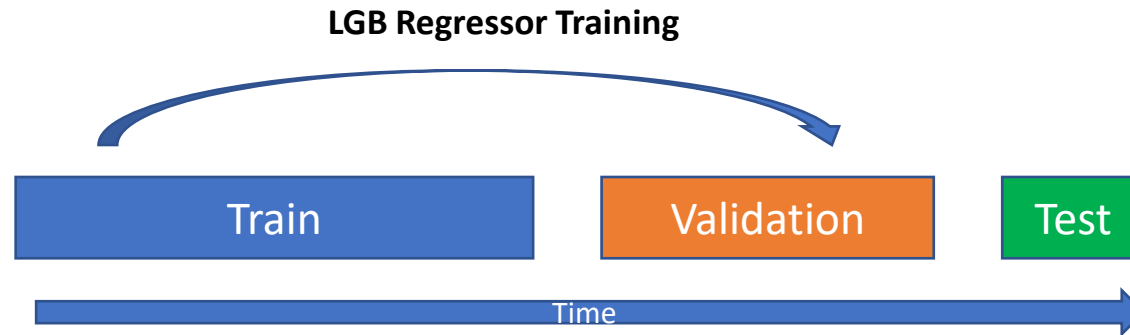
Task 2. Predict the next place a passenger will hail a cab

Features

- Latitude/longitude-based features
    - Previous pickup/drop-off coordinates
    - Previous trip duration / distance in different slices
- Driver's features
    - Average statistics
    - Number of trips before the current one
- Date and time
    - Week
    - Day of Week
    - Hour

Validation scheme
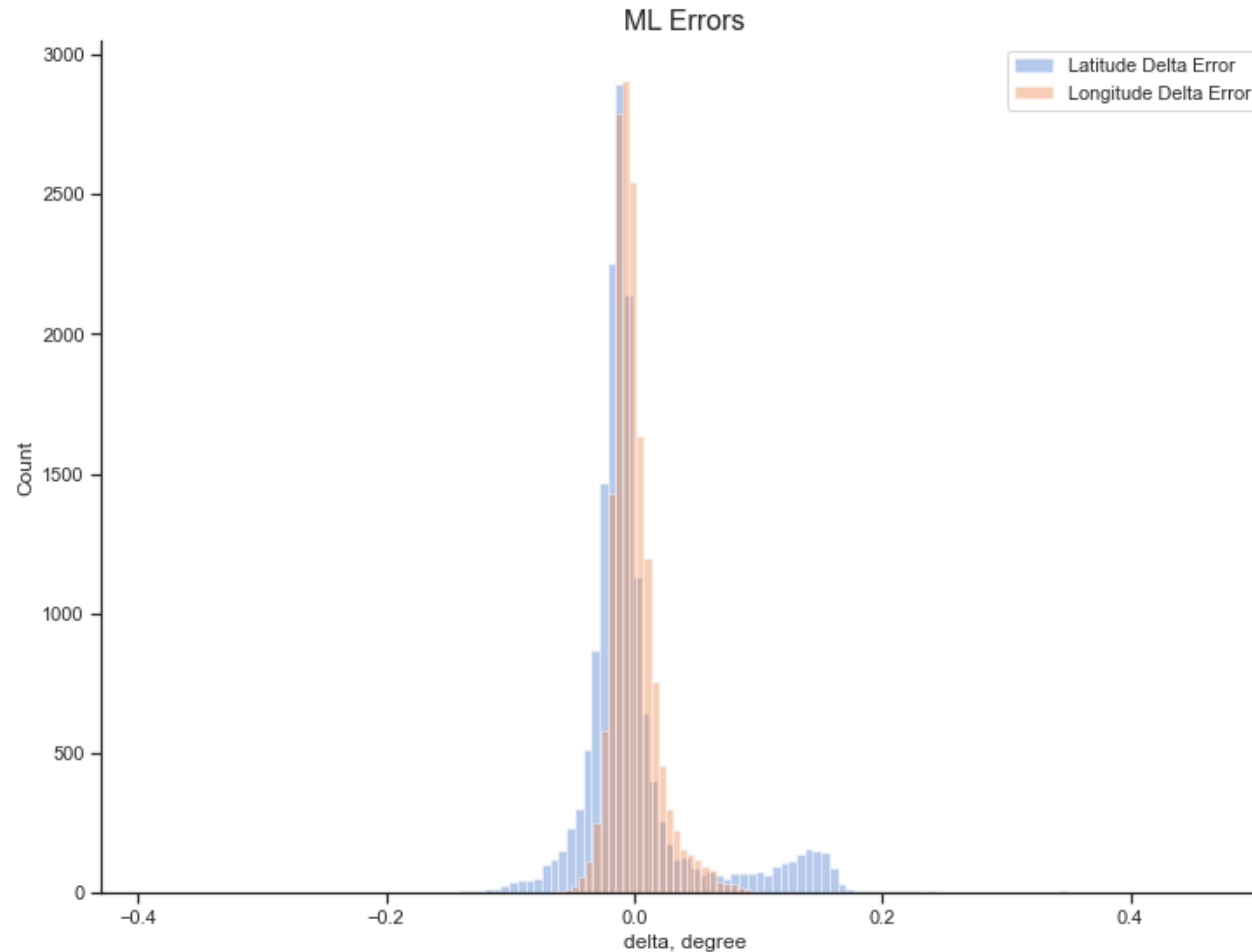
- Time-series validation

LGB Regressor Training

Train	Validation	Test

Time

Loss

- RMSE

Business Impact
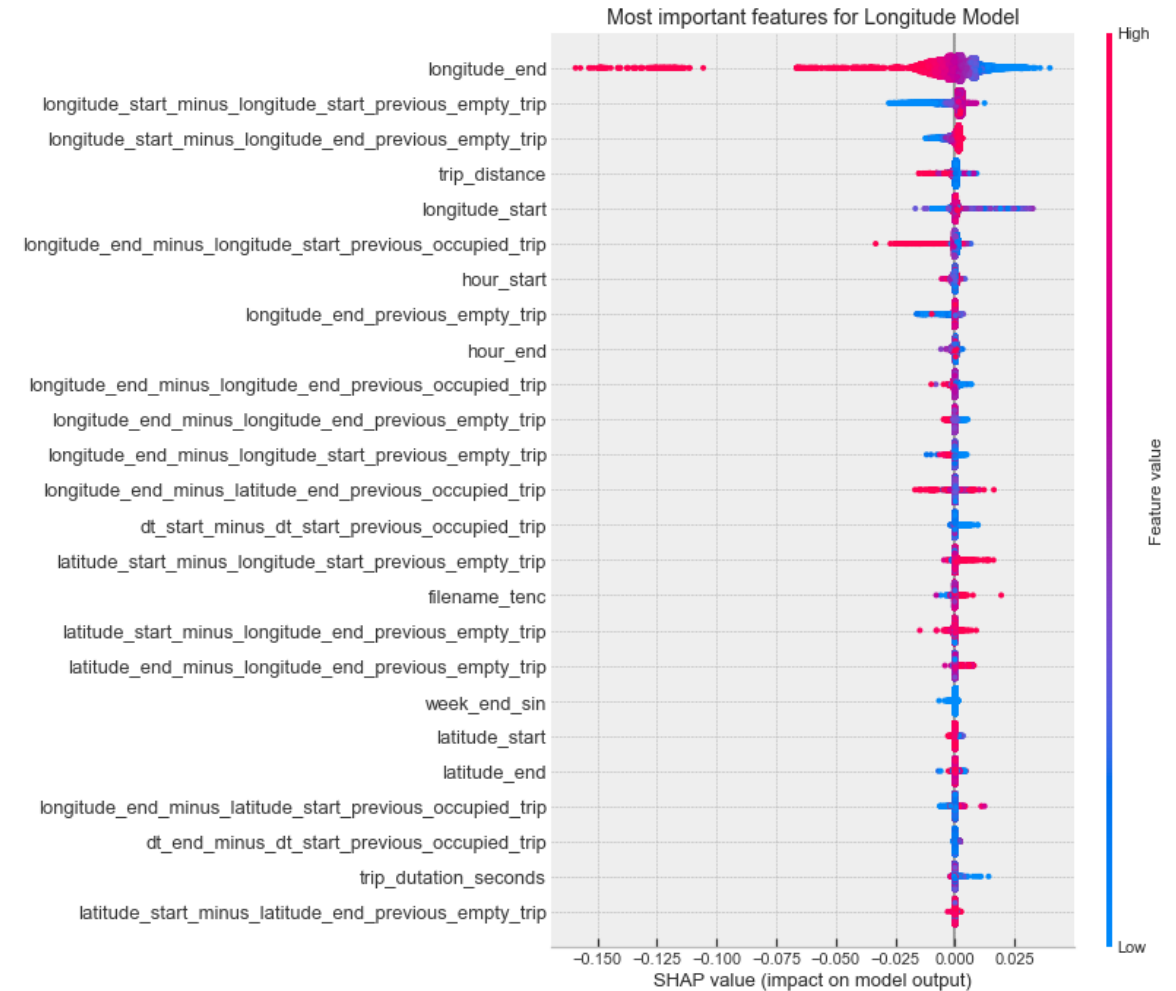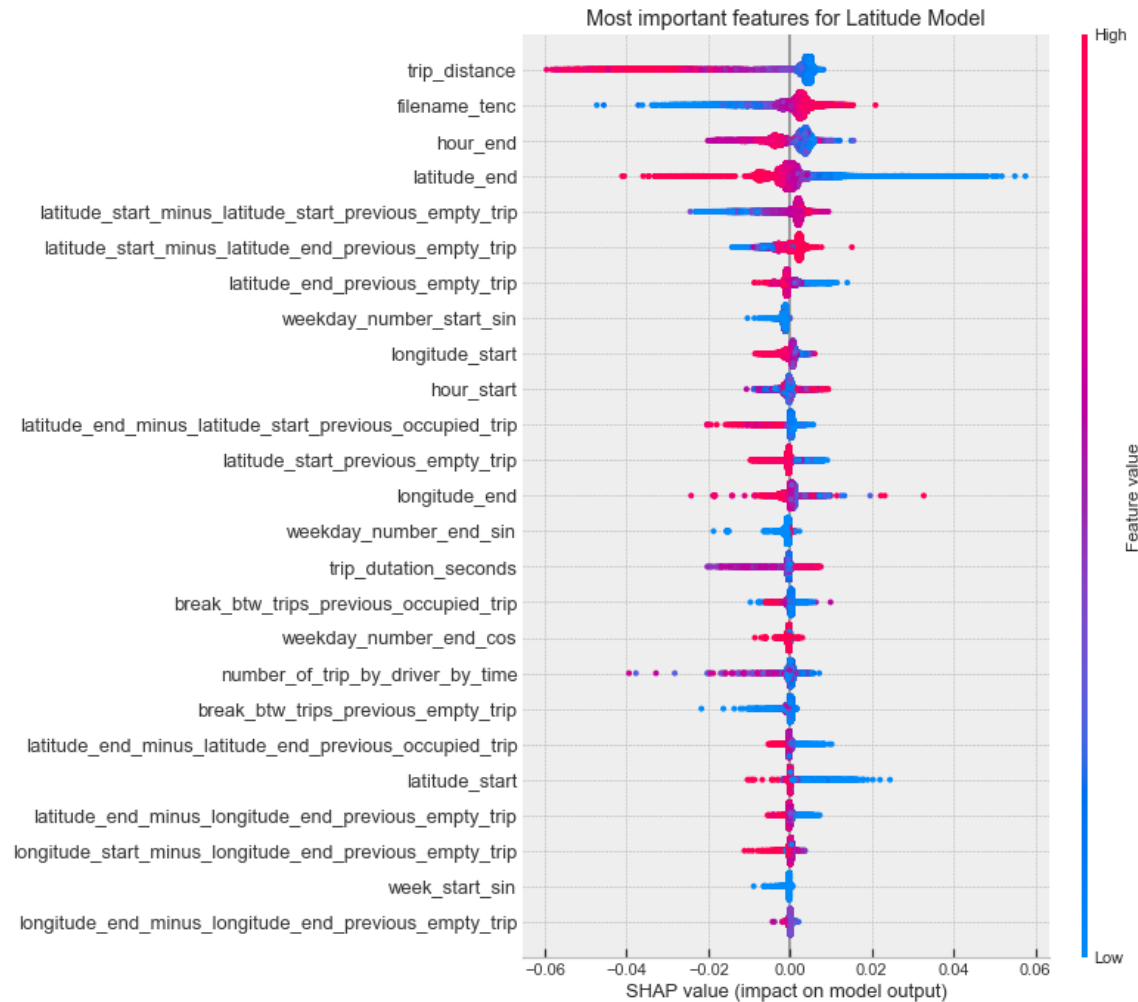
- Distance to get new passenger, in kilometers

Task 2. Predict the next place a passenger will hail a cab

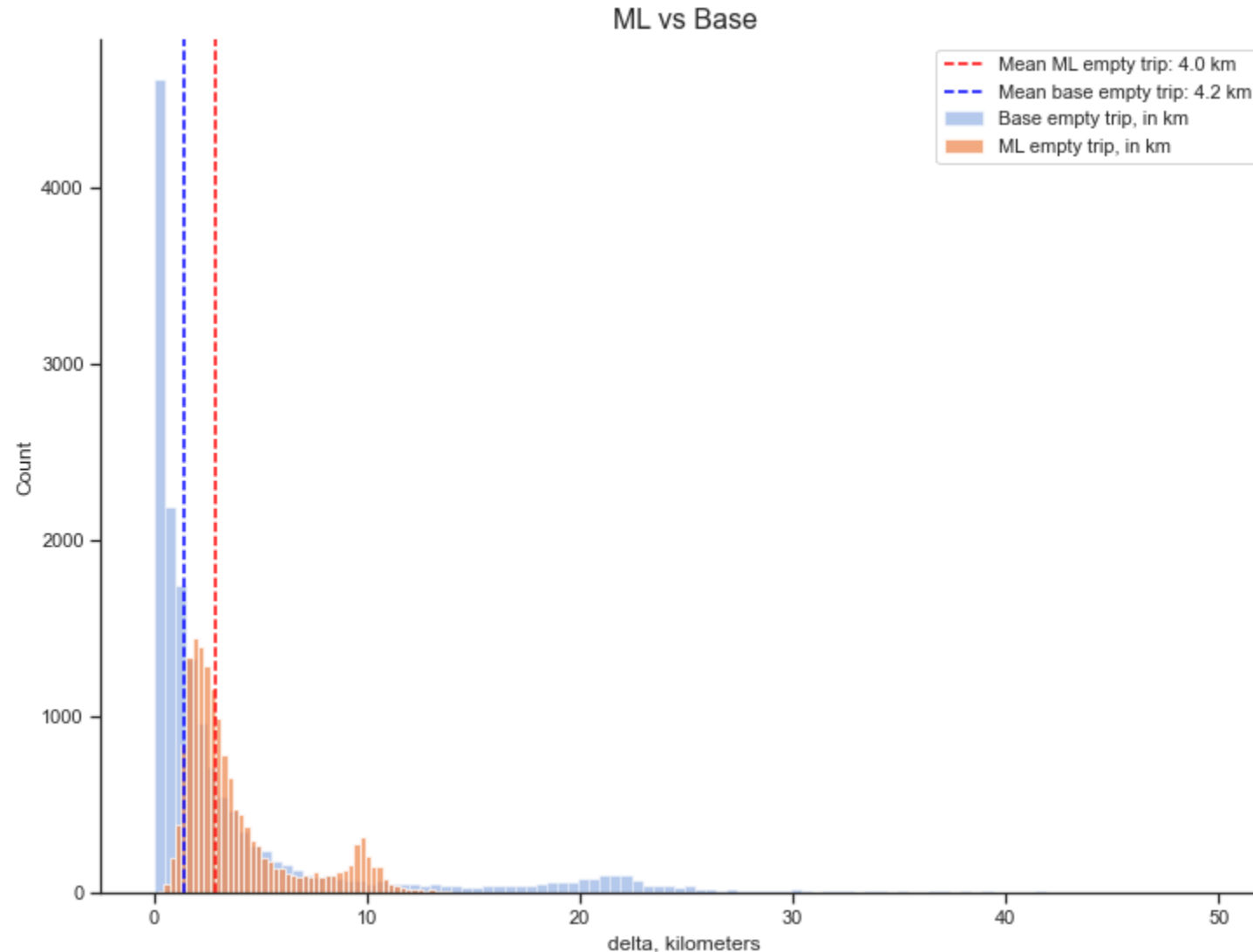- Longitude residuals are equally distributed across 0
- Latitude is harder to predict
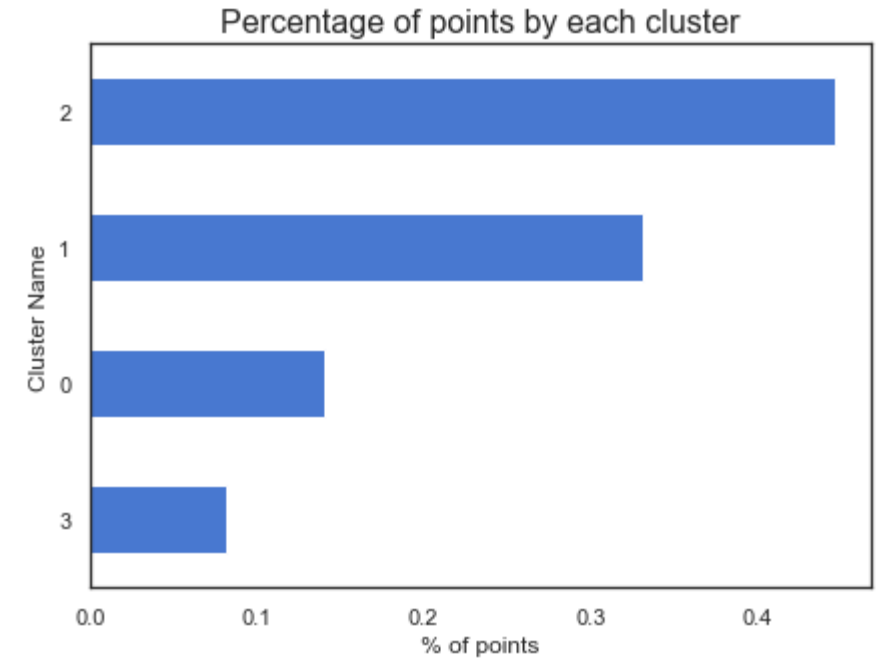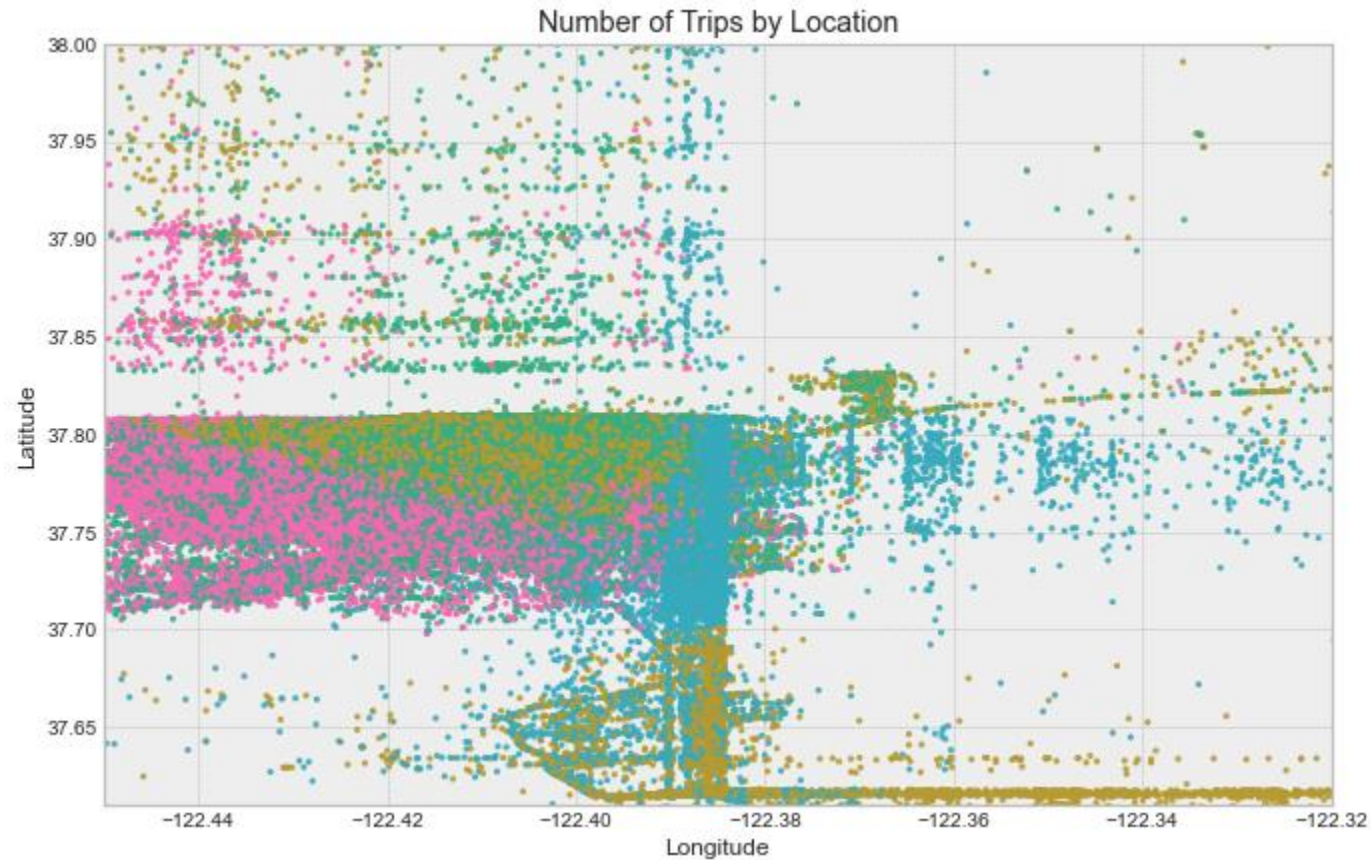
Feature Importance

Task 2. Predict the next place a passenger will hail a cab

Potential reduction of empty mileage is about 7% (based on predictions for trips after 2008-06-08)
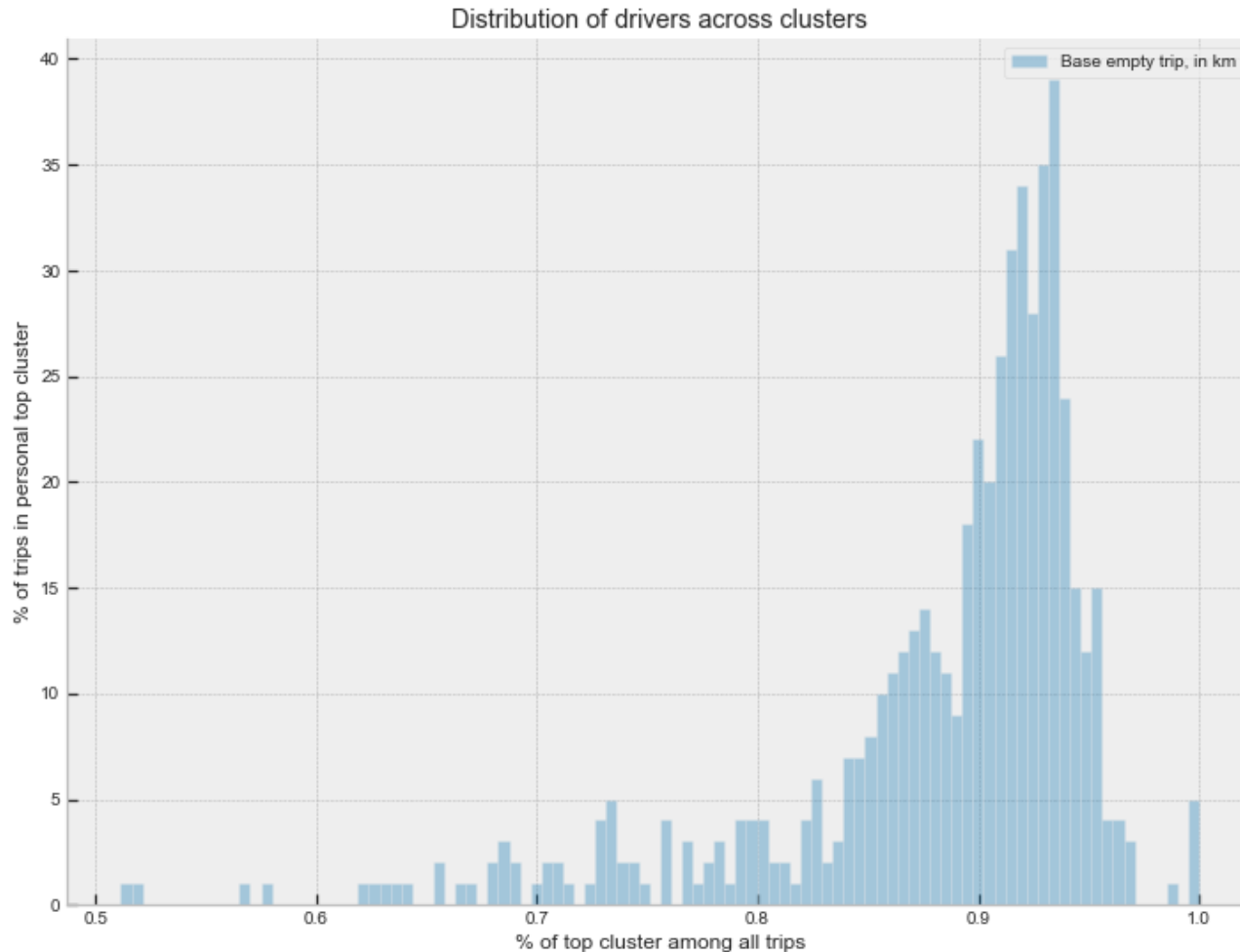

ML vs Base

Task 3. Identify clusters of taxi cabs

- KMeans clustering based on latitude and longitude (both pickup and drop-off) for all trips by 4 clusters

Task 3. Identify clusters of taxi cabs

- Every driver has 4 clusters
- But 90% of trips by driver belong to only one cluster
- This one cluster can be used as a main one in each case

filename	cluster	occupancy	occupancy_share	rank
new_abboip.txt	0	0.0	0.000000	4.0
new_abboip.txt	1	3.0	0.003135	3.0
new_abboip.txt	2	893.0	0.933124	1.0
new_abboip.txt	3	61.0	0.063741	2.0

Questions

Thank you!