NYC Taxi Trip Duration Prediction Report

1. About Data and Competition

In this competition, Kaggle is challenging you to build a model that predicts the total ride duration of taxi trips in New York City. Your primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

1.1. Data Fields Description

- **id** - a unique identifier for each trip
- **vendor_id** - a code indicating the provider associated with the trip record
- **pickup_datetime** - date and time when the meter was engaged
- **dropoff_datetime** - date and time when the meter was disengaged
- **passenger_count** - the number of passengers in the vehicle (driver entered value)
- **pickup_longitude** - the longitude where the meter was engaged
- **pickup_latitude** - the latitude where the meter was engaged
- **dropoff_longitude** - the longitude where the meter was disengaged
- **dropoff_latitude** - the latitude where the meter was disengaged
- **store_and_fwd_flag** - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- **trip_duration** - duration of the trip in seconds

1.2. Weather Data for New York City

This data is not part of the datasets provided by the competition. Somebody shared it with people enrolled in the competition for sake of help and I found it useful.

1.3. Weather Data Content

Weather data collected from the National Weather Service. It contains the first six months of 2016, for a weather station in central park. It contains for each day the minimum temperature, maximum temperature, average temperature, precipitation, new snow fall, and current snow depth. The temperature is measured in Fahrenheit and the depth is measured in inches. T means that there is a trace of precipitation.

2. EDA Findings Summary

2.1. The Target Feature (trip duration)

After applying the log scale to it, it looked like a gussian distribution - Figure 1 .
log scale(base 10) was applied to better show the distribution as outliers exist.

Most of the trips are between about 3 to 52 minutes.

We can see the prescence of outliers by noticing the long right tail in the distribution and the huge values that extend beyond the upper whisker in the box plot below.

And based on this, I applied some more investigations to better select a threshold for removing part of these outliers with keeping quite conservative to not loose much data.

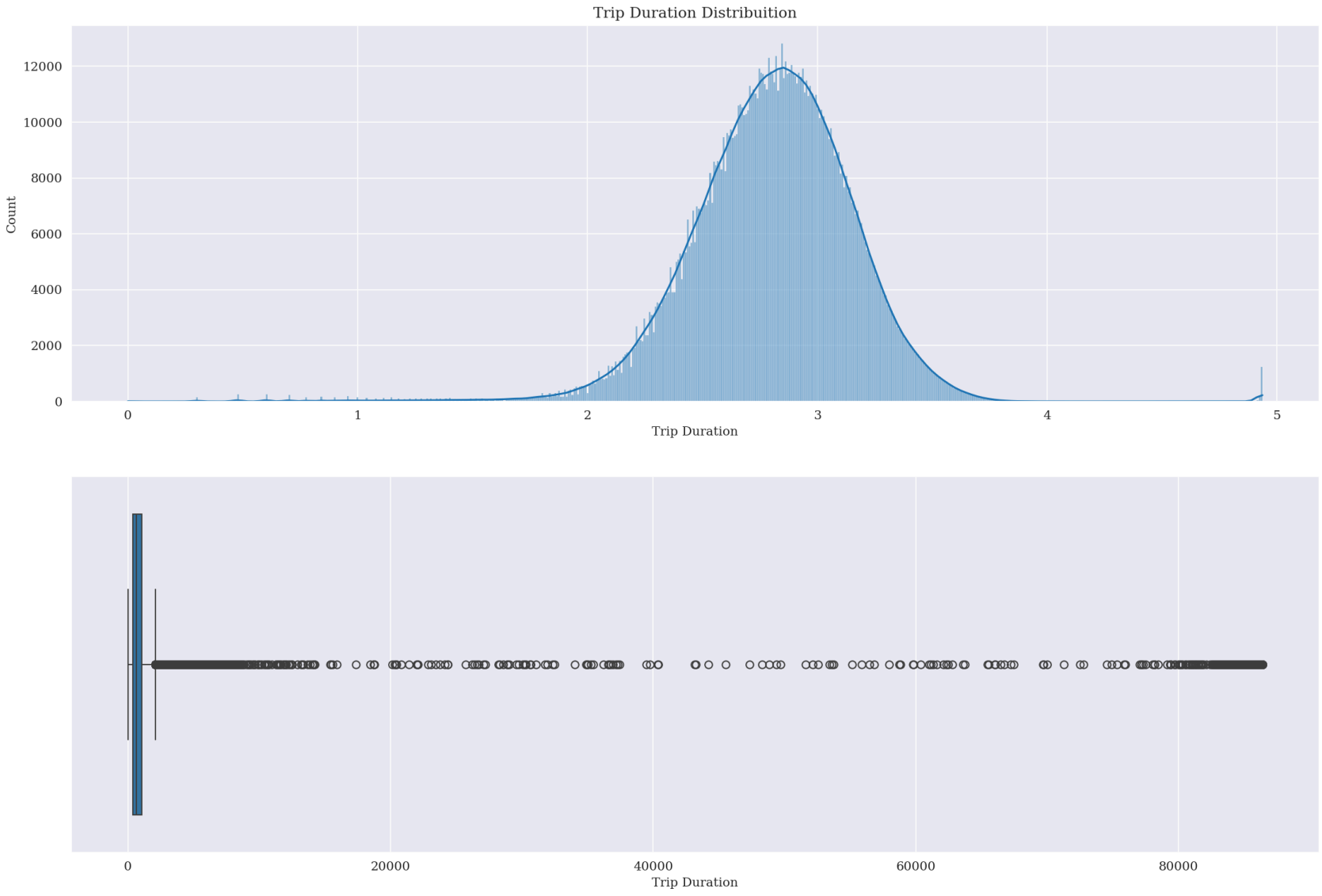


Figure 1

I plotted the distribution of trips above 1 hour and I had this interesting pattern - Figure 2 .

Number of trips from the beginning of the histogram(above 1 hour) and up to the point 3.8 on the x-axis (1.75 hour) was 7911 .

between the two points 3.6 and 3.8 (1h to about 1.75h) there are about 9800 trips.
It's not a small number in absolute but if compared to the total trips number it's too small.

it's not reasonable to have such long taxi durations in NewYork in my opinion. but as I said, they are small number compared to tha total. so, they may be special cases like long distance trips while so heavy traffic or mistakes in data.

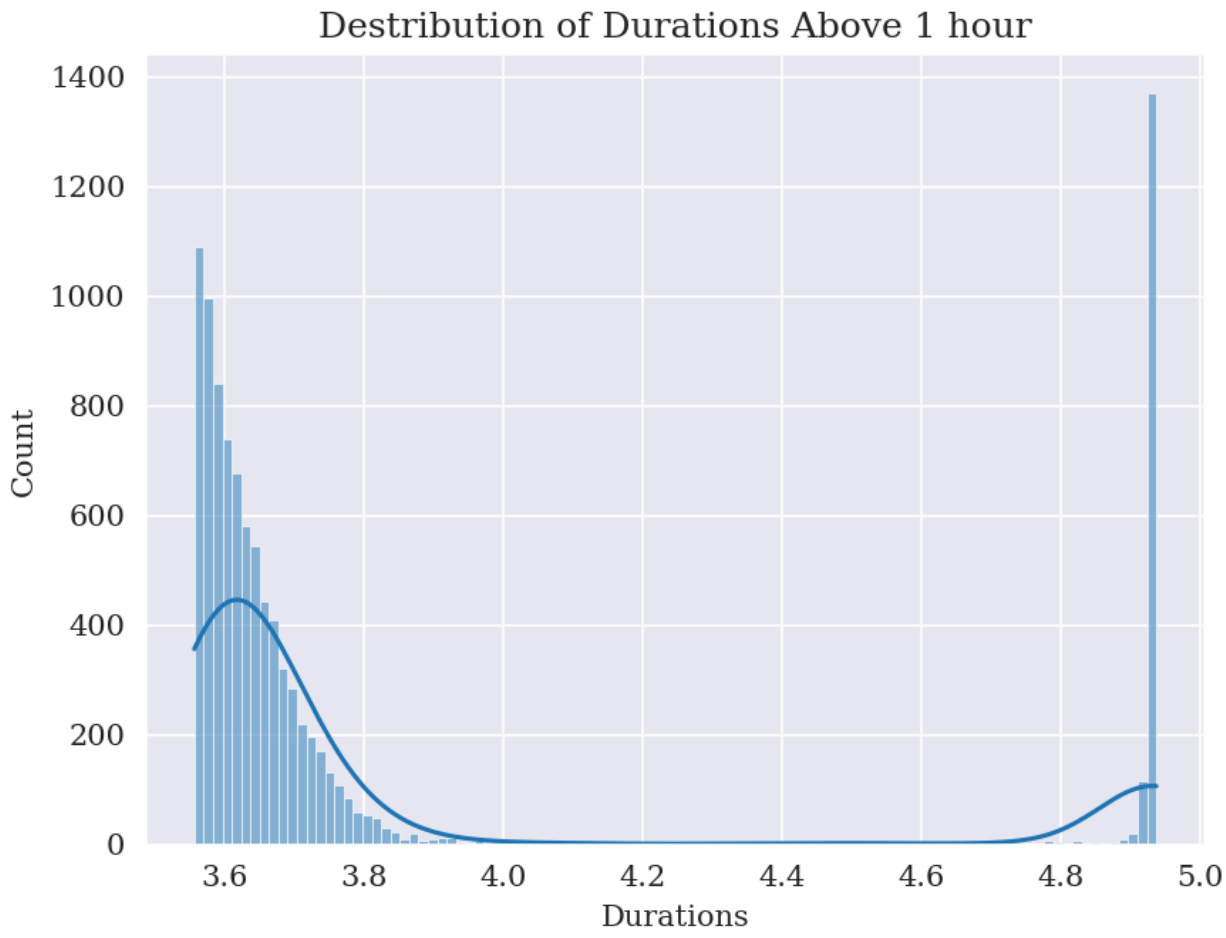


Figure 2

2.2. Discrete Features

The majority of trips with 5 and 6 passengers are with vendor 2 (the left plot at Figure 3).
This may be because vendor 2 vehicles' have bigger size or vendor 2 offer vehicles dedicated for such number of passengers.

As seen in the right plot at Figure 3, there no noticeable trend in the average trip duration between different passenger counts.

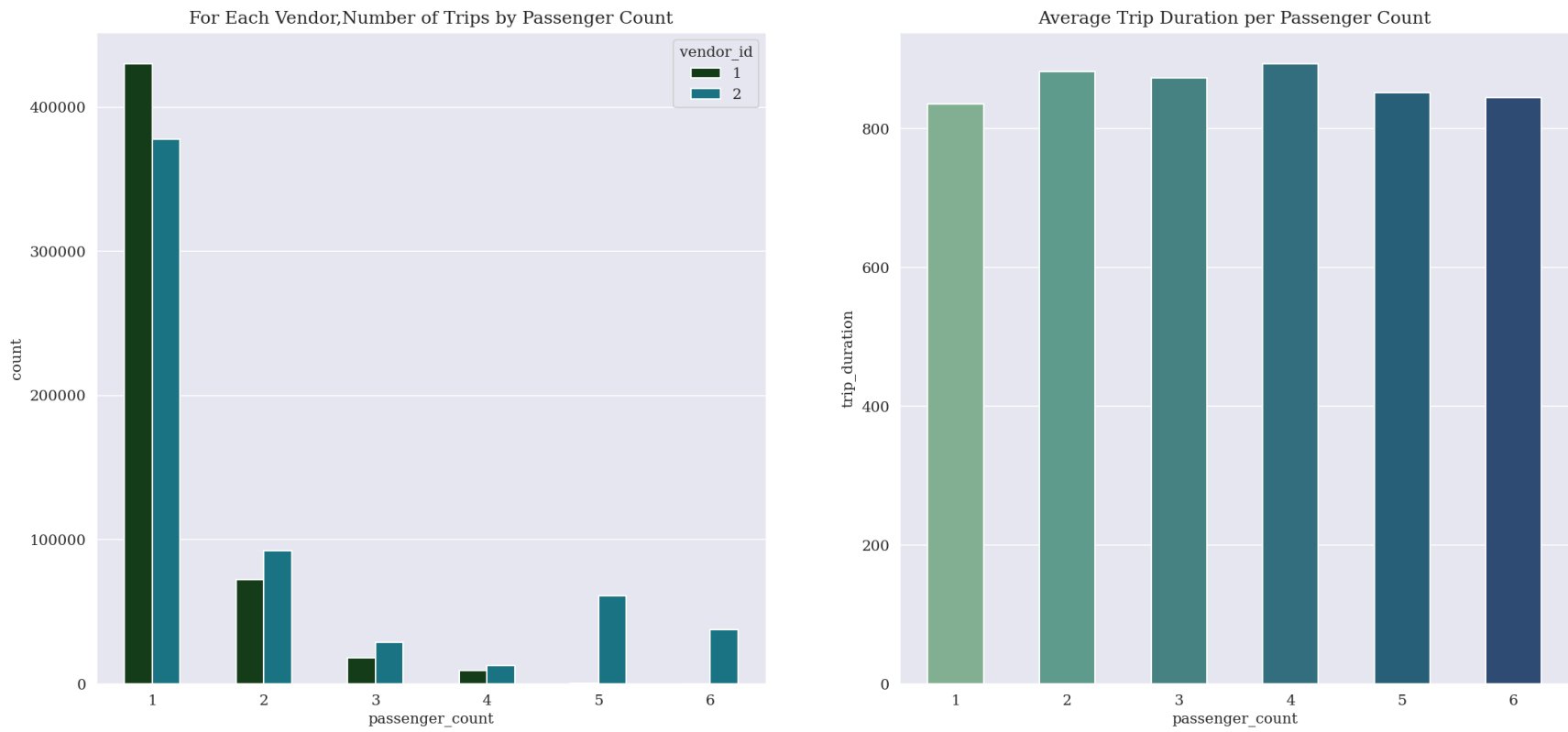


Figure 3

2.3. Weather Features

I found that temperature change has its effect on the average duration.
As seen at Figure 4, as temperature increases, the average duration increases.
We may assume that people tend to ride taxies (or even their own cars) when it's hot. Hence, the traffic gets more busy.

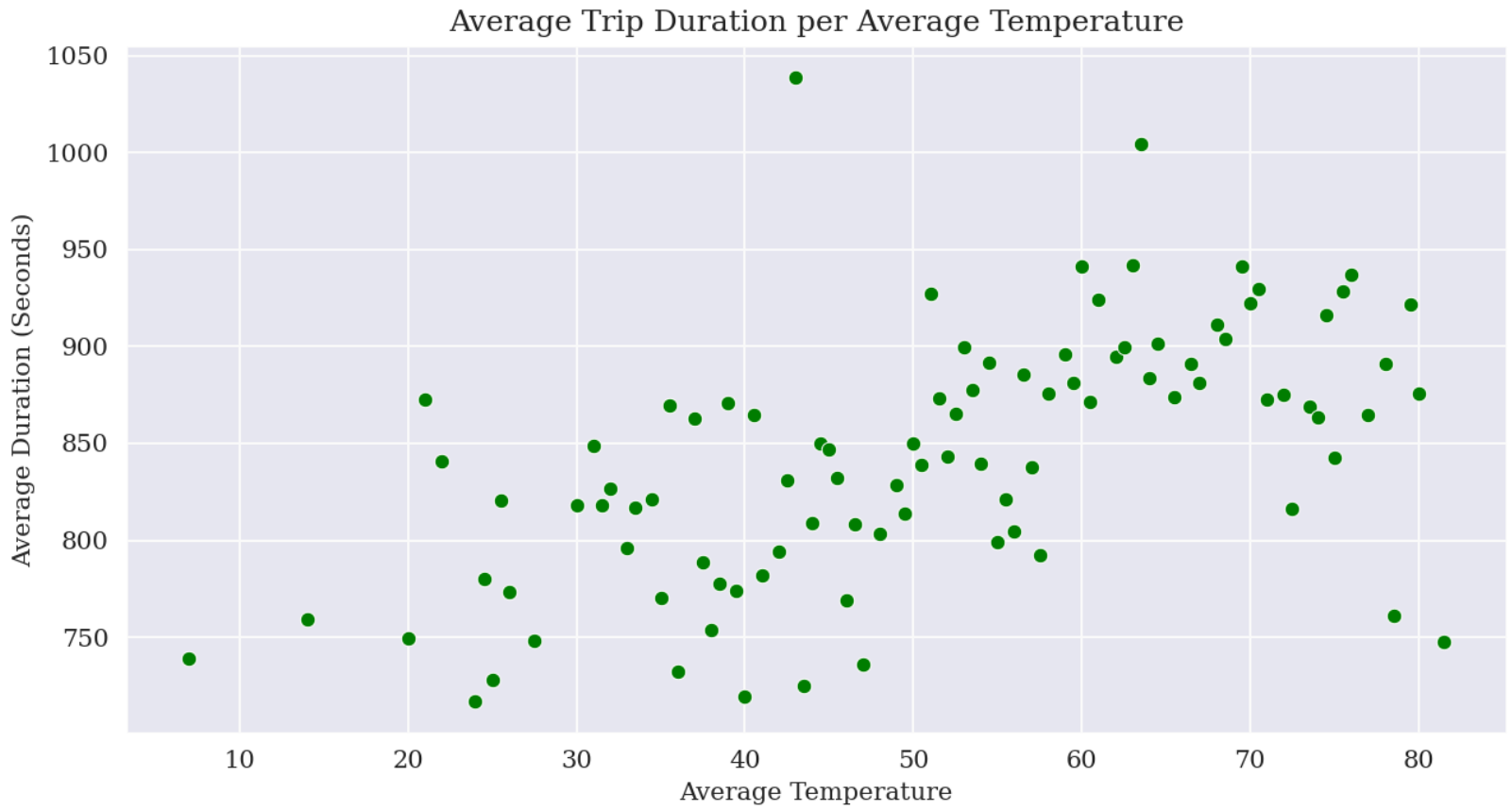


Figure 4

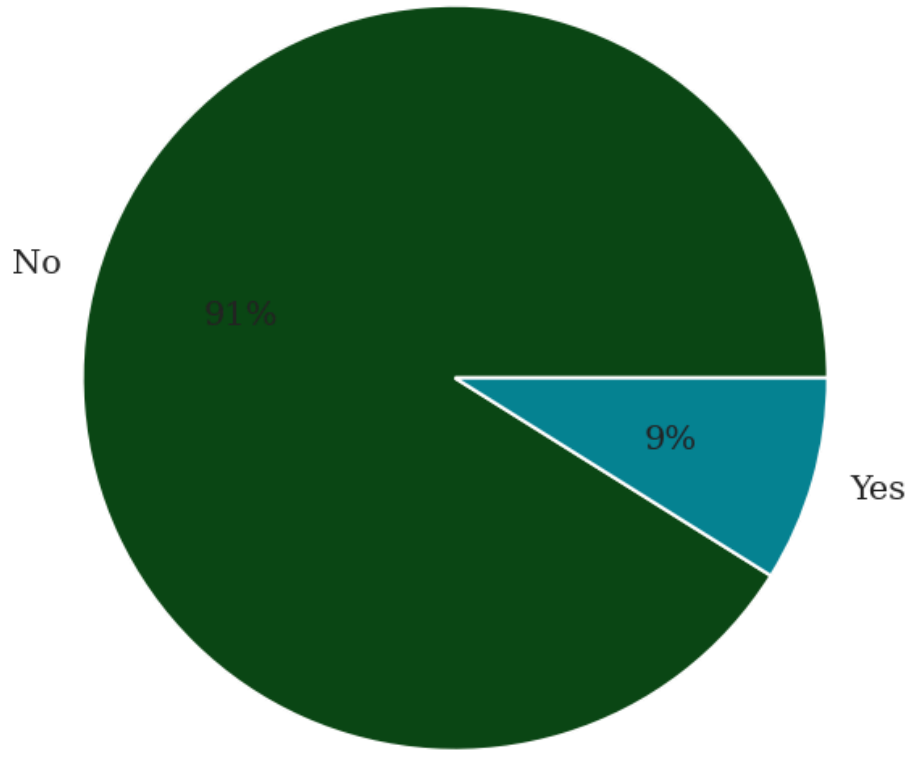


Figure 5

Initial investigation about snow effect on the average duration time didn't reveal any remarkable trend.

Number of snowy days are much less than the other days - Figure 5 . but both has almost the same median and range as shown in Figure 6

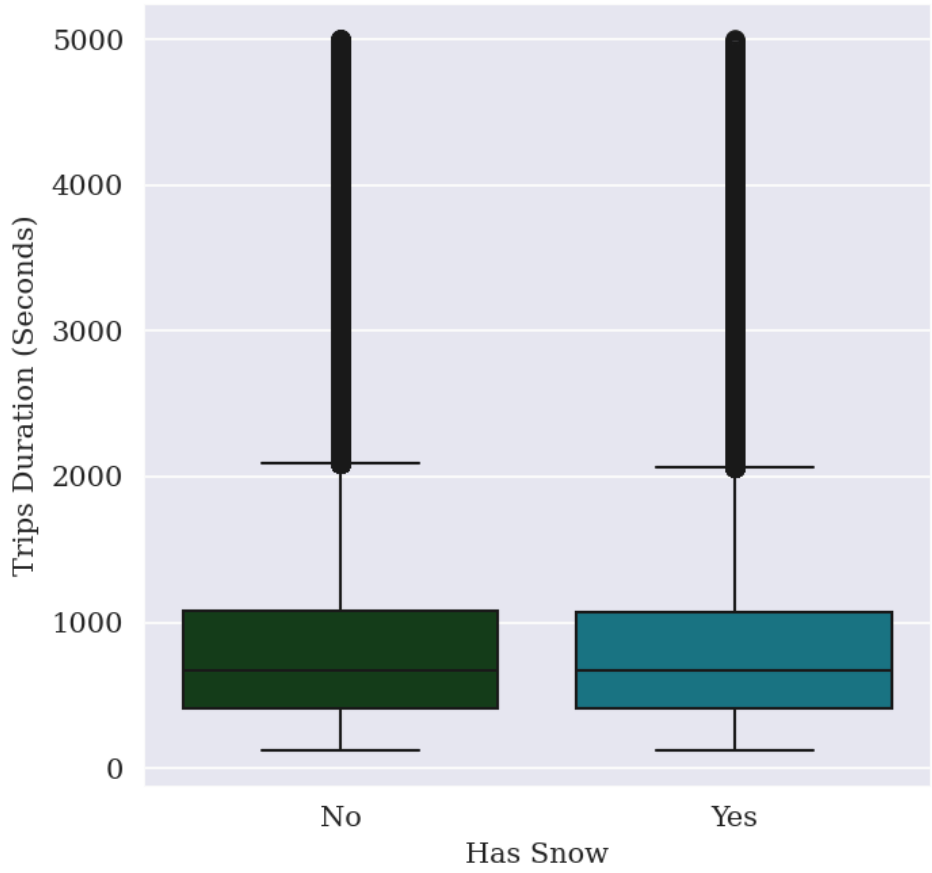


Figure 6

2.4. Date-Time Features

First, the average duration for each week day shows a clear pattern - Figure 7.
starting from Monday,The average duration is about 14 minutes and continue in an increasing pattern till Friday. which is the last official working day.

Then, we notice a decline in the average duration through the days of the weekend,Saturday and Sunday.

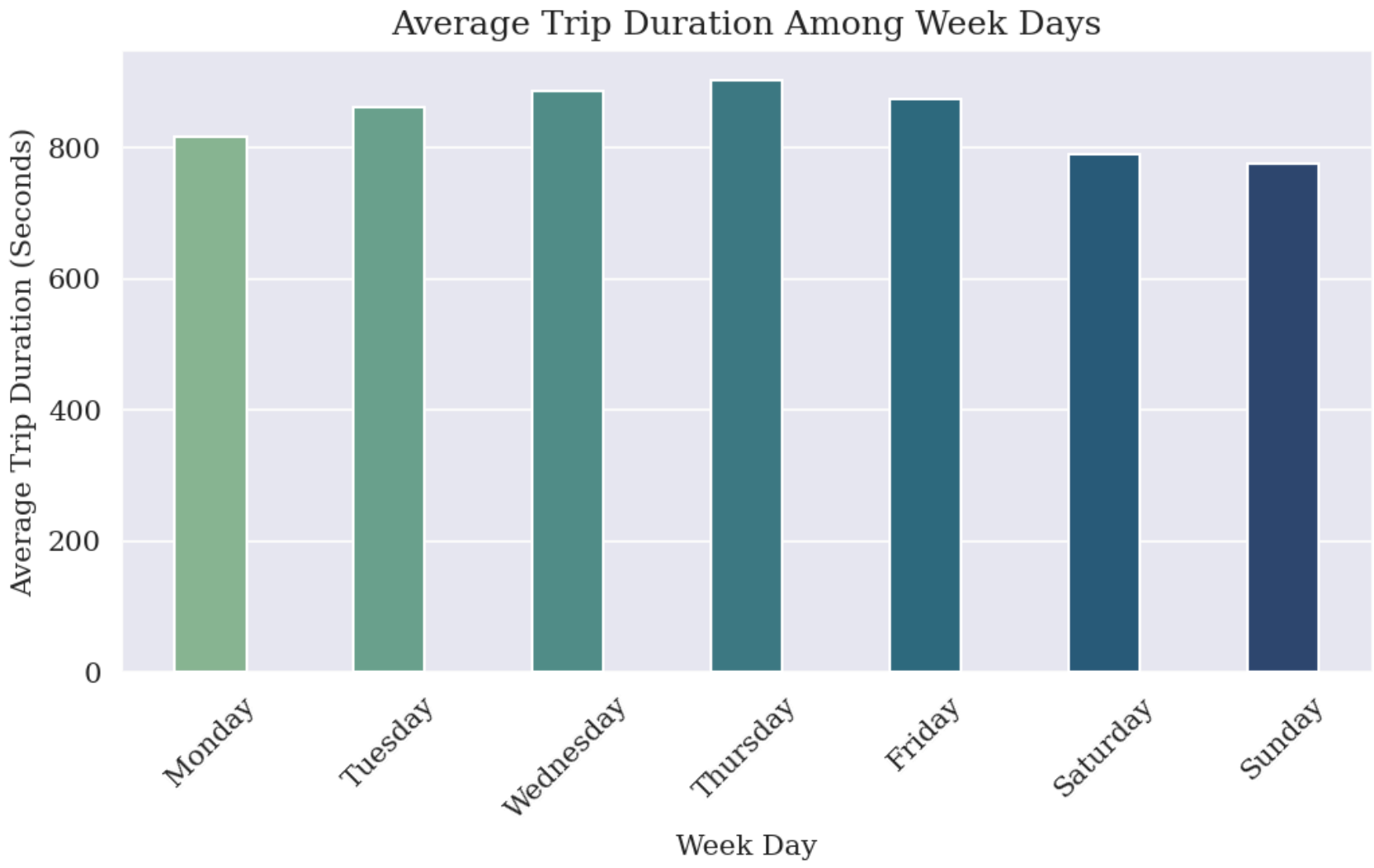


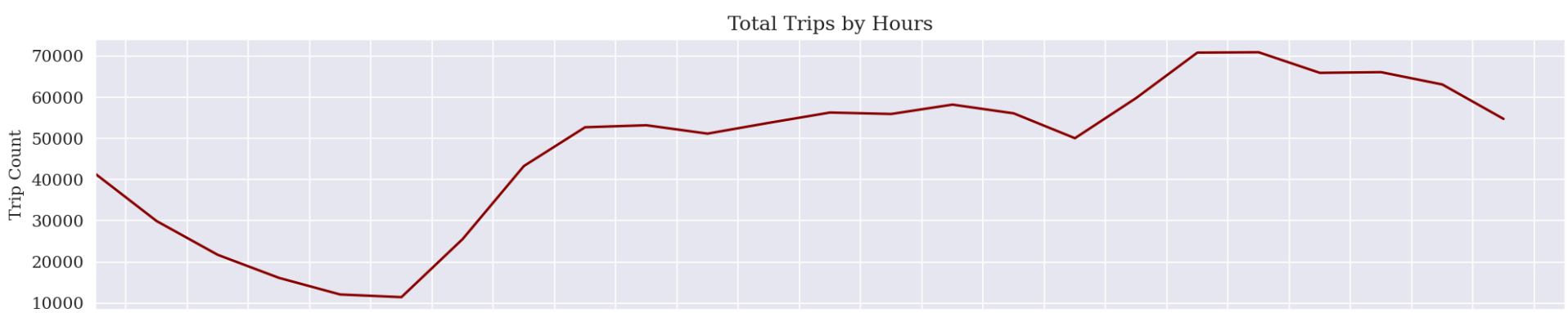
Figure 7

From Figure 8 below, where we investigate the trips activity on an hourly level, we can notice that the trips number at night are much less than in morning.

There is an increasing pattern in the trips number from the night to morning. And the peak hours are from 6 PM to about 10 PM at the days from Monday to Friday.
Specifically, the Thursday at 9 PM has the most rides count with 11117 trips.

And while the work days (Monday to Friday) have few trips at the beginning hours of the day (12 AM and up), Saturday and Sunday are a little busy at these hours in a remarkable way.

These insights help us understand how people use taxis throughout the day and week.



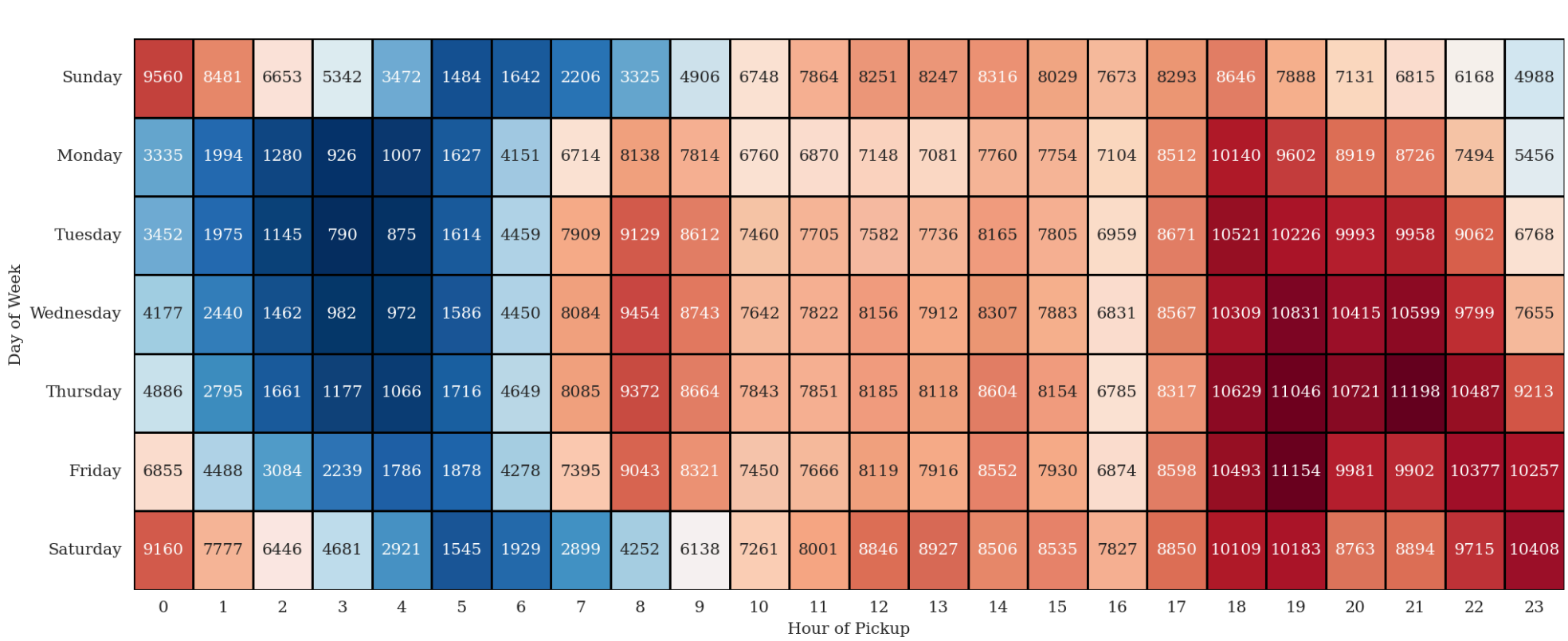


Figure 8

2.5. Distance Feature

For the choice of the type of distance to calculate, I chose the manhattan distance(L1 distance) because I see it’s more realistic for the trips in city like New York (and most of cities generally)

The distribution of the distance (scaled by log base 10) has outliers that goes up to 50 KM and above.

The outliers here is expected like I found outliers in the duration distribution.

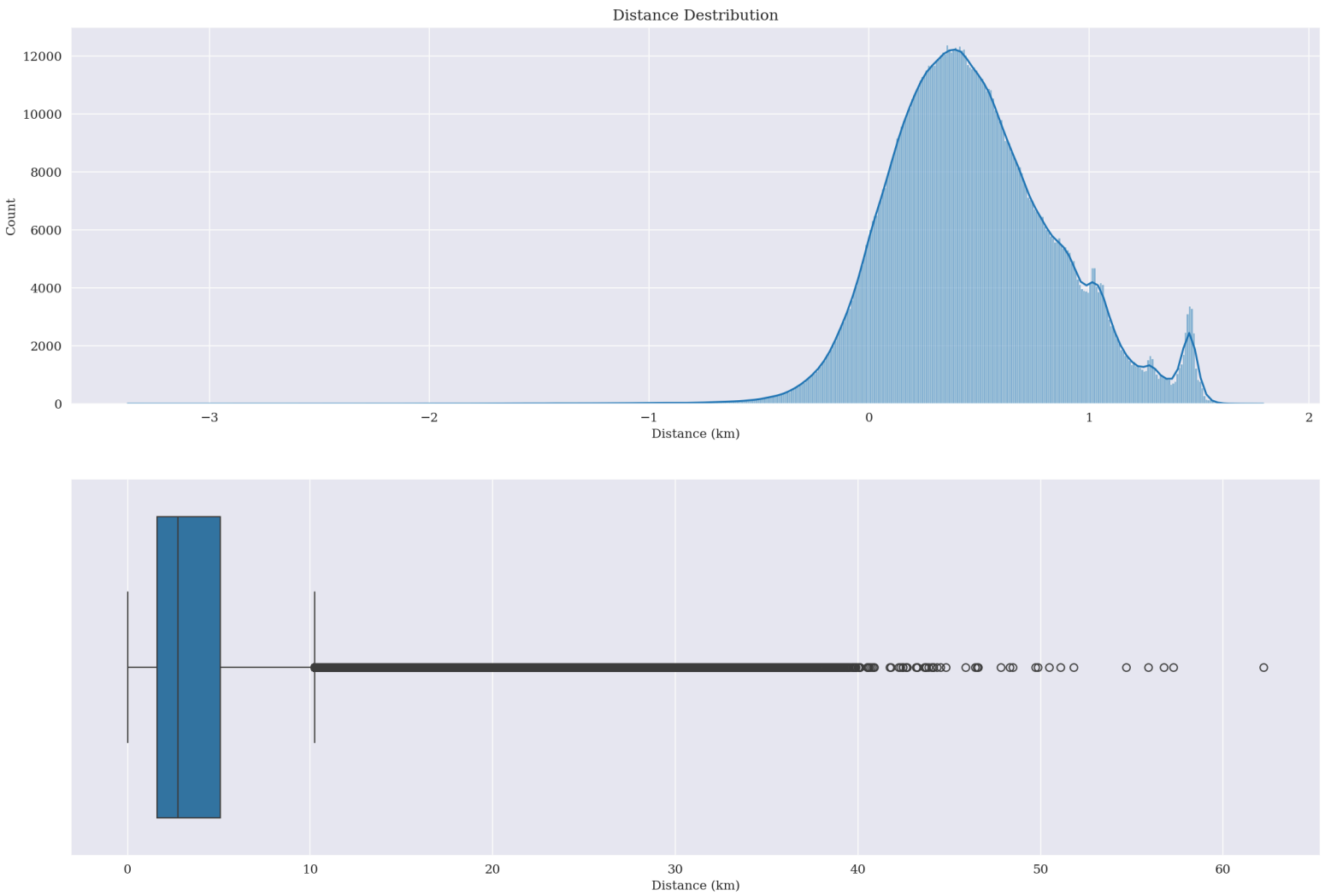


Figure 9

I applied further investigations to make sense of how much these outliers abnormal - Figure 10 .

This pickup coordinates map shows that many of the trips started much far from the geographical boundaries of New York.

which may be a problem in the coordinates.
Either in specifying them or when entering them to the system.

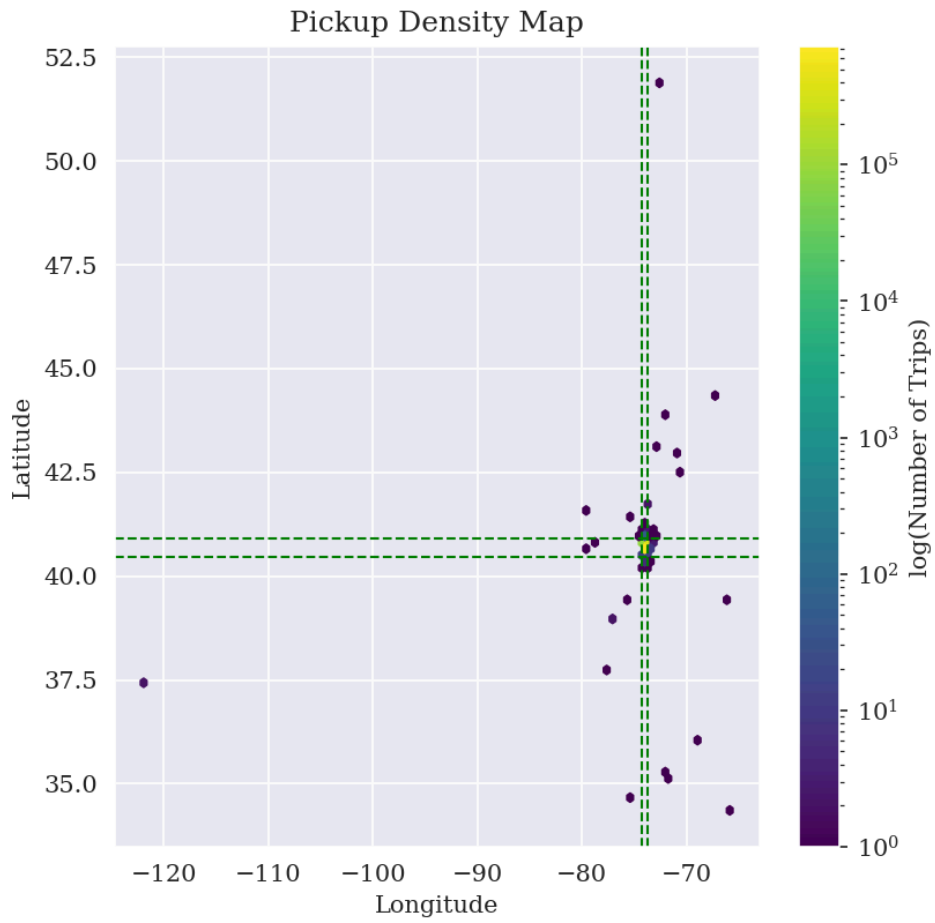


Figure 10

Plotting the trips that have distances above 45 KM against their durations - Figure 11 - shows that thir durations aren't as big big as their distances.

The maximum duration of them is about 70 minutes and it's corresponding distance is about 57 KM. which isn't sensible at all. which suggests that their coordinates are not correct.

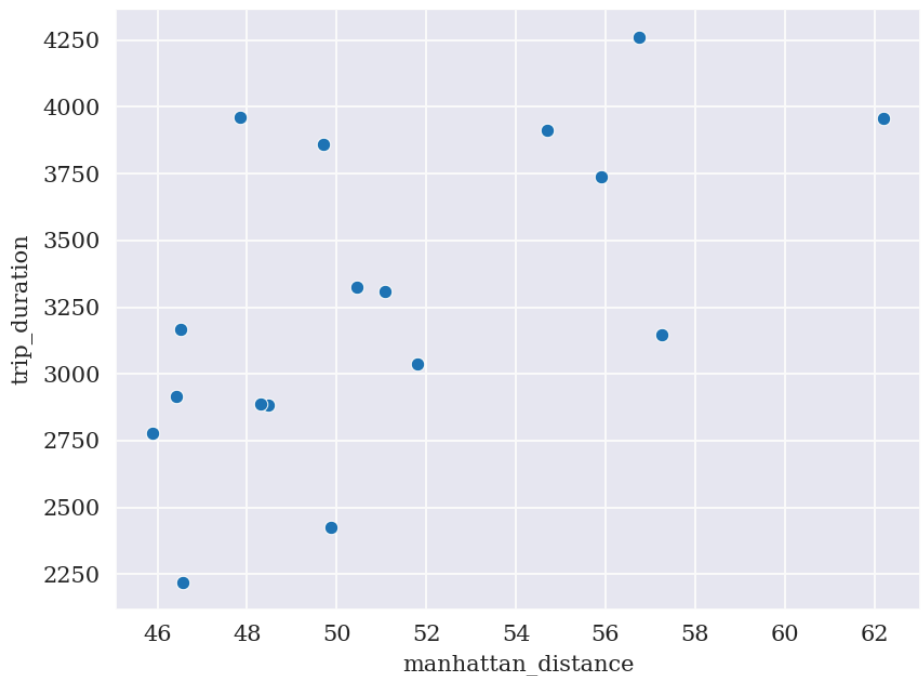


Figure 11

2.6. Changes in Data

Through the EDA journey on this data, here are the changes I made to it :

- Replaced the 'T' value in the columns **precipitation**, **snow fall** and **snow depth** with the smallest value at each of them
- Removed trips with durations below 2 minutes and above 5000 minutes (about 1h 22m)
- After checking the coordinates of the trips, I removed the trips that has a pickup coordinates out of New York city.
  The limits that were used are:-
  - Latitude: 40.4774° N to 40.9176° N
  - Longitude: −74.2591° W to −73.7004° W
- Based on investigation for the outliers in distance column, I removed the trips with distances above 30 KM and these that have 0 distance and non-zero duration.
- I extracted the following features :
  - **Day :** The day of month.
  - **weekday :** The name of the day of the week.
  - **hour :** hour of the day.
  - **manahattan_distance :** The manhattan distance between pickup and drop off locations.
  - **rush_hour :** A boolean feature that indicates whether or not this hour is considered a peak hour (based on investigating the trips number and the average duration on an hourly level).
- I removed these columns :
  id , pickup_datetime , dropoff_datetime , store_and_fwd_flag , maximum temperature , minimum temperature , snow fall , day

2.7. Modeling

2.7.1. Data processing

For the numerical features, I applied the minmax scaler and applied one-hot encoding on the categorical features.
But before scaling the numerical features I applied polynomial features of degree 2 for introducing some variance in the model(and it improved its results already).

Also,I applied the log scale(base e) to the target feature before training.

2.7.2. The Used Models

I trained 2 models, Ridge regression and multilayer neural network (MLPRegressor), then I picked the best of them based on the validation results.
The neural network is the one which showed better results.

2.7.3. The Used Metrics

The metrics that I used are the RMSLE and the R^2 score.
I see these are good enough metrics for measuring the average relative error of the model and get a sense of how well the model explains the variance in the data.

2.7.4. Train and Validation Results

		Ridge	**MLPRegressor**
Training	RMSLE	0.515	0.413
	R^2	0.452	0.645
Validation	RMSLE	0.7	0.545
	R^2	0.226	0.531