Project: NYC taxi trip duration prediction

To improve the efficiency of taxi dispatching systems for ride-hailing services, it is important to be able to predict how long a driver will have his taxi occupied. If a dispatcher knew approximately when a taxi driver would be ending their current ride, they would be better able to identify which driver to assign to each pickup request. The objective is to build a predictive model using the NYC taxi dataset.

Exploratory Data Analysis

After reading the NYC taxi dataset, I explore the dataset with respect to shape, datatypes and missing values in order to proceed with the preprocessing. There are ~730,000 rows and 11 columns. There are no missing values -

```
In [8]: print(missing_values)
id                    0
vendor_id             0
pickup_datetime       0
dropoff_datetime      0
passenger_count       0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude     0
dropoff_latitude      0
store_and_fwd_flag    0
trip_duration         0
dtype: int64
```

Initially I tried to use the interquartile range to look at outliers across all the numerical columns, however, it turns out to be not such a good idea for the longitude columns as all entries show up as outliers. We can address the outliers later for the fields that I'm mostly going to use.

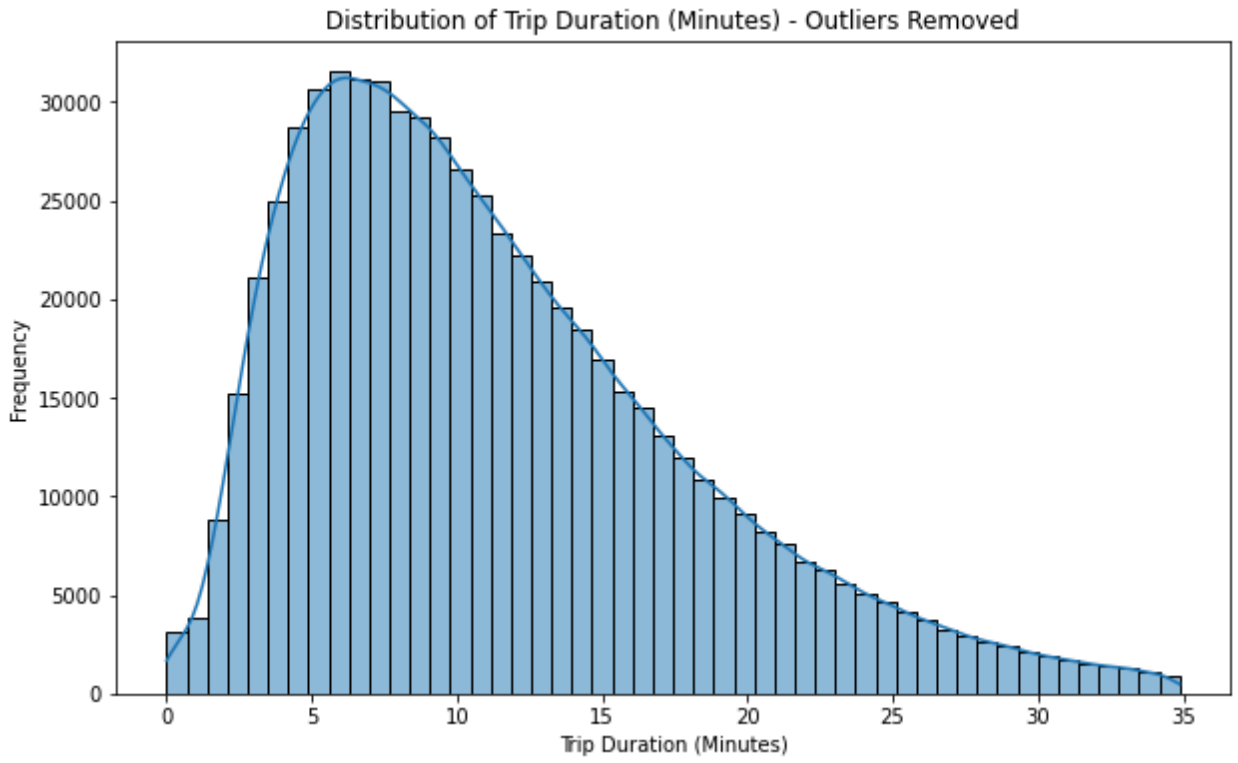
After using the to_datetime functionality in pandas for the pickup time and drop-off time columns, I have separated the details of the hour of the day, day of the week and month of the year into separate columns.

There is a "store_and_fwd_flag" in Y/N terms and hence I convert that into Boolean type.

The trip_duration field is in seconds, which I convert to minutes. I can see some very large numbers , which could be either wrong data or long ride taxi bookings and hence at least not the right data for ride hailing services. Similarly, some very short trips lasting for a fraction of a minute, which doesn't need to be considered.

```
In [24]: df_taxi['trip_duration(mins)'].describe()
Out[24]:
count    729322.000000
mean         15.870486
std          64.410437
min           0.016667
25%           6.616667
50%          11.050000
75%          17.916667
max       32328.933333
Name: trip_duration(mins), dtype: float64
```

After removing the outliers for trip duration, the distribution of the trip duration in minutes is plotted below which shows a lognormal shape. This would make sense as the duration can't go negative. Majority of the trips seem to last between 5-10 minutes.



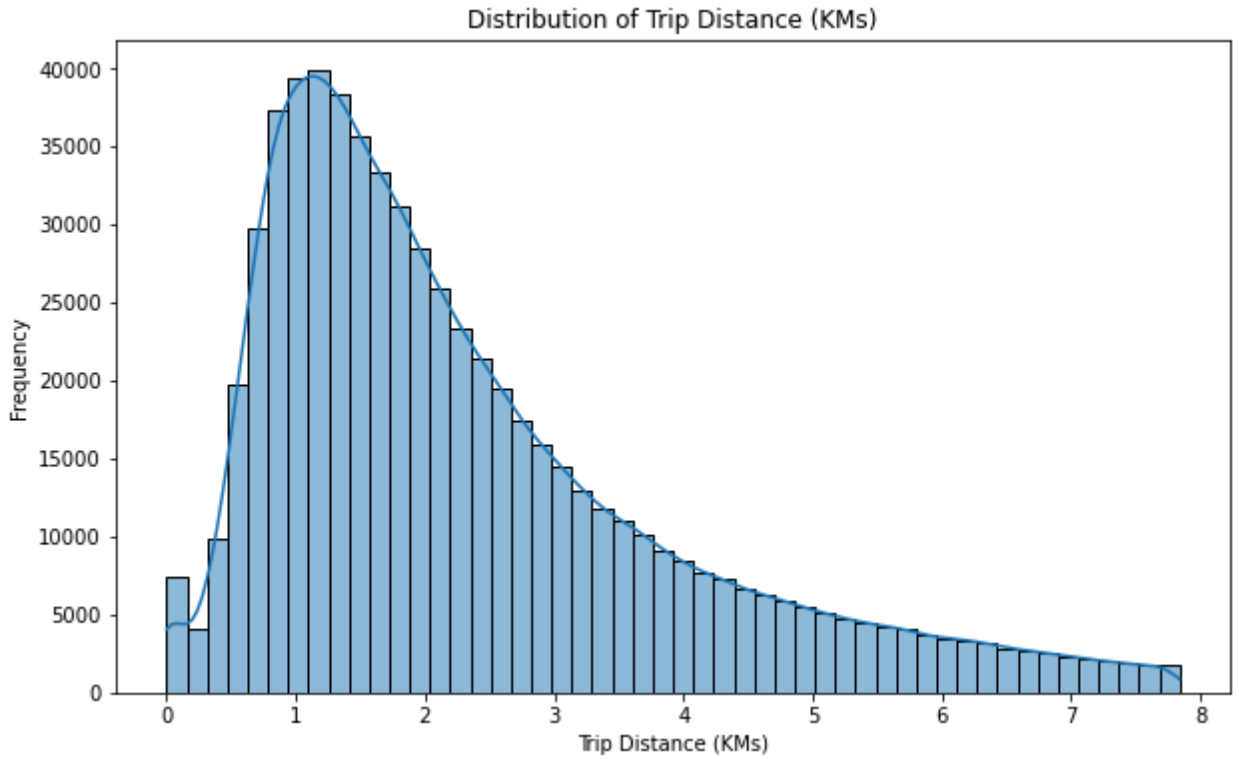
For the distance travelled on those trips, I have used a "geopy" package to convert the pickup and drop-off latitudes and longitudes into distance in kms. Similar to trip duration, I observe many outliers on both ends which don't make a lot of sense -

```
In [28]: df_taxi['distance_km'].describe()
Out[28]:
count    729322.000000
mean          3.442464
std           4.356668
min           0.000000
25%           1.233153
50%           2.096073
75%           3.875828
max        1240.510256
Name: distance_km, dtype: float64
```

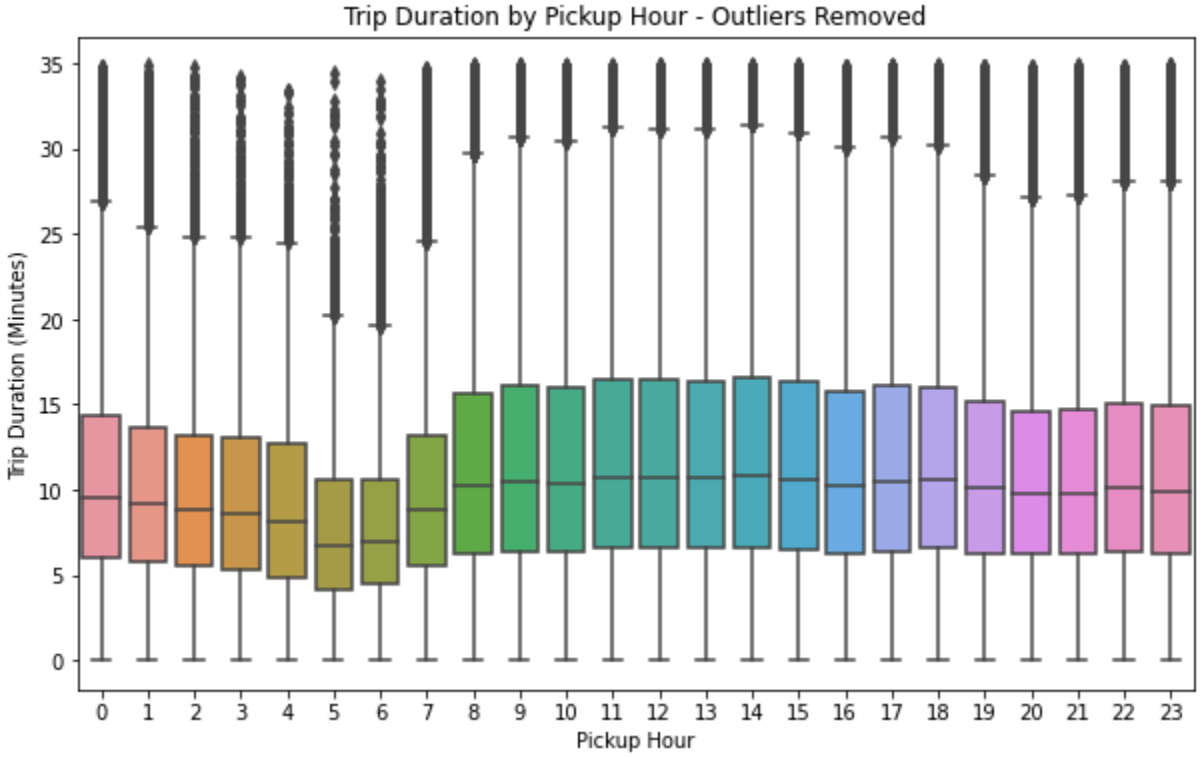
After removing the outliers for the distance covered, I get a filtered dataset for further analysis. Printing the shape of the new dataset, I see the original data gets reduced by approximately 80,000 data points after removing outliers -

```
In [40]: print(f"Original data size: {df_taxi.shape[0]}")
    ...: print(f"Filtered data size: {filtered_data.shape[0]}")
Original data size: 729322
Filtered data size: 649950
```

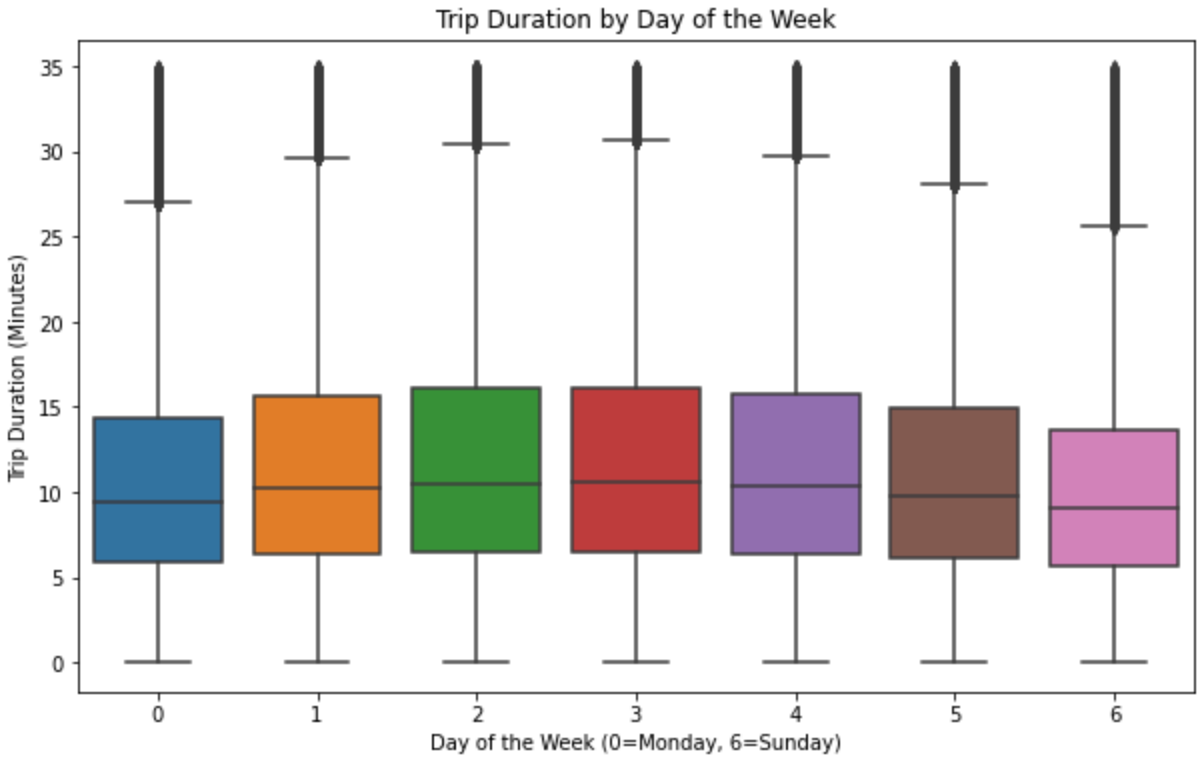
Similar to trip duration, the distance covered distribution plotted below shows a lognormal shape, with majority of the trips covering between 0.5km to 2kms.



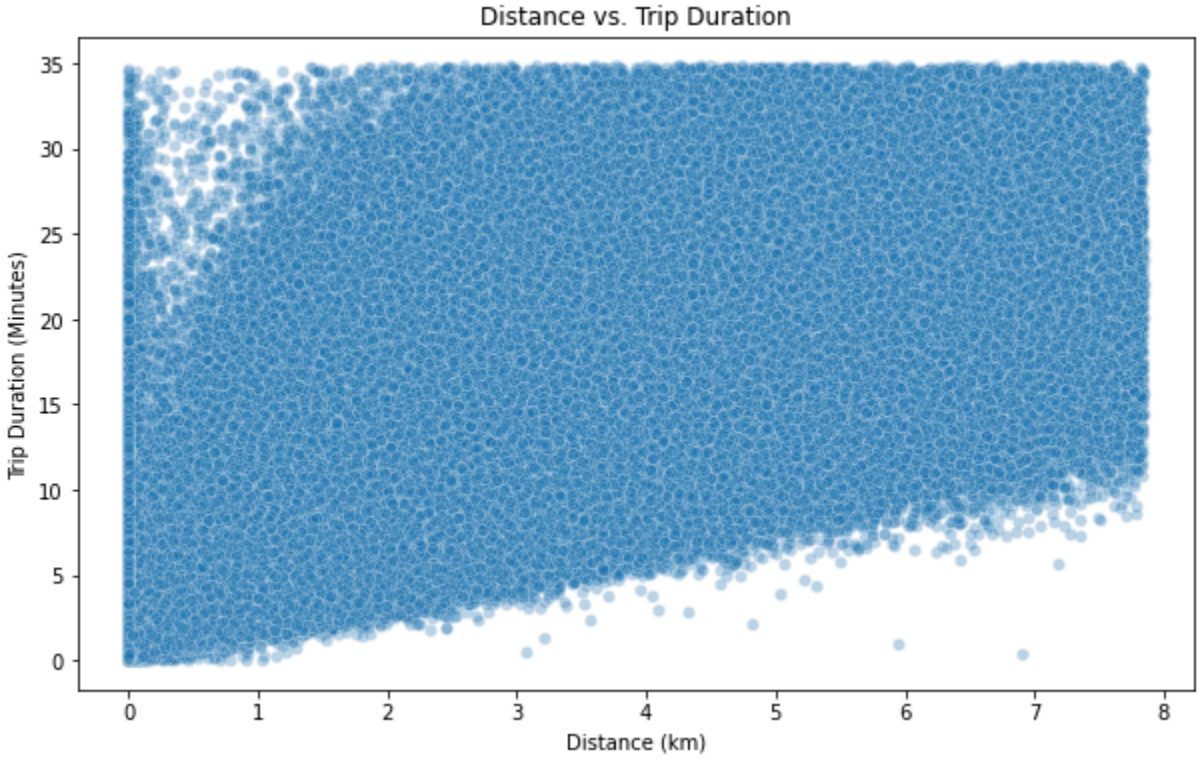
The boxplot below visualizes the distribution of trip duration (in minutes) for each pickup hour (from 0 to 23). The horizontal line within each box which depicts the median trip duration, shows that across most times of the day, the trips range between 10-15 minutes, whereas between 2-6am, they tend to be slightly shorter between 5-10 minutes. The height of each box which represents the Inter-quartile range,i.e. 25th to 75th percentile, seems lower again between 2-6am while for the rest of the busier hours, we can see a wider range. During late night and early morning (0 to 6 AM), trip durations are shorter and have less variability, probably reflecting faster travel due to less traffic. Similarly, higher variability during other hours may be due to higher traffic.



Another boxplot below visualizes the distribution of trip durations (in minutes) for each day of the week (0 = Monday, 6 = Sunday). The median trip duration remains relatively consistent across all days, around 10 to 15 minutes, suggesting that average trip times are not strongly affected by the day of the week. The whiskers and outliers show that the longest trips (approximately 30–35 minutes) occur consistently across all days. No significant difference is observed between weekdays (Monday to Friday) and weekends (Saturday and Sunday) in terms of trip duration distribution, although one could say Sundays have a slightly lesser variability compared to other days of the week.



The scatter plot below visualizes the relationship between distance traveled (in kms) and trip duration (in mins), which shows a clear positive correlation between distance and trip duration.



A small number of trips have unusually short durations for long distances or unusually long durations for short distances. These could be outliers due to data entry errors, unusual traffic conditions, or other factors may be playing a role. There is a clear upper bound around 35 minutes for trip duration, indicating trips rarely exceed this time.

To further delve into some of these unusualities, I look at the long distance with short durations and vice versa outliers -

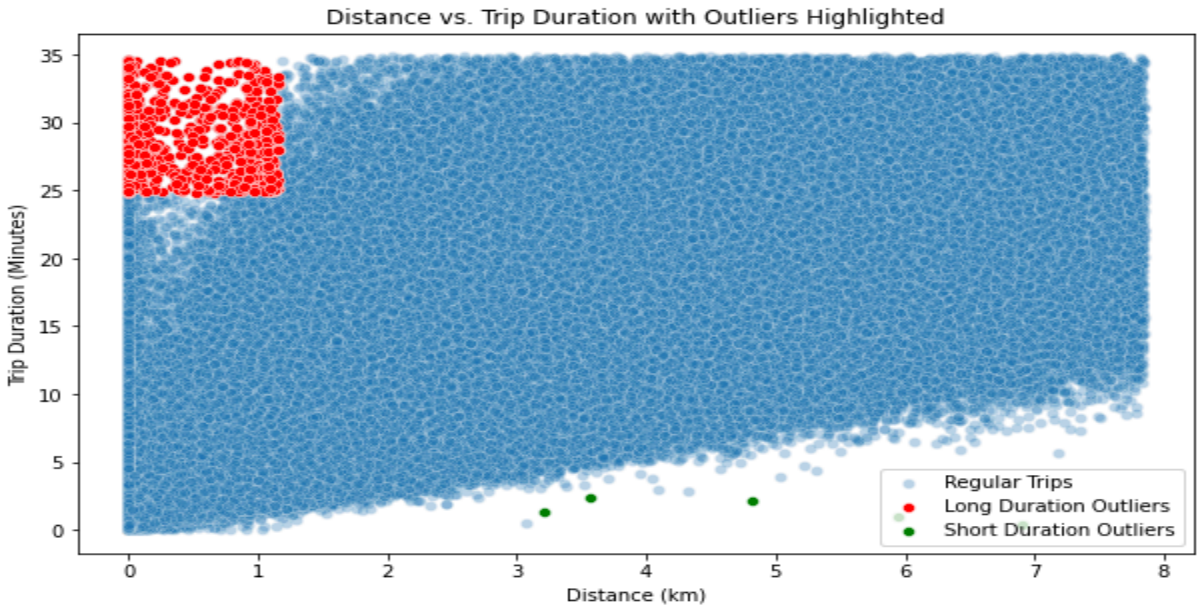
Summary of Long Duration Outliers (Short Distance):

	distance_km	trip_duration(mins)
count	676.000000	676.000000
mean	0.511053	28.634467
std	0.428672	2.704175
min	0.000000	24.866667
25%	0.045549	26.316667
50%	0.514431	28.158333
75%	0.937324	30.783333
max	1.156742	34.633333

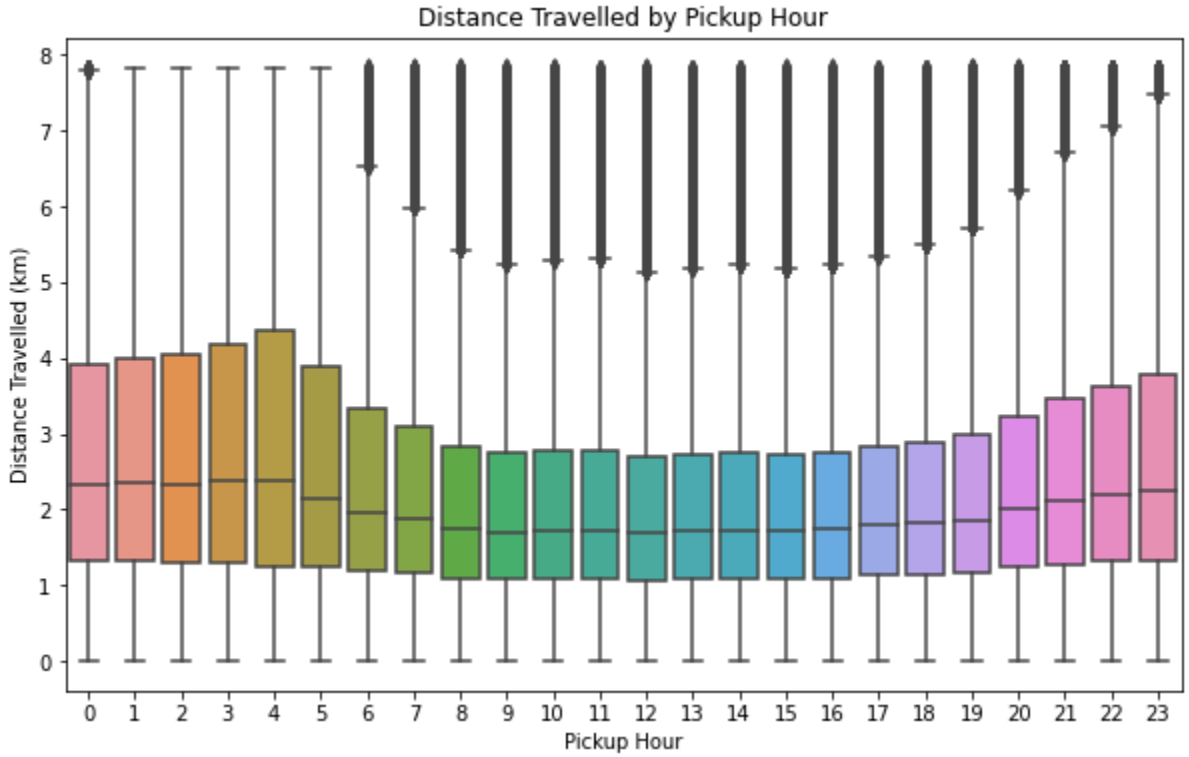
Summary of Short Duration Outliers (Long Distance):

	distance_km	trip_duration(mins)
count	5.000000	5.000000
mean	4.888316	1.416667
std	1.561128	0.817432
min	3.209599	0.333333
25%	3.567779	1.000000
50%	4.816561	1.316667
75%	5.941183	2.083333
max	6.906459	2.350000

These outliers are visualised in the plot below -



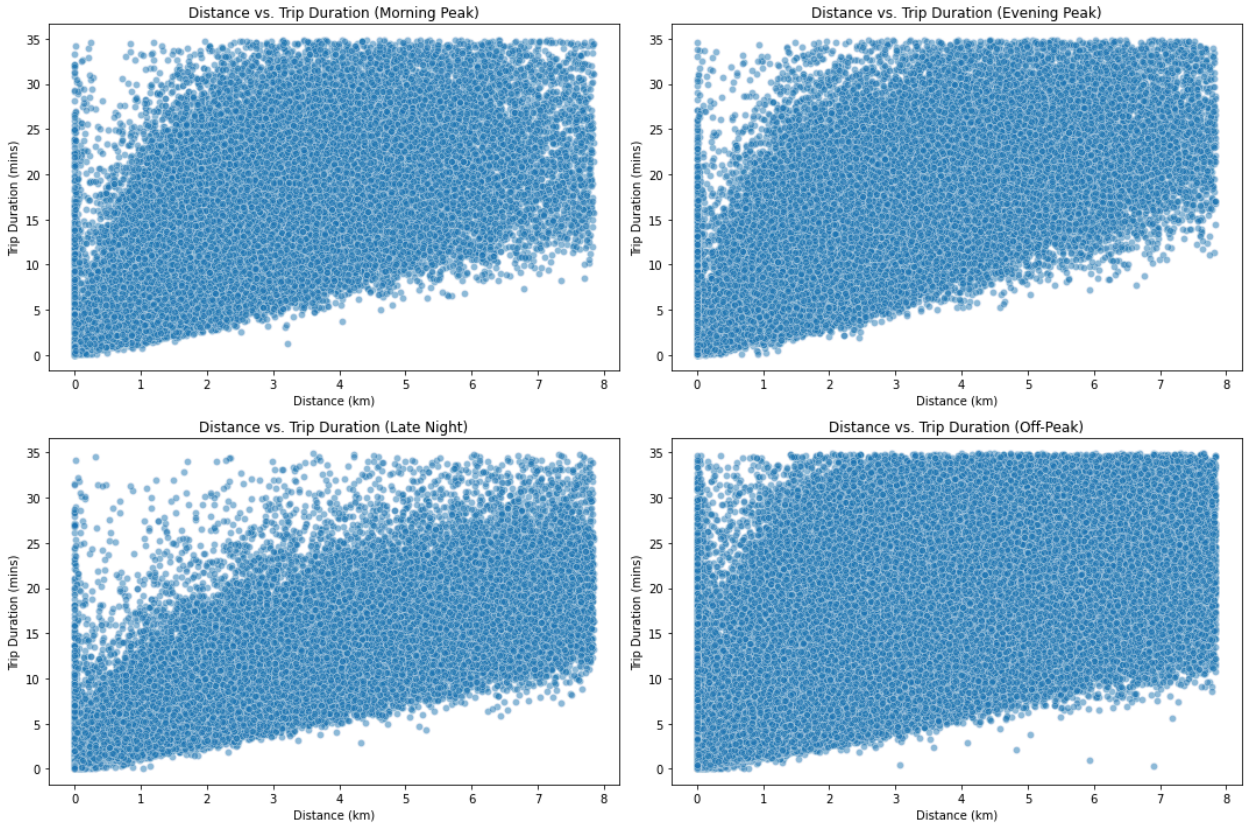
The boxplot shown below visualizes the distribution of distance traveled (in kms) across different pickup hours (0 to 23). The median distance is relatively consistent across all hours, typically around 2–3 kms. In the early hours between 12am to 5am, we see larger IQR, suggesting more variability in trip distances during that time. The slightly higher medians during these hours suggests lower traffic may facilitate longer trips (one of them could be trips to the airports, which are usually in the outskirts of the city). During peak commuting hours (8–10 AM and 5–7 PM), the distance distribution narrows slightly, with fewer long-distance trips compared to off-peak times. This could indicate that taxis tend to cater to shorter trips during these hours due to high demand.



To look at the correlation between trip distance and duration by time windows, I categorised the times of the day into 4 different categories - Morning peak (8am-10am), Evening peak (5-7pm), Late night (12am-5am) and Off-peak for all other times. The correlation for each of these windows is calculated as shown below -

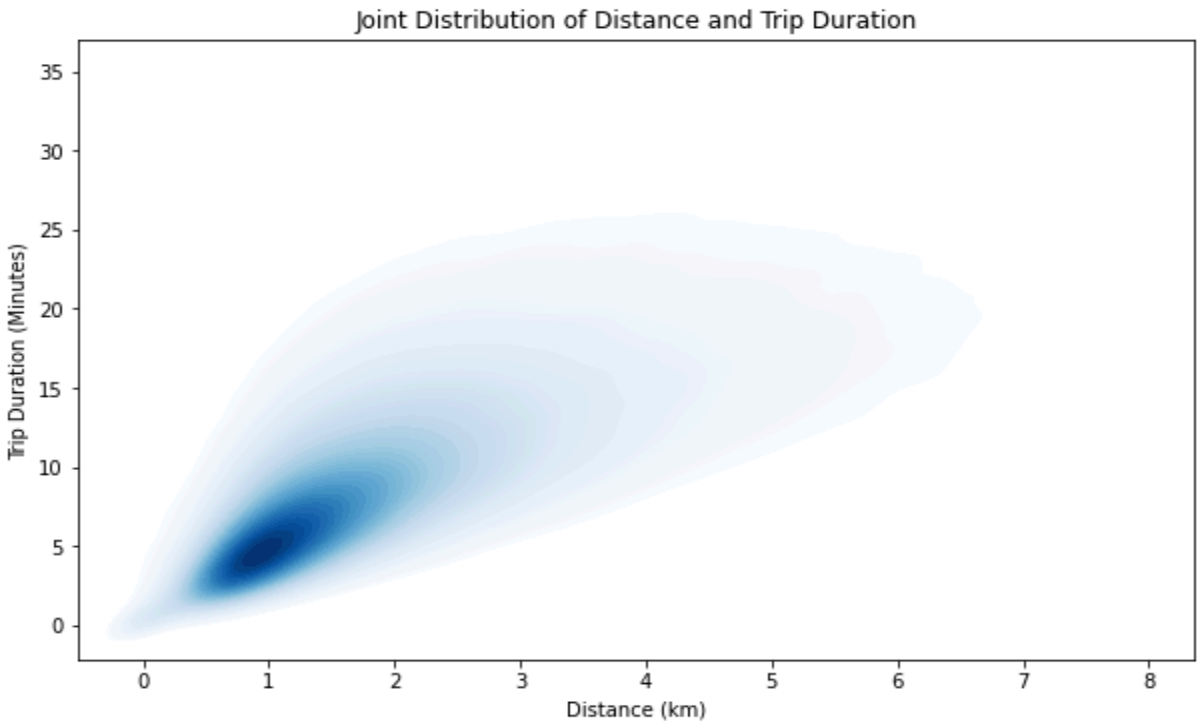
```
Correlation Between Distance and Trip Duration by Time Window:
Evening Peak: 0.74
Late Night: 0.77
Morning Peak: 0.68
Off-Peak: 0.70
```

A stronger correlation as shown above indicates that trip distance is a reliable predictor of duration during that time window. The scatter plots below provide a segmented view of the relationship between distance traveled and trip duration during specific time windows-

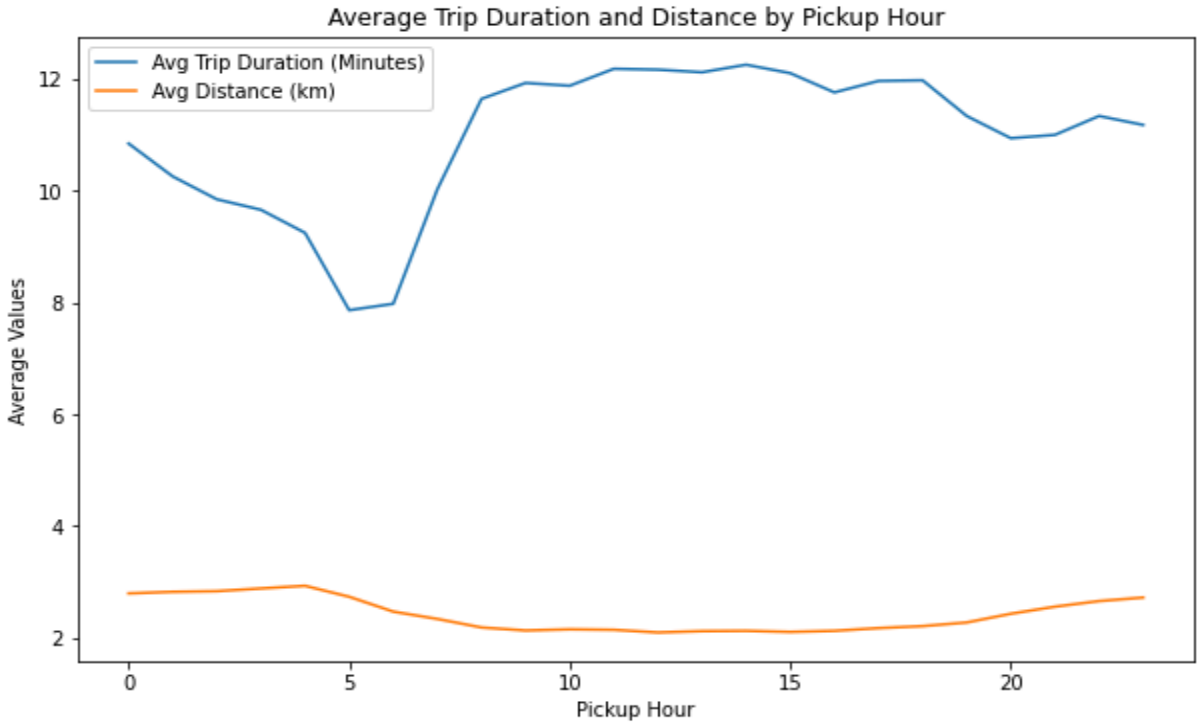


Both for the morning and evening peaks, the scatter plots above show a moderate positive trend, indicating that as distance increases, trip duration also increases. However, there is significant variability seen for both short and long distance. For the “late night” category, the relationship between distance and duration is more linear compared to the peak hours and variability is comparatively lower. This can also be verified from the highest correlation in the different categories, i.e. 0.77 for the late night category.

The joint distribution plot of distance traveled and trip duration using a density visualization (KDE plot in seaborn) shows the darkest area at the bottom left indicating the highest concentration of trips. These trips have short distances (0–2 km) and short durations (0–10 minutes), reflecting a common pattern of short taxi rides within urban areas. As the distance increases (beyond 2 km), the distribution becomes wider and less dense. This indicates more variability in trip duration for longer distances, likely due to factors like traffic, road conditions, or route choices. A clear positive trend is visible, i.e. as the distance increases, the trip duration also increases. The distribution has a clear boundary for trip durations. Most trips, regardless of distance, fall below 35 minutes. This could be helpful in predicting when a driver will eventually become free for other trips.



The below plot between the average trip duration and average distance traveled for trips, segmented by pickup hour (0 to 23) is similar to some of the box plots we have seen earlier, which mostly validates some of the findings.



The trip duration starts high during late-night hours (12 AM to 4 AM), then dips to its lowest at around 5–6 AM, likely due to minimal traffic during early morning hours. It sharply increases from 6 AM to 9 AM, peaking during the morning rush hours (8–10 AM), which reflects increased travel time due to congestion. The average trip distance remains relatively consistent throughout the day, hovering between 2–3 kms. There is a slight increase during late-night hours (12 AM to 4 AM) and evening hours (7–10 PM), indicating that trips during these times are slightly longer on average , likely due to less congestion in traffic.

Predictive modelling

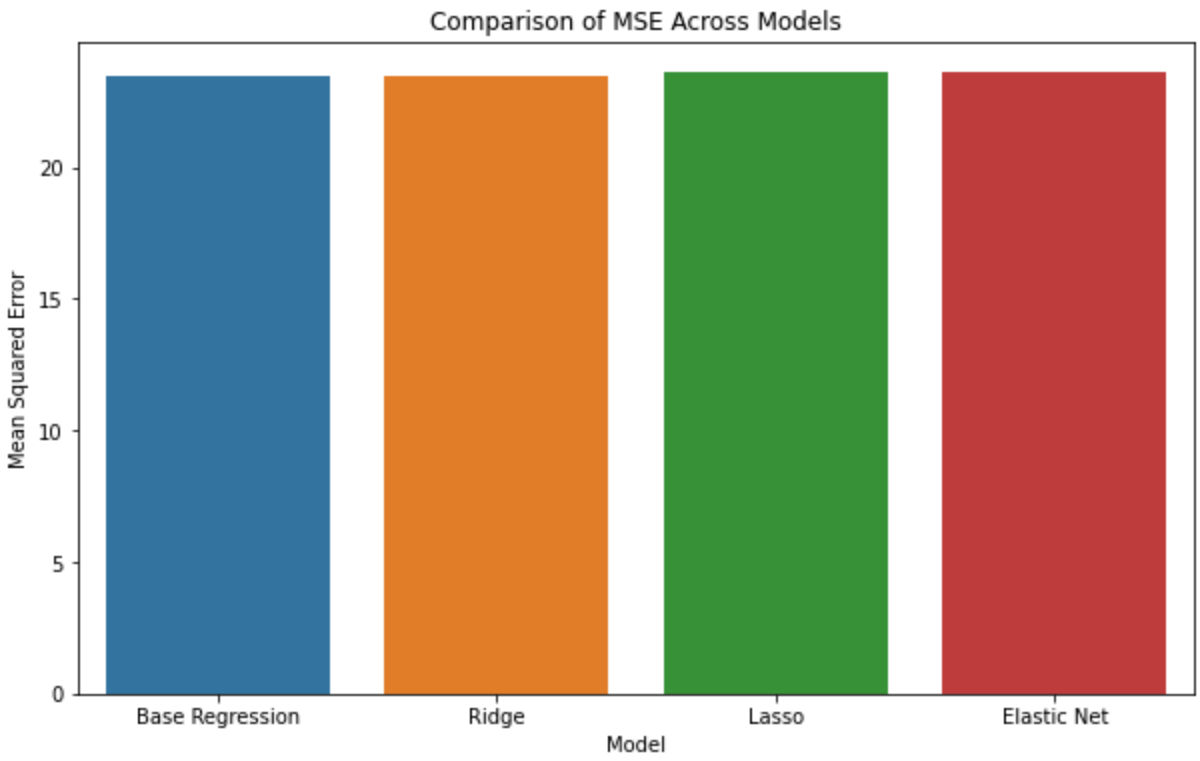
I drop the “id” column , which isn’t of much use in the analysis for the purpose of this project. For the purpose of predicting the trip durations, I separate the dataset into the independent variables, while the duration being the “dependent” variable. Since we have the other features such as hour/day/week, I drop the pickup, drop-off time and time-window columns before using scaling for the X variables and then split the data into training and testing sets using a 70-30% split.

Initially I use a simple linear regression as the base model to predict the trip duration for the test dataset. The mean squared error for the base model is at 23.45, while the R-squared is at 0.49.

Then I try to introduce the 3 different types of regularisation, L1 (Lasso), L2 (Ridge) and Elastic net with a 0.5 ratio between L1 and L2. For the alpha values for regularisation, I use cross validation for different values of alpha ranging from 0.1 to 20. By displaying the different R-squared and MSE values across different values of alpha across the 3 types of regularization models, I choose the best alphas for each type for further analysis. The results across the 3 different models in comparison to the base model, i.e regression with no regularization is shown below in a table -

Model	**Best Alpha**	**MSE**	**R-sqd**
Base Regression	N/A	23.45	0.49
Ridge Regression	20	23.45	0.49
Lasso Regression	0.1	23.59	0.487
Elastic Net Regression	0.1	23.60	0.487

We can see from the above table, there is hardly any impact from including regularization and the MSE actually slightly increases by using Lasso or Elastic net regularization, although the differences are visually insignificant.

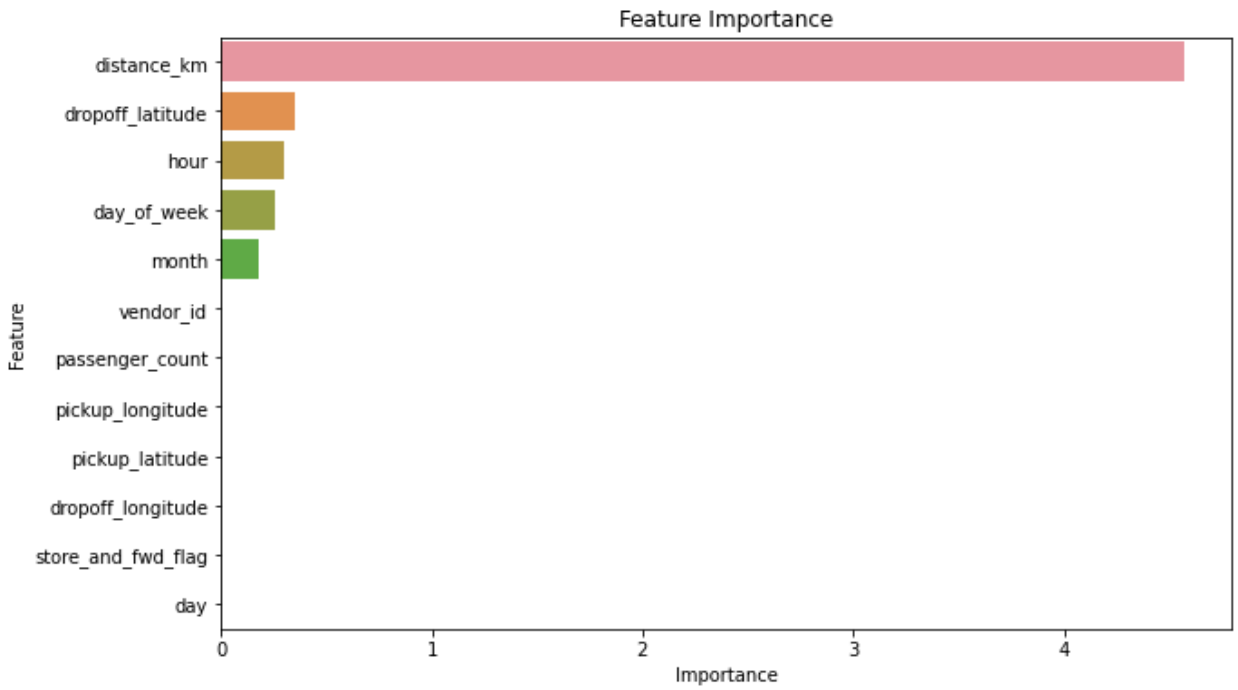


Finally I try to include feature selection by order of importance to see if it results in a better predictive model. The most important features are shown below -

```
Feature Importance from Lasso:
              Feature  Importance
11        distance_km    4.566807
5    dropoff_latitude    0.350671
7                hour    0.299355
10        day_of_week    0.258589
9               month    0.179289
0           vendor_id    0.000000
1     passenger_count    0.000000
2    pickup_longitude    0.000000
3     pickup_latitude    0.000000
4   dropoff_longitude    0.000000
6  store_and_fwd_flag    0.000000
8                 day    0.000000
```

Using a threshold of 0.01 for importance, the selected features are -

```
Selected Features based on importance (Threshold = 0.01):
['distance_km', 'dropoff_latitude', 'hour', 'day_of_week', 'month']
```



Using the above features only and the best alphas for each category of regularization, I re-fit the data for both Ridge and Lasso regressions. However, the results are hardly different from those earlier shown. In fact, we observe an insignificant increase in the MSE for both models, thus suggesting it won't be any different for the Elastic net type either.

Model	**Best Alpha**	**MSE**	**R-sqd**
Ridge Regression	20	23.53	0.49
Lasso Regression	0.1	23.73	0.484

This suggests that regularization does not improve model performance significantly, likely because the features do not exhibit multicollinearity or overly large coefficients. From the above, we can conclude either other forms of predictive modelling rather than regression could be better suited for improving accuracy of prediction or there could be additional features that might help in predicting better.