What Determines the Duration of a Cab Trip in New York?

2nd Capstone Project Milestone Report

1. Background, problems and questions to be answered

Cabs, buses, trains, any forms of public transportation are the integral and important parts of life of any city. This is especially true if we talk about large cities, such as New York, Chicago or Los Angeles. People drive or take public transport to go to work places and airports, visit friends and attend events. And everyone, before embarking on a trip, asks one and the same question: Do I depart early enough to be on time? And very often, not one factor is necessary to consider to answer this question. The duration of any trip, be it public or private, depends on the level of traffic congestion, day of the week (workday or weekend), hour of the day, weather conditions determined in part by the time of the year, state of the road, and even the unexpected accidents that slow down traffic. The influence of some factors is obvious and known well in advance allowing to make almost certain conclusions. For example, everyone knows that roads are less cluttered during weekends compared to workdays, meaning that a trip during weekend will be faster with very high probability (trips to special events such as Super Bowl are definitely exceptions). The role of other factors may not be that trivial and the majority of people overlook them. Suppose someone wants to take a cab to airport in August and tries to estimate the duration of the trip based on another business trip that was made in June. The time of the flight is approximately the same and the summer weather is excellent. But does this person take into account the fact that more people tend to go on vacation in August rather than in June, resulting in busier roads to airport and hence slower traffic? It is doubtful that many people go that far in their analysis. The abovementioned example strongly suggests that there are many factors, often hidden and intertwined, and thus difficult to analyze rigorously. That is why it is important to perform comprehensive statistical analysis of all available data that can affect the time it takes for a passenger to travel between two destinations.

In this project we will analyze just one large dataset that contains the continuous variable of duration of the trip, as well as several other variables such as pickup time, drop-off time, geo-coordinates of pick-up and drop-off, number of passengers (total 11 variables). The dataset was prepared by the NYC Taxi and Limousine Commission and can be downloaded from:

https://www.kaggle.com/c/nyc-taxi-trip-duration

By solving regression problem, we would like to answer one major question: Which features influence most the duration of the taxi trips in New York City, and what is their relative importance? Other questions we would like to explore are:

a) How correlated is the trip duration with the distance between pick-up and drop-off, as well as other variables?

b) How are the average speeds of the trips distributed?

c) Where do the fastest and slowest trips originate?

d) When do the fastest and slowest trips originate?

2. Potential clients

There are two groups of clients that could be interested in the findings from this project.

a) Taxi companies are interested in optimizing the trips within the city limits as well as between cities. In other words, they are interested in knowing which trips, originated at different locations and times, are likely to be faster given the same distances between origins and at destinations.

b) Citizens who use cabs frequently are themselves interested in knowing where and when it may be convenient to catch a cab in order to save time.

3. Data preparation and exploratory data analysis

3.1 Data preprocessing steps

a) We first read-off the dataset. Since, we are not participating in the Kaggle competition, we will be using just the training set that contains the target variable, namely the trip duration. We see that our dataset is quite large containing 1458644 entries. In addition to the trip duration (the target variable), the dataset includes pickup datetime, drop-off datetime, geo-coordinates of pick-up and drop-off, number of passengers, and several other variables (total of 10 feature variables).

b) Albeit we don't expect any missing values in the Kaggle dataset, it is useful to ascertain that each column (out of 11) contains zero number of missing values. We then look at the 'id' and 'vendor_id' columns. The column 'vendor_id' contains just two values 1 and 2, and we will keep it. From the other side, all values in the 'id' column are different, meaning that this column can’t be anyhow useful and should be dropped. 'store_and_fwd_flag' column contains two categorical variables 'Y' and 'N', so we convert them to 1 and 0.

c) We then create the columns giving the year, month, day, hour and minute of pickup time. All data pertain to 2016 year, so we also drop this column. Also, because we have the duration of the trip in seconds, we drop then the 'dropoff_datetime' as well as 'pickup_datetime' columns.

d) Next, we create the special column containing the geographical distance in kilometers between the pick-up and drop-off places. We use the so-called Vincenti formula for the distance that is accessible through the "geopy" module. One should notice that it takes almost 8.5 minutes to do the calculation for 1458644 observations.

Strictly speaking, we need the driving distances instead of geographical ones. Importation of the driving distances, however, requires using the geolocation-python module which uses the Google maps API. Free usage of this application is limited to only 2500 requests per day and hence impractical for large datasets. We will see that engineering the geographical distance column will be already useful.

e) Finally, we convert the trip duration to being presented in hours rather than seconds. The corresponding column will be renamed to 'trip_duration(hrs)'. We also create another feature column 'effcetive_speed(kmph)', given by the ratio of the value in 'geographical_dist(km)' to that

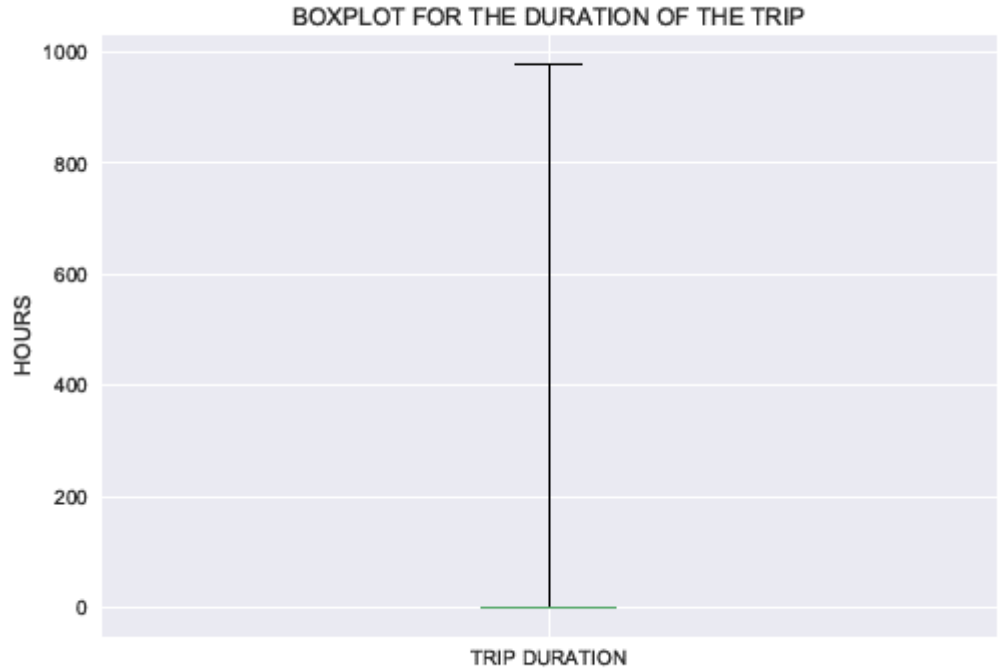
in 'trip_duration(hrs)'. This is an important column in exploratory data analysis, but we can not use it in predictive modelling because it was itself obtained using the target variable.

The new working dataframe contains now 15 features, one of which 'trip_duration(hrs)' will be our target variable. The feature columns will be the 'vendor_id', 'passenger_count', 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude', 'store_and_fwd_flag', 'pickup_month', 'pickup_day', 'pickup_hour', 'pickup_minute', 'pickup_day_of_week', geographical_dist(km)' and 'effective_speed(kmph)' columns.

3.2 Exploratory Data Analysis

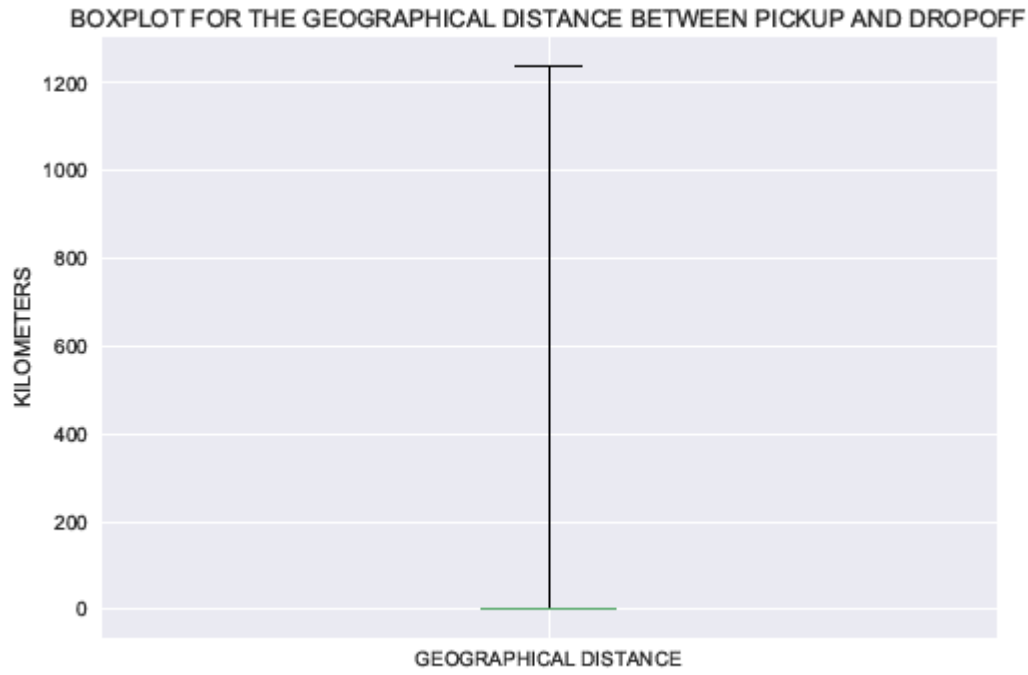
Looking at our columns to start exploratory data analysis (EDA), we assume that 'vendor_id', 'passenger_count' and 'store_and_fwd_flag' columns are rather trivial and of little interest to visualize, so we focus on the data in other columns. We first make the boxplots for 'trip_duration(hrs)', 'geographical_dist(km)' and 'effective_speed(kmph)'.

a) Let's first make the boxplot of and apply the describe methods to the 'trip_duration(hrs)' column. We see that the maximum value in the dataset is as high as 979 hours, while the minimum values is around 1 second.



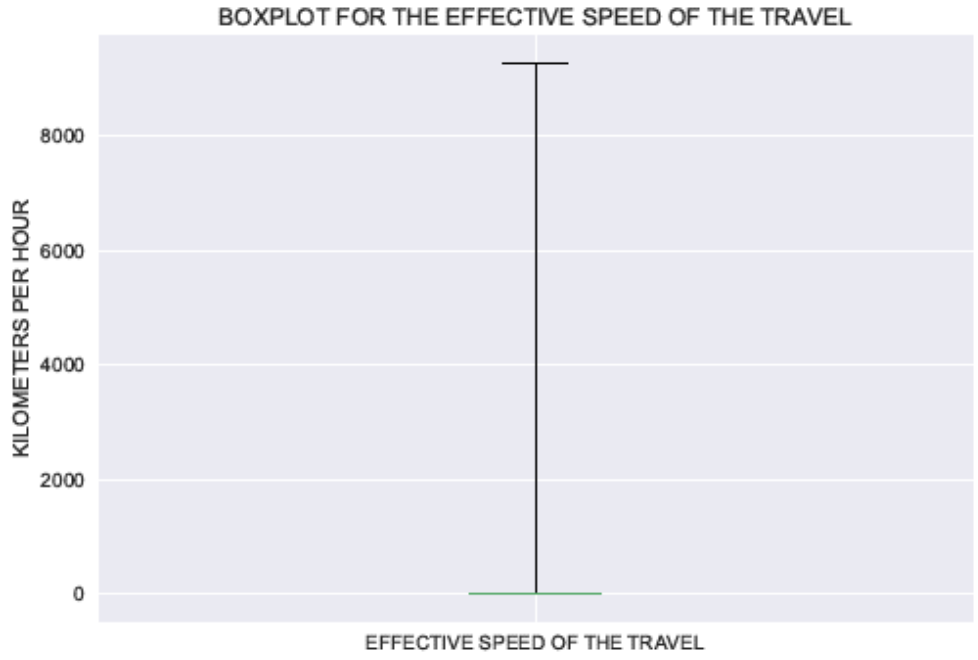
From our point of view, the durations that last more than 16 hours do not reflect a real single trip and could be either the result of data errors, or some combinations of actions or events that simply can’t be interpreted as a normal trip. There are 1968 such trips.

b) Then we make the boxplot of and apply the describe methods to the 'geographical_dist(km)' column. We see that the maximum value in the dataset is as high as 1240 km, while the minimum values is just 0 km.



It is clear that the entries having 'geographical_dist(km)' equal to zero mean the trip did not occur at all. There are 5897 such entries. There are also 13 trips between locations separated 150 km apart. It is difficult to say whether these trips were real, but these data are anomalous from the point of view of cab drivers and passengers.

c) The data for the 'effective_speed(kmph)' column (the artificially engineered effective speed of the travel feature) should be handled in a similar way. The minimum effective speed is 0 kmph, while the maximum is 9297 kmph.



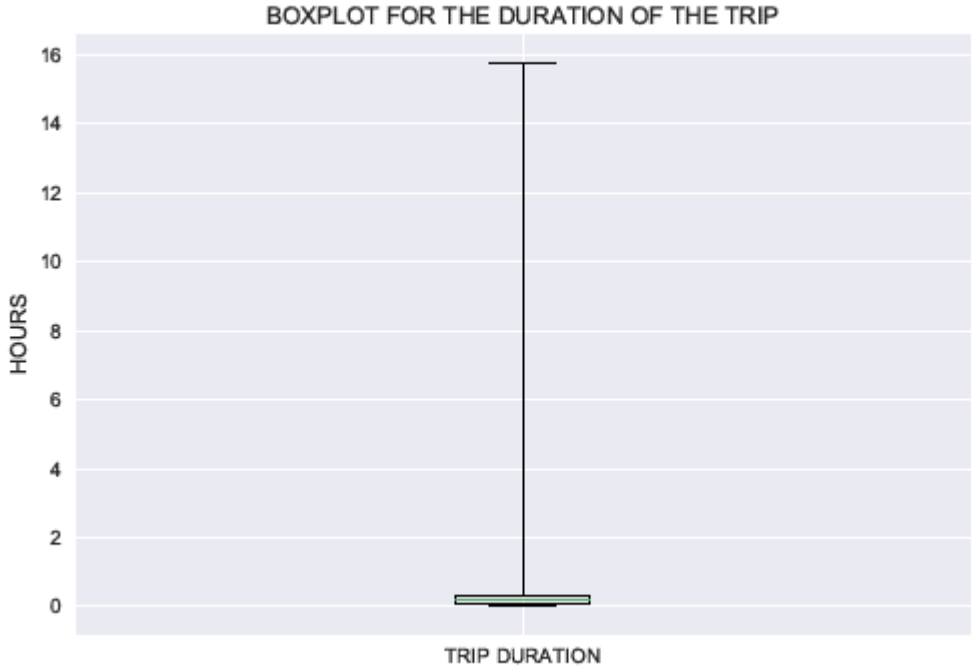
The speed limit in New York is 55mph, or 88.5 kmph. There are 205 trips that resulted in effective speed greater than the speed limit.

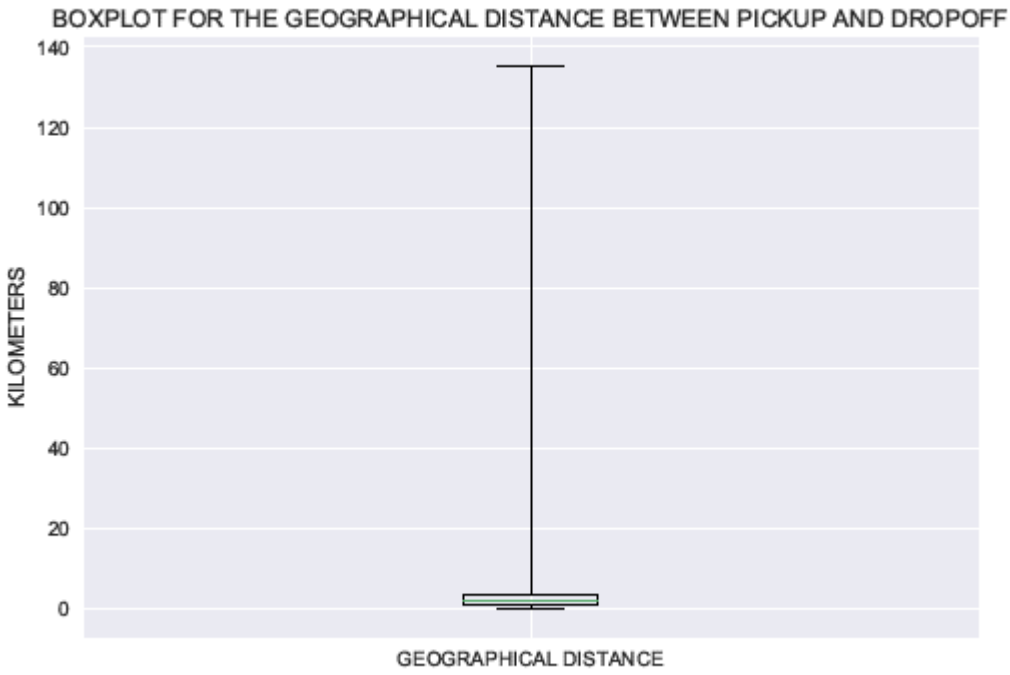
e) We remove from the data: (i) trips that lasted more than 16 hours, (ii) trips with the geographical distances of either 0 km or more than 150 km, (iii) trips where the effective speed was greater than the speed limit.

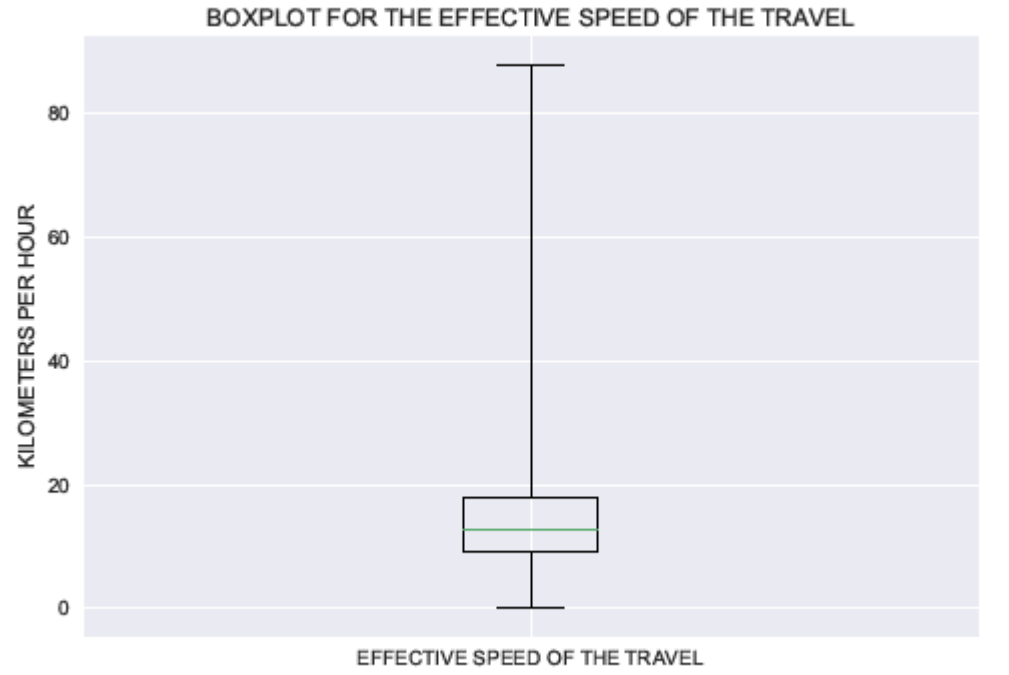
f) We also remove two entries with suspicious longitudes from the data-frame (entries 1177854, and 1062842), since it is unlikely that those trips occurred in New York, and these points may negatively affect predictive data analysis. Our column will then contain 1450573 instances.

g) The mean value of the trip duration is 0.23 hour, while the standard deviation is 0.2 hour. The mean value of the geographical distance between pickup and drop-off is only 3.45 km with the standard deviation being 3.95 km. We see that the trip durations and geographical distances are quite broadly distributed. We see also that the average effective speed for the taxi trips in New York is rather small (14.4 km per hour which is smaller than 20 km per hour); this is quite reasonable for a busy megapolis. Larger speeds are likely to occur for the intercity trips in which the high-speed highways were used.

h) We then make the boxplots of the the 'trip_duration(hrs)', 'geographical_dist(km)', 'effective_speed(kmph)' once, what we believe the anomalous data has been removed. The plots reveal that there is still a small number of trips that lasted either very long, or resulted in large travel distances. This means that the data for the trip duration as well as for geographical distances have very long tails in their distributions.

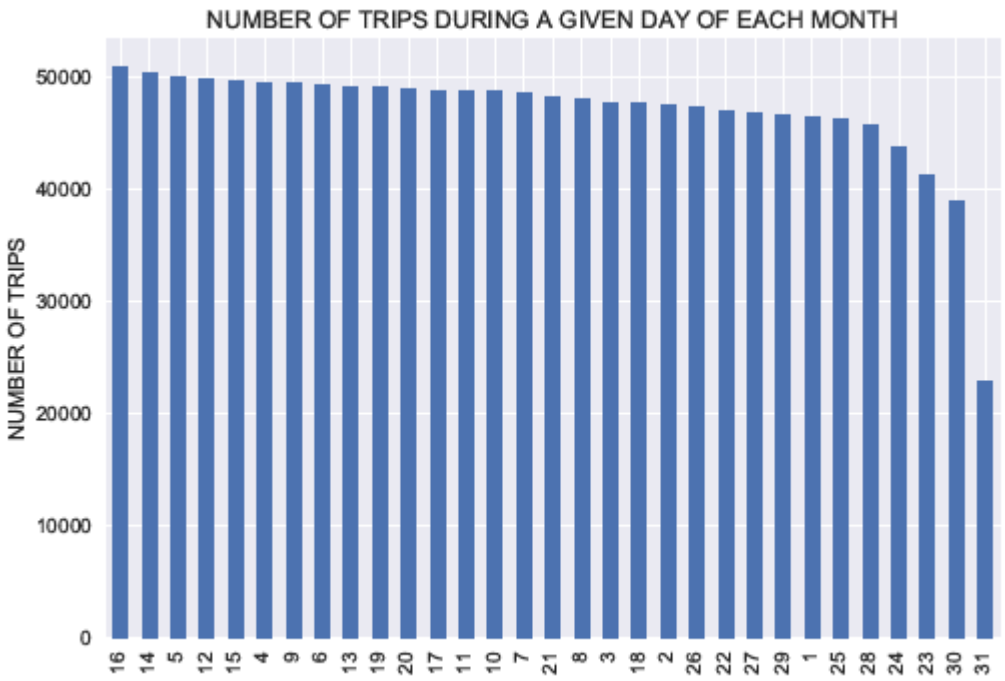


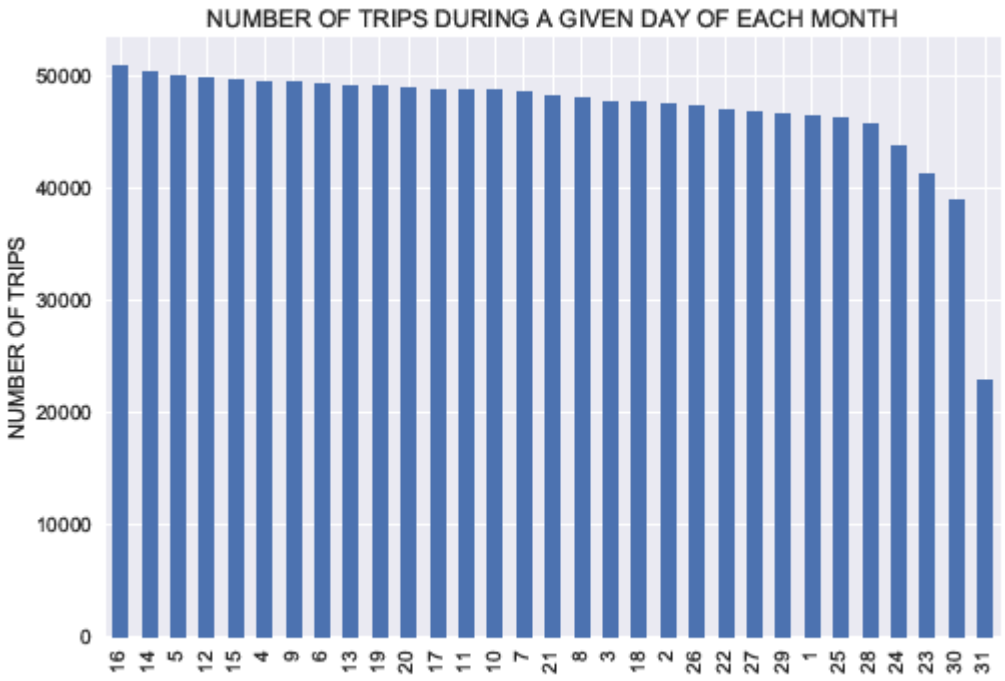


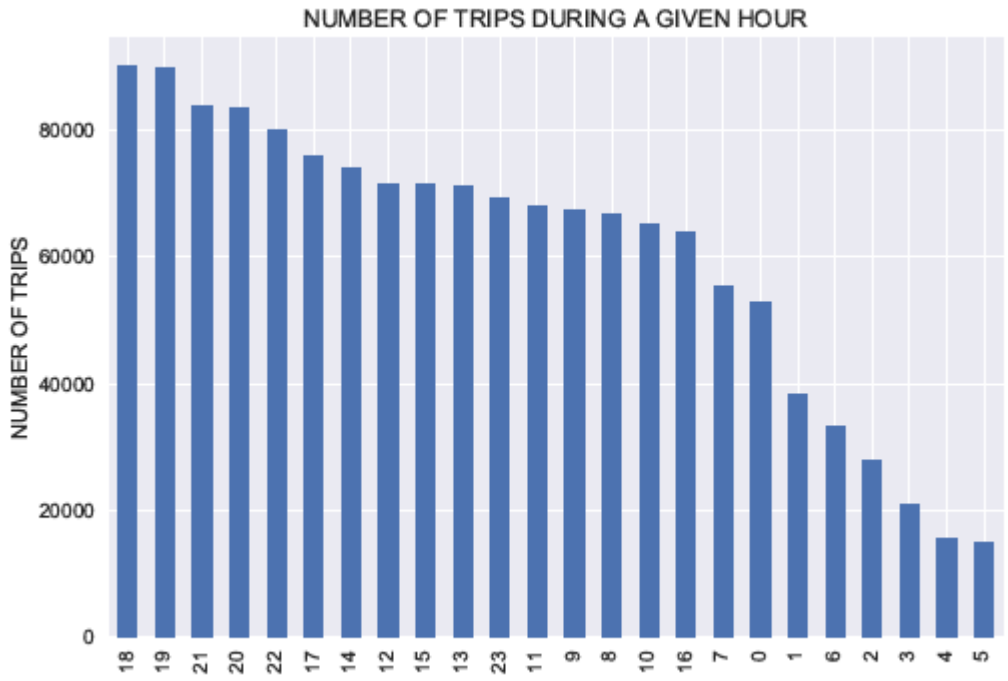


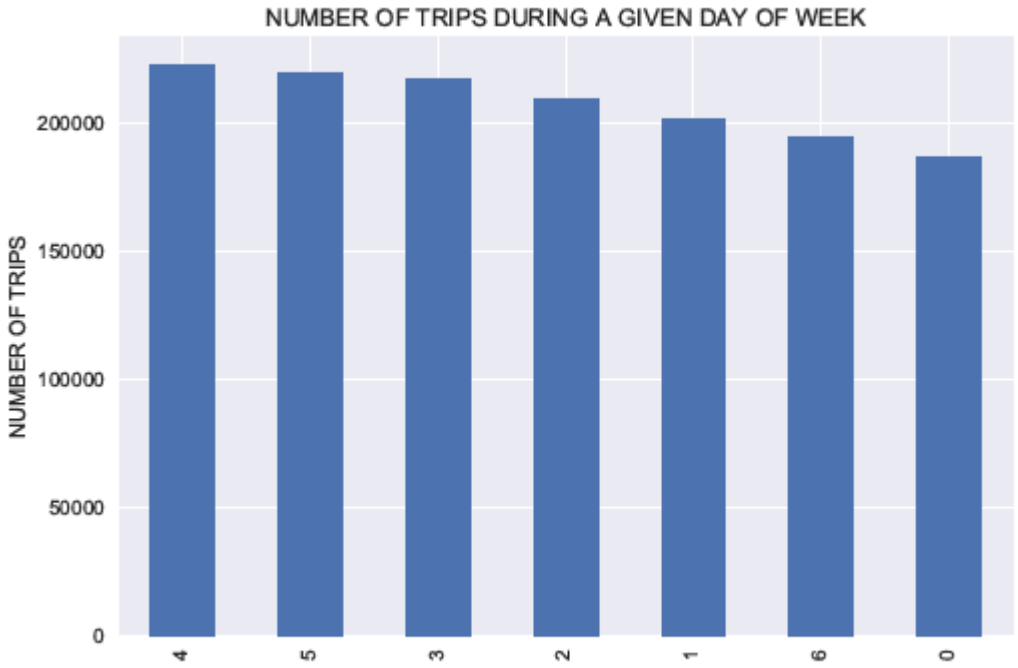
i) Let's make four bar plots for the trip numbers for given values in the 'pickup_month', 'pickup_day', 'pickup_hour' and 'pickup_day_of_week' columns. The first bar plot suggests that the number of trips does not vary significantly with month, while the plot displaying the number of trips per given day of each month suggests that people are slightly more reluctant to take cabs during the third decade of each month. The number of trips made on the 31st day is roughly half the similar number for any other day, which is expected. The third plot, that shows the hourly distribution of the number of trips, reveals that 6,7,8,9 and 10 pm are the busiest hours, while

late night and early morning are the least busiest ones. Finally, the last plot suggests that people are least likely to take cabs on Sundays, Mondays and Tuesdays.



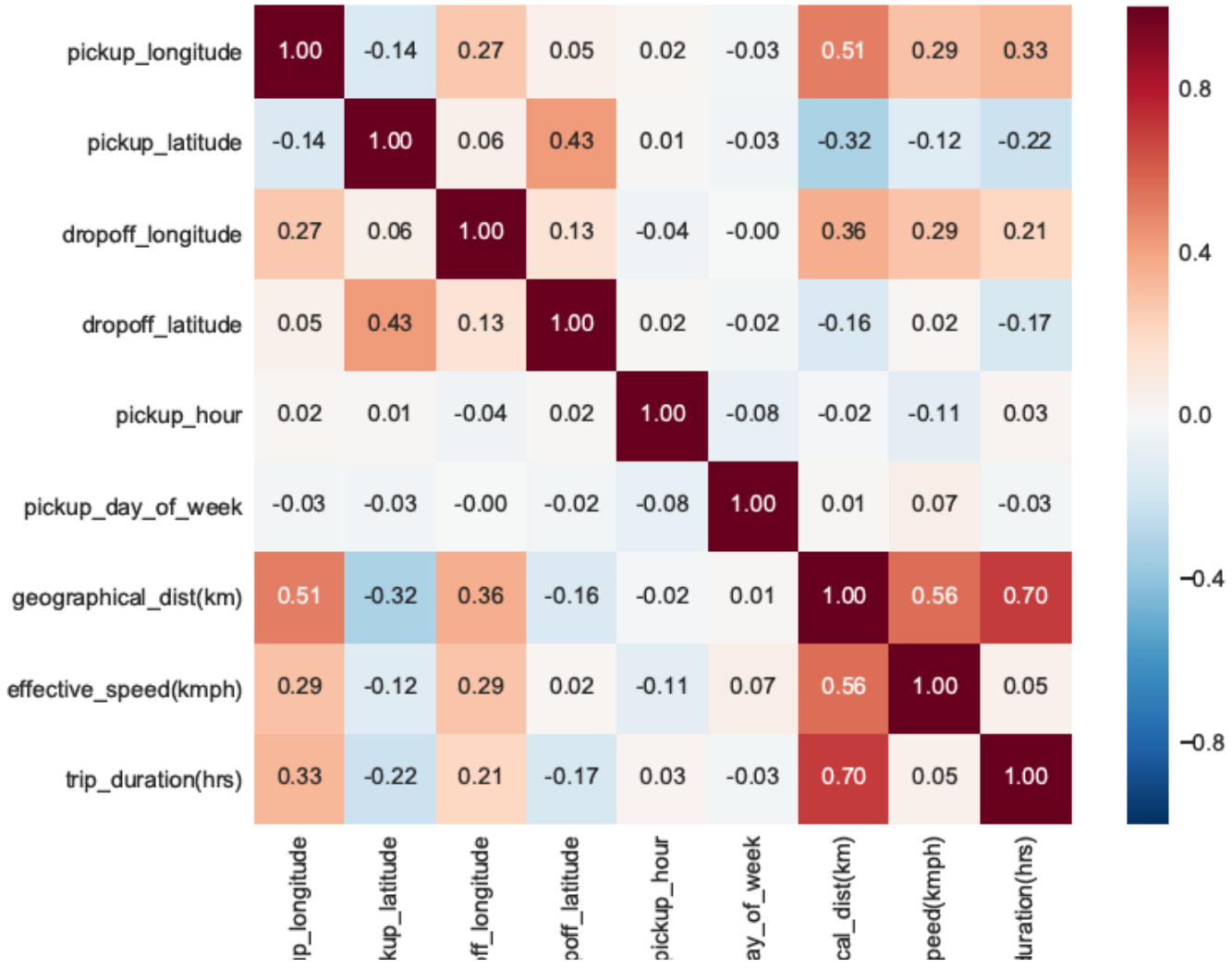






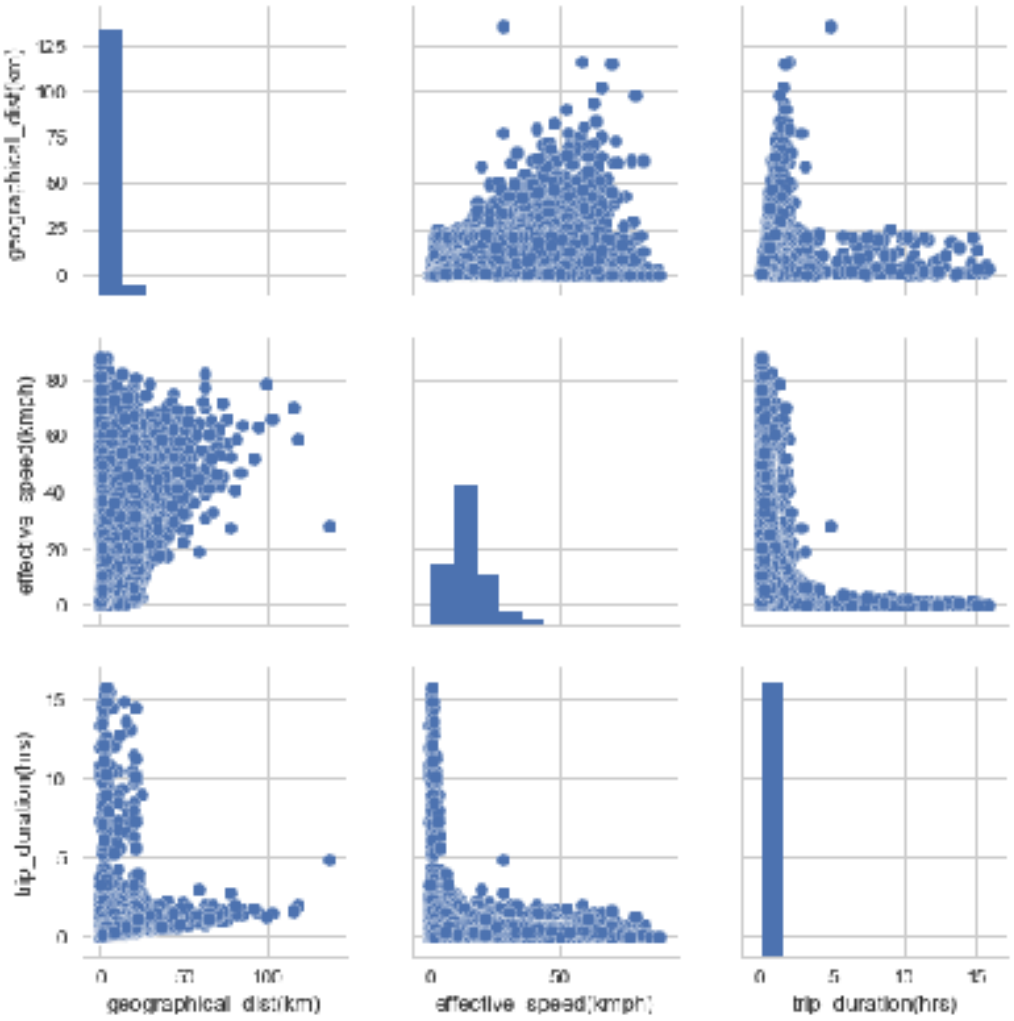
j) Now we plot the heat map of the correlation matrix array choosing the following columns: 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude', 'pickup_hour', 'pickup_day_of_week', 'geographical_dist(km)', 'effective_speed(kmph)' and the target column 'trip_duration(hrs)'. Values in the 'pickup_month'and 'pickup_day' columns, having almost uniform distributions, are almost not correlated with anything else, and we are not plotting these columns. We see that the drop-off and pickup longitudes are rather strongly correlated with each other; the same can be said about the drop-off and pickup latitudes. There is also a notable

correlation between the values in the 'geographical_dist(km)' and 'effective_speed(kmph)' columns, which is not unexpected. There is also weak but far from zero correlation between the values in 'geographical_dist(km)' column and those in the 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude' columns respectively; this latter fact can be explained by the presence of the sea coast near New York city. There is a strong correlation between the 'trip_duration(hrs)' values and the values in the 'geographical_dist(km)' column. The 'trip_duration(hrs)' values are only a little bit correlated with the values in the 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude' columns.



k) Finally, we make the seaborn pair plot involving the columns 'geographical_dist(km)', 'effective_speed(kmph)', and 'trip_duration(hrs)'. It follows from the plot that geographical distances between the pickup and dropoff locations are narrowly distributed with a small number of large distances. The same is true for the trip durations -- there are few instances with very large time intervals. Large trip durations probably mean substantial waiting periods during the trips as follows from the scatter plots 'trip_duration(hrs)' vs 'geographical_dist(km)' and 'effective_speed(kmph)'. One should pay attention to the broad distributions of effective speeds. It follows from the data that low effective speeds correspond to the presence of notable waiting

periods during which there was no motion of the cab, but taximeter kept counting. One can also conclude that the larger geographical distance the higher the effective speed is. Also, the larger trip durations correspond, on average, to smaller effective speeds, as expected.



We have now the modified data-frame that is ready for predictive modelling using certain methods. We convert it back to the csv file ('NYCTripDuration_modified.csv') to use it in another program.