Capital One Data Science Challenge: Trips!

For this challenge, we use data collected by the New York City Taxi and Limousine Commission about Green Taxis. We use data from September 2015.

Question 1

- *Programmatically download and load into your favourite analytical tool the trip data for September 2015.*
- *Report how many rows and columns of data you have loaded.*

The dataset was downloaded and imported into Jupyter Notebooks using Pandas. The input dataset ('green_tripdata_2015-09.csv') must be placed in the Jupyter working directory prior to running this notebook.

Number of Rows: 1,494,926 Number of Columns: 21

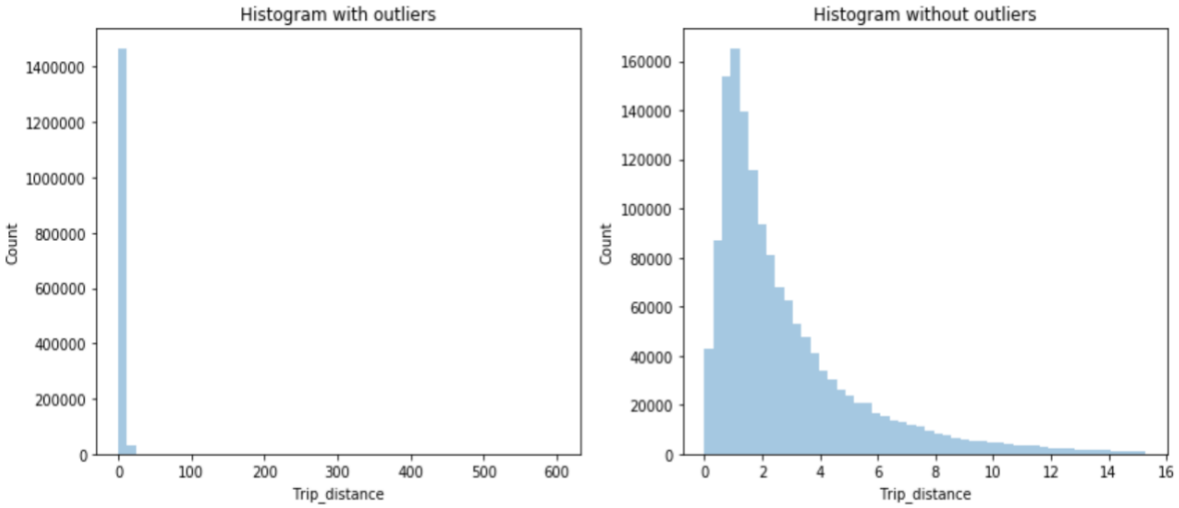
Question 2

- *Plot a histogram of the number of the trip distance ("Trip Distance").*
- *Report any structure you find and any hypotheses you have about that structure.*

	Trip_distance
count	1.494926e+06
mean	2.968141e+00
std	3.076621e+00
min	0.000000e+00
25%	1.100000e+00
50%	1.980000e+00
75%	3.740000e+00
max	6.031000e+02

Initially, I used the describe() function on the Trip distance variable in the data to study it further. We can see that the median is about 2 mi and the mean is 2.98 mi. This highest point of 603 mi is clearly an outlier. Usually, it is safe to assume that points 3 std above mean as outliers. In this case, however, this meant that even trips with trip distances of 11 miles were being considered outliers. Since 11 miles is a reasonable trip distance (especially for airport trips), we can remove anything above 4 standard deviations above the mean.

I then plotted the histogram of Trip distances. This is the graph I obtained:



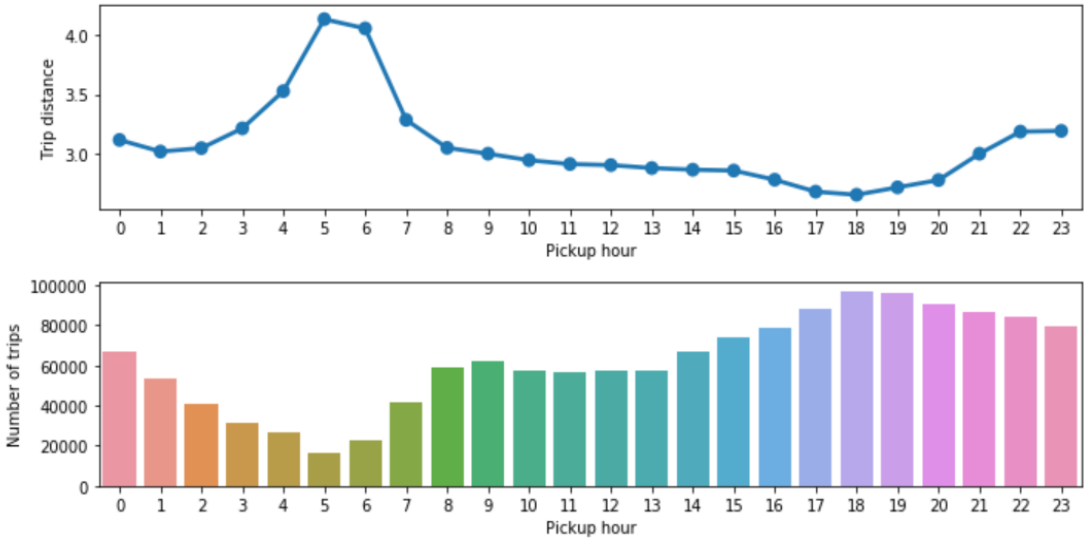
Structure and Hypothesis

The distribution seems to be a right skewed distribution with positive skewness. It seems to fit a lognormal distribution.

The lognormal fit makes sense because trip distance cannot be negative.

The lognormal fit shows that the trip distance is not completely randomly distributed (does not follow a normal distribution). Rather people seem to take the taxi much more for short distances (below 20 mi). This may tell us that people prefer to take taxis for smaller distances, and use other forms of transport for longer hauls. This is due to the fact that taxi fare is based on the distance travelled, and may get prohibitively expensive as distances increase. As distances increase, it makes more sense to use public transport/personal transport or rent transport by time.

Below, I look at the mean Trip distances and number of trips across hours of the day to figure out other patterns.

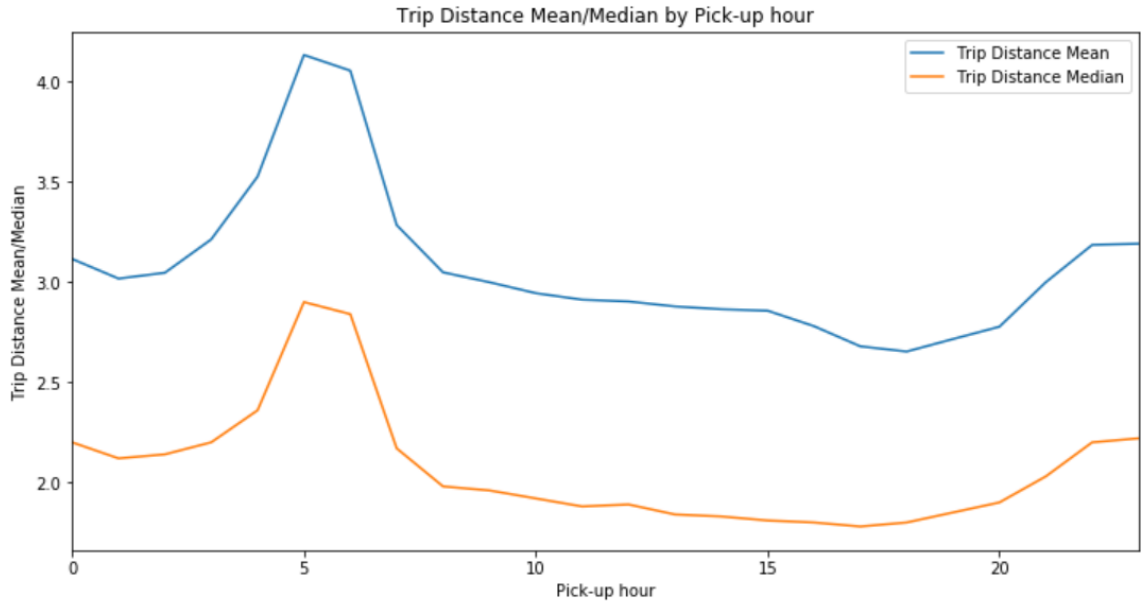


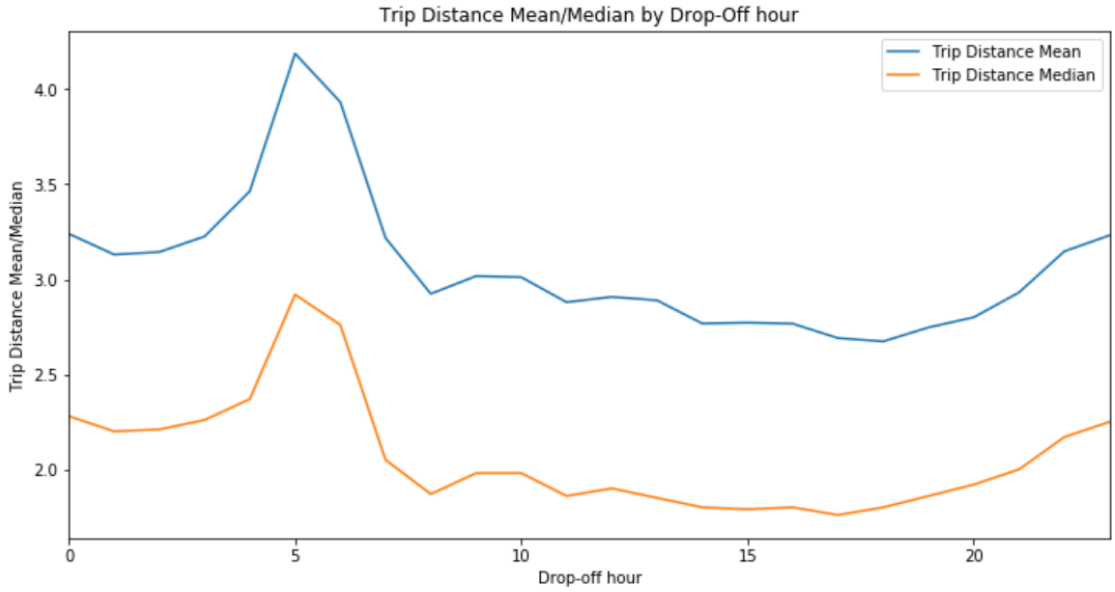
We can see that trip distances are high during the early hours of the morning (peaking at 5 am), and are lowest during the rush hours both in the morning and evening. We can also see that the number of trips is highest (understandably) during the rush hours in the morning around 9 am and the evening around 6-8 pm. This shows that a lot of people seem to be taking taxis during rush hours, and these trips are usually short. This could mean that people use these taxis to travel to and from work, people prefer to take subways or other forms of transport when the roads are congested during rush hours if they have to travel long distances.

Question 3

- *Report mean and median trip distance grouped by hour of day.*
- *We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fare, and any other interesting characteristics of these trips.*

For this question, I plot the mean and median distance by pickup and drop off hours. The graphs are below:





From the above graphs, we can see that the maximum trip distance seems to be travelled around 5 am. After this peak, the trip distance reduces, and reaches a low during the afternoon times. We can see an increase in the trip distance again towards midnight.

Airport trips

From the dictionary of variables, we can see that the RateCodeIDs of 2 and 3 correspond to JFK and Newark respectively. However, this RateCodeID leaves out an important airport in NYC, LaGuardia.

So we can choose an area around each airport, construct polygons on shapely and go from there. We can see how many trips originate or end within these polygons to determine the total number of airport trips.

I use the API on http://www.get-direction.com/ and Google Maps for the airport latitudes and longitudes, for the purpose of putting in an approximate buffer around the airports

The number of JFK and EWR trips using the RateCodeID field is 5552. This number seems to be low compared to the total number of taxi trips. Therefore, I decide to use Shapely for all the airports (JFK, La Guardia and EWR)

Using this approach, I create variables called 'Is_JFKPickup', 'Is_JFKDrop', 'Is_NewarkPickup', 'Is_NewarkDrop', 'Is_LGAPickup', 'Is_LGADrop' which denote airport pickup and drop-off trips.

Once this is done, I combine all these variables into one variable, 'airport', which denotes whether a trip involved an airport pickup or drop-off.

After this, I analyse the airport trips in detail.

The following are the total pickups/drops and average fare of each airport.

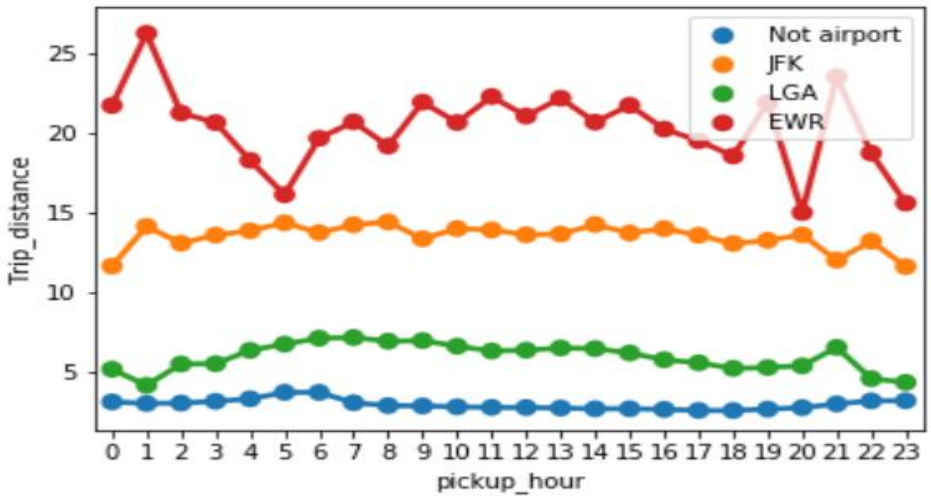
Airport	Total Pickups/Drop-offs	Average Fare
JFK	7814	41.3
La Guardia	16103	21.04
EWR	738	75.68

I also looked into some interesting characteristics of airport trips

	Average Fare	Average Tolls
Airport Trips	29.1	1.36 (EWR is 9.57)
Non-airport Trips	12.27	0.10

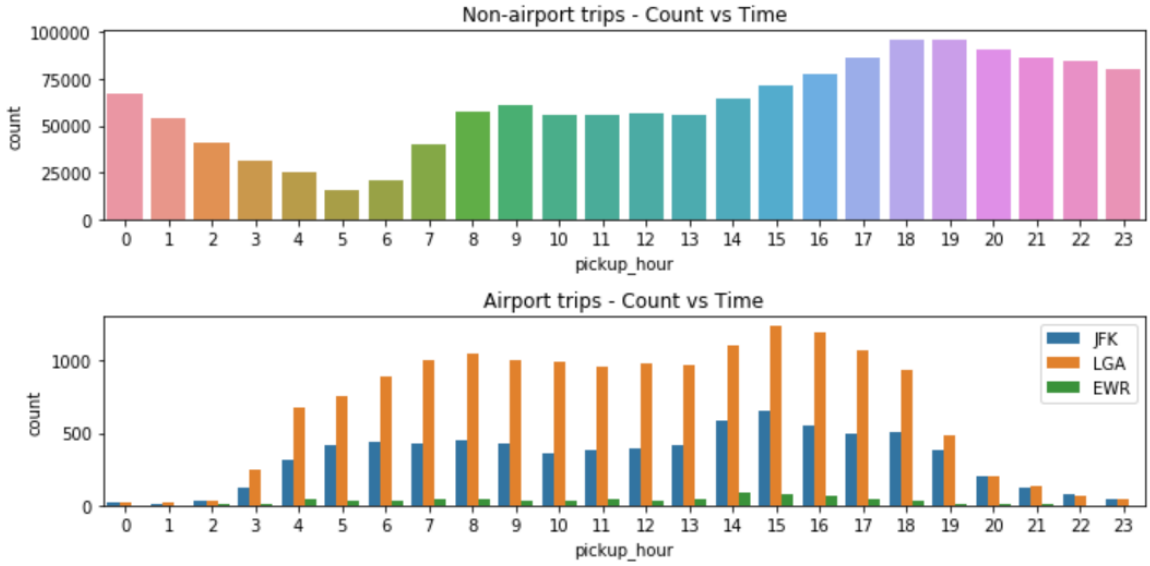
As we can see, the average airport fare is much higher than the normal non-airport trip fare. This is understandable because airports are typically farther away from the places of living and cost more compared to a trip to work.

When it comes to airport tolls, EWR tolls are higher compared to other airports which makes sense since driving from NJ to NYC or vice versa involves turnpike tolls, Lincoln tunnel tolls and so on. Next, I look into the mean Trip distances of airport and non-airport trips across various pick-up hours



As we can see, trip distances are more across pickup hours from and to airports. Newark airport trip seem the longest, understandably. La Guardia, being a local airport, has shorter trips, and JFK distances are between LGA and Newark

Finally, I examine the timing of the airport rides.



From the above graphs we can see that the busiest times to and from the airport are not the same as rush hours for normal trips.

The trends across all airports remain the same in terms of busy times. 3 pm seems to be the busiest time for airport trips.

To make sense of trip timings, we can take JFK airport as an example. 1pm-4pm are when most Europe flights arrive at this airport. Also, 5pm - 9pm are when most Euro flights depart.

Since people usually reach airports ahead of time, the large number of trips from 2-5 pm makes sense. Similarly, most domestic and Asia-bound flights from JFK fly from 6-10 am, which justifies the large number of early morning riders we observe.

Question 4

- *Build a derived variable for tip as a percentage of the total fare.*
- *Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). Provide an estimate of performance using an appropriate sample, and show your work.*

We can build a derived variable for tip as a percentage of the total fare.

```
In [148]:  1 data['Tip_Percent'] = (data.Tip_amount/data.Total_amount)*100
```

Predictive model for tip as a percentage of the total fare

To build a model to predict the tip percentage, I would like to follow these steps:

1. Data cleaning
2. Feature Engineering
3. Exploratory Data Analysis
4. K-Fold Cross Validation
5. Feature Importance Determination
6. Hyperparameter Optimization using GridSearch CV
7. Performance Analysis of the final model

Data Cleaning

In this step, I clean the input data by handling missing values of data, data which is not reasonable because of possible errors and so on.

1. I removed E_Hail fee. This column has a majority of NaNs and can be removed.
2. 'Trip type' has 4 missing values. I replace them by the most frequent trip type
3. Looking at the RateCodeID column, there seem to be 6 observations with a RateCode ID of 99. I replace them by the most frequent value of RateCodeID
4. Fare amount cannot be negative, 0 or lesser than minimum amount. The minimum amount is given in this article : http://nymag.com/nymetro/urban/features/taxi/n_20286/

I replace negative or lesser than minimum values of the fare amount with the mean value of the fare amount. Another approach would be to regress the fare amounts on trip distance to figure out values for these. The same is done for the 'total_amount' variable. Tip amount, surcharge, Extra and Tolls amount cannot be negative. I replace negative values of these variables with the mean values of these variables.

The data looks good now. As a final step, we can drop the Is_airport_pickup and Is_airport_dropff variables as this information is captured in the 'airport' variable

Feature Engineering

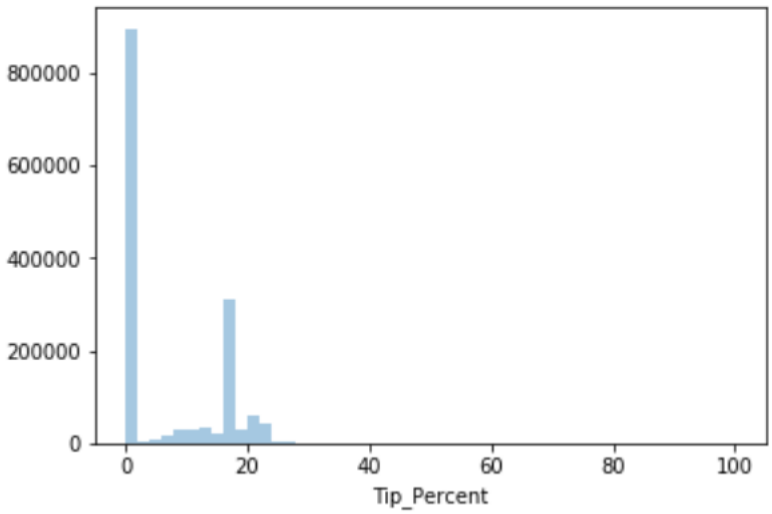
In this section, I add additional features to the data, which might be relevant during tip prediction

1. Day of the trip and trip duration could be possible features. I add these into the dataframe
2. Another feature could be 'is_PayDay'. I would like to look at whether people tip more on their paydays. 9/11/2015 and 9/25/2015 were paydays according to the calendar in NYC: http://www.nyc.gov/html/opa/downloads/pdf/2015_Pay_Calendar.pdf
3. One-hot encoding the Store and Forward Flag

4. Another possible feature could indicate whether the trip was in daylight or the dark. This could be another possible feature. We can look at the pick up time of the trip for this purpose. Drop off time might be in daylight, but we want to examine when the driver picked up the passenger.
5. A very important feature which could influence tips is the speed of the taxi.
6. One feature I had tried to add was the borough in NYC each trip originates and ends in. However, because of time considerations, I decided against this approach. I have included the code for this approach in my final analysis.

Exploratory Data Analysis

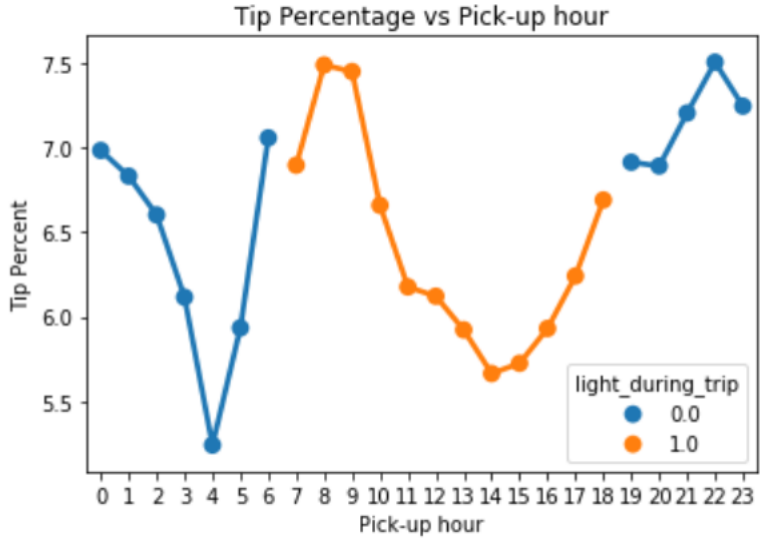
The following is the histogram of the tip percentage



From the histogram as well as describe() a large proportion of tips are seen to be 0%. I found that the number of trips with 0% tips constituted 60% of the total trips!

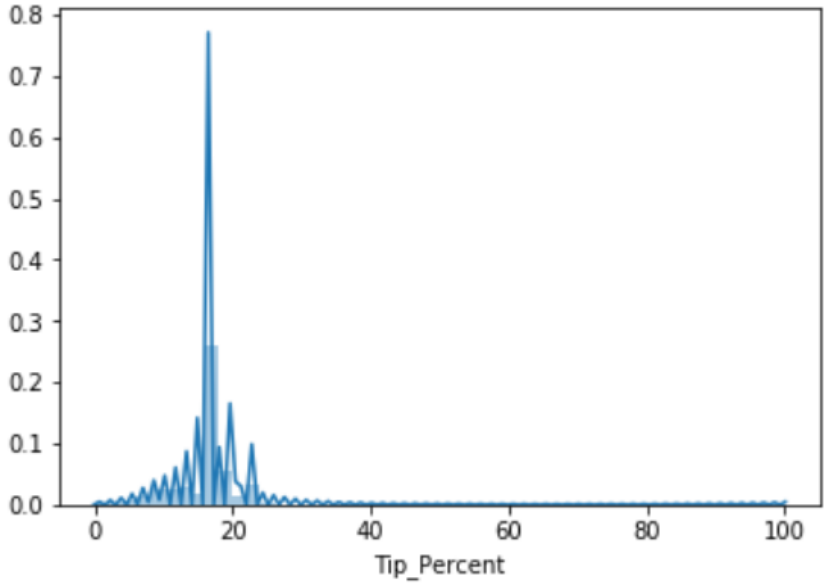
From this point in the analysis, I look at the relationship between the Tip Percentage (target variable) and several of the other variables in the data.

1. First, we can look at the relationship between tip percentage and pick-up hour



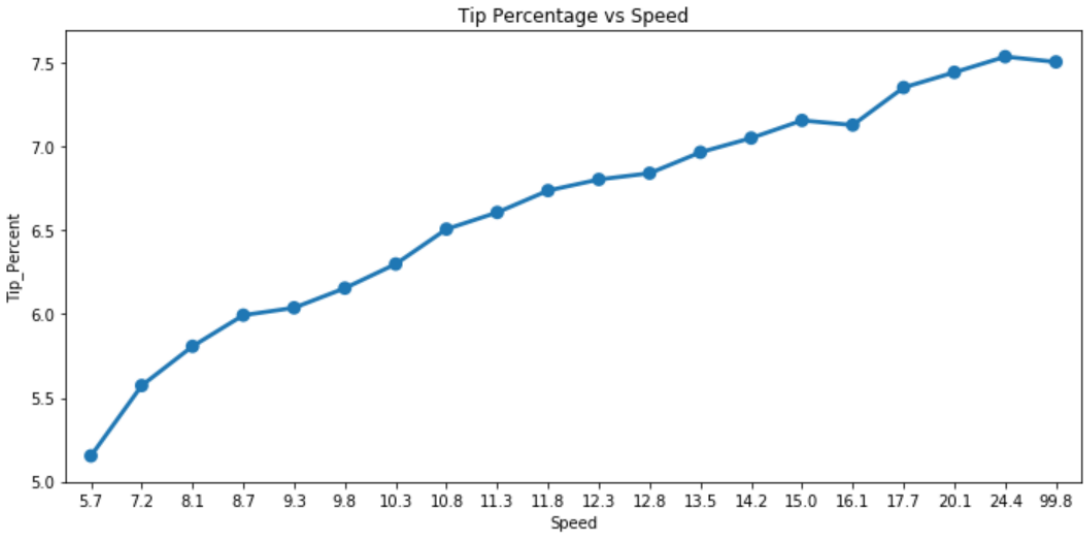
We can see that the tips seem to be on an upward trend during night. Also, tips seem to be high during morning rush hours, and relatively high during evening rush hours. It seems like people might be tipping more when they want to travel to office and so on. Tips go down once rush hour ceases.

2. Next, we look at the distribution of tip percentage among transactions that yielded a tip:



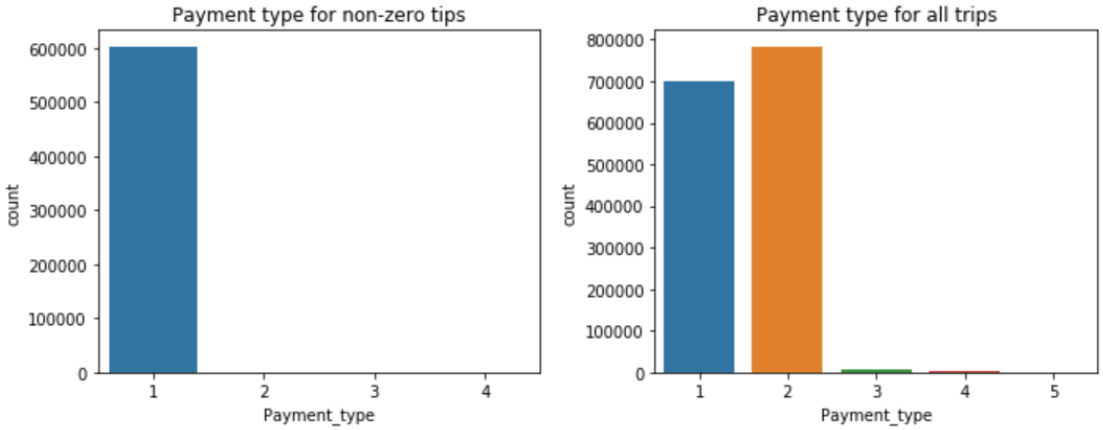
The mean tip percentage among transactions that yielded a tip was 16.45%.

3. Next, I analyse the relationship between tip percentage and Speed. For this purpose, I divide the speed into 20 quantiles and then plot Tip percentage vs Speed



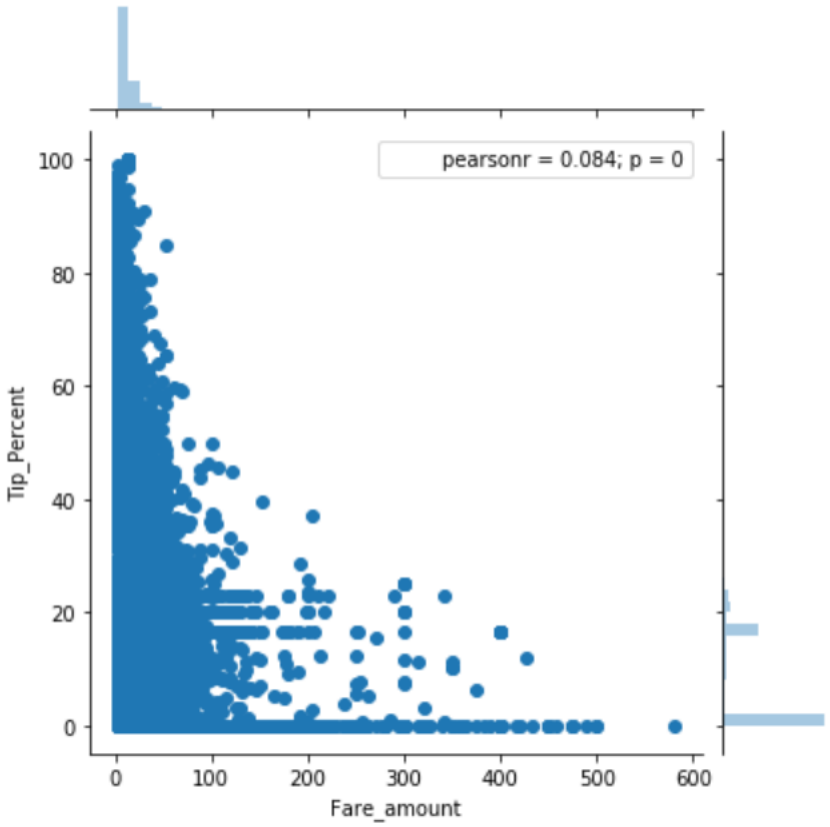
We can see that the tip percentage seems to increase with speed. This could be because people might want to reward their drivers for driving faster, till a certain point. After a certain speed is hit, the tip percent seems to start decreasing. This could be because people do not want the taxis to go too fast, and might decrease the tips if they do. Overall, there seems to exist a strong relationship between the tip percentage and speed.

4. Another important relationship could be between payment type and tip percentage



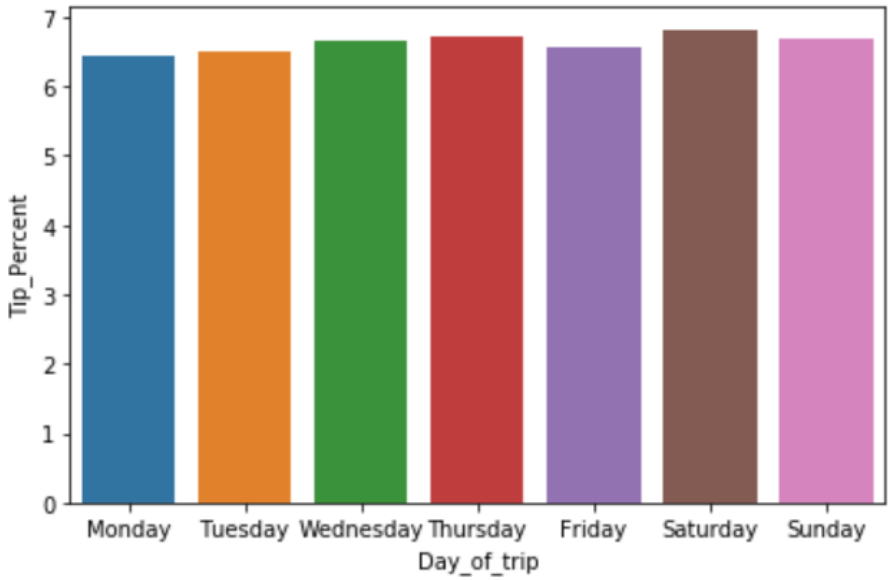
One observation is that almost all the transactions which include tips have been done using credit cards. Though a large number of payments have been made using cash, not many of these yielded a tip. This shows that the payment type could be a very important indicator of tip percentage, with credit card transactions having a large proportion of non-zero tips. One reason for this could be that drivers might not have recorded cash tips in the system, whereas tips using credit cards are automatically entered in the system.

5. Fare amount could have a relationship with tip percentage. To check, I plotted a joinplot between Fare amount and tip percentage



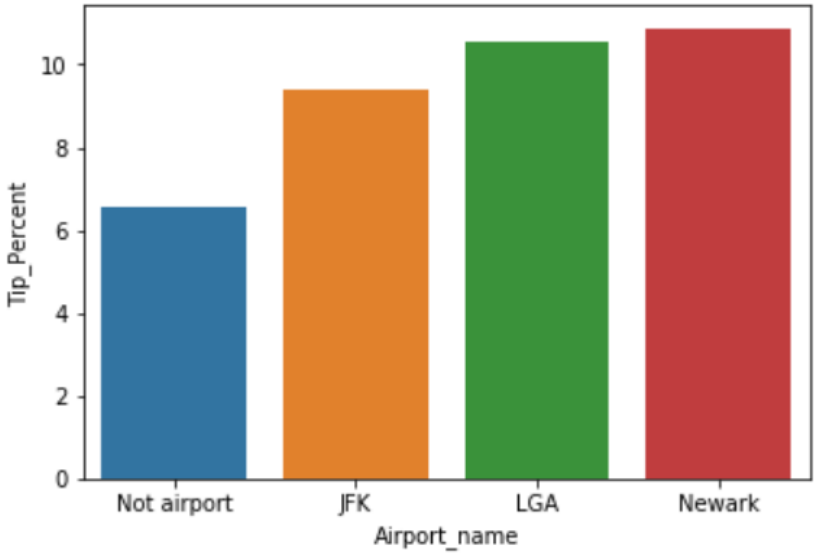
Tip percent decrease as fare increases. One reason could be that people probably do not want to pay large tips on top of high fares. Also, some people may pay the same tip amount each time, and the tip percent seems high when the fare is smaller as compared to tips on large fares for the same tip

6. Day of the Week vs. Tip Percentage:



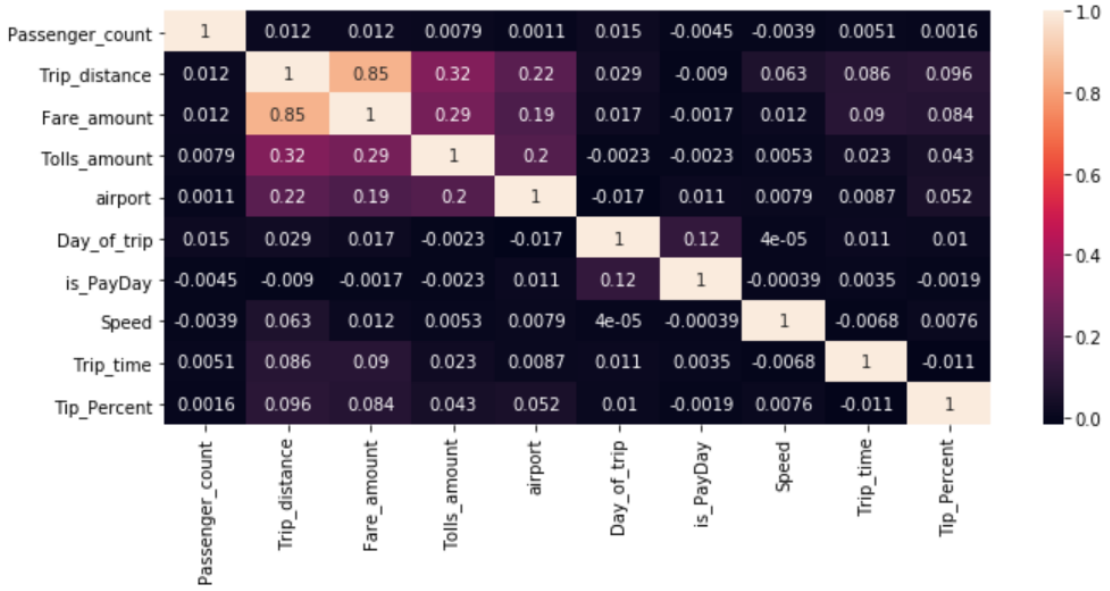
As expected, the tip percentage is slightly higher on weekends.

7. It would be interesting to look at whether airport rides pay more tips as compared to non-airport rides. I plotted a graph to deduce this:



We can see that airport rides seem to have more tips compared to non-airport rides. So this variable is correlated with tip_percent.

8. Finally, we can study the correlations between the available variables and the target variable, as well as between variables using a heatmap:



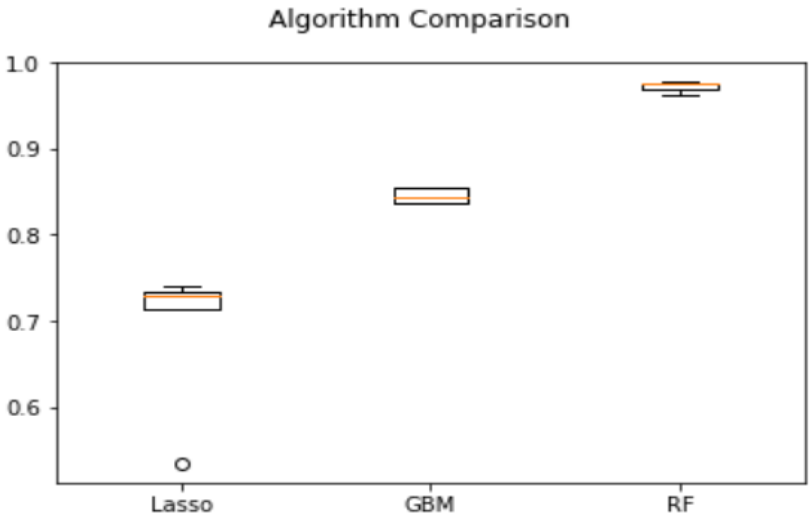
There seem to be a few major correlations between the variables. First, between Trip distance and Fare amount, which is to be expected. While selecting features to build the final model, it would be adequate to use just the Fare amount as a feature.

K-Fold Cross Validation

We need to build a regression model for predicting the tip percentage. As the first step in this process, I perform K-fold cross validation to select a model which has a high accuracy. For this step, I consider three potential models - A Lasso regressor, Gradient Boosting Machine using the XGBoost library and a Random Forest Regressor. I consider only a subset of the training set (50000 samples) to save time

The following is the graph comparing the performance of these three algorithms

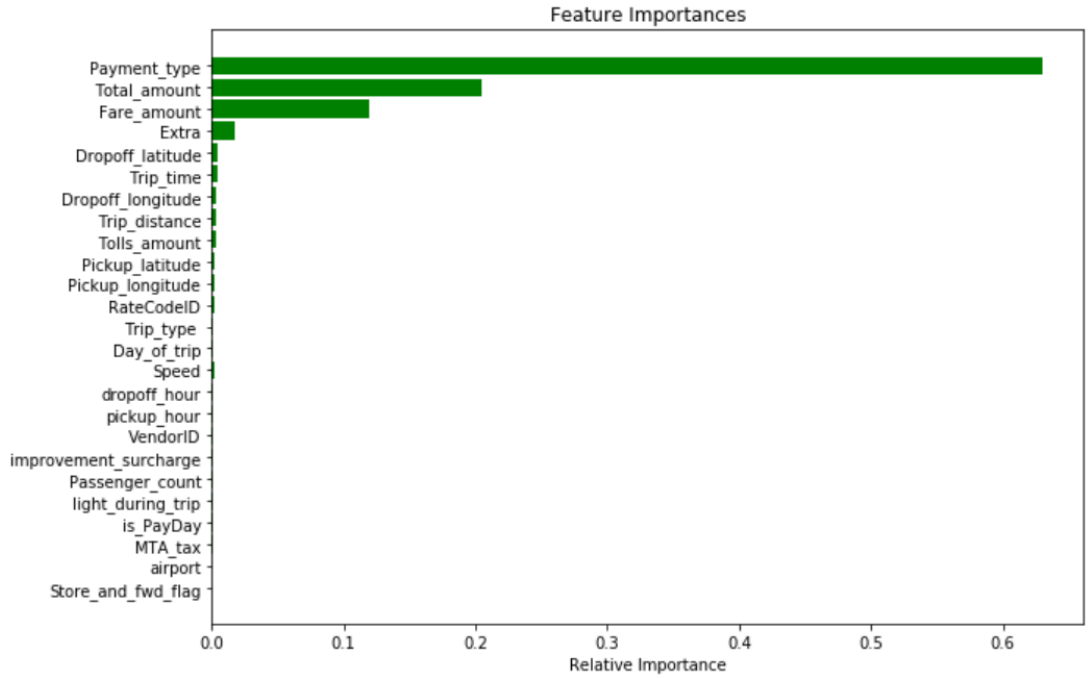
```
Lasso: 0.690686 (0.078367)
GBM: 0.845420 (0.007700)
RF: 0.971616 (0.005609)
```



As we can see, the Random forest regressor seems to have the highest accuracy. I will use a Random Forest Regressor to predict the tip percentage.

Feature Importance

I will use a random forest regressor to derive the feature importance and use the top 10 features in my final model. Below is the graph depicting this feature importance:



Based on the above feature importance figures, I can pick up ten important features. These are - Payment_type, Total_amount, Fare_amount,Dropoff_latitude,Extra, Trip_time, Trip_distance, Speed, Dropoff_longitude, Tolls_amount. Since we saw that there was a high correlation between the Trip Distance and Fare amount, I will use only the Fare amount as a feature for tip prediction and leave out Trip Distance.

Hyperparameter Optimization using GridSearch CV

To fit the final model, I use train_test_split() to partition the entire dataset into training (80%) and test data (20%) and use the training data to fit the final model.

Using GridSearch CV, I tune the parameter 'n_estimators' which denotes the number of estimators in the final Random Forest Model

The following was my Grid Seach CV output:

```
Fitting 3 folds for each of 3 candidates, totalling 9 fits
[Parallel(n_jobs=-1)]: Done   7 out of   9 | elapsed:  5.4min remaining:  1.5min
[Parallel(n_jobs=-1)]: Done   9 out of   9 | elapsed:  7.4min finished
Out[341]: GridSearchCV(cv=3, error_score='raise',
       estimator=RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
           oob_score=False, random_state=None, verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=-1,
       param_grid={'n_estimators': [10, 15, 20]}, pre_dispatch='2*n_jobs',
       refit=True, return_train_score='warn', scoring=None, verbose=2)
```

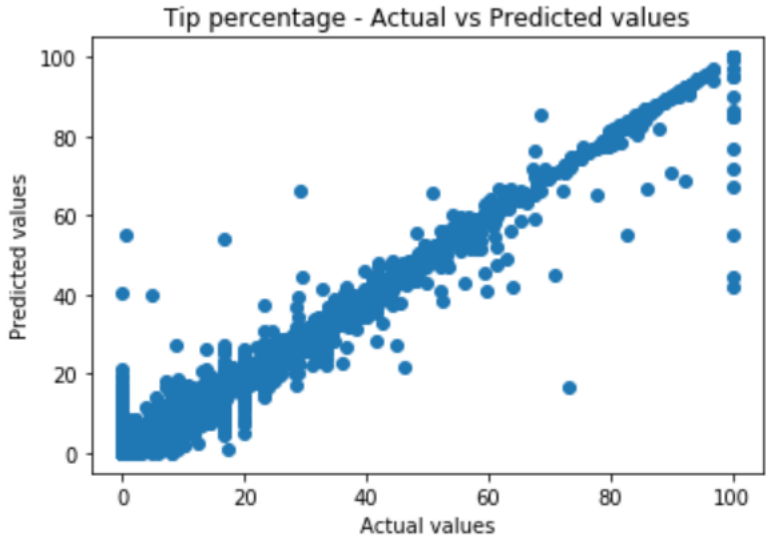
Finally, we get the final model from the Grid Search CV output.

```
#Getting the optimized random forest after hyperparameter optimization

optimized_rf = grid_search.best_estimator_

```

Performance of the final model

To measure the performance of my final model, I looked at key metrics like R^2 error and Mean Squared Error. I also plotted the scatterplot of the actual vs predicted values to analyse performance.

1. The R-Square of the model is 0.997. This is very high, and indicates great model performance.
2. The Mean Squared Error on the testing set is 0.20 and the MSE on the training set is 0.048. The MSE on the training and test set are very small, and this is a good indicator.
3. Finally, I look at the scatter plot of the actual vs predicted values.



Conclusion

The above Random Forest Regression model is very highly accurate, as can be seen from the R^2 of 0.997. The mean square error of this model is also very small. The scatter plot of the actual vs predicted tip percentage values show that all the points are very close to the regressed diagonal line. From these key performance metrics, we can say that the model is very successful in predicting the Tip Percentages using New York Green Taxi data.

Question 5

- *Build a derived variable representing the average speed over the course of a trip.*
- *Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?*
- *Can you build up a hypothesis of average trip speed as a function of time of day?*

I have already created the derived variable Speed, which was instrumental in the prediction of the tip percentage

Testing to determine if average speeds were materially the same across all weeks of September

First, I remove outliers in Speed. Speeds cannot be very high, as the speed limit in NYC was 25 mph in 2015. To determine if the average speeds are materially the same across the weeks of September, we can perform a Oneway ANOVA test. The following was the output from the ANOVA test

```
#Performing the ANOVA test
import scipy.stats as stats
stats.f_oneway(speed_week1, speed_week2, speed_week3, speed_week4, speed_week5)
```

```
F_onewayResult(statistic=949.4715066784031, pvalue=0.0)
```

As the p-value is very small (0), we can definitely reject the null hypothesis. Therefore, the average speeds are materially different across the weeks of September. I found the average speeds across the weeks of September as follows:

```
Average speed in week 1: 13.273665649
Average speed in week 2: 12.6121427438
Average speed in week 3: 12.6114066918
Average speed in week 4: 13.0857226418
Average speed in week 5: 12.4190777996
```

We can see that the average speed decreased from Week 1 to 2 and stayed almost the same in Week 3. After this, the speed increased in Week 4, finally dropping to the lowest speed among all the weeks in Week 5.

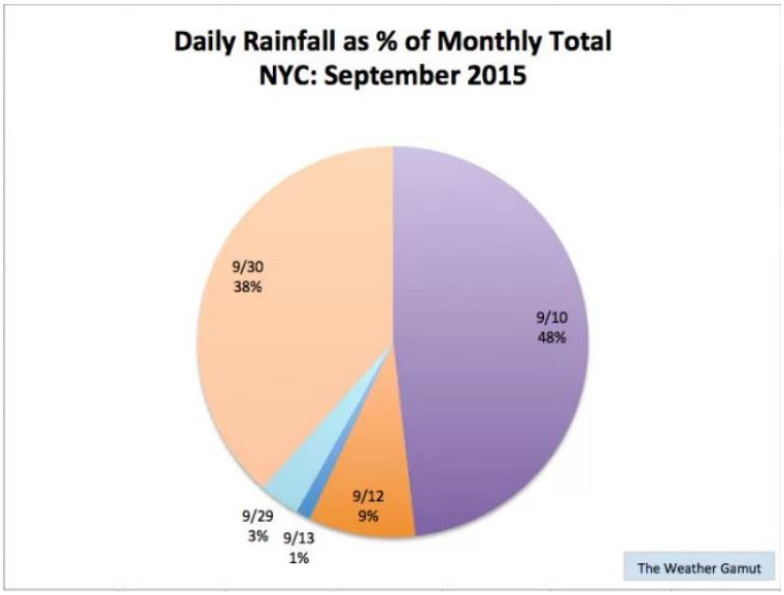
Hypothesis for differing mean speeds across the weeks of September 2015

Speed can be different on different weeks because of factors like important events happening in NYC around different weeks of September.

One example of this is the US Open which was hosted in NYC in September 2015. It continued till 13th September. This might have let to the congestion of roads as more people travel to NYC to watch this event and avail taxis, and hence taxi speeds could have decreased towards the end of the event, around the second week.

Another key factor which could be influencing the taxi speeds is the weather across the weeks of September 2015 in NYC. The following website details the weather across the weeks of September 2015: http://www.weathergamut.com/2015/10/01/nyc-monthly-summary-september-2015/

As we can see, in weeks in which heavy rainfall was recorded, we can see lesser average taxi speeds, probably because drivers want to drive more safely in bad weather and rain. In weeks in which the weather was sunny, like the first and fourth week, speeds appear to be more.

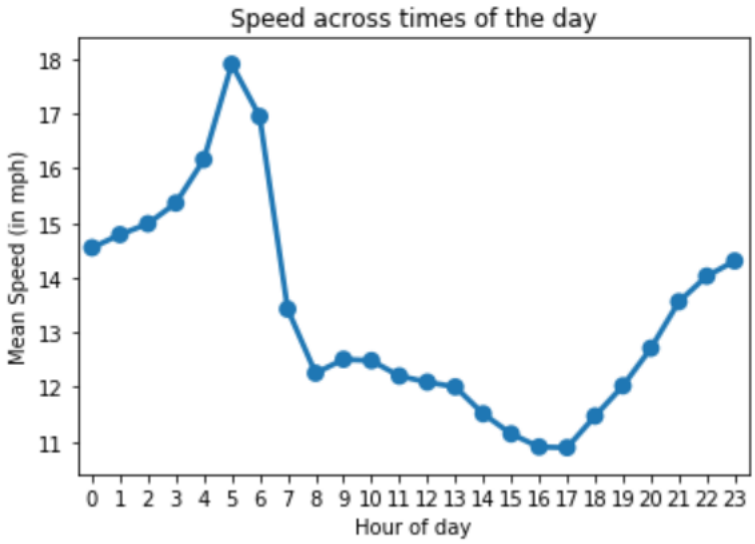


Average Trip Speed as a function of the time of day

I perform the One-Way ANOVA test again, for speeds across the hours of the day. The following is my output

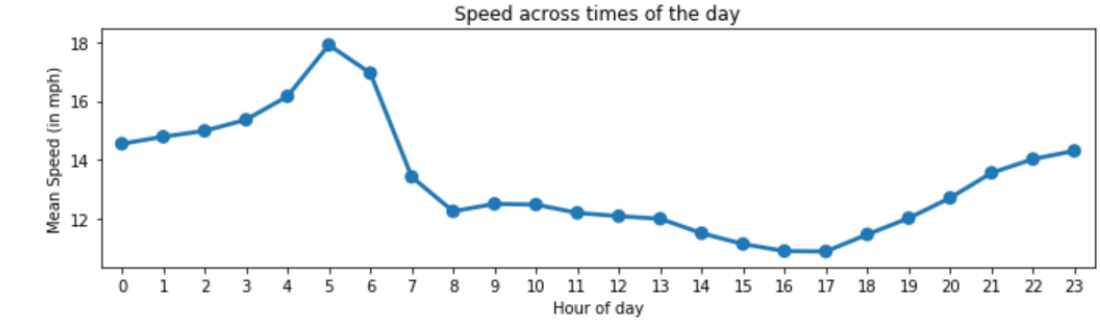
```
F_onewayResult(statistic=4585.9597987349352, pvalue=0.0)
```

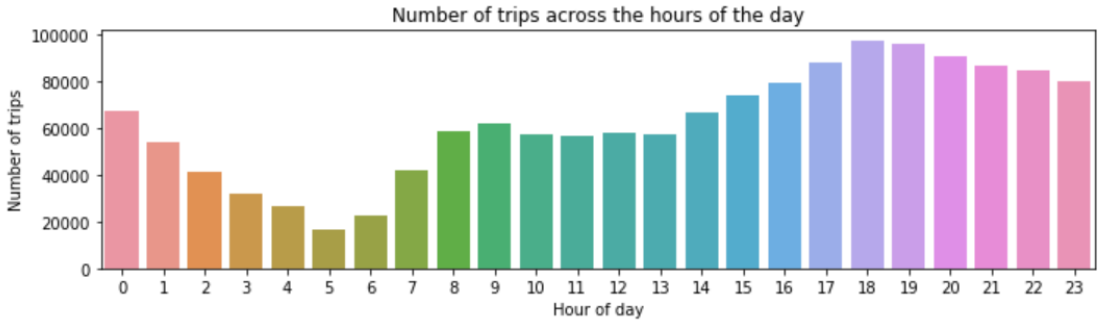
Again, the zero p-value indicates that we can reject the null hypothesis and conclude that the speed varies materially across the time of the day. I plot the average speeds across the hours of the day



Hypothesis of varying speed across times of day

1. Effect of rush hours on Speed





We can very clearly see the effect of rush hours on speeds. The average speed is very low during the morning rush hours (8-10 am) and during the evening rush hours (5-7 pm). This might be because a lot of people will be travelling to and from work during these times, and so the roads might be very congested. Therefore, the average speed decreases.

During the less busy afternoon hours, the speed slowly picks from the morning rush hour speed only to decrease again towards the evening rush hour. As night approaches, the roads clear out again, and the speeds pick up. The highest speeds seem to occur at 5 am in the morning. We can see that this is also the time when the least number of trips occur. Therefore, the roads are very empty and people can travel faster.

Second, the speeds during morning are greater than the evening and night speeds. This might probably be because of greater visibility in the morning and passengers wanting to travel slower at night to be safe

2. Effect of airport trips
   As a part of the answer for question 3, we could see that a lot of airport trips are undertaken early in the morning around 5 am in NYC. Since airport trips are mostly along highways and Freeways, the average speed of these trips might be much higher than normal trips. Since many of these trips are around 5 am, we might be seeing higher speeds at this time. The average non- airport trip speed is 15.54 and airport trip speed is 30.53

To conclude, we can say that the average speeds of taxi trips differ materially across the weeks of September as well as across the hours of the day.