SUBMITTED BY:

JAGPREET SINGH SETHI

Table of Contents

Question 1:

Part 1: Programmatically download and load into your favorite analytical tool the trip data for September 2015.

To download it programmatically, I have used urllib library to download files from Amazon S3 server.
https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015-09.csv

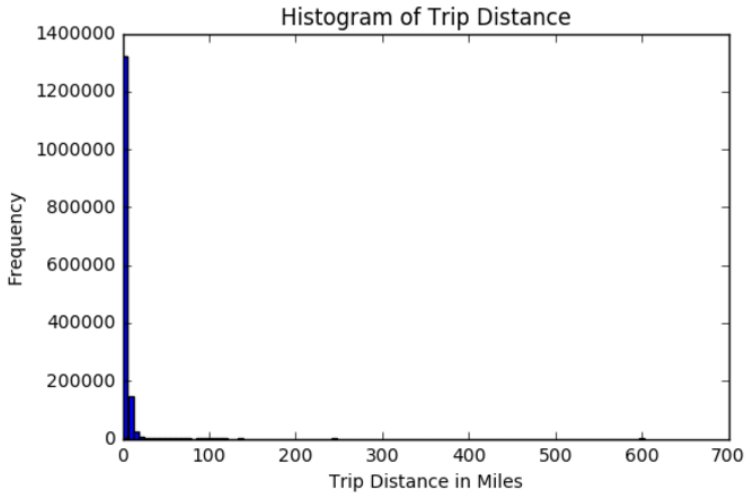
Further it was loaded in Jupyter Notebook to perform further analysis.

Part 2: Report how many rows and columns of data you have loaded.

NYC Green Taxi - Sep'15 Dataset contains 1494926 rows and 21 columns

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.



Histogram of Trip Distance feature is highly right skewed. This gives us an understanding that there are few records whose trip distances are exceptionally above mean. Therefore, there is good likeliness that these records could be outlier or incorrect entries. So, we should remove such records as these will affect histogram range.

To identify incorrect enteries that could affect the Trip Distance field we'll be consider few cases

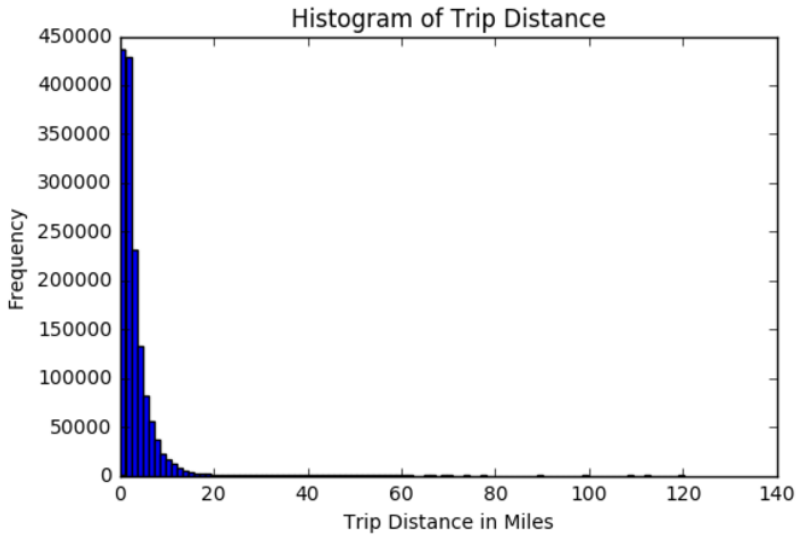
Case 1: As per the NYC road limits and the fact that NYC Green Taxi operate in city, it would be safe to assume that they won't exceed 40 miles per hour or 0.011 miles per second. Using this fact, we'll create a new field 'Speed' and filter records that have MilesPerSecond less than 0.011.

Case 2: Any record with Trip Time equal to 0 are likely to be incorrect because customers won't pay if they haven't book a ride. Hence, such records are meaningless and should be removed.

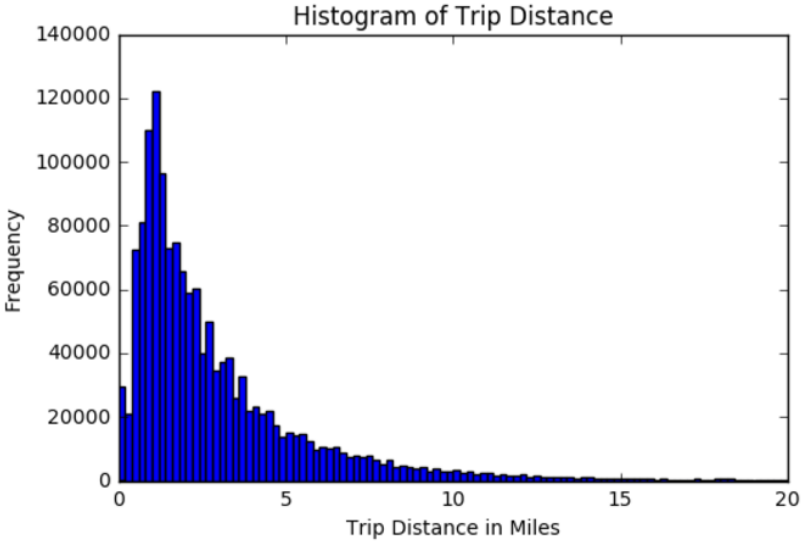
Case 3: It is always the customer who pays to the driver for the ride service provided. Therefore, Total_amount should always be greater than zero. We shall remove all records where total_amount is negative.

Case 4: As observed in the histogram, there are few cases where Trip_distance and Fare_Amount is zero.

Based on this, we create a more practical histogram



Let's zoom the area by limiting the x-axis range to 20 to understand better the distribution of most frequent Trip Distances.

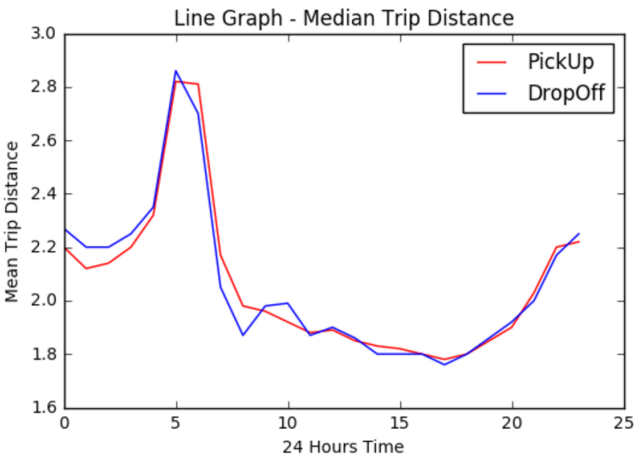


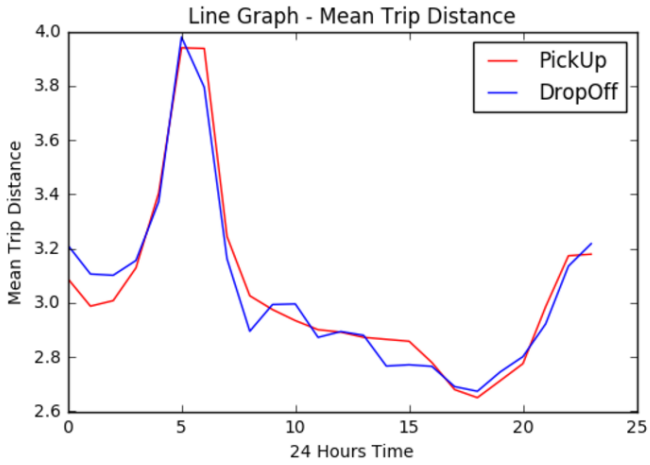
***Interpretation**:*

1) There are rides when NYC Green Taxi was used to travel as long as 140 Miles.
2) Most of the trip rides were short distanced from 1 mile to 4 miles. In other words, from one place to another within city or borough.
3) 95% of the rides were with 10 mile radius.
4) From the Fare Amount histogram, we get an understanding that despite the fact Trip_distance is equal to Zero, customers were charged reasonable fare amount. These could be the cases when Green Taxi was booked and they made them wait for long time.

Question 3

Part A: Report mean and median trip distance grouped by hour of day.





Interpretation:

1) Maximum distance travelled by commuters is between 4 AM to 7 AM. Such commuters most likely stay at the outskirts of NYC and use Green Taxi to reach nearest public transportation so that they can reach their work place.

2) During office hours like 10 AM to 4 PM, people are at their work place which is often in an area with lot of restaurants and shopping malls. So, incase someone hires Green taxi during these hours they are likely to hire it for short distances as almost everything would be in their vicinity.

3) In the evening at around 5PM, people like to travel back to their home which is at an considerable distance from work place.

Note: Mean metric is very sensitive to outliers and hence, we see mean trip distance at 5AM to be higher than median trip distance at 5AM. Here, for our analysis we are using MEDIAN as a metric for analysis as it is robust to outliers.

Part B: 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 fair, and any other interesting characteristics of these trips.

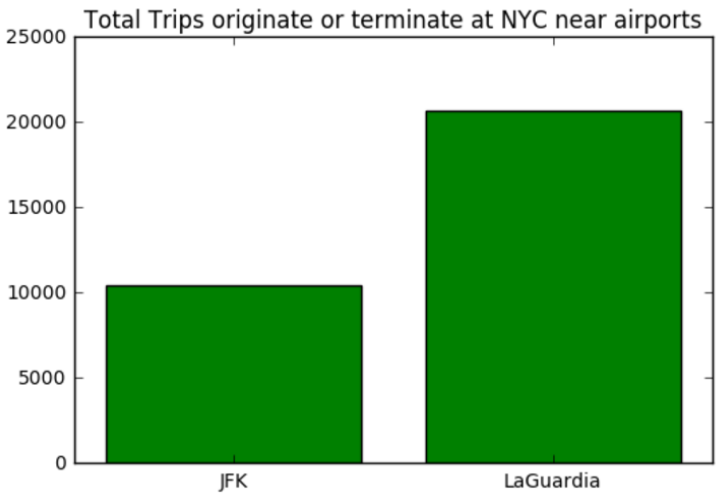
There are two prominent airports in the area of New York:
1) JFK Airport (JFK)
2) LaGuardia Airport (LGA)

Newark Airport is considered to be in New Jersey, So I am excluding that out.

NYC Green Taxi Dataset's variable description is provided in the document: http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf. According to this document, JFK Airport is labeled as 1 in RateCodeID column.

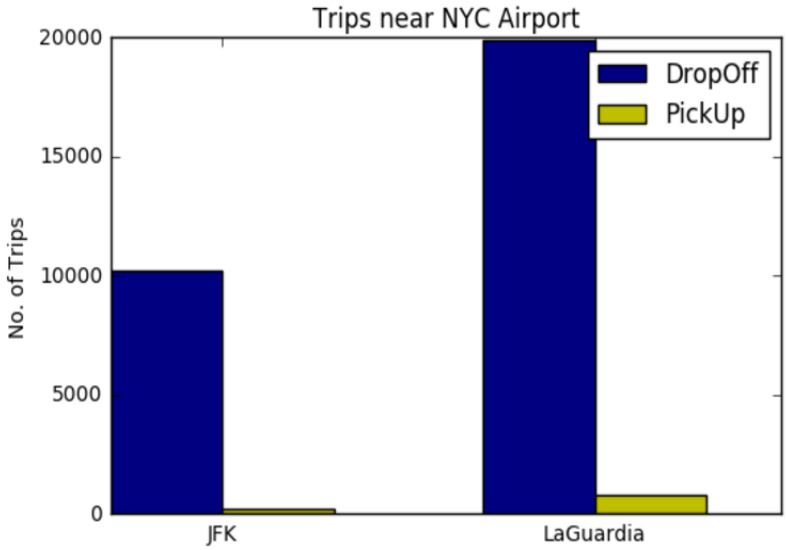
However, no information is provided for the other two airports i.e. LaGuardia Airport and Long Island MacArthur Airport.

To figure out if there was a dropoff to or pickup from LaGuardia Airport, we'll be measuring distance between LaGuardia Airport's geospatial coordinates (Latitude: 40.771765, Longitude: -73.872434) and dropoff or pickup coordinates. To measure this distance we are using 'Great-Circle distance' calculated using Haversine formula.



Interpretation

1) The count of originate or terminate trips for LaGaurdia airport is much higher than JFK airport.



Interpretation

DropOffs are much higher than PickUps for both the NYC airports. This signifies people do prefer Green Taxi from the city to reach airport. All these trips needs to be pre-booked and hence, we see only the drop-off traffic.
However, neither can customers hail these green taxi's at the Airport areas nor can they pre-book these taxis from different city. That's why we see only the DropOffs for both the airports.

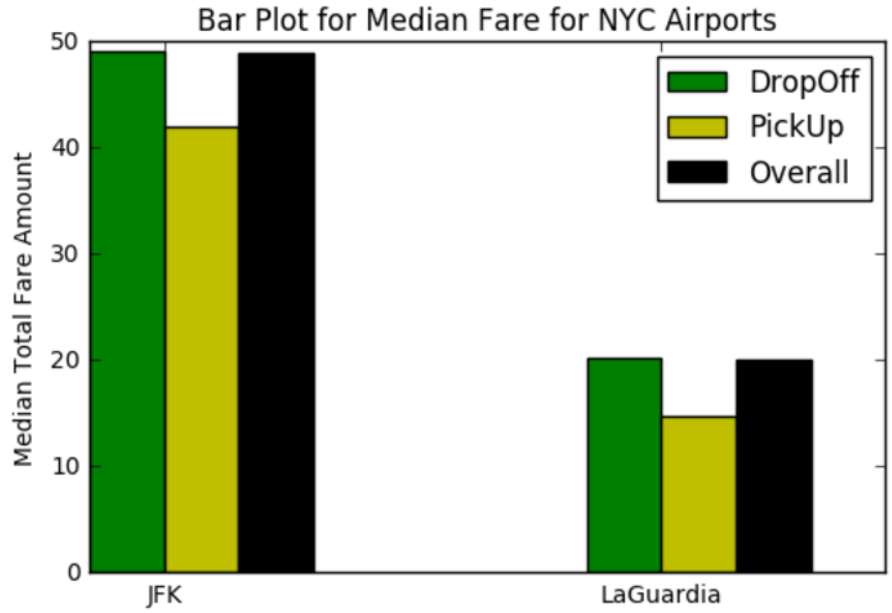
JFK AIRPORT:

Number of DropOff at JFK Airport is 10225
Number of PickOff at JFK Airport is 235
Total Number of trips that originate or terminate at JFK Airport is 10460

LAGUARDIA AIRPORT:

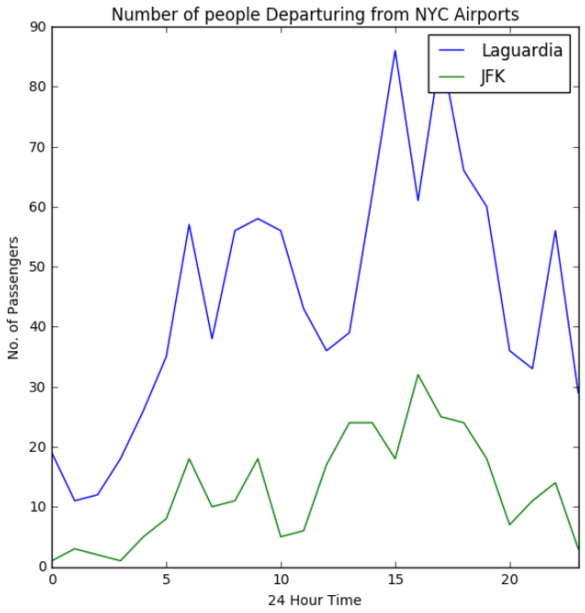
Number of DropOff at LaGuardia Airport is 19889
Number of PickOff at LaGuardia Airport is 799

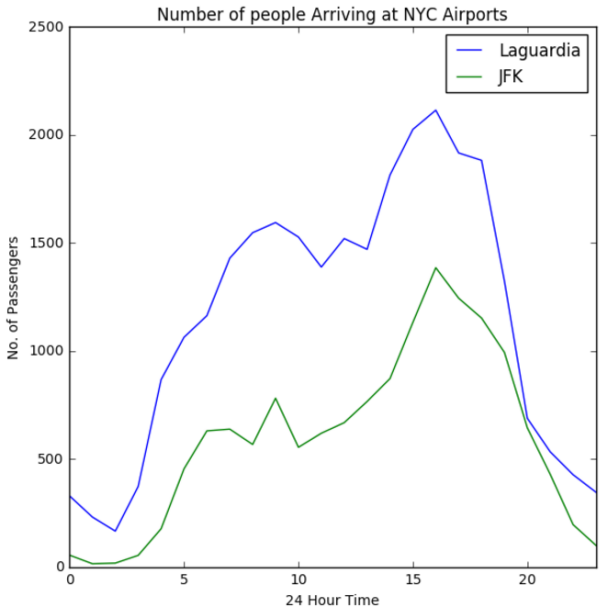
Total Number of trips that originate or terminate at LaGuardia Airport is 20688



Interpretation:
Median total fare amount for DropOff or PickUp trips to JFK Airport is roughly double as compared to Laguardia airport.

Interesting Question: At what time maximum number of people transit from these airports

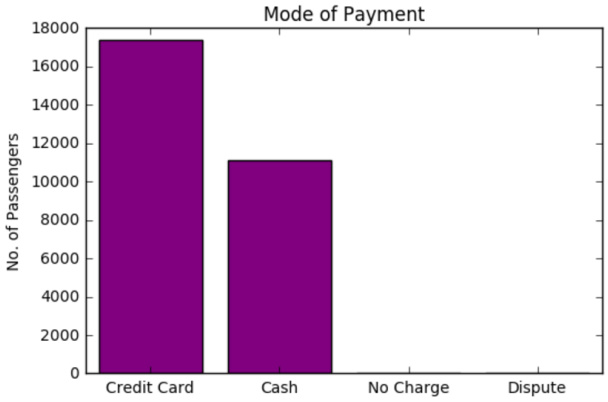




Interpretation:

- Both the Arrival and Departure graph show sudden surge in number of people during afternoon hours
- Later in the evening, their demand drastically decreases. This could is because they receive lot of hails from street passengers and don't prefer airport pre-bookups any more. This fact is well justified from the graph previous built up.

Interesting Question: What is the most preferred mode of payment by customer

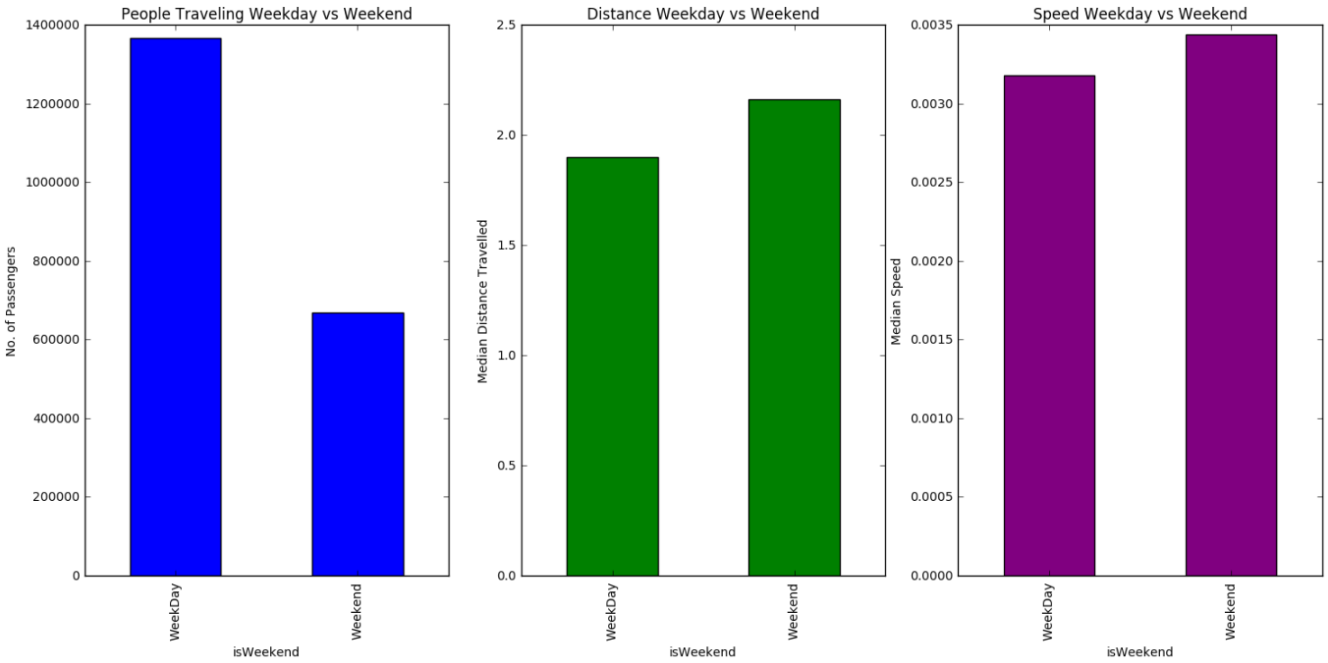


***Interpretation**:*

- Significantly good number of people use credit card to make the payment and this is very true as plastic money is preferred over paper money.
- 54% more customers prefer Credit card over cash payments

Interesting Question: Let's analyze the effect of Weekend and Weekday on

1) Number of passengers planning to commute via Green Taxi
2) Distance travelled during Weekdays vs Weekends
3) Speed of the Green Taxi on Weekends vs Weekdays



***Interpretation**:*

1) Lot more number of people travel travel on Weekdays than on weekend. This is because lot of people hail Green Taxi's to reach their work place and relax at their home on weekends.
2) On Weekends, median distances are more as they probably travelled inter-borough.
3) As the streets are less crowded, Taxi drivers are able to drive at relatively higher speed on weekends.

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). We will validate a sample.

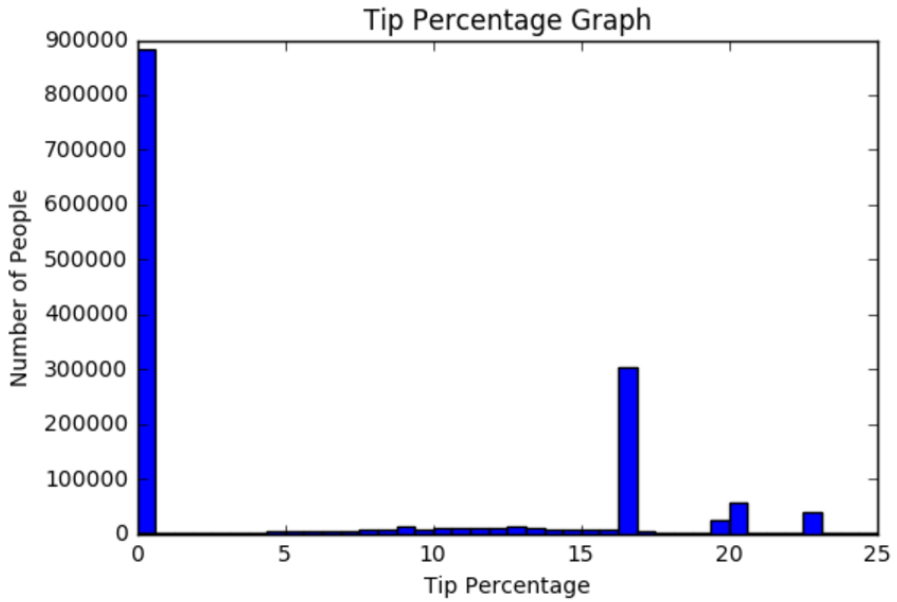
Predictive Modeling:
Model Building process involves following critical steps:

- Data Cleaning
- Exploratory Data Analysis
- Model Selection
- Model Fitting
- Model Evaluation

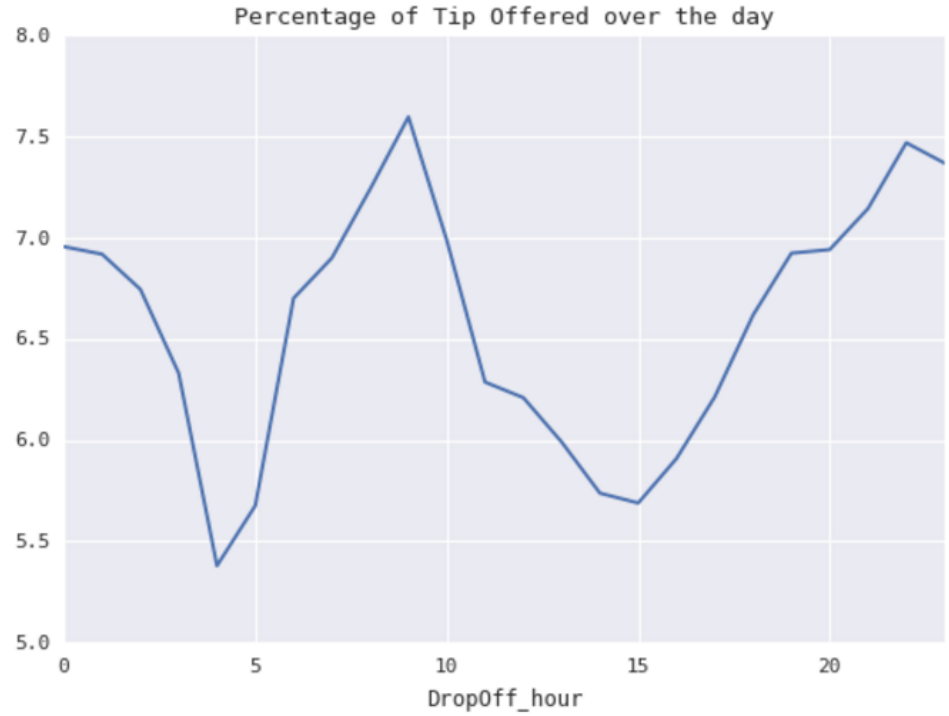
We are asked to build a model to predict the Tip as a percentage of the total fare. Therefore,

- Response or dependent variable is **Tip_Percent**
- Predictors would be all the other features.

As Tip_Percent variable is created from Tip_amount and Total_amount, there is very high correlation between two. We should remove either of these to make the model practical.
In our case, we'll be removing Tip_amount from the dataframe.



***Interpretation**:*
In most of the trips customers didn't pay any tip. However, 2nd most top tip percent is 16.6%.

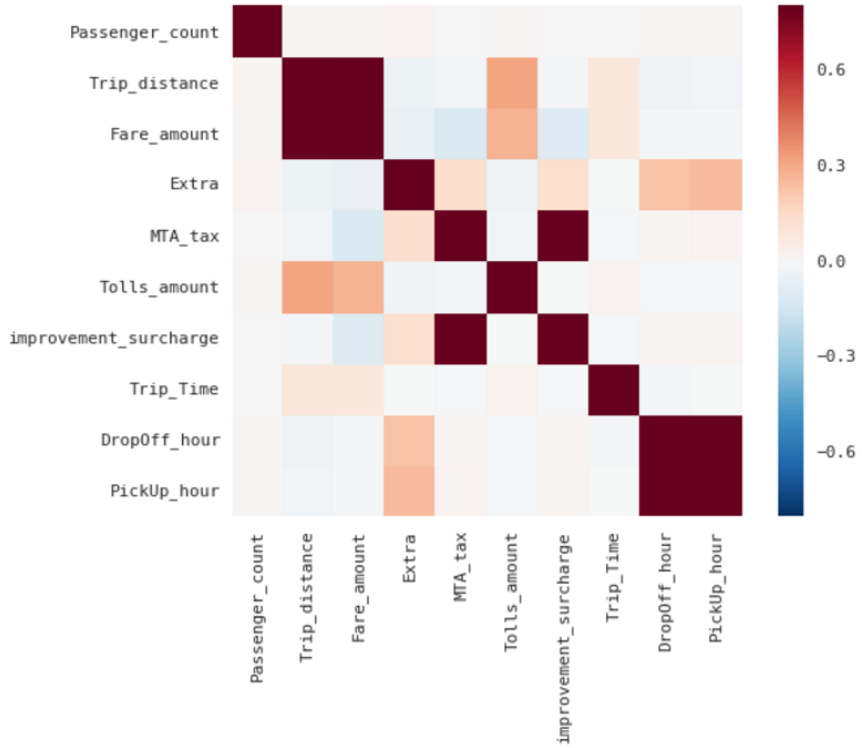


***Interpretation**:*
Early morning and late evening is when passengers prefer to give tip and that too in decent amount.
Probably because Taxis make passengers reach their destination on time and they are thankful for this. So, very genorously they pay good tips.

Data Cleaning and Model Preparation:

Considering our objective to predict the Tip Percentage amount, we can ignore following variables as they won't help in prediction and unnecessary make the model complex. The decision has been made having understand the business requirement.

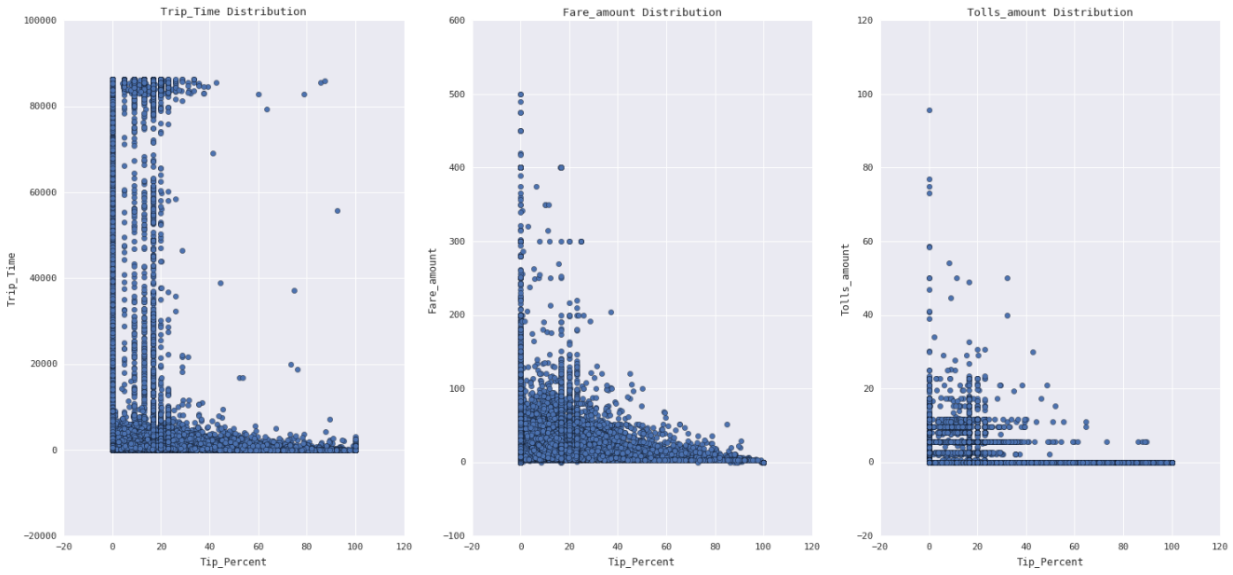
- lpep_pickup_datetime
- Lpep_dropoff_datetime
- pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- LaGuadia_Pickup_Distance
- LaGuadia_DropOff_Distance
- LongIsland_Pickup_Distance
- LongIsland_DropOff_Distance
- JFK_Pickup_Distance
- JFK_DropOff_Distance



Interpretation:

- Trip_distance and Fare_amount are highly correlated. This is because, if distance increases, fare_amount will increase. In other words, they are proportionaly related. Also, correlation coefficient is above our threshold limit of 0.85. Therefore, we'll remove one variable Trip_distance and keep Fare_amount.
- There seems to be strong relationship between MTA_tax and Improvement_surcharge as well. Strong positive relation signifies that MTA_tax increases with increase in improvement_surcharge. Also, the correlation coefficient is above our threshold limit of 0.85. Therefore, we'll remove improvement_surcharge and keep MTA_tax.

Following scatters plots are being build to understand if there is any linear relation between predictors and response variable as this is one of the important asssumption for linear models.



Approach 1: Predict Tip Amount in Numerics

BASELINE MODEL:

Using Baseline model of mean value, we get mean_squared_error of 78.37

MODEL: LINEAR REGRESSION

- Linear Model has an **assumption** that there should be no outlier. In the beginning itself, we have taken care of it and removed all the possible outliers out of it.
- Secondly, it assumes that its predictors have normal distribution. We have performed KS-test on each variable and checked the results. Q-Q Plot could also be used here to perform the normality check.
- Thirdly, linear models assumes that there should be no collinearity between the predictor variables. To do so, we have removed all such variables which have more than 0.9 correlation coefficient.
- Fourthly, linear model assumes that there is no auto-correlation between variables. In our cases, each ride and it's fare is independent, and has no correlation with past rides.

Hence, all our assumptions are fulfilled and we are good to go with Linear Regression model.

Lot of variables are skewed and thus, we took log10 transformation.

Cross Validation Results:

- Without log tranformation, the model gives mean MSE of 35
- After making log transformation, we observe mean mean MSE of 27.45

Analyzing Coefficient of Determination (R2) which gives explains the amount of variance in the response variable. Here, we are considering Adjusted R2 as it considers the effect of extra and irrelevant variables in the model.

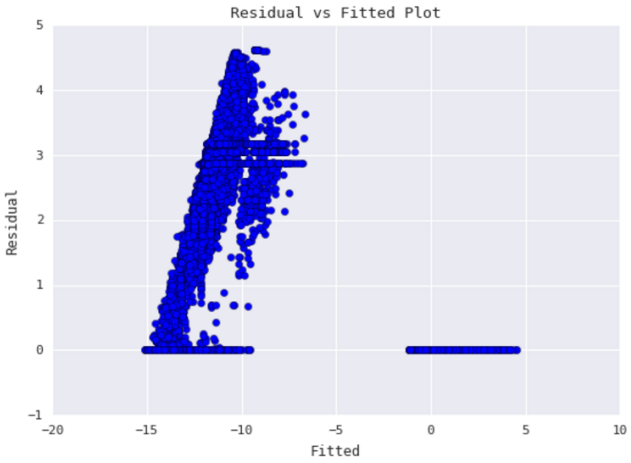
- Without log transformation, R2-Squared value was roughly 56%
- After making log transformation, R2-Squared value has increased to 64%

Conclusion: Log10 Transformation has significantly improved the results.

Using BoxCox transformation, we get to know about transformation that should be used on response variable.

On performing natural log transformation, our results improve further and we get R2-Squared value of 75%.

Final Residual vs Fitted Graph is as follows:



Final Conclusion:

We observed that after performing **Natural Log transformation** on Response variable. Our R2-Squared value has increased from 63% to 75%, which is a significant improvement.

Also, the residual vs fitted plot is still not normal but better than previous. Not the ideal one anyway.

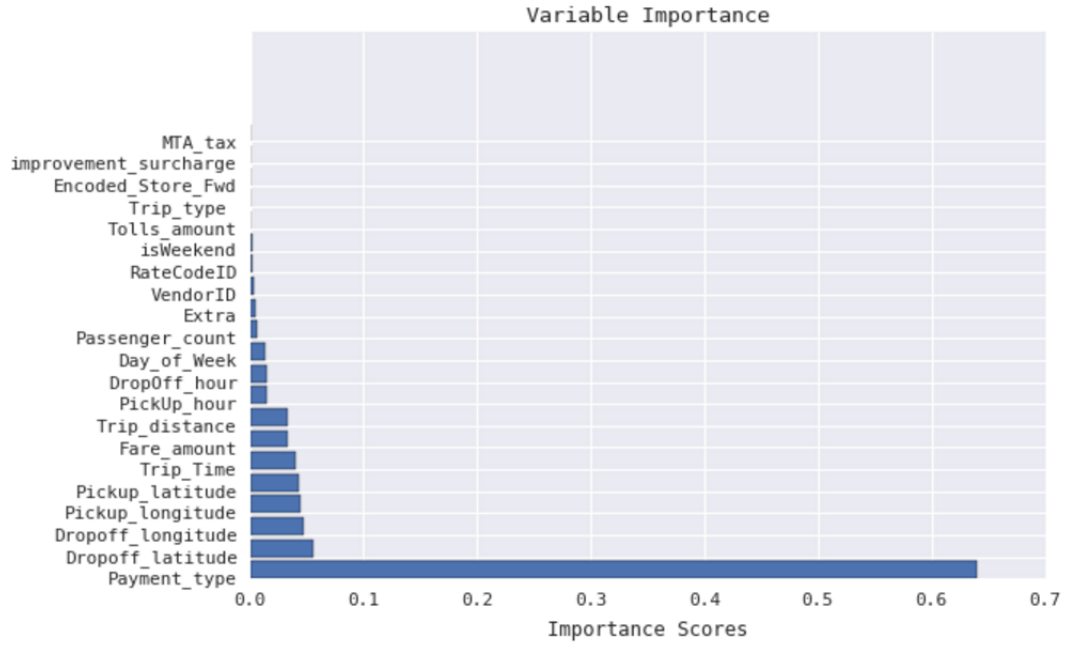
There is scope of improvement with transformation like log1p, exponentials etc.

There is non-linearity in the predictor variables as well, which suggests us to move ahead with non-linear model such as Random Forest Regressor.

MODEL: RANDOM FOREST REGRESSOR

Random Forest makes better predictions than naive Linear Model. It is because there is some non-linearity in the dataset and this is handled by Random Forest better than the linear models.
Also, ensembling multiple decision tree is giving much better results in the end.

- Payment_type is the most important variable among all. This can be justified with the fact that people prefer paying Tip using Credit Card rather than Cash.
- Latitude and Longitude has given significant information, in predicting tip amount.



Approach 2: Predict Tip Amount Bin

Four Tip bins are created

Bin1: Tip between 0 and 10

Bin2: Tip between 10 and 20

Bin3: Tip between 20 and 30

Bin4: Tip between 30 and above

And we try to predict the bin number using Random Forest.

For the evaluation purposes, we are using FScore, Precision, Recall and Support metrics.

	FScore	Precision	Recall	Support
Bin1	0.994407	0.991755	0.997073	187232
Bin2	0.971261	0.993181	0.950287	89514
Bin3	0.914833	0.861452	0.975265	18476
Bin4	0.792631	0.661151	0.989387	848

Here again, we see Payment_type as the most significant variable. FScores for each of the bin are respectable. Considering these scores, we can conclude that it is possible to predict the bins more correctedly rather than predicting the exact tip amount.

Question 5: Anomaly Detection

- What anomalies can you find in the data? Did taxi traffic or behavior deviate from the norm on a particular day/time or in a particular location?
- Using time-series analysis, clustering, or some other method, please develop a process/methodology to identify out of the norm behavior and attempt to explain why those anomalies occurred.

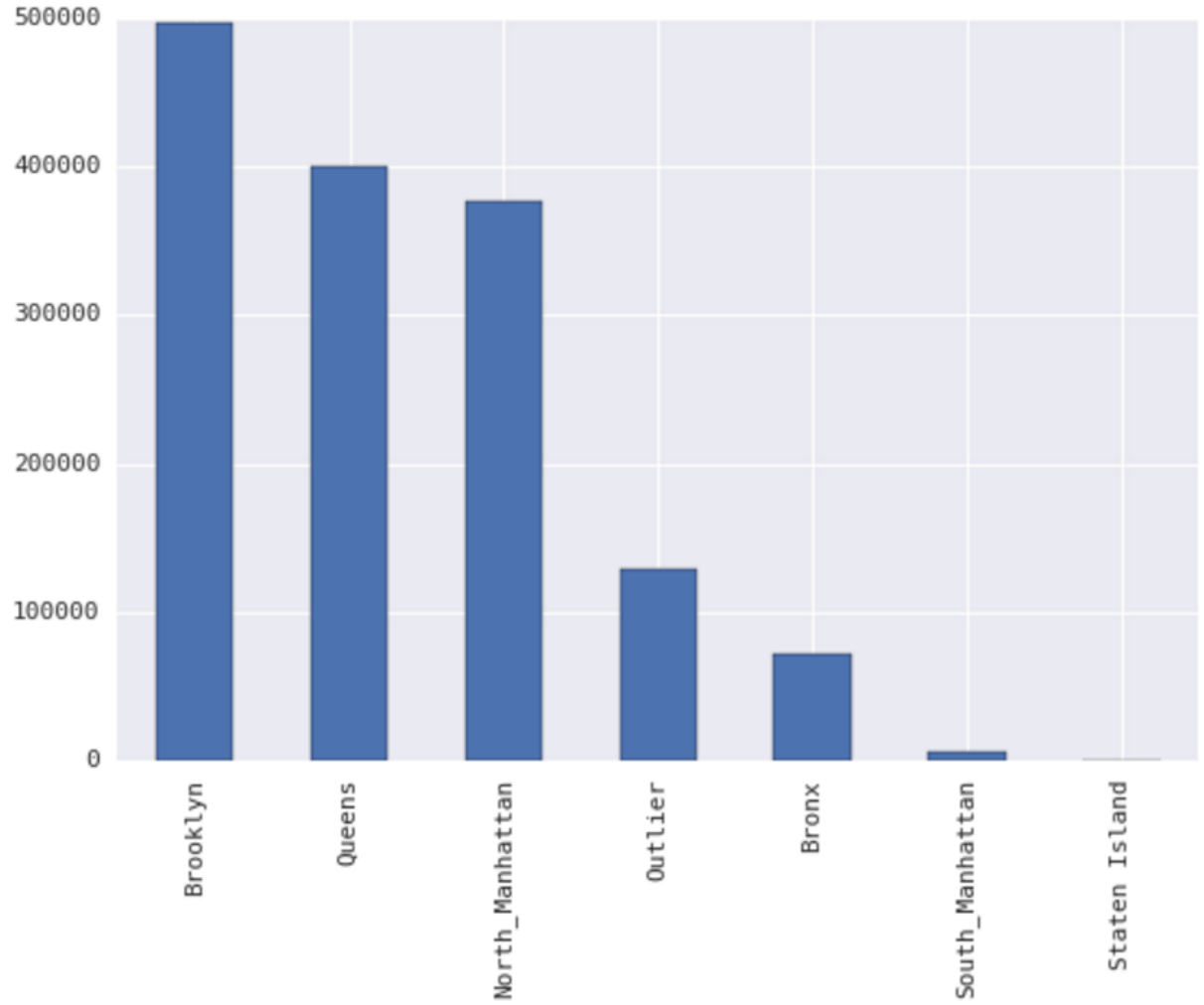
Approach 1:

As per the website http://www.nyc.gov/html/tlc/html/passenger/shl_passenger.shtml

- **Fact 1:** Boro Taxi drivers can pick up passengers from the street in northern Manhattan (north of West 110th street and East 96th street), the Bronx, Queens (excluding the airports), Brooklyn and Staten Island and they may drop you off anywhere.
- **Fact 2:** Boro Taxi drivers can be dispatched to pick you up in northern Manhattan, the Bronx, Queens, Brooklyn and Staten Island and at the airports, but may not pick up any trips – pre-arranged or street hail – in the Manhattan exclusionary zone.

According to this information, I have developed a geocoordinates polygon for each of the area in consideration. This includes:

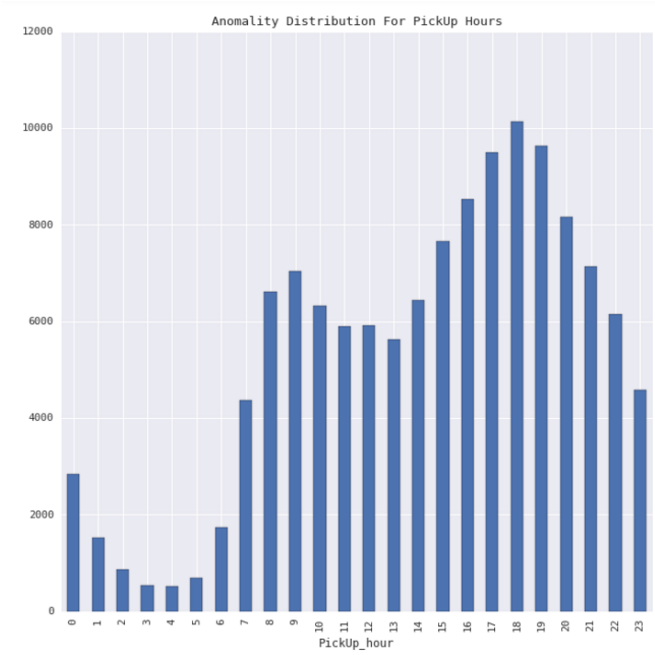
- Brooklyn
- Bronx
- North_Manhattan
- South_Manhattan
- Staten Island
- Queens

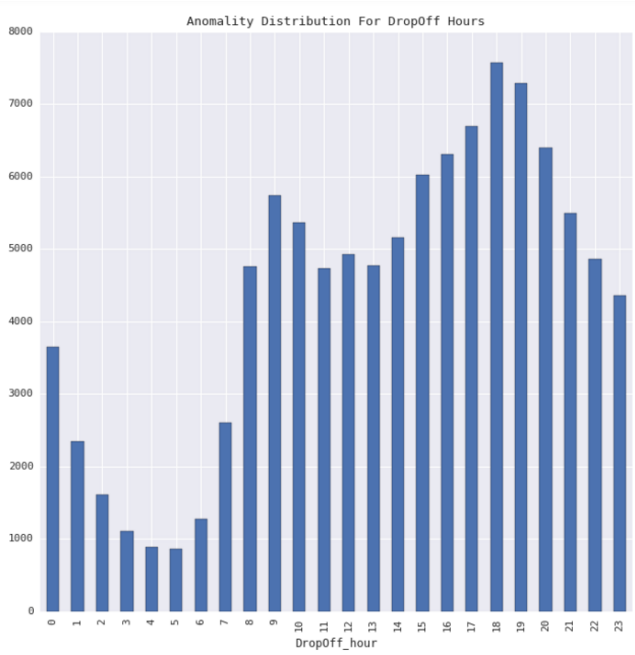


Interpretation:

South_Manhattan areas would be the restricted ones and NYC police department can impose fine on those taxi drivers. Other than this, there are significant number of Outlier, which possible means went out of NYC city or out of these boroughs.

Number of taxis went out of authorized areas: 128433
Number of Taxis that did pick up from Restricted Manhattan area is 5486

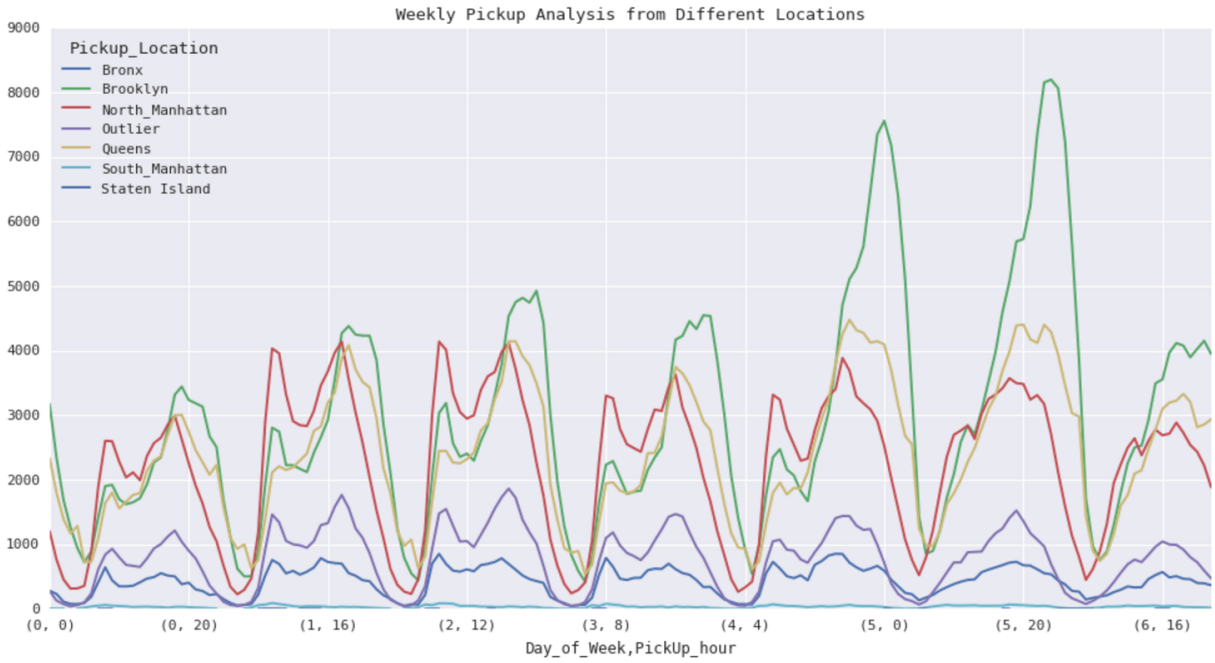




Interpretation:

If you observe it carefully the overall distribution is very similar. This can be justified with the situation where passenger hailed the taxi in one of boroughs and took it over to a place where Green Taxi's generally don't operate like New Jersey. It is understood, that driver would again look for a passenger who wants to go one of those boroughs.

Thus, Outlier at a DropOff Hour is followed by Outlier at Pickup hour.



Interpretation:

- From the above graph we can conclude that on Friday's and Saturday's there is a sudden surge in demand for Green Taxis in Brooklyn area.
- Demand from Queens and North Manhattan types area follows relatively same pattern through out week.

Approach 2: Anomaly detection using DBSCAN Clustering

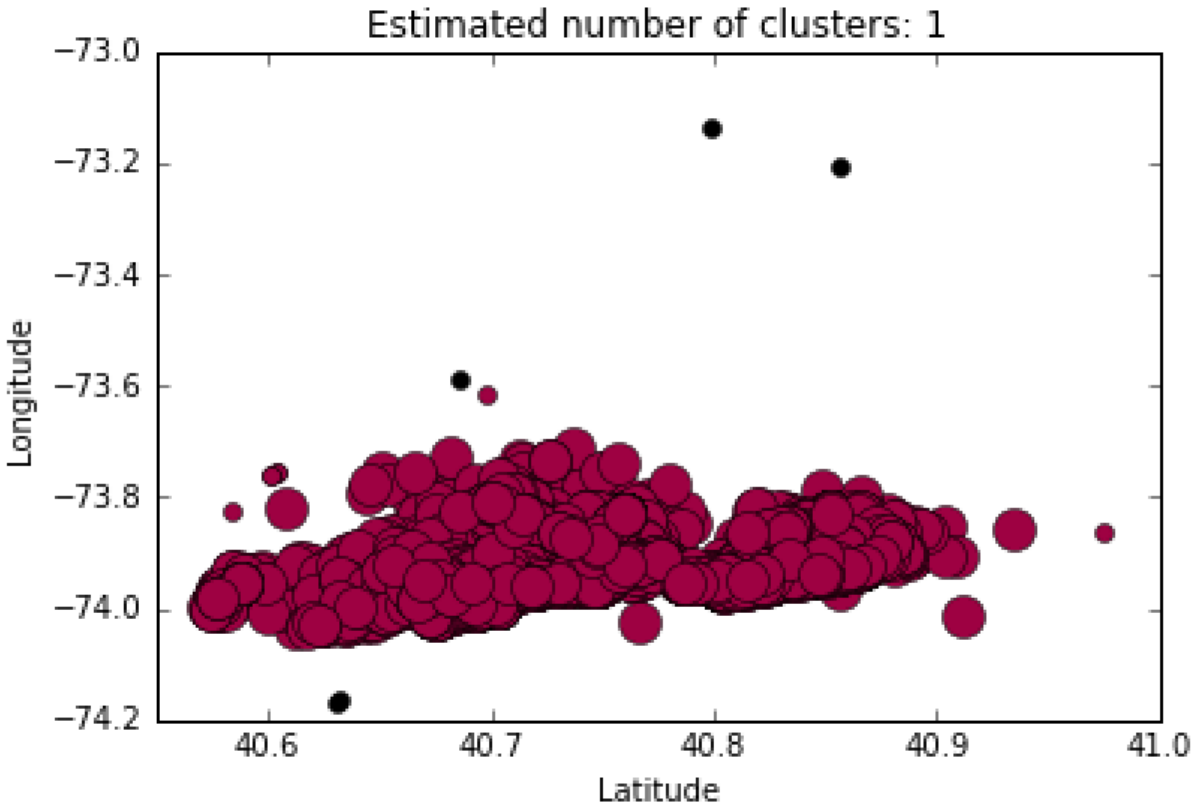
DBSCAN is a density based clustering algorithm which can help us identify anomaly. Well known K-means clustering algorithm cannot be used here as we are primarily dealing with Geo-coordinates to form clusters. K-means uses only Euclidean distance which is not correct when dealing with geo-coordinates.

DBSCAN clustering Algorithm allows us to use Haversine Distance to calculate miles distance between two geo-points and hence, this is the right choice.

Epsilon(eps) metric specifies how close points should be to each other to be considered a part of a cluster.

DBSCAN is a CPU intensive algorithm and it is not possible to run it on a 8GB RAM with i5 Intel processor.

However, I was able to form cluster for random 30000 data points and could observe 5 outliers.



AREAS OF IMPROVEMENT / FURTHER WORK:

- To detect Anomaly, we can try - one-class-SVM model and even unsupervised Random Forest which uses promixity values to find anamolous records
- To predict Tip Fare, I could have created a feature that would take into account the population density. For this, we can create NxN matrix and use geo-coordinates to give us rough estimate of population density.
- To predict Tip Fare, external data saying something about the age of the passenger could also help in predicting tip price. This is because, old people are more generous as compared to young professionals.
- At Last, lot of further exploration and insights could be provided if we could join some external dataset like weather, driver information etc.

SOFTWARE AND DEPENDENCIES

Software:

- Python 2.7
- Anaconda 4.2 Distribution
- Jupyter Notebook

Dependencies:

- Python Scikit-Learn – For Regression and Clustering Models
- Python Matplotlib – For plotting
- Python Pandas – For handing Dataframe