NYC Green Taxi Data Analysis

By Manan Vasani

By
Manan Vasani
Master of Engineering Management,
Duke University.
Contact info: (919)-381-7864
manan.vasani@gmail.com
www.linkedin.com/mananvasani/in

Contents

Introduction

It was a great experience working on the Capital One Data Science Coding challenge. Thanks to Taxi and Limousine Commission of New for making the data open source. It was an extremely rich dataset with data type ranging from categorical, numeric to geospatial and datetime objects. I start by downloading the dataset of Green taxi for the month of September 2015 and importing it into jupyter notebook. Following this I do some basic visualizations and exploratory data analysis to find correlations and dependence of each variable with others. Later I develop a model to predict tip amount that a cabbie could expect for each of his/her trip.

I have used python for all the data analysis and also relied on Tableau to create rich graphs and other visualization materials.

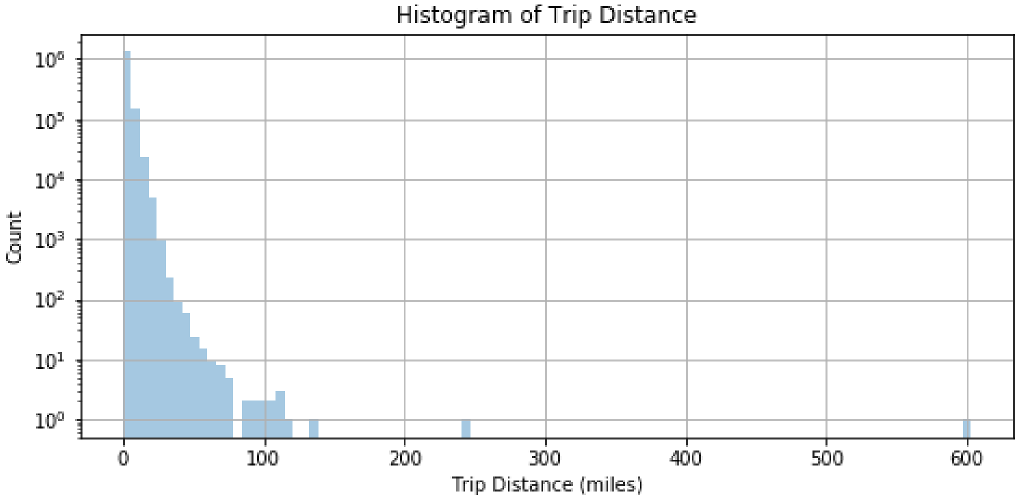
Question 1. Importing Data

I begin by importing data analysis libraries: pandas, numpy and visualization libraries: matplotlib and seaborn. I downloaded the dataset for green taxi and imported it to a dataframe called df and checked the first 5 entries of it to see whether the data was loaded properly or not. Next I checked the size of data frame to learn the number of rows and columns by using the command df.info() or df.shape.

Dimension of dataframe: 1494926 rows x 21 columns

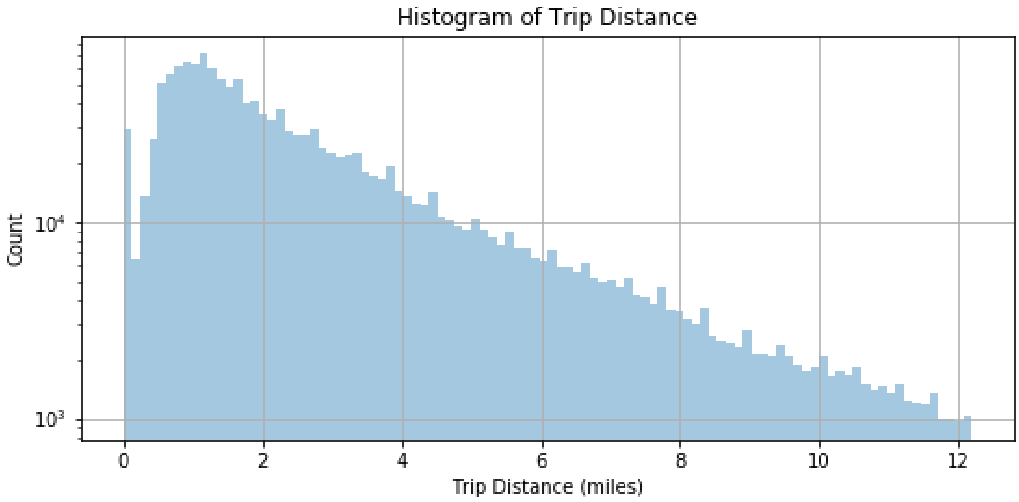
Question 2. Histogram of 'Trip Distance'

First, I plot a histogram of trip distance using all data points available. The plot looks as follows.



It is seen that most trips are less than 50 miles long. However there seems to numerous outliers as evident by some trips being 600 miles long. This seems to be due to erroneous reading captured by the meter because logically a trip of more than 500 miles would eventually take one to Quebec, Canada if traveling north, Detroit if going east and Charlotte, NC if one plans to drive south.. We need to get rid of these outliers to get a more accurate distribution of trip distance.

For this I create a new variable that represents z-score of all the trip distances. I use these z-scores to eliminate entries that lie beyond 3 standard deviations of the mean.

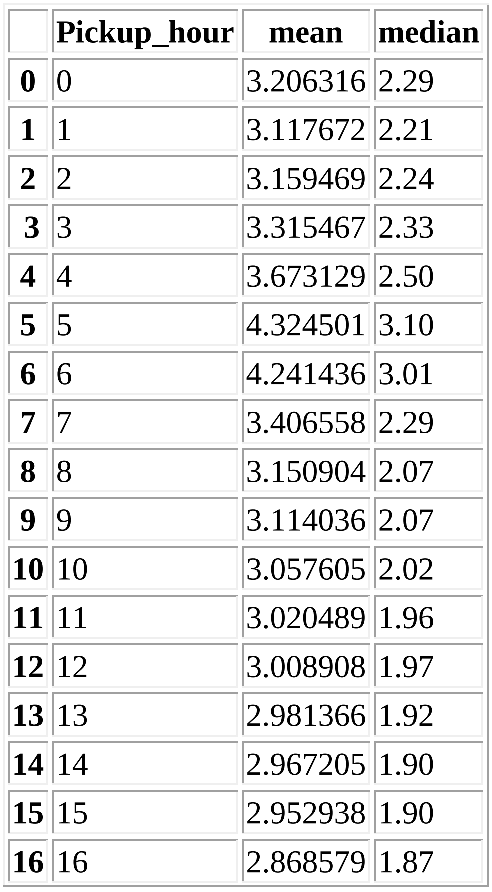


The plot now looks much cleaner, with majority of trips being 4 miles or less. The data now makes sense as the maximum length of a trip is a bit more than 12 miles with a mean distance of 2.7 miles and median of 1.92 miles.

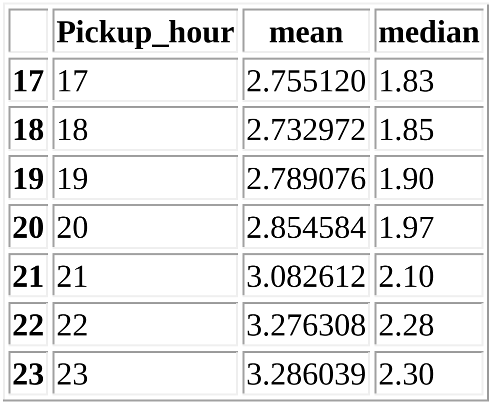
Question 3. Further analysis of 'Trip Distance'

To learn more how does the distance of a trip vary, I first create date time objects from the pickup date provided in the dataset and extract days and hour of day from it. To determine how day of the hour affects a trips distance I create a pivot table containing the mean and median values grouped according to time of the day.

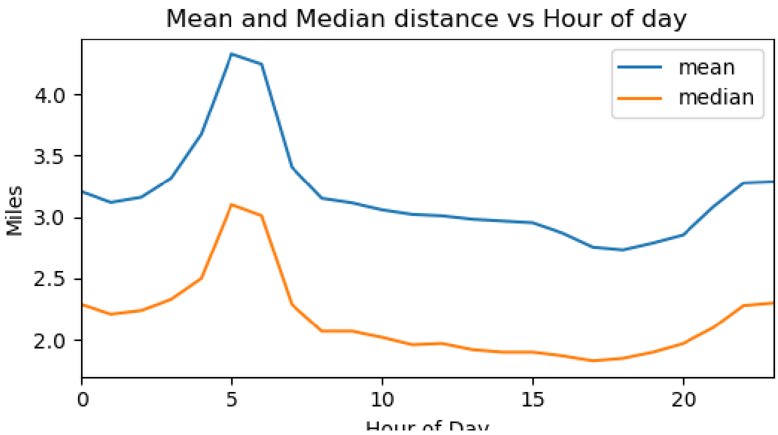
Table of Mean and Median Distance according to hour of day



	Pickup_hour	mean	median
0	0	3.206316	2.29
1	1	3.117672	2.21
2	2	3.159469	2.24
3	3	3.315467	2.33
4	4	3.673129	2.50
5	5	4.324501	3.10
6	6	4.241436	3.01
7	7	3.406558	2.29
8	8	3.150904	2.07
9	9	3.114036	2.07
10	10	3.057605	2.02
11	11	3.020489	1.96
12	12	3.008908	1.97
13	13	2.981366	1.92
14	14	2.967205	1.90
15	15	2.952938	1.90
16	16	2.868579	1.87



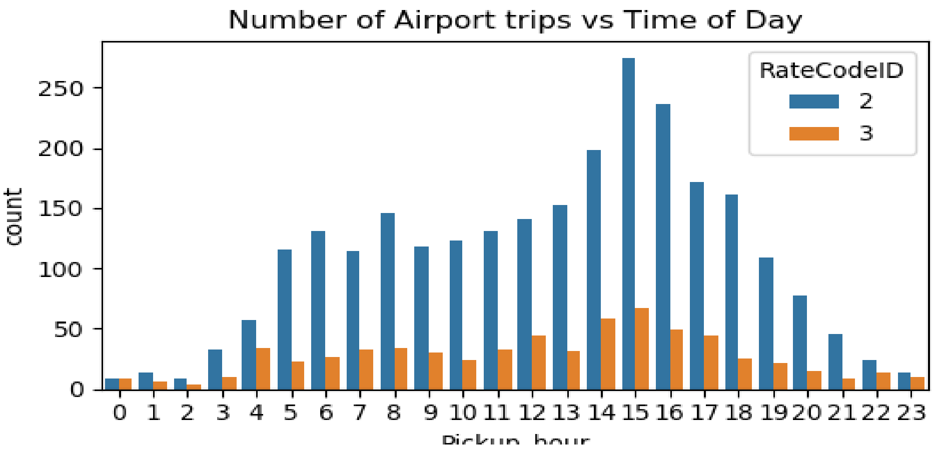
	Pickup_hour	mean	median
17	17	2.755120	1.83
18	18	2.732972	1.85
19	19	2.789076	1.90
20	20	2.854584	1.97
21	21	3.082612	2.10
22	22	3.276308	2.28
23	23	3.286039	2.30



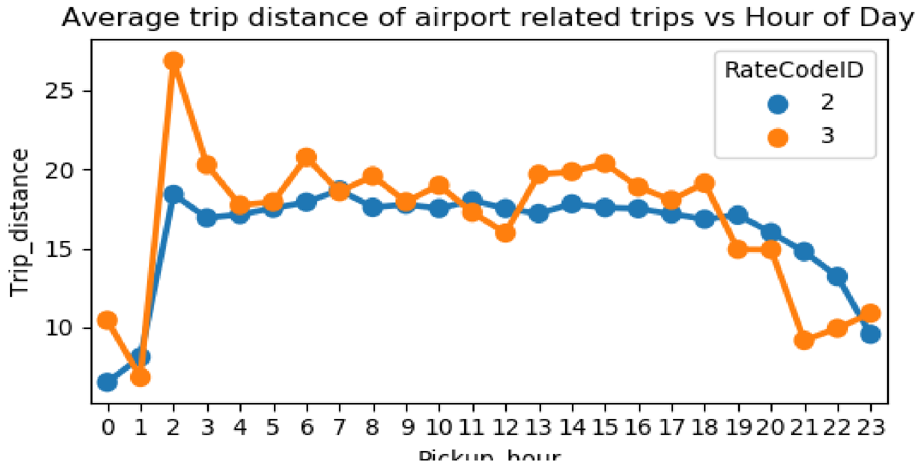
Reading the data dictionary, it becomes evident that datapoints with variable ‘RatecodeID’ set as 2 or 3, denotes trip either originating or terminating at JFK or Newark.

There were **5552** such entries. Hence It makes sense to create a new dataframe called df_airport_trips, with information just pertaining to airport trip. I created new variables such as tip percentage, whether it is weekday or weekend, etc to get a better idea of what variables and factors affect the trip distance.

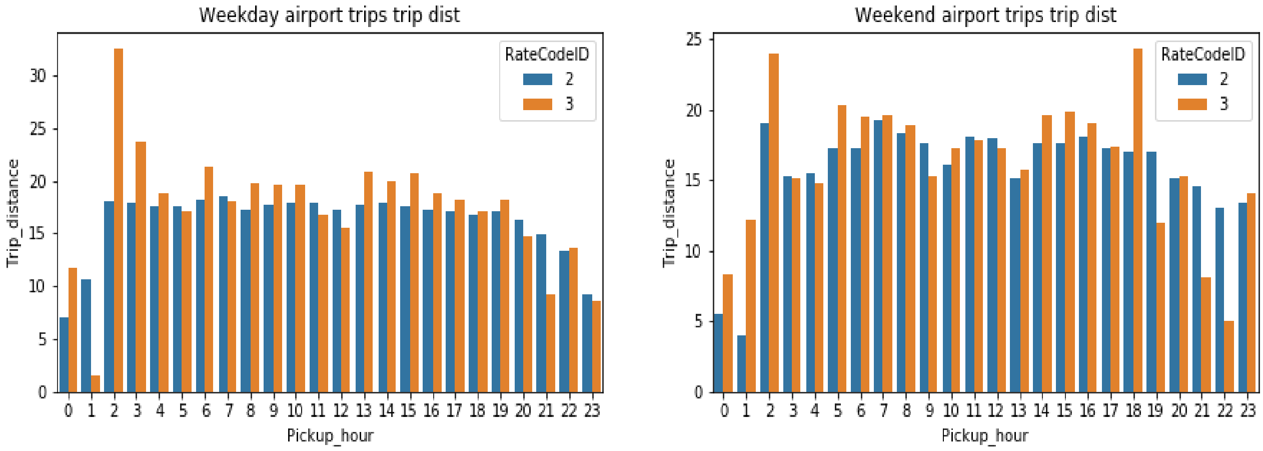
Hypothesis 1: People tend to travel more frequently during evenings. This seems logical because if someone is travelling for a business meet he/she would prefer to take flight on previous day evening and return on next day evening. Hence, we see a spike in the number of trips taken to/from aiport between 3 pm and 6 pm.



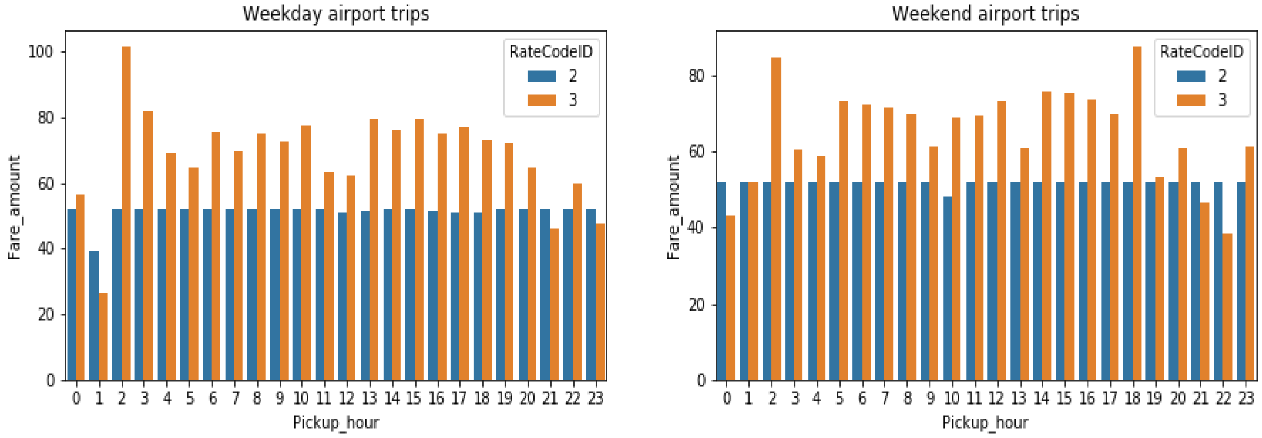
Hypothesis 2: Since Newark airport is far it is logical to think that the trips related to LaGuardia would be longer that those to/from JFK. It is clear in the plot below.



Hypothesis 3: There seems to be negligible difference in the average trip distance taken over a week day and a trip taken over weekend.



Hypothesis 4: Plotting the average fare for trips to/from airports reveals that fare related to JFK seems to be constant whereas the fare to Newark airport seems to be dynamic in nature. Hence, we can hypothesise that a trip to JFK is a flat fare affair with fare fixed at nearly $51.75. While the average fare for LaGuardia is $70.5 which makes it possible that there might be a surcharge for inter-city travel.



Question 4. Analysis of Tip Paid

Before carrying this out, I have added various other parameters that would help in an to better understand the tipping behaviour of customers. I made following derived variables:

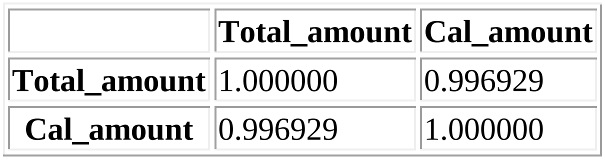
1. Pickup hour – Depicts the time of day when the trip was taken
2. Pickup Day – Day when the trip was taken
3. Pickup day name – Which day of the week was it (Monday, Tuesday, etc)
4. Pickup day type – Whether it was weekday or weekend (Mon-Fri: Weekdays, Sat-Sun: Weekends)
5. Airport trips – Denotes 1 if an airport trip else 0
6. Trip Duration – Denotes the length of a trip
7. Speed – Speed of the trip

Before building a model to predict tip data had to be cleaned off anomalous entries. I sequentially analysed each column to get rid of logically and literally impossible data.

1. The latitude bound of NYC is (40.63, 40.85) and the longitude bound is (-74.03, -73.75). Thus, it makes sense to check whether our data is restricted within this limit or not. Also it is seen that numerous latitude and longitude values are 0. The only logical reasoning behind this would be that the device was unable to make an entry into the system because latitude = 0 means a pickup from some place on equator which is impossible.
2. A number passenger counts are 0 which means that the driver charges himself/herself for the trip. Hence replacing 0 count with median value.
3. Many entries in the Fare amount column are negative which does not make sense. So first I converted these to its absolute values. Further reading the rules and regulations on TLC website, it becomes clear that the minimum fare amount for a taxi ride is 2.5 .Any value below this is replaced by the median value.
4. Trips less than 2 minute of duration seems hard to believe. Therefore, I dropped those.
5. Researching a bit, I found that the speed limit in NYC is 50 mph. Hence, I dropped all the data beyond 70 mph so as to prevent model overfitting.
6. Similarly, various other data points in columns Tip amount, Toll amount, Trip Distance, etc were found to be negative. Hence converted those to its absolute values and replaced others with the median value.

Having many outliers affect the mean of a variable in great proportions whereas median remains unaffected. And since this dataset has huge number of outliers I replace the unwanted values with its median values. After cleaning the dataset of unwanted values, I create two more features.

1. Cal amount: The total amount paid by customer. Created by adding fields of fare amount, toll amount, extra, MTA tax, improvement surcharge and Tip amount. Here it is seen that the total mount provided in dataset is highly correlated with the calculated total amount. Thus, I drop erroneous data provided by Total amount and replace it with Calculated amount.



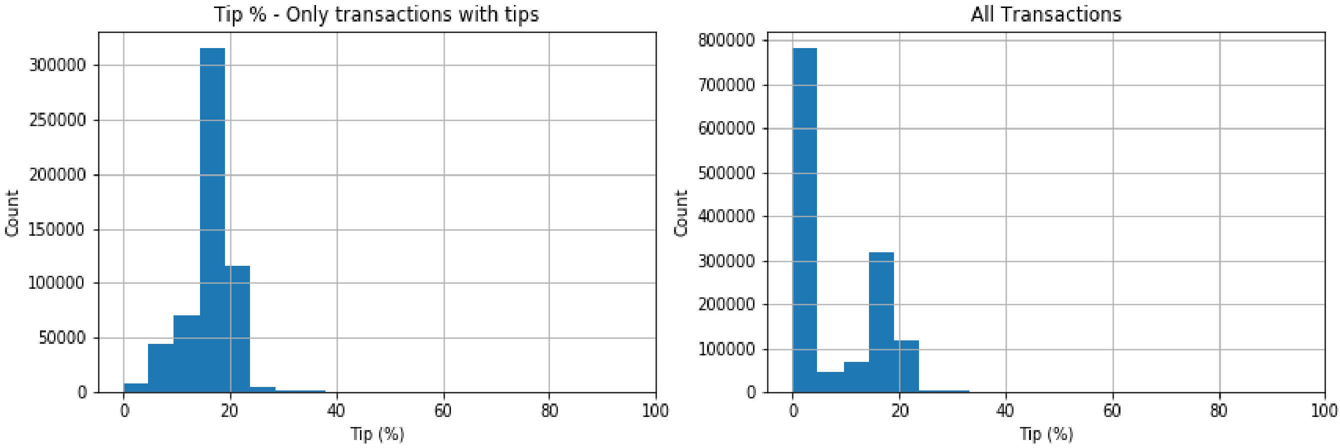
	Total_amount	Cal_amount
Total_amount	1.000000	0.996929
Cal_amount	0.996929	1.000000

Correlation between Calculated amount and total amount

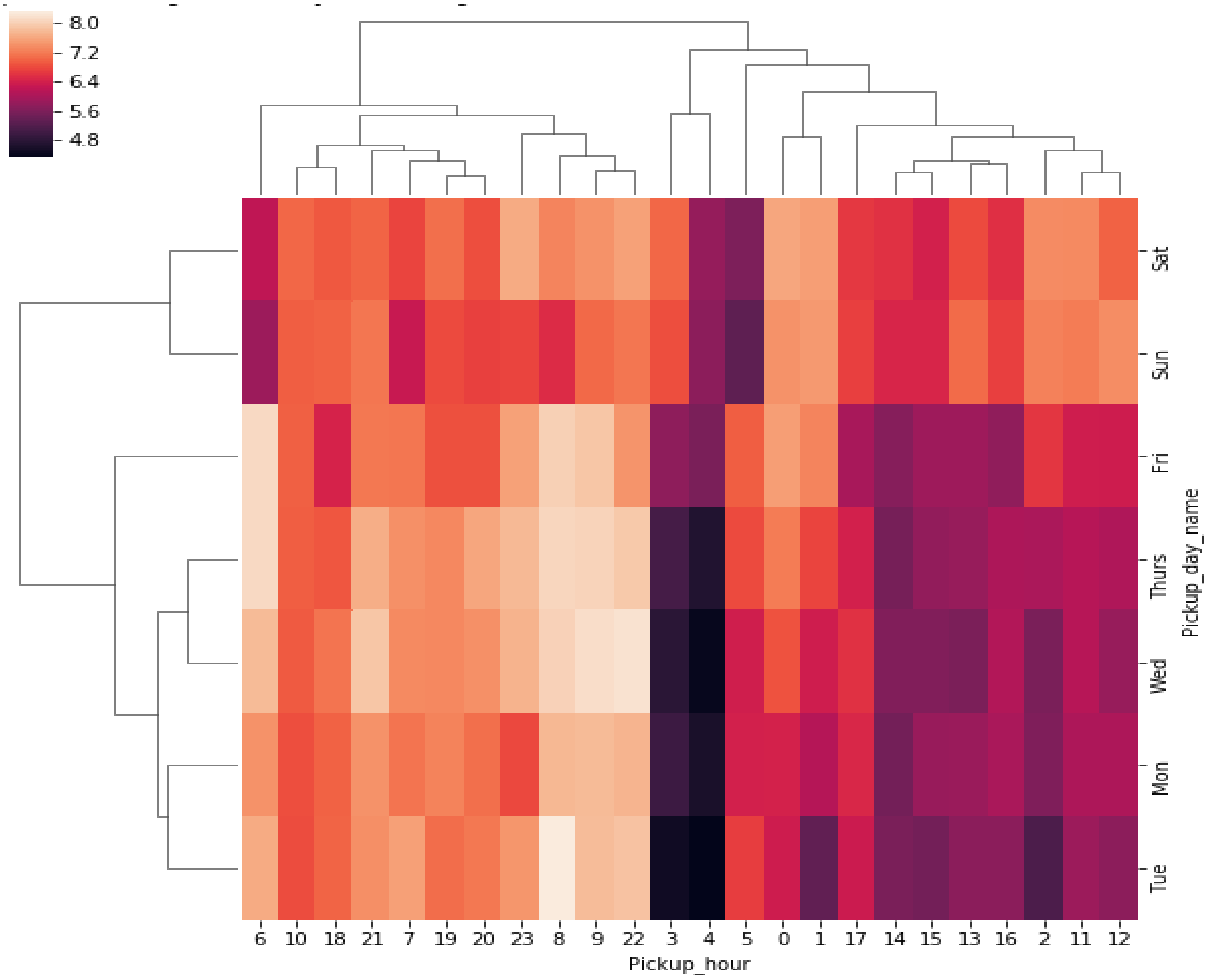
2. Tip percent: Tip paid as a percent of the calculated amount

Furthermore, I also drop various other columns that are irrelevant and store it into a new dataframe called df_raw to avoid losing data for future reference. The dimension of new data frame after feature engineering is 1337111 x 26.

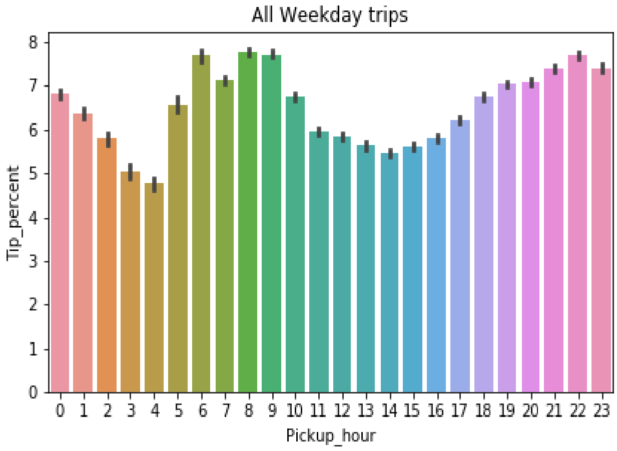
Before building a model, lets just analyse the normal tipping behaviour of people and learn what factors affect it the most. We see that 57.9% of all transactions to do not involve a tip. Only 42.1% of the trips involve a tip being paid. Of the transactions involving a tip, we see that the average tip paid is 16.31%.

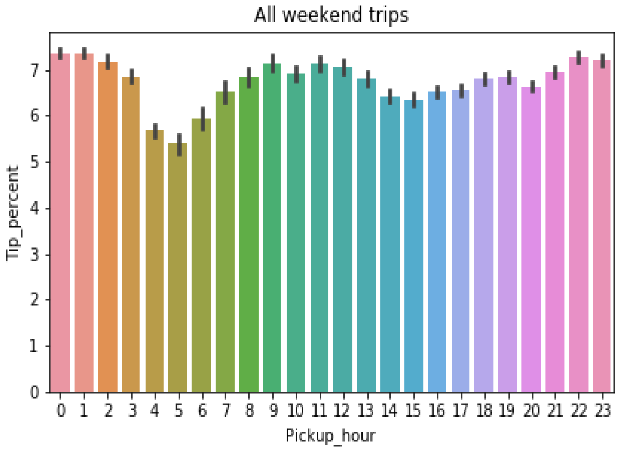


Having a look at the clustermap that represents average tip paid depending on what day of week it is and what hour of day it is, gives a rough idea about the tipping pattern.

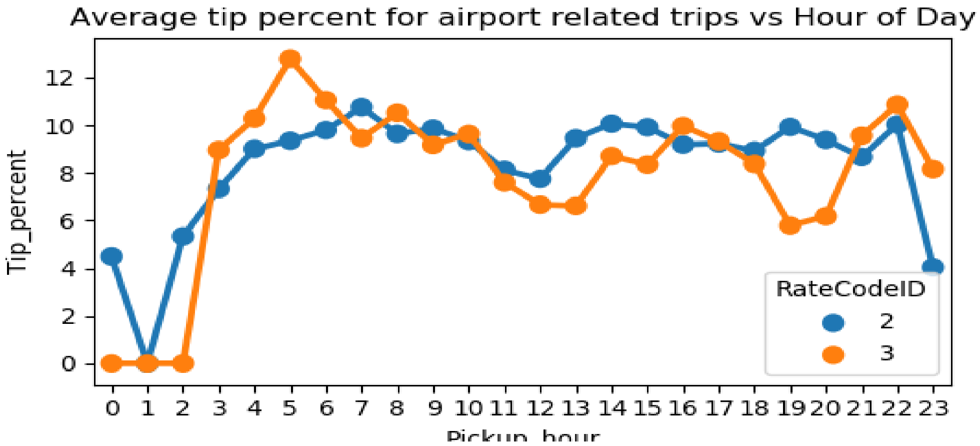


Also tipping behaviour depends whether it is a weekday or weekend. It is seen that people are more generous in the morning when they are off to work and tip cabbies heavily for dropping them on time. However, there is a sharp decline in tip paid after lunch hours and again rises as people leave work for their home. On the contrary tipping behaviour seems to be more uniform during weekends except for hours in night.

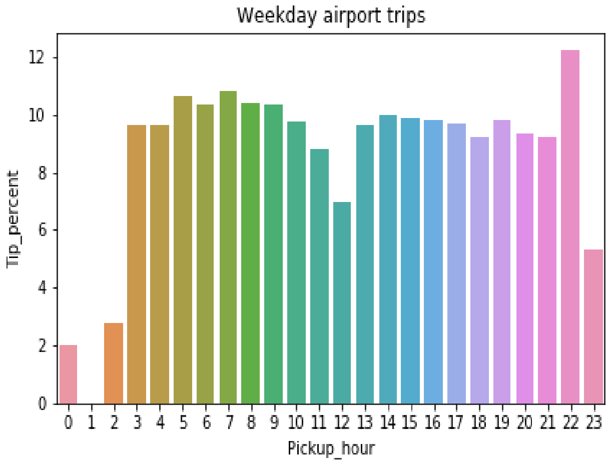


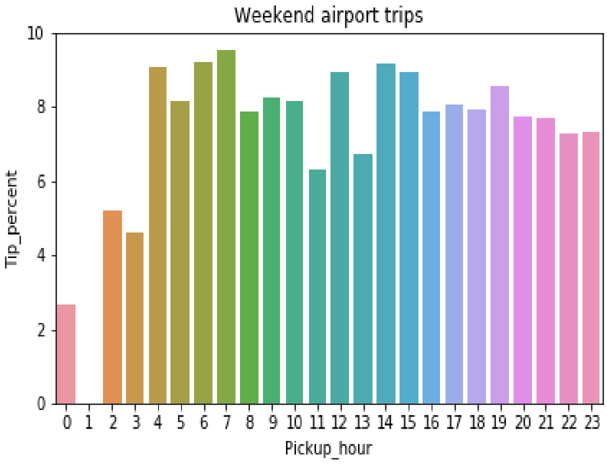


It is seen that the amount of tip paid also depends if trip is airport bound or coming from it. Something odd about this plot is that no one pays tip when the cab is hailed at 1 am in the night.



Also amount of tip varies whether an airport cab is booked on weekdays or weekend as seen in the graph below.





We can conclude that tipping pattern is highly skewed and depends on a lot of factors. Therefore, I divide my model development into two parts.

1. To identify whether tip would be provided or not.
2. If classified as a transaction with tip, predict how much tip would be paid

Part 1:

I plotted numerous columns against tip percentage in Tableau to understand how each affects the tip percent and tipping behaviour. Learning the important parameters that affect our target variable, I applied logistic regression to predict the that whether tip will be paid or not. I achieved an area under ROC of 0.9713 which is pretty good.

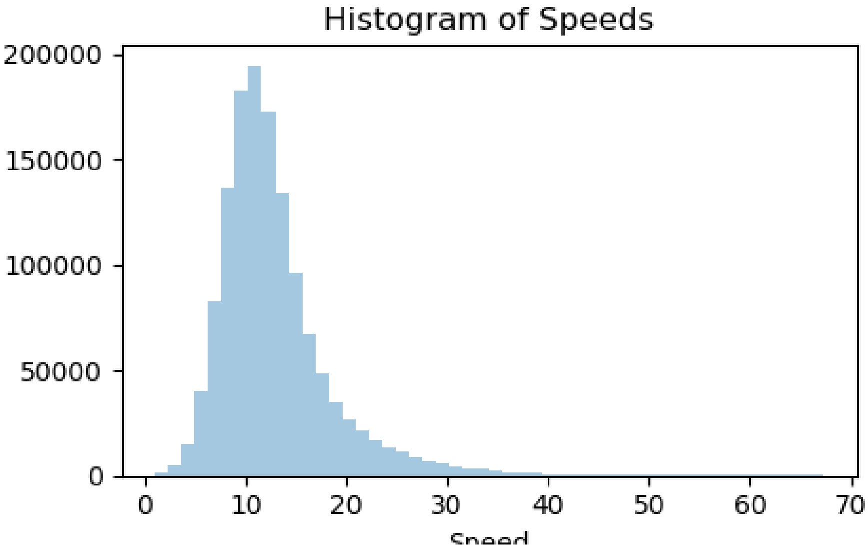
Part 2:

After figuring out whether a tip will be paid or not, I run linear regression to predict the amount of tip that the driver could expect. However, currently our data resides in a dataframe and we have datatypes that are numeric as well categorical. Numeric data is fine but to deal with categorical data we would need to use one-hot encoding process from sklearn library which works only on numpy arrays. Hence, we convert relevant columns to numpy arrays for further analysis and encode them to one-hot vectors.

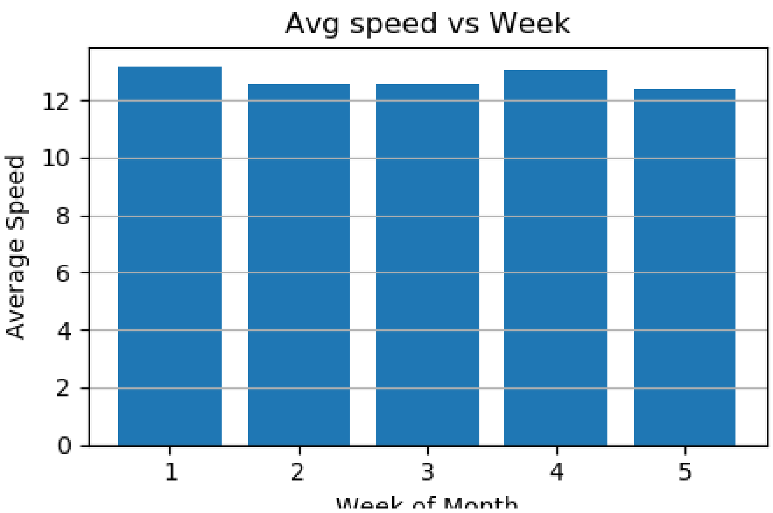
The result obtained after running linear regression algorithm on modified data is as follows: Mean squared error was 13.64$ and the variance score was 0.8165. Considering that a variance score of 1 is the best a model can achieve, this model performs really good.

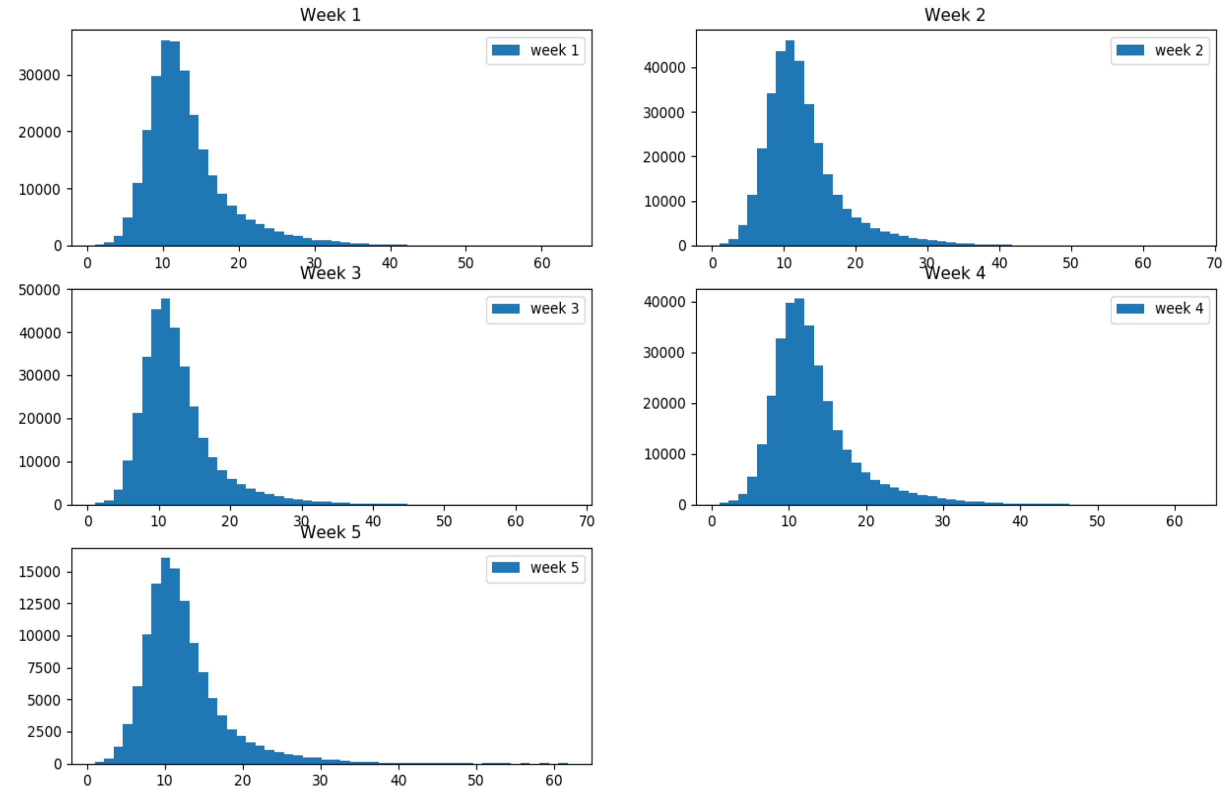
Question 5. Average Speed Distribution

I already created feature of speed and used it for my predictive model. Logically thinking, it does not make sense to have a trip with speed less than 1 mph for entire duration and trips with speed greater than 70mph as limit is 50 mph in NYC. Hence, I dropped those values. To calculate average speed over a week I create a derived feature called Week that determines which week of the year it is. Plotting a histogram of speed shows that most cabs run at an average speed somewhere between 8-16mph.



A plot of average speeds per week is made to see the variations. Average speed per week is as follows: Week 1: 13.179632, Week 2: 12.565661, Week 3: 12.566875, Week 4: 13.046631, Week 5: 12.372110





Histogram of speeds observed according to time of day for each week

Now to check, if this variation in speed is just by chance (Null Hypothesis) or it has some statistical significance (Alternative hypothesis) ANOVA analysis has been done. We get:

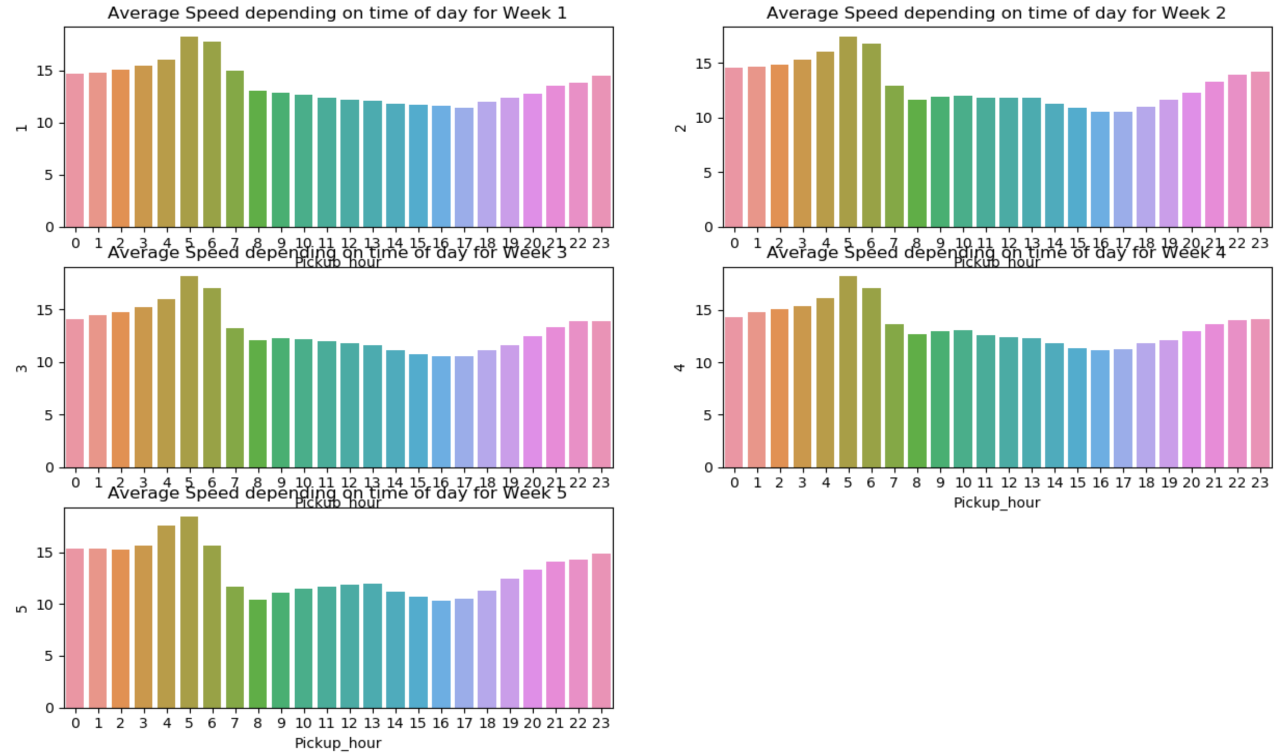
F_onewayResult (statistic=1002.9721898951893, pvalue=0.0)

That means we can reject the null hypothesis and conclude that this variation in average speed is not by chance. The ANOVA test indicates a large f-value and a small p-value, therefore we reject the null hypothesis and we conclude that the differences between the groups are statistically significant which implies that the week of the month does seem to be related to the average speed.

Similarly, applying ANOVA analysis on the data grouped by hour, we get:

F_onewayResult (statistic=5401.881835007106, pvalue=0.0)

The ANOVA test for sets partitioned as per the hour of the journey also gives a high f-value and p-value of 0, implying that there are statistically significant differences in the data sets considered



Average Speed depending on time of day for Week 1

Average Speed depending on time of day for Week 2

Average Speed depending on time of day for Week 3

Average Speed depending on time of day for Week 4

Average Speed depending on time of day for Week 5

Average speed observed during each hour of day per week

This figure makes a lot of sense. As followed by logical thinking average speed at night is more as compared to daytime. This makes sense as during night there is less traffic and fewer people on road. We see a spike in average speed during 4-5 am in the morning and it falls drastically as people get up and get going with their routine. As expected the average speed is lowest in the morning between 8-10 am when people are going to work.

Further Work

It was an amazing experience to work on this extremely rich dataset. However due to constraint of time there were many ideas that just remained in tits initial stages and could not be implemented. As future work and to fine tune my model I would love to do the following:

1. Learn more about how to work with geographic data and how to incorporate it within machine learning models. I would love to see if there is a difference in tipping pattern if the cab is hailed from one of the five different boroughs of New York. To do this I plan to create clusters of different boroughs and analyse tipping behaviour in each individually.
2. Another aspect that could provide interesting insights is inter-city trips. It would be exciting to see if tipping pattern change depending on intra-city or inter-city travel.
3. Since this is a huge dataset with 21 features to engineer, one way to approach it could be to try to reduce dimensionality by applying LASSO regularization.
4. Linear regression seems to a decent job of predicting tips. However, it is seen that many variables depict a non-linear trend and thus would love to try polynomial regression to better fit the model onto data points.
5. Implement Gradient boost to determine which features are most important and which could be dropped.