Manhattan Green Cab Report

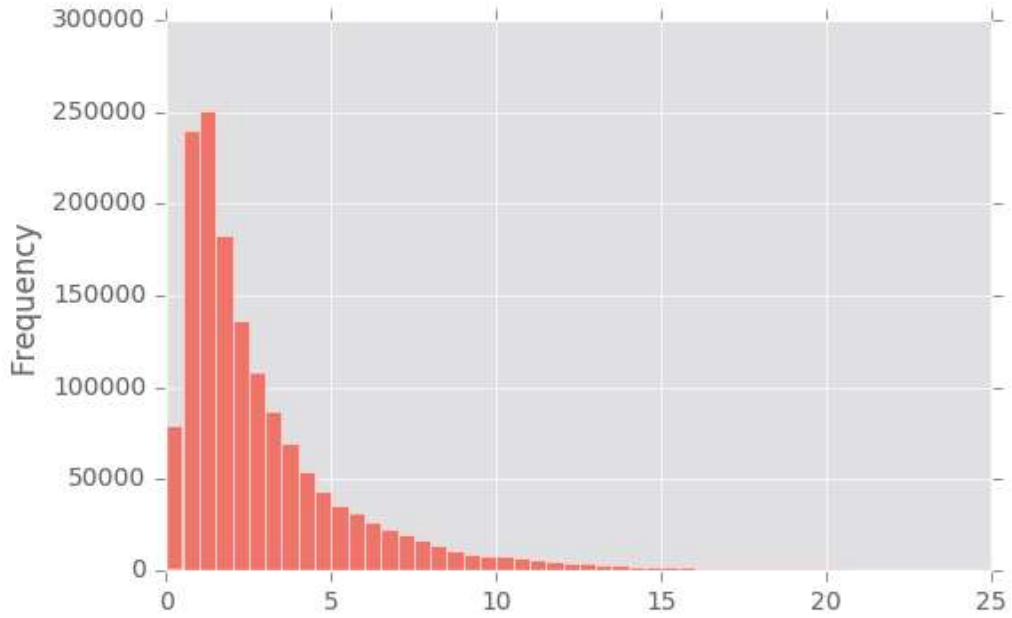
Part 1. Introduction

In this report, I used Python as my analytical tool to research on characteristics of New York Boro Cab("Green" taxi) in September 2015. The main findings are as below:

1. There are approximately 1.5 million trips by green cab in Sept. 2015.
2. Over 75% of the green cab rides have a distance less than 4 miles.
3. Trip distance varies a lot between different hours. Surprisingly, people travel longest distance by green cab at around 5-6 a.m.!
4. Around 11% trips originated or terminated at JFK or LGA. (I will present the exact number in Part 1) The average fare of "Airport trips" is $24.5, $11 higher than "In-Island" trips.
5. The average tip rate of "Airport trips" is 1.2% higher than "In-Island Trips".
6. Around 23% trips happened during late night. (23 p.m.-6a.m.) I derived this variable to try to find out more about what influenced the tip rate.
7. The average tip rate is 6.7%. This surprised me because I always pay at least 15% tips! (Well, although I know the actual tip rate should be tip amount divided by the total fare amount deducted by tip)
8. The number of passenger, trip distance, whether it is an airport trip, average speed of the trip and the amount of total fare all have significant impact on the tip rate, while whether it is a late-night trip does not significantly influence the tip rate.
9. The average speed for all the trips is 12.67 miles per hour.
10. The average speed for different weeks in September vary significantly.
11. I built up a function to try to predict the trip speed as the time of day.

Part 2. Data Description

The histogram of trip distance is as below:



Graph 2.1 Trip distance distribution

From this histogram, it is obvious that over 80% of all the trips are shorter than 5 miles. Around 1/3 of all the trips are between 0.5 miles and 1.5 miles.

Thanks to Google map, I made some interesting assumption and calculation:

The total are in Manhattan where green cab can run is 96E Street to 110W Street which is approximately a 2.5 mile * 0.8 mile square, plus LGA and JFK if booked.

There are 15 streets and 14 avenues in this area, meaning the total length of roads in this area is 2.5*15+0.8*14=48.7 miles.

The max distance in Manhattan should be 2.5+0.8=3.3 miles take 4 to consider some specific situation.

The MAX distance from LGA to “Green-cab” area is 9.6 miles and MIN distance is 7.4 miles.

The MAX distance from JFK is 20 mile and min distance is 18 miles.

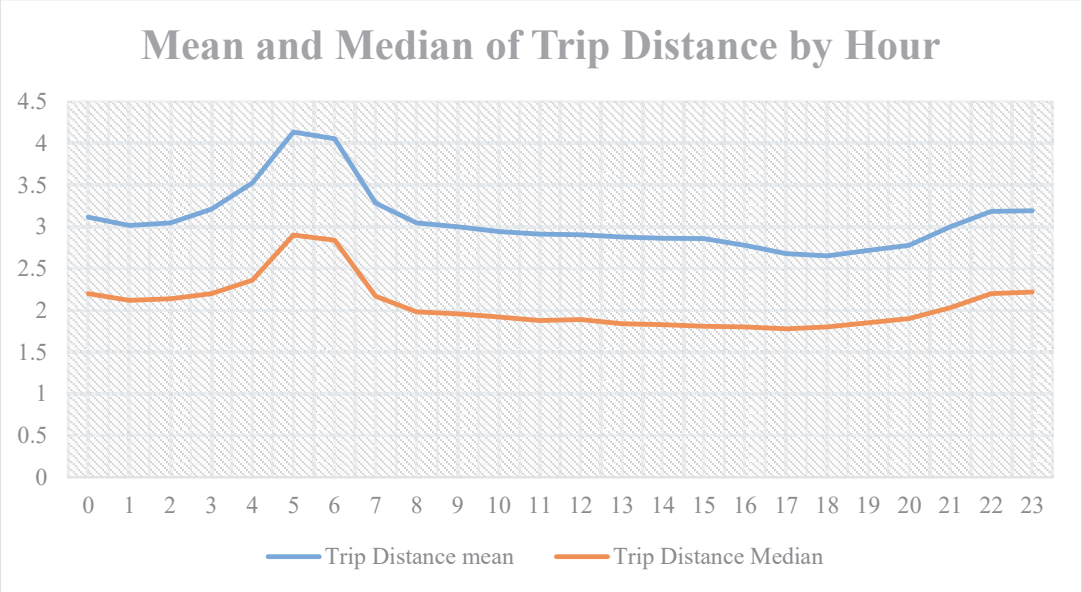
In this way, we can know that approximately 10% of the trip originated/terminated at the airports, (trip distance larger than 7.5 miles), which is proved by what I found when using longitude as the precise index to find the number of "airport-trips".

Considering there are about 75% of the trips are shorter than 4 miles, there should be 15% of the trips are between 4-7.5 miles, meaning either the passengers have some special request or the "one-way" rule makes the taxi driver take a detour. (Well, there is also possibility that the driver take the detour deliberately, but, let's just ignore it.)

The table and the graph below show you the mean and median trip distance grouped by hour of day.

Table 2.1 Mean and Median of Trip Distance Grouped by Hour

Hour	Mean	Median	Hour	Mean	Median	Hour	Mean	Median
0	3.12	2.2	**8**	3.05	1.98	**16**	2.78	1.8
1	3.02	2.12	**9**	3.00	1.96	**17**	2.68	1.78
2	3.05	2.14	**10**	2.94	1.92	**18**	2.65	1.8
3	3.21	2.2	**11**	2.91	1.88	**19**	2.72	1.85
4	3.53	2.36	**12**	2.90	1.89	**20**	2.78	1.9
5	4.13	2.9	**13**	2.88	1.84	**21**	3.00	2.03
6	4.06	2.84	**14**	2.86	1.83	**22**	3.19	2.2
7	3.28	2.17	**15**	2.86	1.81	**23**	3.19	2.22



Graph 2.2 Mean and Median of Trip Distance Grouped by Hour

Noticed that the medians are always smaller than mean. This is because the distribution of the trip distance is highly skewed to the right.

I noticed that in the data set, there are data of drop-off and pick-up coordinates, which in this way I can precisely know how many of transactions are between the airports and the Manhattan Island. From the Google map, I get the information about the coordinate of JFK and LGA. Although the latitude cannot show us clearly between the airports and the Island, longitude does help much. Any longitude smaller than -73.9 implies a place on the Island while those larger than -73.9 imply either JFK or LGA. In this way, if

[pick-up longitude-(-73.9)]*[drop-off longitude-(-73.9)] <0

It is sure that the trip originated or terminated at the airports.

In this way, I found out there are 161446 trips originating or terminating at the airports, accounting for 11% of all the trips.

The mean for the tip rate (percentage of the total fare) is 6.65%, while the median is 0.00%. Again this implies a highly-skewed distribution. Most people are not willing to pay high tip, which makes sense.

I used OLS to do this prediction model. Although there maybe non-linearrelationship between the variables, I do not have time to make more complex model.Maybe non-linear model can be a further improvement.

Furthermore, to avoid overfitting and multicollinearity, I tried to use as many variables as I can after filtering. For example, it is obviously that the payment method, tax and surcharge have nothing to do with the tip rate.

The following table shows the result of my first attempt:

Table 2.2 Result of First Attempt

OLS Regression Results

Dep. Variable:	tipPercentage	**R-squared:**	0.017
Model:	OLS	**Adj. R-squared:**	0.017
Method:	Least Squares	**F-statistic:**	4336.
Date:	Sun, 05 Mar 2017	**Prob (F-statistic):**	0.00
Time:	05:04:15	**Log-Likelihood:**	1.5121e+06
No. Observations:	1494926	**AIC:**	-3.024e+06
Df Residuals:	1494919	**BIC:**	-3.024e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	0.0539	0.000	254.216	0.000	0.053 0.054
Passenger_count	0.0001	6.93e-05	1.879	0.060	-5.6e-06 0.000
Trip_distance	0.0028	4.95e-05	57.429	0.000	0.003 0.003
Fare_amount	0.0002	1.36e-05	13.702	0.000	0.000 0.000
whetherNight	-0.0004	0.000	-2.389	0.017	-0.001 -7.47e-05
whetherAirport	-0.0270	0.000	-108.356	0.000	-0.028 -0.027
speed	0.0004	1.42e-05	26.240	0.000	0.000 0.000

Omnibus:	450107.470	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2364188.606
Skew:	1.360	**Prob(JB):**	0.00
Kurtosis:	8.527	**Cond. No.**	72.1

Although most of the variables have significant relationship with tip rate, the R-square is low. I decide to move forward to omit some insignificant variables in my next step.

The following table shows the result of my second attempt:

Table 2.3 Result of Second Attempt

OLS Regression Results

Dep. Variable:	tipPercentage	**R-squared:**	0.343
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	1.564e+05
Date:	Sun, 05 Mar 2017	**Prob (F-statistic):**	0.00
Time:	05:05:00	**Log-Likelihood:**	1.4803e+06
No. Observations:	1494926	**AIC:**	-2.961e+06
Df Residuals:	1494921	**BIC:**	-2.960e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Passenger_count	0.0080	6.33e-05	126.980	0.000	0.008 0.008
Trip_distance	-0.0017	4.72e-05	-36.798	0.000	-0.002 -0.002
Fare_amount	0.0017	1.26e-05	132.898	0.000	0.002 0.002
whetherAirport	-0.0314	0.000	-123.390	0.000	-0.032 -0.031
speed	0.0029	1.01e-05	287.797	0.000	0.003 0.003

Omnibus:	438139.673	**Durbin-Watson:**	1.959
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2563426.334
Skew:	1.286	**Prob(JB):**	0.00
Kurtosis:	8.877	**Cond. No.**	71.5

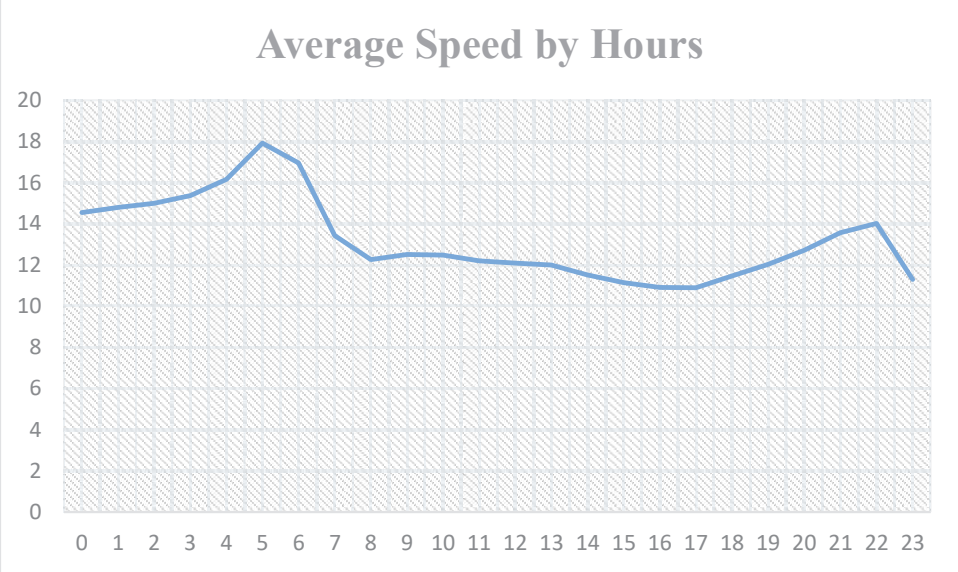
This time, the result makes more sense. More passengers, longer distance, larger amount of fare and higher the speed contribute the higher tip rate. However, it is strange that airport-trips pay lower tip rate. The summary of the relationships is as the following table.

Table 2.4 Relationships between Different Variables with Tip Rate

Variables	Relationship with Tip Rate
Number of Passengers	Positive
Distance of Trip	Positive
Fare Amount	Positive
An airport trip?	Negative
Speed	Positive

A further consideration should be normalize the variables. This should contribute to a more precise prediction model.

The graph below shows the average speed by different hours. Noticing there is an obvious non-linear relationship, I did non-linear model in the following prediction model.



Graph 2.3 Mean of Speed Grouped by Hour

I first generate an index to signal the different weeks of the trip. Then I used F-test to test whether the average trip speeds are the same in different weeks.

I used scipy.stats to help me with the test and the test result is:

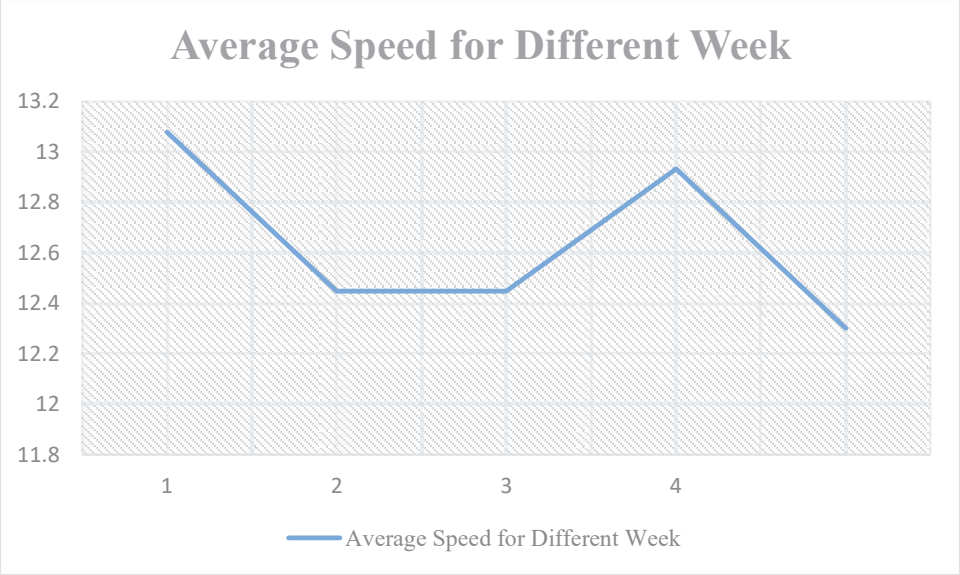
```
speedWeek1=greenFinal.speed[greenFinal.week==1]
speedWeek2=greenFinal.speed[greenFinal.week==2]
speedWeek3=greenFinal.speed[greenFinal.week==3]
speedWeek4=greenFinal.speed[greenFinal.week==4]
speedWeek5=greenFinal.speed[greenFinal.week==5]
```

```
import scipy.stats as stats
stats.f_oneway(speedWeek1, speedWeek2, speedWeek3, speedWeek4, speedWeek5)
```

```
F_onewayResult(statistic=843.14106813529077, pvalue=0.0)
```

Graph 2.4 F-test Result of Average Speed of Different Weeks

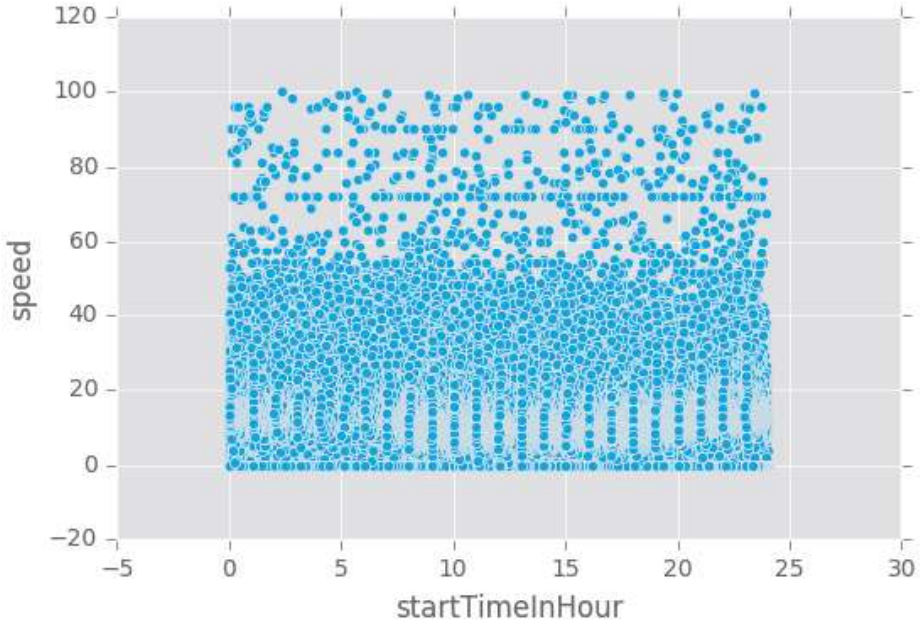
Obviously, the average speed for different weeks are not the same. The following graph shows the difference.



Graph 2.5 Mean of Speed Grouped by Week

As far as I am concerned now, there are several aspects that may influence the traffic speed, which are weather, holidays, special events, etc. So I googled these aspects and found out that the weather between Sept.8th 2015 and Sept. 15th 2015 was terrible. It rained heavily, which I believed is an important contribution to the low speed of week2 and week3. Also I thought some events like 9-11 parades and US-open can contribute to traffic jam, they all happened during the second of week of September.

The following graph is the scatter plots for the speed in different time of a day.



Graph 2.6 Speed of Different Time of a Day

However from Graph 2.3 we know that the average speed of different hours seems to follow a quartic function because it has 2 turning points. Thus I used Python to fit a quartic function between the average speed and the hour. The function should be as below:

$$\mathbf{y = -0.0004527x^4 + 0.02349x^3 - 0.3806x^2 + 1.78x^1 + 13.76}$$

Where x is the hour of the day and y is the average speed.

Part 3. Further Discussion

In this part, I will present the potential improvement of the research.

Firstly, the raw data has outlier points, which I introduced a function called reject_outliers to omit the influence of these points. The method I use in this function is to omit any point that lies outside 6 times standard deviation from the mean. However, I'm not sure whether 6 is a good multiplier for this case.

Secondly, when calculating the duration of each trip, I omitted those trips that started from day t and ended at day t+1 for simplicity. For example, if one trip started at 11:59p.m. 15th Sept. and ended at 0:05 a.m. 16th Sept. I just omitted this record. These kind of record account for 1.7% of the total trips. Although it will not have significant impact on the final result, I will try to improve it once I have more time.

Thirdly, when doing linear regression to explore the relationship between tip rate and other variables, I did not normalize all the variables, which will have bad influence on the final result. We can see from the JB test result that the data is enormously deviate from normal distribution.

The last point is the model I used to construct the predictive model for tip rate is linear while I do think there will be non-linear factors. For example, it is not necessary that the higher the speed is, the higher the tip rate is for the reason that passenger may feel sick if the cab driver drives too quickly and in this way, the passenger maybe less willing to pay tips.