NYC GREEN TAXI SEPTEMBER 2015 DATA ANALYSIS

Question 1

a) **Programmatically download and load into your favorite analytical tool the trip data for September 2015.**

Soln: urllib used for this. See code

b) **Report how many rows and columns of data you have loaded.**

Soln:

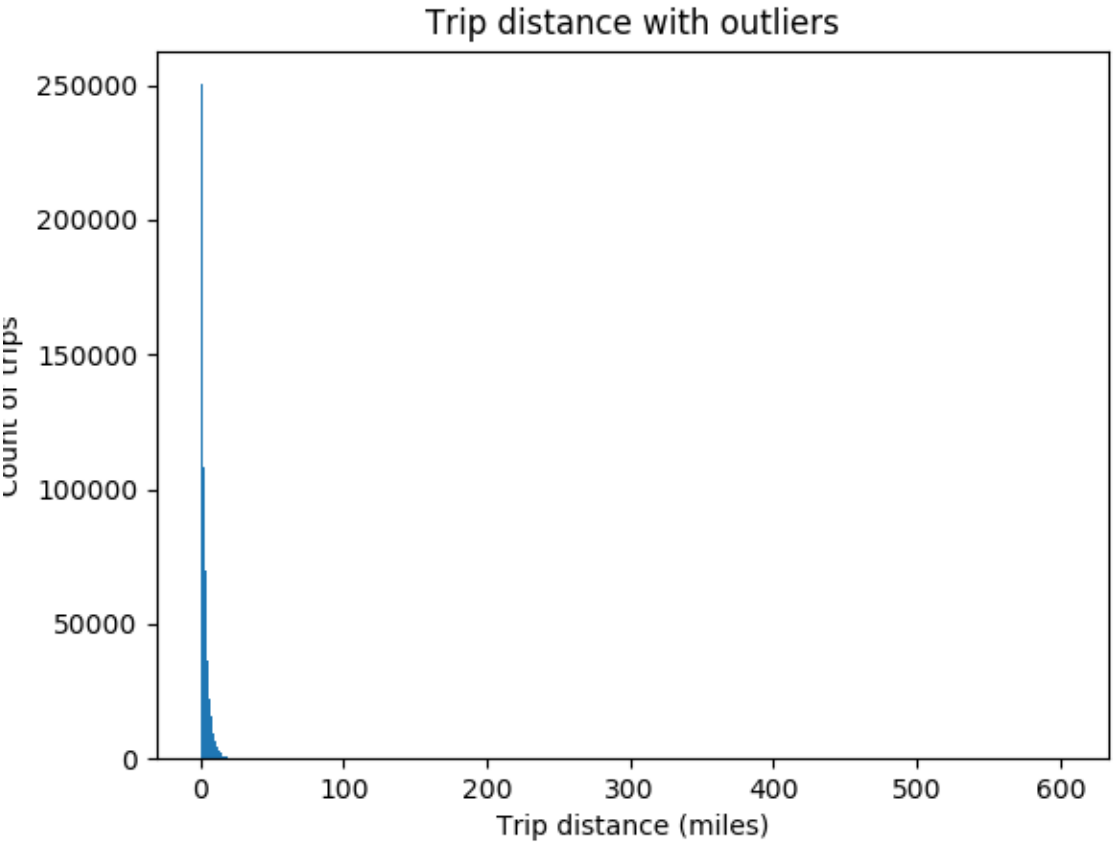
ROWS: 1494926
COLUMNS: 21

Question 2

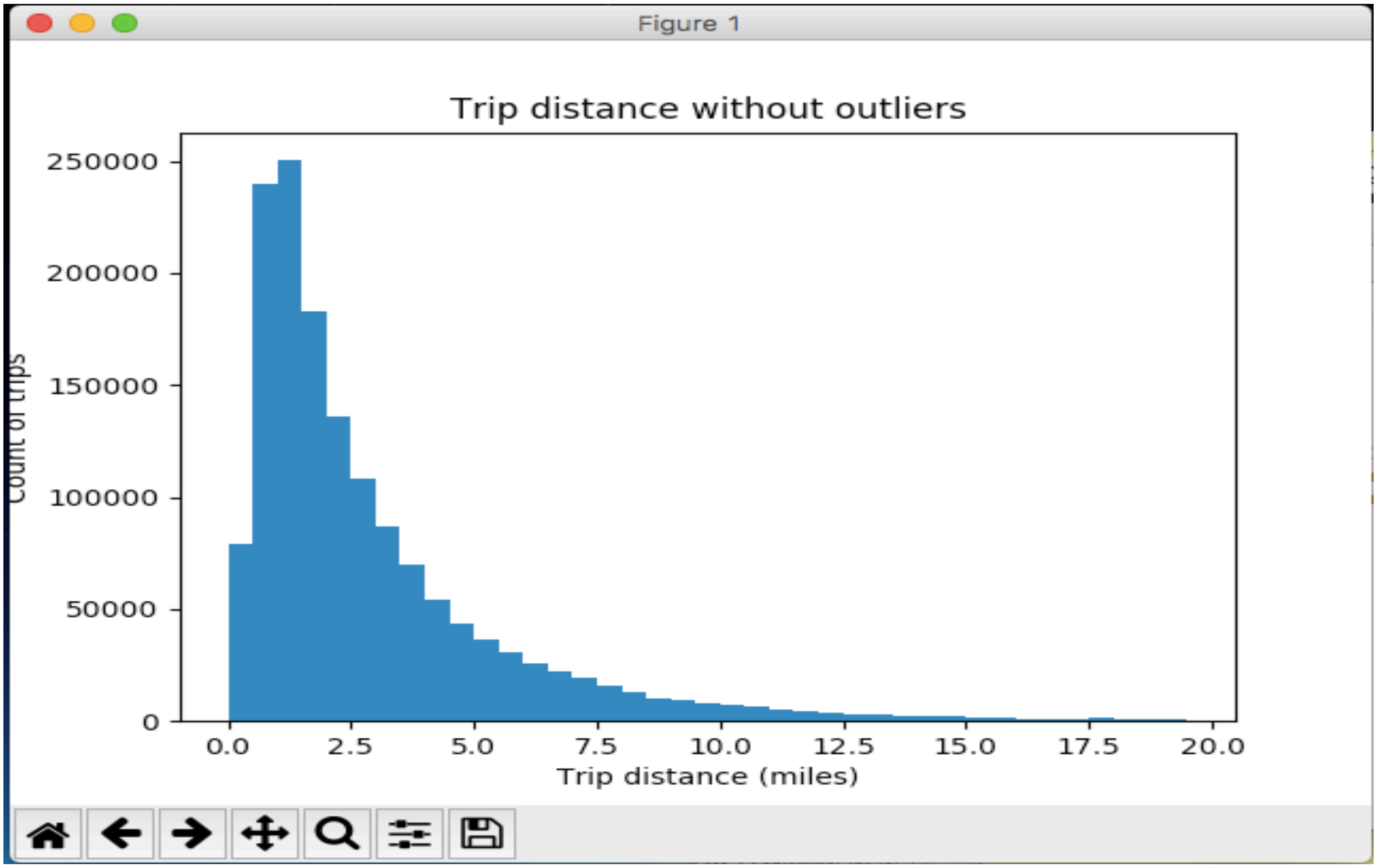
a) **Plot a histogram of the number of the trip distance ("Trip Distance").**

Soln:

Histogram Including Outliers



Histogram Excluding Outliers



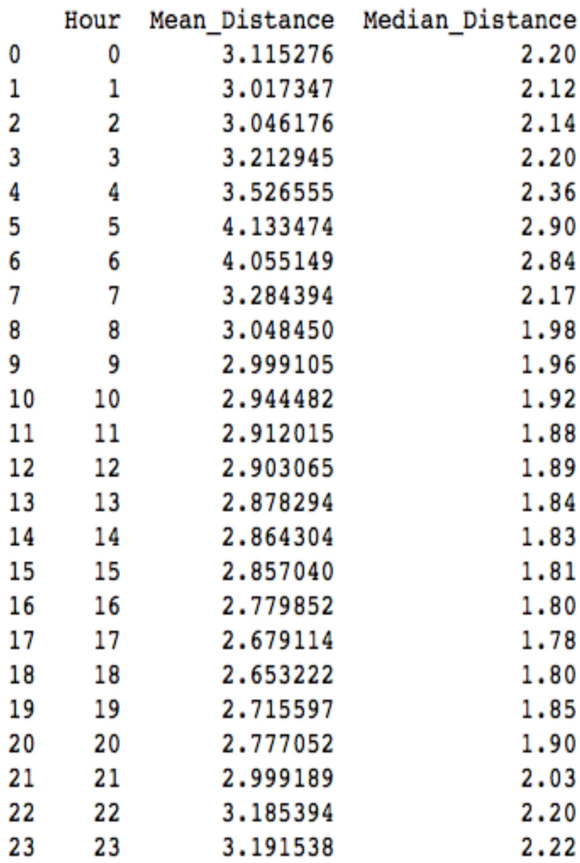
b) **Report any structure you find and any hypotheses you have about that structure.**

The above diagram shows that a majority of trips lie between 0.5 m to 4 miles. People prefer to take a green taxi for shorter distance. This could be due to the following reasons:

1) Taking green taxi for short distances is affordable (in terms of cost) for people while long distance may not be.
2) People may take a cab in the morning for going to work and offices are generally located close to people's home

Question 3

a) **Report mean and median trip distance grouped by hour of day.**



```
    Hour  Mean_Distance  Median_Distance
0      0       3.115276             2.20
1      1       3.017347             2.12
2      2       3.046176             2.14
3      3       3.212945             2.20
4      4       3.526555             2.36
5      5       4.133474             2.90
6      6       4.055149             2.84
7      7       3.284394             2.17
8      8       3.048450             1.98
9      9       2.999105             1.96
10    10       2.944482             1.92
11    11       2.912015             1.88
12    12       2.903065             1.89
13    13       2.878294             1.84
14    14       2.864304             1.83
15    15       2.857040             1.81
16    16       2.779852             1.80
17    17       2.679114             1.78
18    18       2.653222             1.80
19    19       2.715597             1.85
20    20       2.777052             1.90
21    21       2.999189             2.03
22    22       3.185394             2.20
23    23       3.191538             2.22
```

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.**

Total num of trips to/from any of the airport: **5552**
Average fare of trips to/from airports: **$ 48.976945245 per trip**
Average total fare amount before tipping of trips to/from NYC airports: **$ 57.208420389 per trip**

Question 4:

a) **Build a derived variable for tip as a percentage of the total fare.**

Soln: See Code : df['Tip_Percentage'] contains the derived values.

b) **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.**

Soln: See Code: classifier.pkl contains the classifier

1) RandomForestRegressor was used to model the classifier.

2) Complete data was used while attibutes used were : 'Total_amount', 'Trip_duration', 'Avg_Speed_mph'
3) 5-fold cross validation gave following results: RandomForestRegressor test Avg mean squared error: 12.18
4) Number of estimators used were = 200. Different values (50,100,150,200,250) were tested for Mean Squared error. The Value improved as n_estimators increased. However, the improvement was insignificant beyond 200.
5) An interesting thing was approximately 60% people gave no tip. I could have also formed a binary classifier to classify whether a tip will be given or not. However due to time constraints could not.

Question 5 - Option A: Distributions

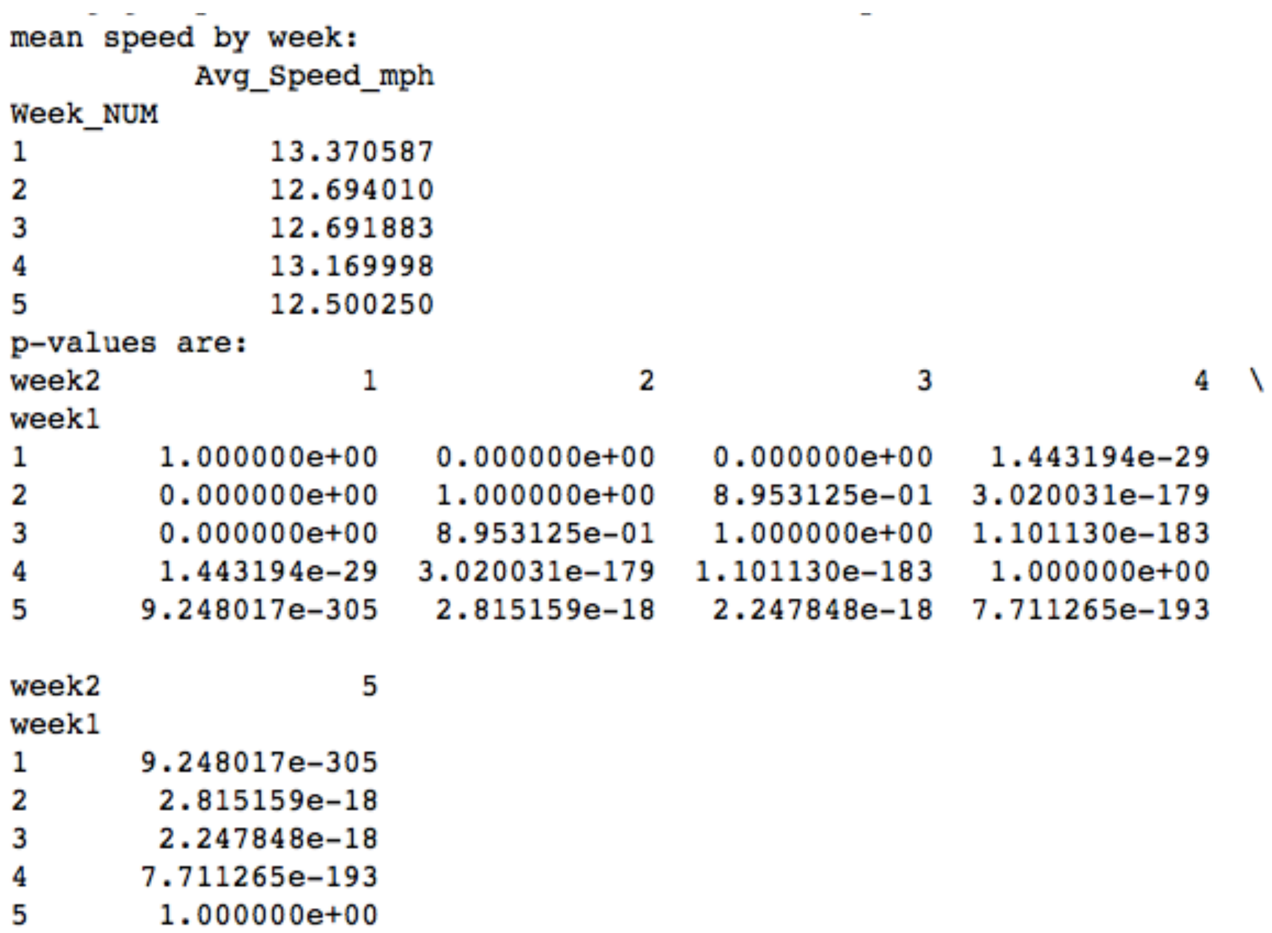
a) Build a derived variable representing the average speed over the course of a trip.

Soln: See Code: Avg_Speed_mph contains this

b) 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?

Soln: I have used Paired t-test to check this.

Null Hypothesis: **Average trip speeds are materially the same in all weeks of September.**



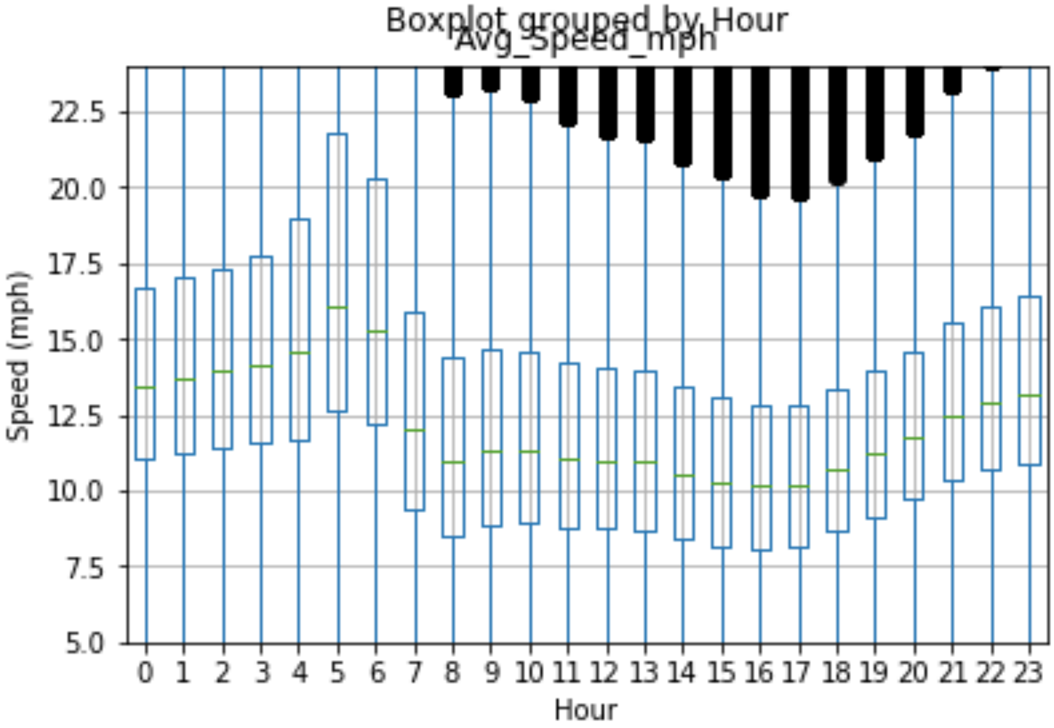
```
mean speed by week:
          Avg_Speed_mph
Week_NUM
1              13.370587
2              12.694010
3              12.691883
4              13.169998
5              12.500250
p-values are:
week2              1              2              3              4  \
week1
1       1.000000e+00   0.000000e+00   0.000000e+00   1.443194e-29
2       0.000000e+00   1.000000e+00   8.953125e-01  3.020031e-179
3       0.000000e+00   8.953125e-01   1.000000e+00  1.101130e-183
4       1.443194e-29  3.020031e-179  1.101130e-183   1.000000e+00
5      9.248017e-305   2.815159e-18   2.247848e-18  7.711265e-193

week2              5
week1
1      9.248017e-305
2       2.815159e-18
3       2.247848e-18
4      7.711265e-193
5       1.000000e+00
```

It can be seen that accept for **week 2 and week 3** which has confidence level of 91%, p-values are very small. Thus we can reject the null hypothesis.

c) Can you build up a hypothesis of average trip speed as a function of time of day?

Let us form a boxplot of **average trip speed as a function of time of day:**



The plot reveals that green taxi's tend to have higher speed during early morning. This may be due to less traffic on road early morning while as the traffic increases, average speed of taxi's decrease. Therefore we can form a Hypothesis as:

Hypothesis: Roads have less traffic during early hours(morning) of the day. Hence **average trip speed for green taxis** reaches maximum early morning and continues to decrease till late evening.