**Green taxi code challenge**

**Question 1**

Approach: Load the data set to python’s pandas data frame

Outcome: The row count is 1494926 and column count is 21

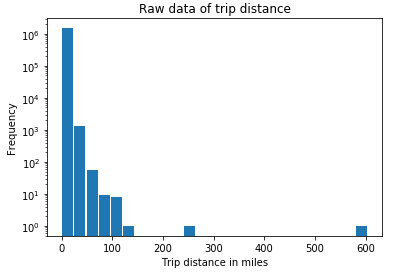
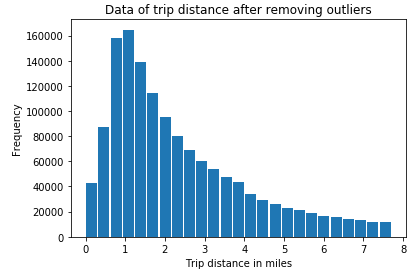
**Question 2**

Approach: Plotted the histogram using plot function in the matplotlib library

Outcome:

After completing this experiment and generating the histogram I was able to find the following things:

1. When I first plotted the histogram for “Trip Distance” from raw data, I got a weird looking histogram and trip distance value ranged from 0 to 600 which directed me to check for outliers using box plot and guess what there were many **outliers** which I successfully removed
2. After removing outlier, I was able to generate a required histogram and found out that the shape of data is not symmetric that is it was **right-skewed**.
3. Now it can be clearly observed that which **range** of trip distance has the maximum and minimum trips
4. **Hypothesis**: Highest number of trips have trip distance close to 1 mile and lowest number of trips have trip distance around 6 miles which may be related to fact that people take a taxi when they want to travel for short distance, and they don’t prefer taxi for a longer distance

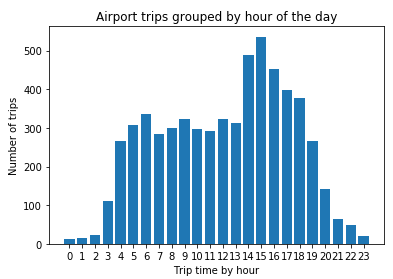
**Question 3**

Approach 1(failed): Calculate the geospatial distance between pickup, drop off location data and Airport coordinates and check whether the distance is less than 1 miles and classify it as airport location or not. This part of the code is commented because time complexity to perform this was high.

Approach 2(Passed): Dropped pins on google maps around JFK area to find out the coordinate range and checked whether pickup and drop off coordinates fall in this range and if they do classify them as airport trip and if they don’t classify them as other trips.

Outcome:

1. A rough count of JFK airport trips: 5996
2. Average fare of JFK trips: 41$
3. I did some more analysis on the JFK airport trips and found the approximate count of airport trips grouped by hour of the day as seen below in the graph and discovered that very high number of trips where done in afternoon that is around 2pm to 4pm by which we can infer that the frequency of flights arrivals and departures are high around these timings.



**Question 4**

Approach: Build the Predictive model in the following phases.

1. Data Cleaning: The code for this part is well commented and self-explanatory.
2. Feature selection: To determine required predictors I tried two methods i.e. wrapper and filter and finally choose filter method as a wrapper (code is commented) is memory inefficient. I dropped a few features that I felt where not required for training the model.
3. Model Selection: I tried four different model namely XGBoost, XGBoost with K-fold cross-validation (failed attempt due to memory error), Stochastic gradient descent and random forest. I finally choose two random forests and XGBoost.
4. Model implementation: Random forest takes more time to train but has less RMSE and XGBoost is faster but has more RMSE than random forest. I have kept the code for XGBoost commented as running both models were causing memory error in my system. Please keep one model commented while testing other to avoid memory error.
5. **Please be patient while running the models as it may take little long based on machine configuration.**
6. Model evaluation: Both models were evaluated based on RMSE and I found out Random forest RMSE was better at 0.4 compared to XGBoost that is around 1.584. But there is a tradeoff between two for running time and accuracy. If I had more time, I could have tried improving the performance of XGBoost by parameter tuning.

Outcome: Successfully built and tested two models for predicting tip amount for the given rides which can be easily used to calculate tip percentage.

**Question 5 – Option A: Distributions**

Approach: I performed the following steps.

1. Firstly, some data cleaning to keep only the required features for hypothesis testing and remove unnecessary values that may spoil the future results
2. Performed ANOVA hypothesis testing on average trip speed based on a week of the month
3. Visualized data to build a hypothesis for average speed based on the hour of the day

Outcome: Successfully performed hypothesis testing and proved the results with visualization of distributions.

**References**

**Finding outliers**

(1)<https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba>

(2)<http://blog.minitab.com/blog/3-things-a-histogram-can-tell-you>

**Feature Selection**

(3)https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

**Building Model**

(4)<https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af>

(5)<https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>

(6)<https://www.quora.com/What-are-some-good-machine-learning-algorithms-for-predicting-a-continuous-variable-with-N-number-of-inputs>

(7)<https://datascience.stackexchange.com/questions/23789/why-do-we-need-xgboost-and-random-forest>

(8)https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80

**Testing Hypothesis**

(9)<https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>

(10)<https://machinelearningmastery.com/parametric-statistical-significance-tests-in-python/>

(11)https://towardsdatascience.com/hypothesis-testing-in-real-life-47f42420b1f7

(12)<https://www.sciencebuddies.org/blog/a-strong-hypothesis>

**END**

**Thank you**