Group Zymic:
Michael Chon, mc2153@georgetown.edu
Yuan-Yao Chang, yc704@georgetown.edu
Zheng Chai, zc104@georgetown.edu
Chong Zhang, cz211@georgetown.edu

Week Status Report #2

The list of jobs we have done so far and needed to be done:

1. Took out the outliers
2. Expanded the dataset
3. Performed statistical calculations
4. Optimized the fare and tip group range
5. Divided the taxi pick-up times into weekday and weekend
6. Divided the day into 2 hour range
7. Histogram:
   fare_amount(5), tip_amount(1),
8. Heat Map:
   Pickup and dropoff location- for the highest and lower fare and tip
9. Compare between the seasons on average fare and tip
10. Using Spark MLlib machine learning
11. Start our presentation slide (google site) to put our ideas and results together

Summary

In order to make our analysis more accurate, we decided to increase the size of the dataset from four months to one year because we realized that four months are not large enough to see the patterns of the taxi movement around the city. Also, we have taken out several outliers of the fare and tip amount to make the dataset sound more reasonable because the total fare amount per one particular month is around $800,000, which is relatively high compared to other months.

We have some interesting results from the 1 year range dataset, such as the average tips of each weekday, the average fare of each zone, etc. To convey the results in a clear manner, we will visualize the results using appropriate graphs such as histogram and line graph.