ANLY 502 Spring 2017 Final Project Proposal
NYC Taxi

TEAM:
Yuan-Yao Chang (yc704@georgetown.edu)
Zheng Chai (zc104@georgetown.edu)
Chong Zhang (cz211@georgetown.edu)
Michael Chon (mc2153@georgetown.edu)

Data Science Research Question:
What strategies can NYC taxi drivers use to make more money?

We are going to look into NYC taxi data for two distinct seasons, (Summer 2016: June, July, August, Winter 2015: December, Winter 2016: January, February) to find out how NYC taxi drivers can make more money. Using data analytics, NYC taxi drivers can find different patterns in fares and using these patterns to maximize their profits. Also, using visualizations, we can easily explain the patterns and "Taxi Economics" in NYC.

Dataset:
A total of 6 months
Winter 2015: December,
Winter 2016: January, February
Summer 2016: June, July, August

NYC yellow taxi: approximate 2GB per csv.
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2015-12.csv
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-02.csv
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-06.csv
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-07.csv
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-08.csv

The attributes of the dataset are:
"VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,pickup_longitude,pickup_latitude,RatecodeID,store_and_fwd_flag,dropoff_longitude,dropoff_latitude,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount".

The reference to the attributes is:
http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Smaller Questions:
1. In comparison of the two seasons, which season is more profitable for NYC taxi drivers?
   - We can do this running statistical analysis to find a 5 point summary on the "total amount" column on each season and compare the results to each other.
2. In which neighborhoods do customers tend to tip more for NYC taxi drivers?
   - We can distinguish customers into two categories, where one category is that customers tip and another category is that customers do not tip. Within the category that

customers tip, we can break it down and find out a best pick-up location where customers tend to tip more.

3. Where can NYC taxi drivers pick up customers quick?
   - We can answer this question by finding out most frequent customer pick-up locations in various neighborhoods in NYC.
4. Around what time can NYC taxi drivers pick up customers easily?
   - We can answer this question by finding out the most frequent of customer pick-up times on the entire data or the two different seasons.
5. How can NYC taxi drivers maximize their profits while minimizing cost?
   - We can find a number of trips that contain relatively shorter routes but results in higher total amounts.

Data Preparation:

The dataset is relatively clean. In addition to the given attributes in the dataset, we are going to build one attribute called “the name of neighborhood” based on latitude and longitude by using python package “geopy”. Also, one additional attribute called “fare_amount_group” will be built to break the “total amount” column into groups so that we can use association rule on these two attributes, “fare_amount_group” and “district” and the another column “total_amount”.

Cluster:

We are going to experiment a few clustering analysis with the dataset:

K-mean

Random Forest: We are going to build a random forest model to predict a pick-up density on an average weekday.

ANALYSIS TOOL:

Scala, Python
Hadoop, Pyspark, Spark, AWS, EMR, EC2, S3

VISUALIZATIONS:

Heat Map of the routes and pick up drop off spot.
Pick up density.
Frequently used route.
Histogram of DIstricts and fare amount.
….etc.

WEEKLY SCHEDULE AND MILESTONES:

March 28
— Final Project Group Proposal
April 10
— Cleaning the data. Deciding Model. Read References. Finalize the decision of the goal.

April 17
— Setting up hypothesis. Coding. Train the model. Analysis the results. Visualization.
April 24
— Tuning the parameters of model. Prepare report and presentation.
April 27
— Final project slides due.
May 1
— In class final project presentation.
May 10
— Final project paper submitted

FINAL GOALS:

From this project, we have numerous recommendations to NYC taxi drivers to make more profits by exploiting the unhidden patterns of the dataset.