David Stern
IS 602 Project Part 2

In my project, I want to explore the use of Boro (green) taxis in NYC since 2013. This data is available on the NYC Open Data website in separate data sets for individual years. I will attempt to access each data set - 2013, 2014, 2015 - through the website’s API and display my work in an IPython Notebook. I expect to use pandas to store the dataset and analyze the relation between the many variables.

I am interested in exploring the number of green-taxi rides by month, to see how the program is growing (or shrinking) in response to app-based competition (Uber, Lyft). I am also interested in exploring typical trip distance and speed by hour over the course of a day. I would also like to examine rate type (hail versus hire) by location. Boro Taxis are not permitted to pick up street-hails beneath certain boundaries – W 110th St and E 96th St – so it will be interesting to explore these geographical patterns visually, and determine to what extent these taxis serve citizens in outer boroughs or “cheat” and pick up passengers in excluded areas. I plan to present these findings visually with the matplotlib package (for graphs) and the gmaps package (for maps).

For the maps, the dataset conveniently provides the locations of pick-ups and drop-offs as latitude and longitude coordinates, each split into two fields. I plan to read these coordinate pairs into the gmaps heatmap feature. This feature should allow us to identify hotspots, where activity is highest.

```
data = gmaps.datasets.load_dataset('coordinates’)
map = gmaps.heatmap(data)
        gmaps.display(map)
```

Exploring average trip speed by time of day will require parsing the drop-off and pick-up times with the datatime feature in pandas and finding the difference for each record. We can also find the number of fares per hour by creating a separate column for hours and then using the value_counts() function.

```
df['pickupTime'] = pd.to_datetime(df['pickupTime'],format='[%H:%M:%S]')
df['hour'] = df['pickupTime'].dt.hour
df['hour']. value_counts()
```

I will determine the trip speed by dividing the distance by the total duration of the ride. To plot this against time of day, I will compare it to the time of pick-up rather than drop-off. This plot, and others that explore variables over time will be demonstrated as line graphs in matplotlib.