CSE 519: Data Science Fundamentals Project

**Background**:

Citi Bike is a privately-owned bicycle sharing service located in NYC. When Citbike launched in 2013 there were 332 stations and 6000 bicycles available for use. The service has grown to an annual subscribership of approximately 164,000. Taking into account the number of CitBike users, and the availability of data associated with each bike trip made on a given day, the study of bike usage has, in the past, given great insight into the traffic situation of the city at a given time.

**Problem Statement**:

We intend to predict the number of Citi bike users at a given time based upon traffic congestion. We hypothesize that bike usage will be correlated with the level of congestion. Relatedly, there is some interesting analysis to be done on the effects of bicycling on traffic volume. Because we can look at taxi data before and after the introduction of Citi bikes we can get some insight into how taxi ridership has been affected by bicycling. Further, we have alternate sources of traffic information that will give us further insight into traffic volume. NYC could use the results of such analysis to improve traffic patterns throughout the city.

**Literature Review**:

The research paper[2] that was talked about in the project description, aimed at a rebalancing method, to try to balance the demand and supply at stations across the city. The interesting takeaway from the paper was the way the authors represented the stations as a set of nodes and edges, and thus finding optimal routes between them. Another interesting research paper[2], aimed at understanding traffic patterns via the number of shared bicycles, not only the Citi Bikes, across the city. It observed the behaviour of each certain stations which tend to be used most often, and based on these observations, tried to predict the amount of traffic its effect on average travel times near these stations. Other online studies such as “A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System”[3] by Todd Schneider, gave us insight into the methodology behind understanding trends in bicycle usage. A similar study by the same author[4], indicated the

**Dataset**:

Citibike data - <https://www.citibikenyc.com/system-data>

The Citi bike dataset is released by Citi bike and contains the following columns: Trip Duration (seconds) Start Time and Date Stop Time and Date Start Station Name End Station Name Station ID Station Lat/Long Bike ID User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member) Gender (Zero=unknown; 1=male; 2=female) Year of Birth

The data that we are primarily interested in is (start station, end station), (start time, stop time) and trip duration.

NYC Taxi Data - <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>

The taxi data contains the following fields: <http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>

We are primarily interested in the following fields: (Dropoff\_latitude, Dropoff\_longitude), (Pickup\_latitude, Pickup\_longitude), (tpep\_pickup\_datetime, tpep\_dropoff\_datetime), Trip\_distance, and Fare\_amount.

Traffic cam data - <http://data.beta.nyc/dataset/nyc-real-time-traffic-speed-data-feed-archived>

The traffic cam data provides Speed (average speed of vehicle traversing link) and Travel Time which provides the average travel time for a vehicle traversing the link. This will give us a second perspective on traffic congestion. We know the speed limit for a given road so we can approximate how close to 100 percent efficiency the road is operating.

Traffic Volume - <https://data.cityofnewyork.us/NYC-BigApps/Traffic-Volume-Counts-2012-2013-/p424-amsu>

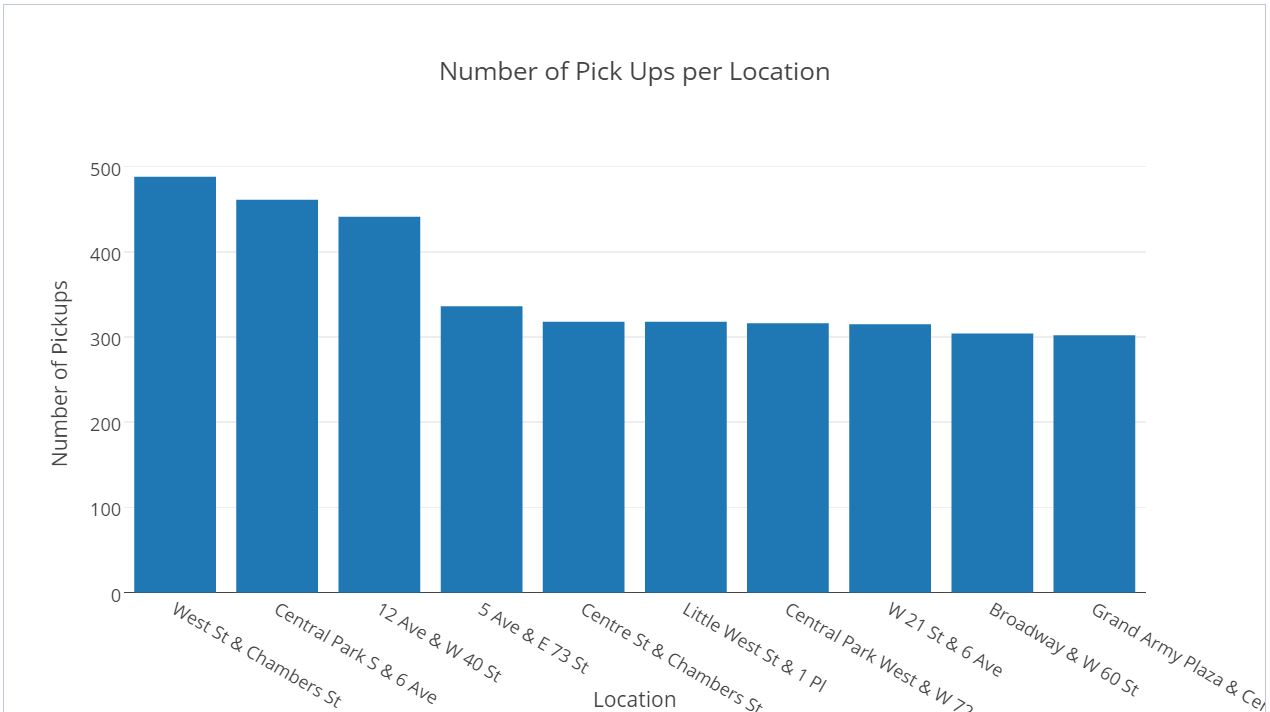
The traffic volume data gives us a third perspective on congestion on the roads. It provides the count of vehicles that cross a road segment by hour.

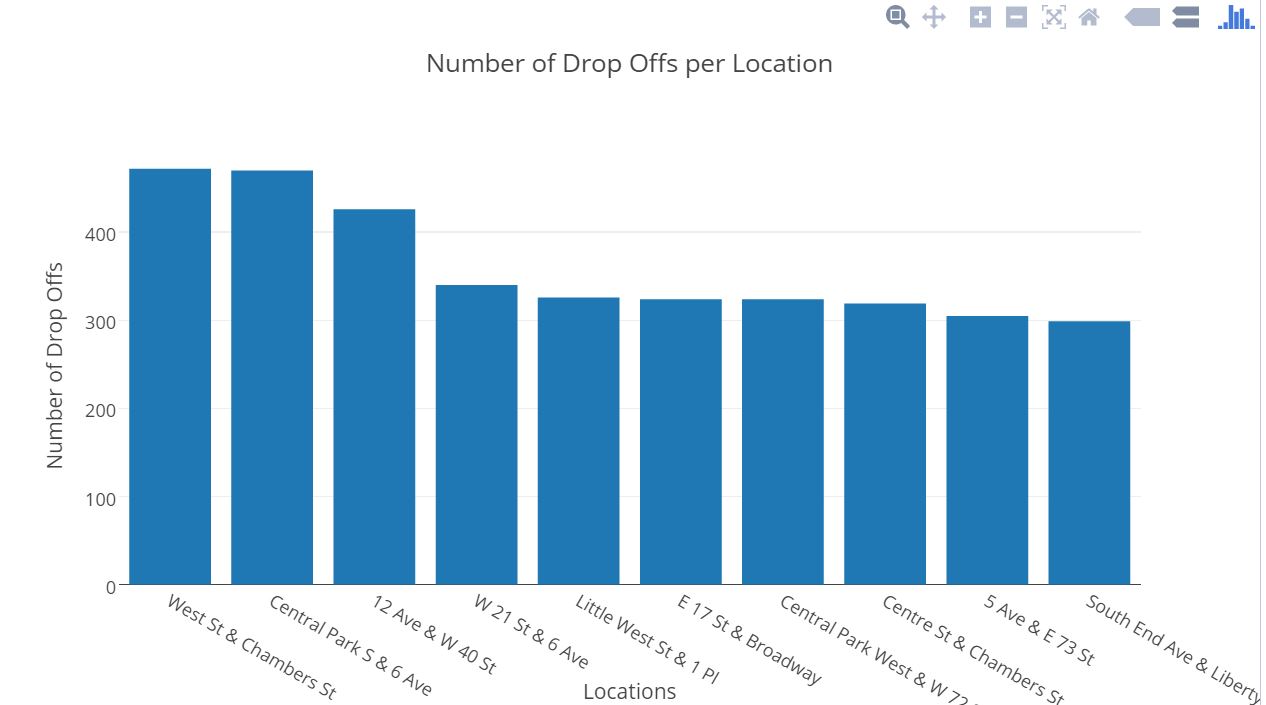
Google Maps - <https://developers.google.com/maps/documentation/directions/intro#traffic-model>

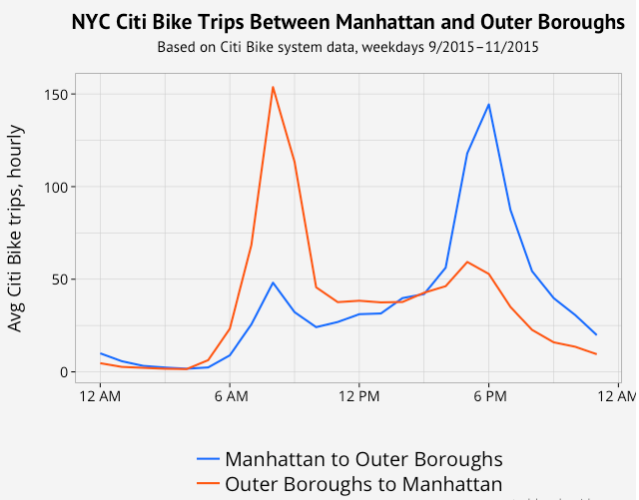
The google maps API provides us with a way to get a google estimated time for a particular route. This will give us a baseline for comparison.

**Methodology**:

1)Upon initial analysis, we observed that certain stations, as per the city bike data have many pickups and drop-offs at a time. For example, consider the plots below, displaying the number of pickups and drop-offs as per locations. West St and St Chambers has many pickups throughout the day, and this trend is also followed by the number of drop offs. As per the study of Todd Schneider [1], the number of trips between the suburbs and the business centres (for example, Outer Burroughs to Manhattan) tend to peak at twice during the day. These hours, are typically when most people tend to leave their homes for work, and when people tend to return to their homes from their workplace. We aim to further investigate this trend at a much deeper level of granularity. For example, our next step will be to investigate the number of pickups and drop offs in a larger geographical area (by combining segments of Central Park etc), to correlate it with the traffic congestion as described by the NYC Can and Traffic Cam datasets, which can be further confirmed by the Google Maps API, which can give us the average time taken via a route depending upon the traffic.







(source: <http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/>)

2)The various data sets we have access to provide different approximations of congestion. The traffic cam data will give us the average speed at a given for a given place. We can create a network of bicycle stations and then identify the most heavily trafficked edges. This will give us an idea of where to look for traffic data for comparison. We can select out and aggregate these trips which will give us Citi bike user volume.

3)We expect that bike and taxi usage will peak at the rush hour points of 7-9 AM and 5-7 PM. This means that we must look to other times to validate our hypothesis because we need to separate out the effect of the commute on the absolute magnitude of users from the effects of traffic. For example, if traffic is high at 12PM on a Saturday then this would be an ideal time to check the bike usage because we would not expect a significant deviation from the norm. But if we do see a substantial deviation we believe that would substantiate our hypothesis. Not only will we be able to correlate congestion with Citi bike usage but we will also be able to validate or invalidate the choice to use a Citi bike because we know the trip times for comparable taxi and bike journeys. We can then use the three different data sets that we are using to approximate traffic congestion and evaluate them to determine the best combination of variables to predict Citi bike usage.

**Considerations**:

1)We will need to adjust usage by number of subscribers -- just because more people are using bikes doesn't mean that there is more congestion. There may be more subscribers.

**Beneficiaries**:

We are positive that this data exploration will result in many interesting findings, which will directly or indirectly benefit the CITI bike program, riders, as well as the New York city planning committee.

2) We need to adjust for days where the number of tourists in the cities tend to increase.

**Team Members**:

1) Michael Liuzzi- 109524485

2) Leena Shekhar- 100944612

3) Sharang Bhat-

**References**

[1] –Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization- Junming Liu1, Leilei Sun2, Weiwei Chen3, Hui Xiong

[2]- Sensing and Predicting the Pulse of the City through Shared Bicycling- Jon Froehlich, Joachim Neumann, Nuria Oliver

[3] <http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/>

[4] <http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/>