**CIS 4560 Undergraduate Class Term Paper Manuscript**

Dat Le, Steven Hoang, Thanh Tu, Rafael Reyes, James Wu

California State University Los Angeles

E-Mail: [dle31@calstatela.edu](mailto:dle31@calstatela.edu), [ttu6@calstatela.edu](mailto:ttu6@calstatela.edu), [shoang5@calstatela.edu](mailto:shoang5@calstatela.edu)

**Abstract:** For our project, we will be focusing on the dataset of New York City Citi Bike to figure out where Bikers ride, the top stations, the trip duration of each station, and so on and so forth. With this information, we can visually analyze the amount of subscribers versus customers who are using the Citi Bikes. This dataset is around 1 gigabyte in size and contains several sources of data inside itself. We will use BigInsights to load the data into clusters and extracting that data into HDFS systems. After loading this data inside the system we will be using a combination of Hive and MapReduce to simplify the dataset. The hive system should also be able to take the simplified data and export the data into excel. By loading the data into excel we can visualize the data into formats that will be easier to both see and read. This will be the process of our data analysis and how we will be visualizing the information. This information will relate from 2013 and 2014 and has information from developers, engineers, statisticians, artists, academics and other interested members of the public.

**1. Introduction**

The purpose of this analysis is to find out data on the New York biking rental service. This service is known as the Citi Bike and it is important to find out the different variables of identifications for the Citi Bikes. The types of identifications that we will be using is the most popular stations, longest trip duration, and the amount of people that use this service. We will also be looking at the data on New York taxi services and compare the amount that people use that instead of the Citi Bike service.

**2. Methods**

The way that we are able to do these types of analysis is through the utilization of both Hive and Pig in order to reduce the amount of data that needs to be interpreted. We would first need to grab the data through the usage of amazon web services databanks. Then we would need to create the tables and use the HDFS systems to put the data inside the system. After that we need to use either Hive or Pig to make the data more presentable and readable towards the readers and business professionals. We would then

take the simplified data and move it towards visualization software like Microsoft Excel’s Power

View and Tableau. With this visualization software we can make a much more detailed analysis based on this simplified data and from then on the analysis can be used for business people and for people who are most involved with Citi Bikes. The way that we are able to do these types of analysis is through the utilization of both Hive and Pig in order to reduce the amount of data that needs to be interpreted. We would first need to grab the data through the usage of amazon web services databanks. Then we would need to create the tables and use the HDFS systems to put the data inside the system. After that we need to use either Hive or Pig to make the data more presentable and readable towards the readers and business professionals. We would then take the simplified data and move it towards visualization software like Microsoft Excel’s Power View and Tableau. With this visualization software we can make a much more detailed analysis based on this simplified data and from then on the analysis can be used for business people and for people who are most involved with Citi Bikes.

**2.1 Pig Data Analysis**

In New York City, millions of commuters are racing to their destination, we wanted to compare the time spent on Citi Bikes versus vehicles. To retrieve the data, we used Pig Analysis to read the trip duration of Citi bikes and found out bikes reach their destination a lot quicker than a vehicle when trying to reach a destination in a short distance. The factors that could cause vehicles to be slower in reaching their destination is traffic which could lead to a dead stop while commuters using Citi Bikes are able to weave through the traffic.  In Figure 1, we visualized the 25 stations that shows the duration spent from a station and reaching their destination in a certain amount of seconds.

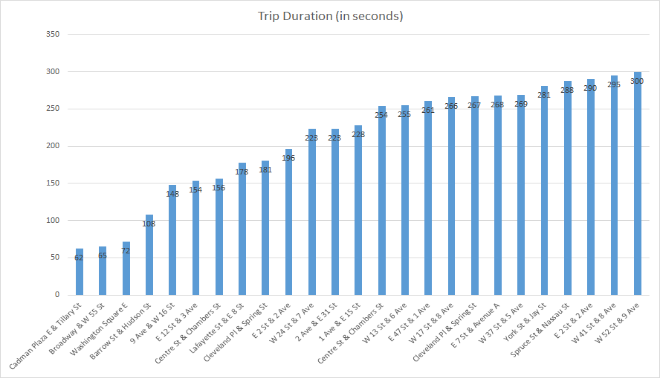


Figure 1. Trip Duration of 25 Station

Citi Bikes have a docking station where they are able to drop off the bikes or retrieve another one. Unlike vehicles where they have to find for a parking spot and having to worry about whether there is enough

time left in the parking meter. In Figure 2, we visualized all the stations from 2013-14.

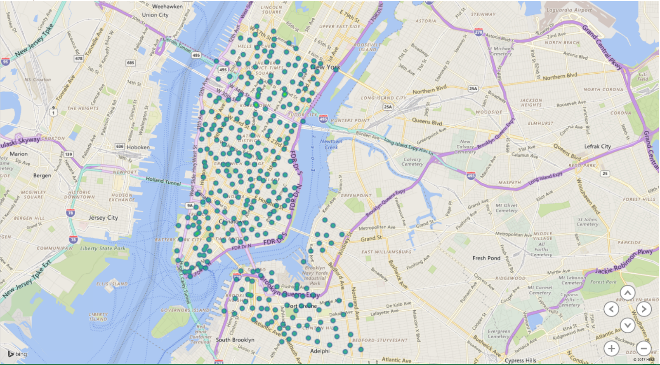
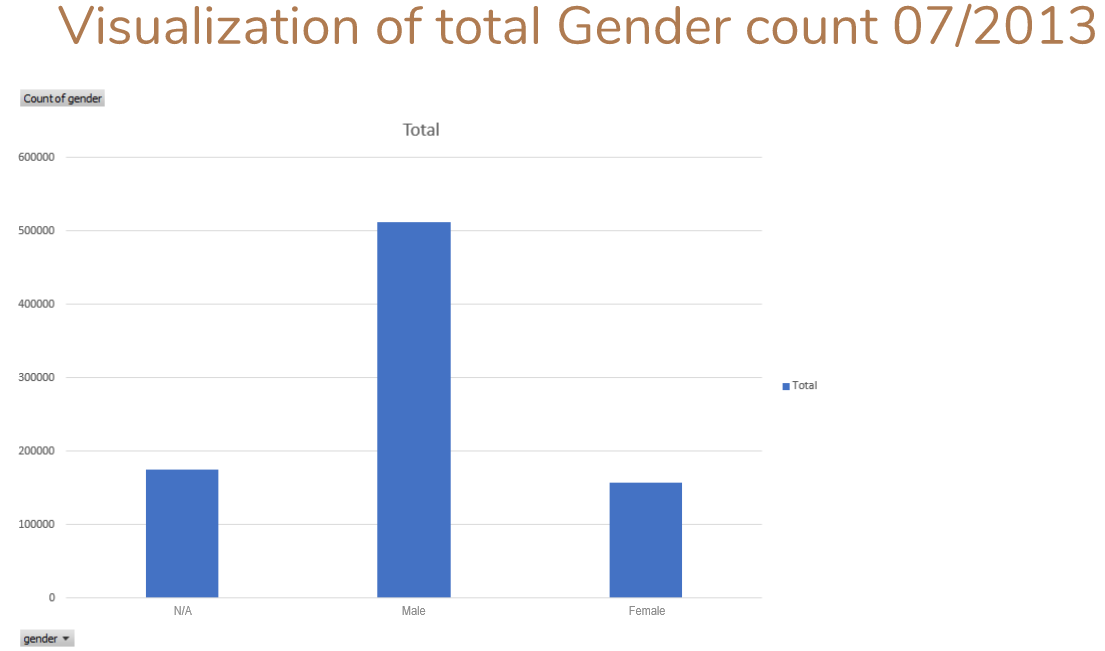


Figure 2. Geospatial Analysis

In Figure 2, the amount of circles represents the stations that are used with the Citi Bike services. The largest concentrations of circles would represent the popularity of the service and how the biking system allows consumers to move around this area in high amounts. Judging by this data we can judge that this service is very popular in New York and it still has numerous customers and subscribers to the service.

**3. Results**

Through our research and analysis, we’ve found a that there were a surprising amount of the people using Citi Bike. This is interesting because of how congested New York is when it comes to motor vehicles. Since we’ve taken a brief look at the registered active amount of vehicles for hires, it’s interesting to see that there’s many more individual users riding Citi Bike than Vehicles for Hire, as seen in Figure 3. It was also shown through our analysis that the users of Citi Bike is predominantly male as 60%-70% make up the total users of Citi Bike as seen in Figure 4. We have also found out that from the span of July 2013 and February 2014, the total users using Citi Bike have dropped significantly, which can be seen by comparing Figure 4 to Figure 5

****

