Exploratory Data Analysis of Rental Bike Service in NYC

Programming and Database Fundamentals for Data Scientists

Submitted by GROUP 13

Vineel Patnana
Rijul
Abhishek Dhyani

Project Guide
Dr. Varun Chandola

ACKNOWLEDGEMENT

We would like to thank to the course instructor Dr. Varun Chandola for letting us undertake the Project, reviewing our work throughout the process of conducting the project and for providing knowledge and material necessary for the project.

We would also like to thank SreeLekha Guggilam, the teaching assistant for taking the time to meet us whenever required and helping in resolving the queries we had throughout the course.

Sincerely,

Vineel Patnana
Rijul
Abhishek Dhyani

Abstract

To understand the behavioural patterns of rental bike service users in New York city. We aim to analyze and interpret duration of bike rides, most popular bike stations, availability of bikes during peak hours, comparison of regular customers vs one time users and other trends in usage of bikes.

Data Set Details

Data is procured from https://www.citibikenyc.com/system-data

The data includes:

- 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 3-day pass user; Subscriber = Annual Member)
- Gender (1=male; 2=female)
- Year of Birth
-

This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of the “test” stations, and any trips that were below 60 seconds in length

Procedure followed for the analysis

The data is procured using Web Scraping where python was used to procure the data and extract an combine the data into one file and pushed into the SQL server for further querying and plotting new features and understanding the data better.

The schema of the data created is as below:

Schema for the tables in MySQL

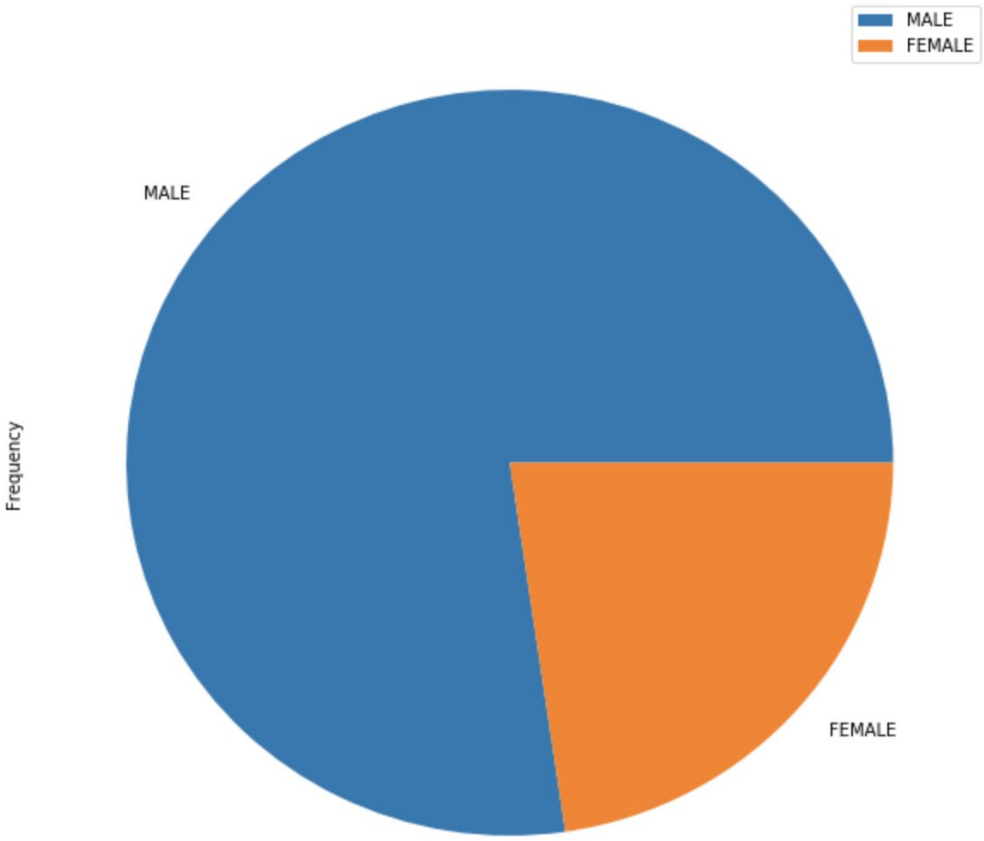
```
DROP DATABASE IF EXISTS Python_Project; CREATE DATABASE IF NOT EXISTS Python_Project; USE Python_Project; CREATE TABLE IF NOT EXISTS CITIBIKE_TRIPDATA( tripduration MEDIUMINT, starttime datetime, stoptime datetime, start_station_id MEDIUMINT, start_station_name VARCHAR(30), start_station_latitude FLOAT, start_station_longitude FLOAT, end_station_id MEDIUMINT, end_station_name VARCHAR(30), end_station_latitude FLOAT, end_station_longitude FLOAT, bikeid MEDIUMINT, usertype VARCHAR(20), birth_year VARCHAR(6), gender TINYINT );

#LOAD DATA LOCAL INFILE 'C:/Users/Vineel/Documents/UB/Python Project/Data/newJC-201703-citibike-tripdata.csv' INTO TABLE CITIBIKE_TRIPDATA FIELDS TERMINATED BY ',' ESCAPED BY '"' LINES TERMINATED BY '\n' IGNORE 1 LINES;
```

Exploratory Data Analysis

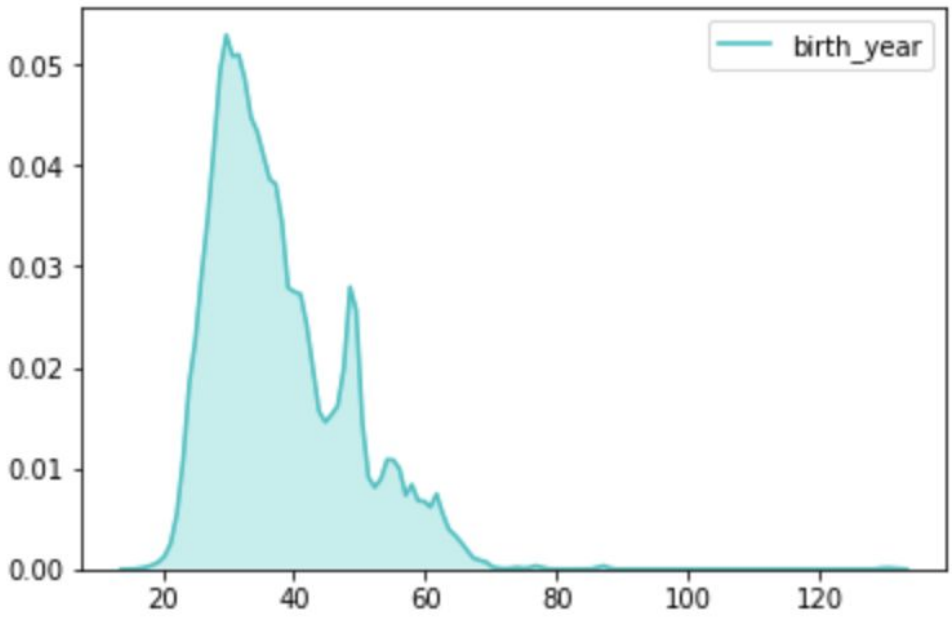
This is an Exploratory Data Analysis for the NYC Citibike data. The data has the following dimensions:

Distribution of the number of Males and Females using Citibike:



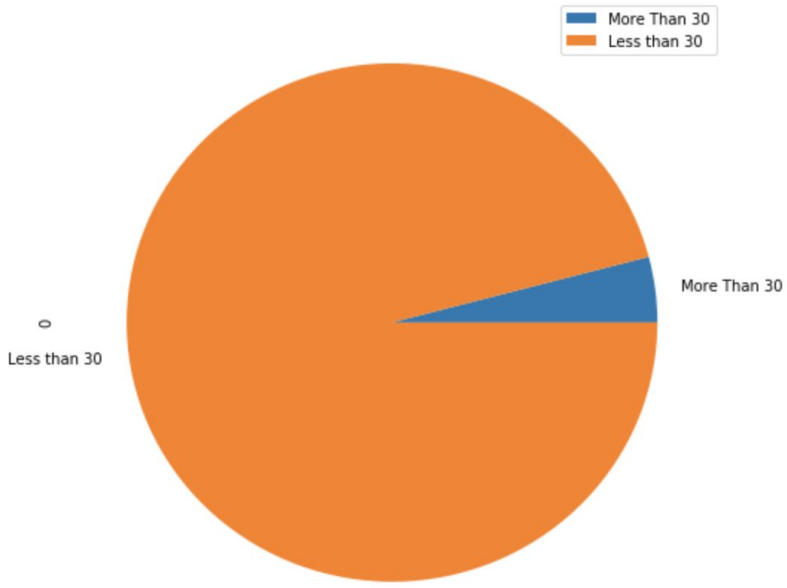
Inference: As we can see from the above graph, the males are approximately 80% of the dataset and females are around 20%. The dataset has variable “Gender” where it can take values 1 and 2. 1 corresponds to Male and 2 to Female.

Now, let us look into the details about the Age of the users in the dataset.



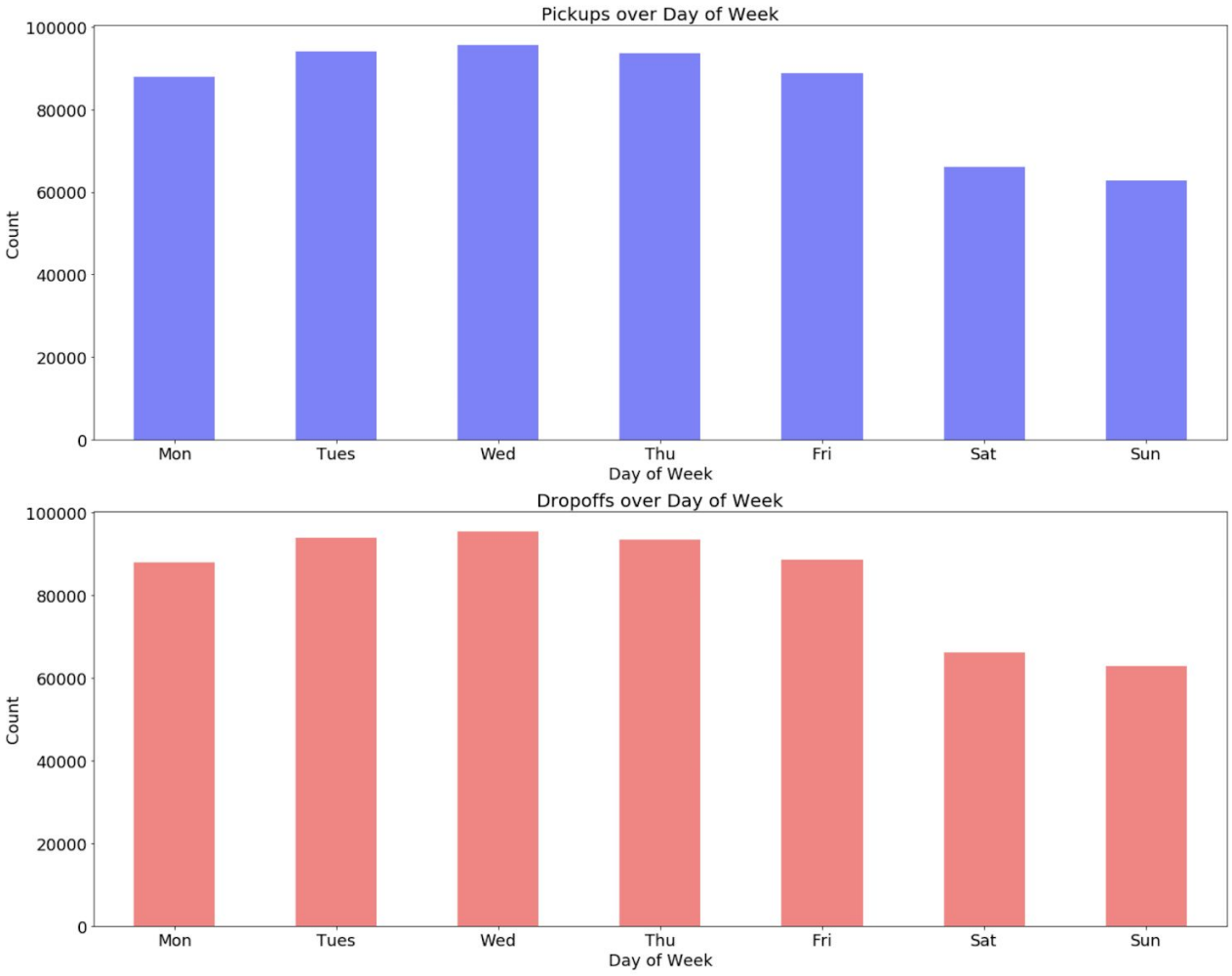
Inference: As we can see from the above graph, mostly user from the dataset lie in the age buckets of 25 to 35 years. So, it can be inferred that, if the Citibike company is planning to expand, it can look for the demographic features of the new locations where the age bucket is same as shown in the graph. Also, the age buckets 45 to 50 also show a spike.

Now, delving deeper into the trip duration of users.



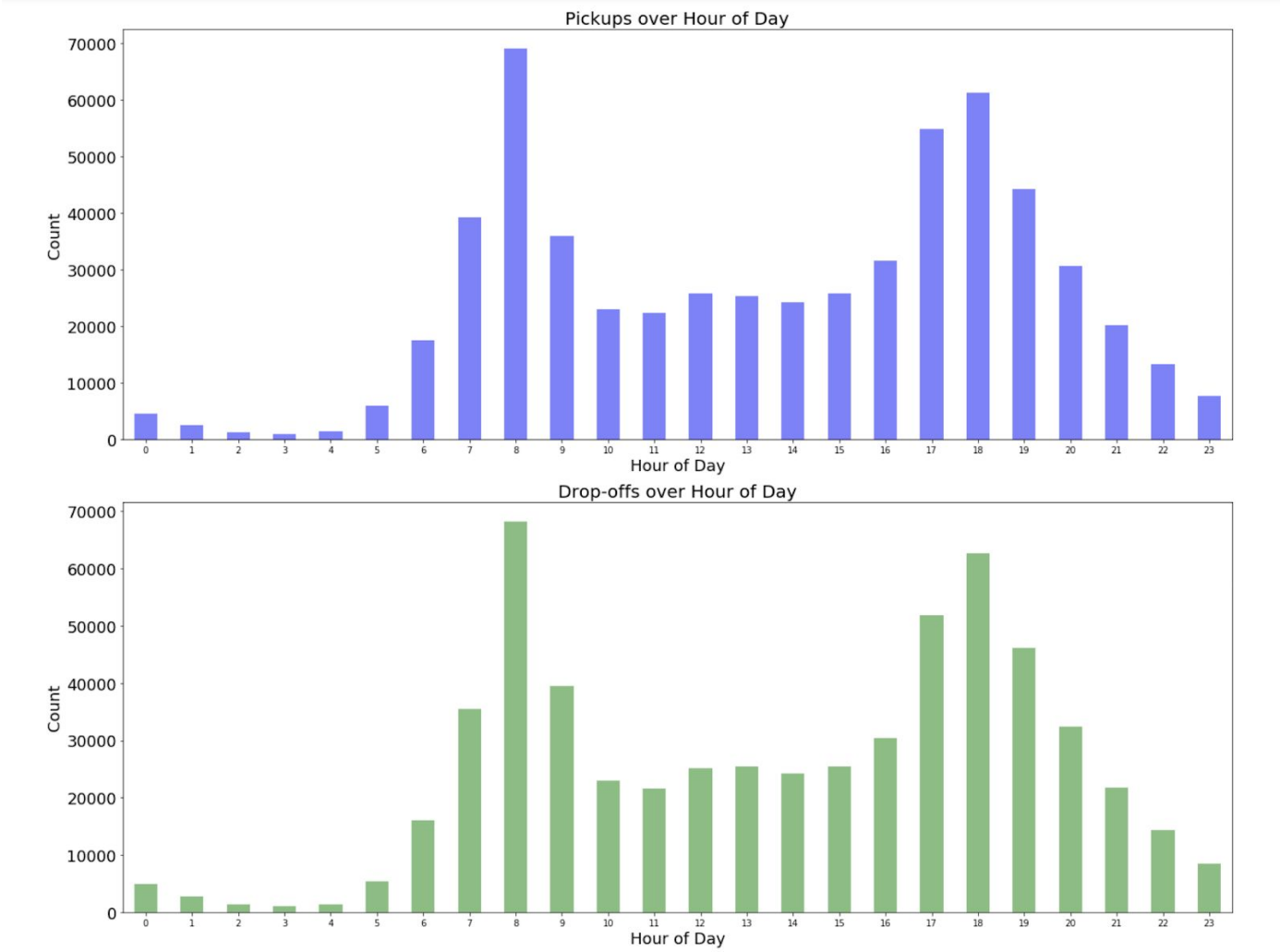
Inference: Here we made the graph for the users who use the Citibike for more than 30 minutes. As we can see from the graph, only 5% of the users travel for more than 30 minutes and 95% of them use it for less than 30 minutes. This tells us that users use the bike for a short span of time like for work etc.

The graph below is made to check the details about pickups. There were a total of 588678 pickups of Citibike for our dataset.



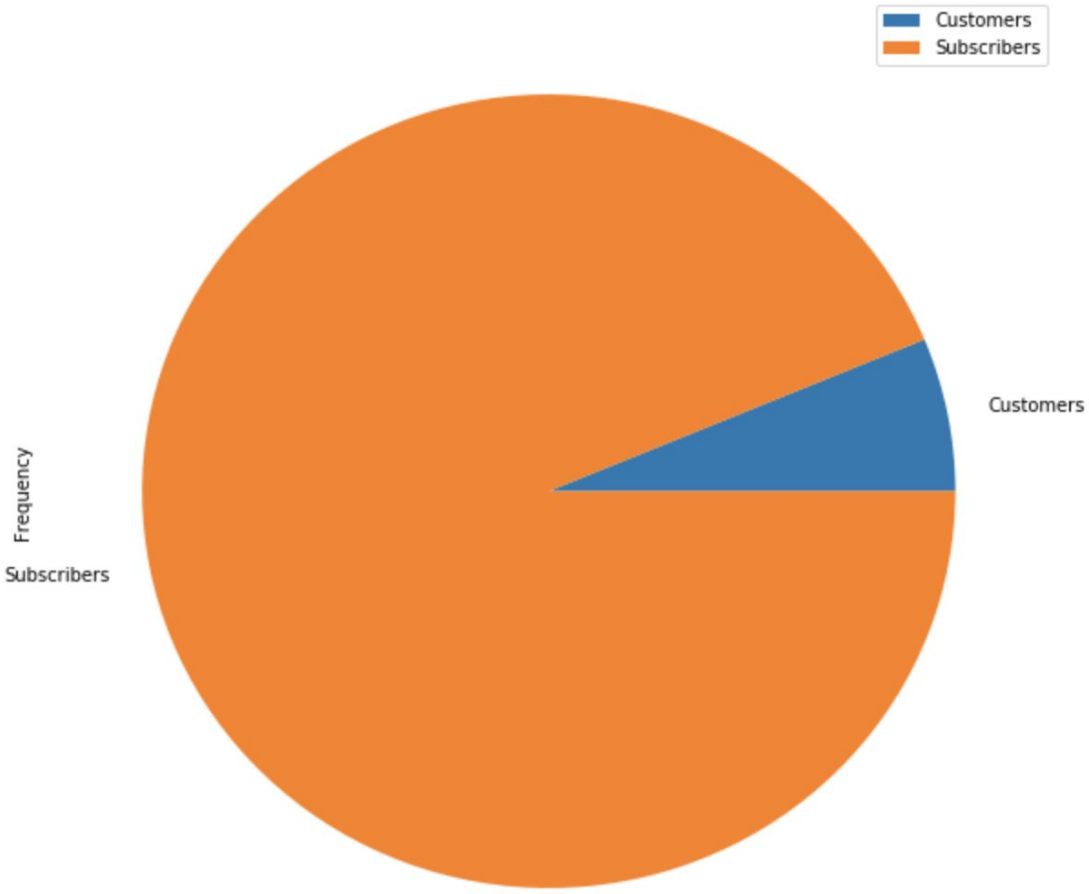
Inference: The greatest number of pickups for the Citibike were made during the weekdays and less on weekends. Here, we can infer that, these bikes were used by the people who are professionals and commute to office using Citibike. And, for the weekends, it may be used more by the travelers or tourists and the professionals may take the days off.

Now, we'll dive further into the timings of the pickups and the drop offs



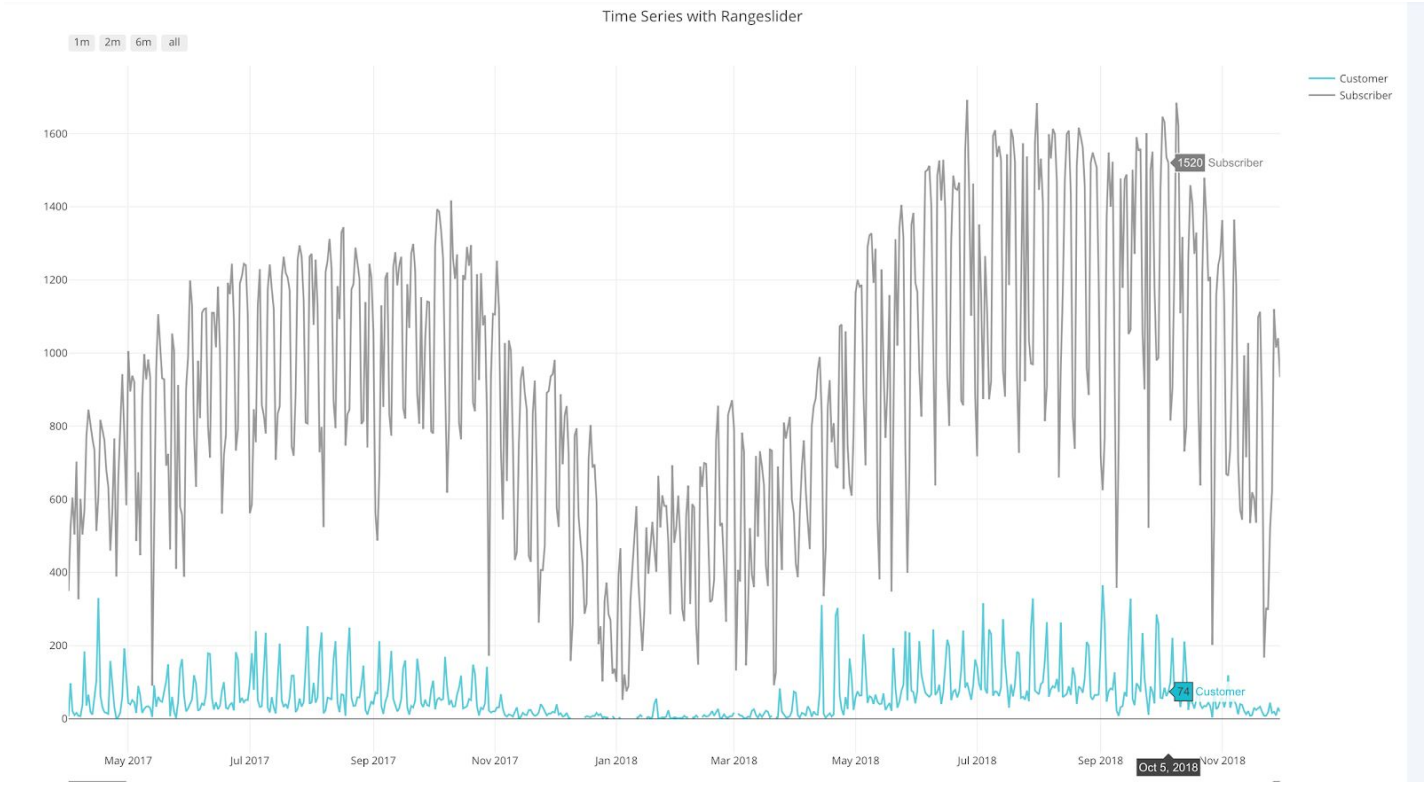
Inference: As we can see from the above figures, the greatest number of pickups were made during 8 am where most of the people commute to work at that time and we can also see that there is peak around 6 pm where the same people come back from the office at that time.

In our dataset, there are one-time customers and regular usage - subscribers. Now, we'll see the details about them.



Inference: The Users in our dataset represents One-time users and Regular customers. As, we can see from the graph that only 5 percent of them are one time users, which maybe tourists and travelers, and the 95 percent of the users are regular customers, which maybe the people who commute to work.

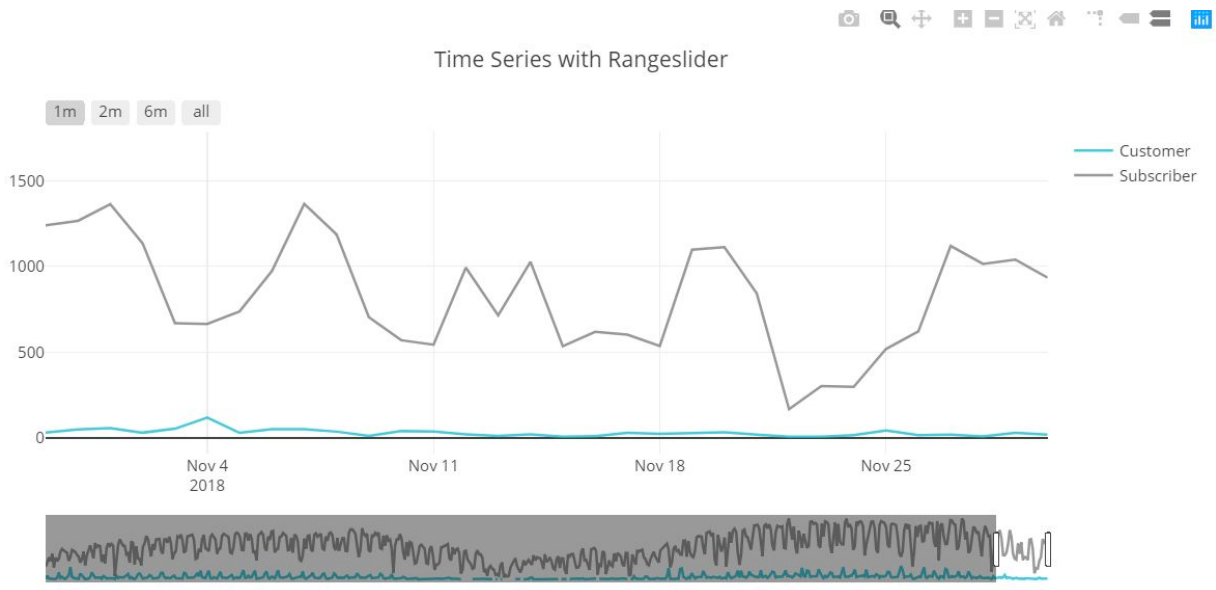
For the next bit we have made a time series plot with range slider for finding the trends in Regular Customers and One-time users.



Inference: As we can see from the above time series plot, there is a trend that, there are recurrent pattern of drops and falls in the number of users.

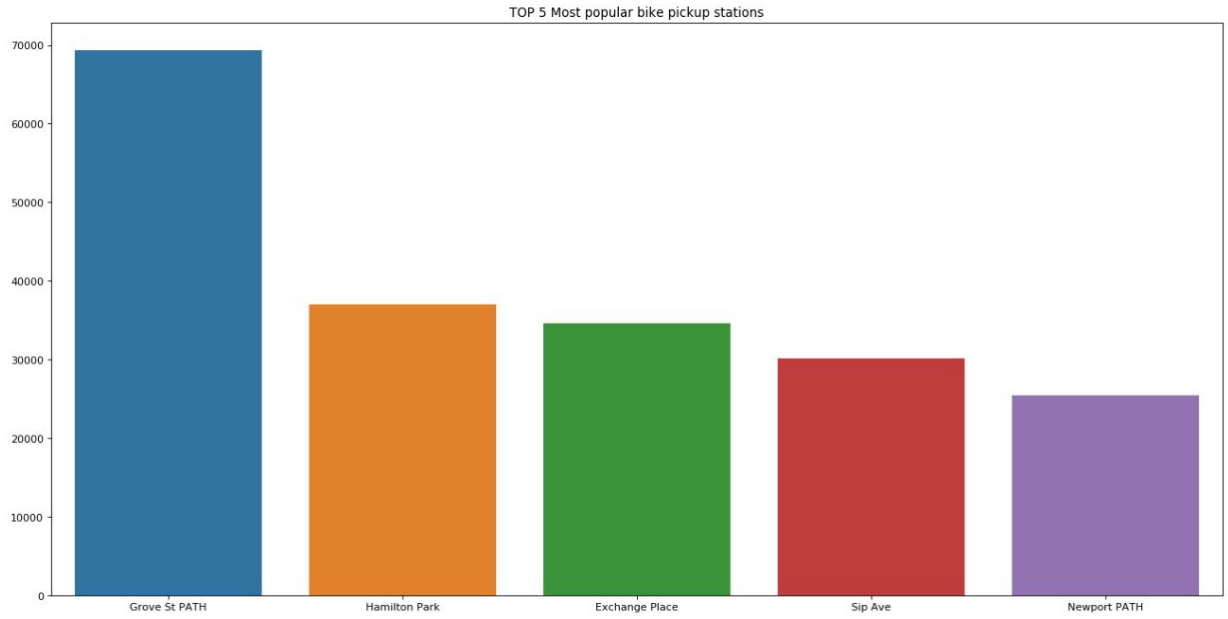
During the holidays, people often spend time at home or do not travel using the Citibike.

We can also look at the latest 1 month, 2 months and 6 month information instead of all the information by clicking on the top left buttons. Below is for the latest 1 month.



Inference: As we can see from the above time series plot for the last one month, there is a trend that, whenever, there is a dip in the graph, it is generally, weekend or some holiday. During the holidays, people often spend time at home or do not travel using the Citibike.

The top 5 most popular bike pickup stations amongst all users is Grove St. PATH.



Inference: Understand the busy trends of these stations to make sure the bike availability is better for Citibike to not lose their customers

For 1 day or 3 day pass customers it can be clearly seen that the start and the end stations are same for most of the cases.

start_station_name	start_station_latitude	start_station_longitude	end_station_name	end_station_latitude	end_station_longitude	frequency
Liberty Light Rail	40.7112	-74.0557	Liberty Light Rail	40.7112	-74.0557	1767
Newport Pkwy	40.7287	-74.0321	Newport Pkwy	40.7287	-74.0321	663
Exchange Place	40.7162	-74.0335	Exchange Place	40.7162	-74.0335	583
Marin Light Rail	40.7146	-74.0428	Marin Light Rail	40.7146	-74.0428	372
JC Medical Center	40.7165	-74.0496	JC Medical Center	40.7165	-74.0496	357
Lincoln Park	40.7246	-74.0784	Lincoln Park	40.7246	-74.0784	353
Exchange Place	40.7162	-74.0335	Newport Pkwy	40.7287	-74.0321	311
Newport PATH	40.7272	-74.0338	Newport PATH	40.7272	-74.0338	289
Newport Pkwy	40.7287	-74.0321	Exchange Place	40.7162	-74.0335	282
Morris Canal	40.7124	-74.0385	Morris Canal	40.7124	-74.0385	257

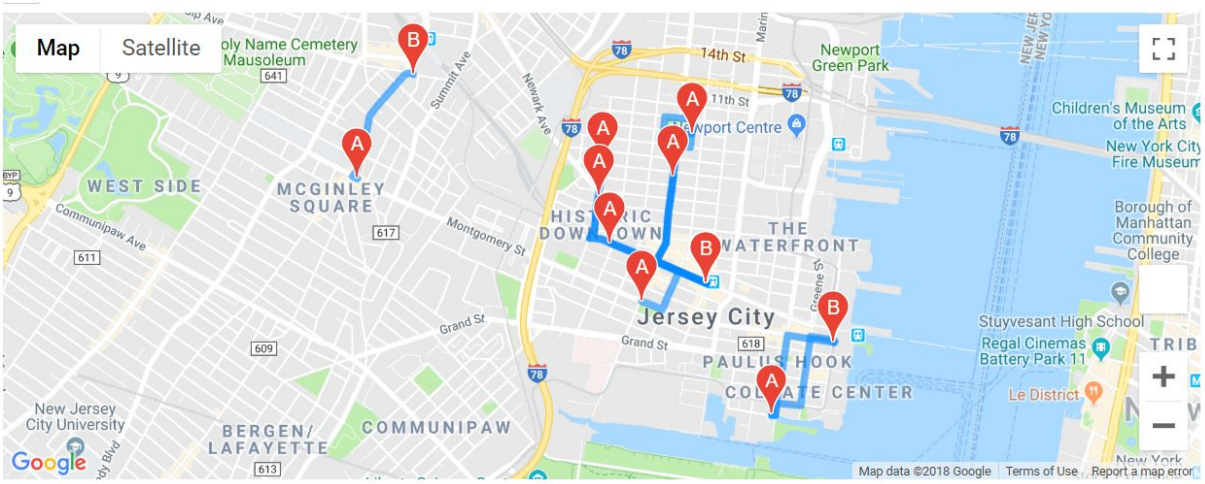
Inference: Tourists who visit New York for 1 day or 3 days tend to use the bikes and hence their pickup and drop stations are same.

For yearly subscribers, it can be clearly seen that the start and the end stations are very different.

start_station_name	start_station_latitude	start_station_longitude	end_station_name	end_station_latitude	end_station_longitude	frequency
Hamilton Park	40.7276	-74.0443	Grove St PATH	40.7196	-74.0431	14320
Grove St PATH	40.7196	-74.0431	Hamilton Park	40.7276	-74.0443	10242
Morris Canal	40.7124	-74.0385	Exchange Place	40.7162	-74.0335	9026
Brunswick St	40.7242	-74.0507	Grove St PATH	40.7196	-74.0431	7061
Jersey & 6th St	40.7253	-74.0456	Grove St PATH	40.7196	-74.0431	6701
Exchange Place	40.7162	-74.0335	Morris Canal	40.7124	-74.0385	6634
Brunswick & 6th	40.7260	-74.0504	Grove St PATH	40.7196	-74.0431	6461
Van Vorst Park	40.7185	-74.0477	Grove St PATH	40.7196	-74.0431	5810
McGinley Square	40.7253	-74.0676	Sip Ave	40.7307	-74.0638	5515
Dixon Mills	40.7216	-74.0500	Grove St PATH	40.7196	-74.0431	5436

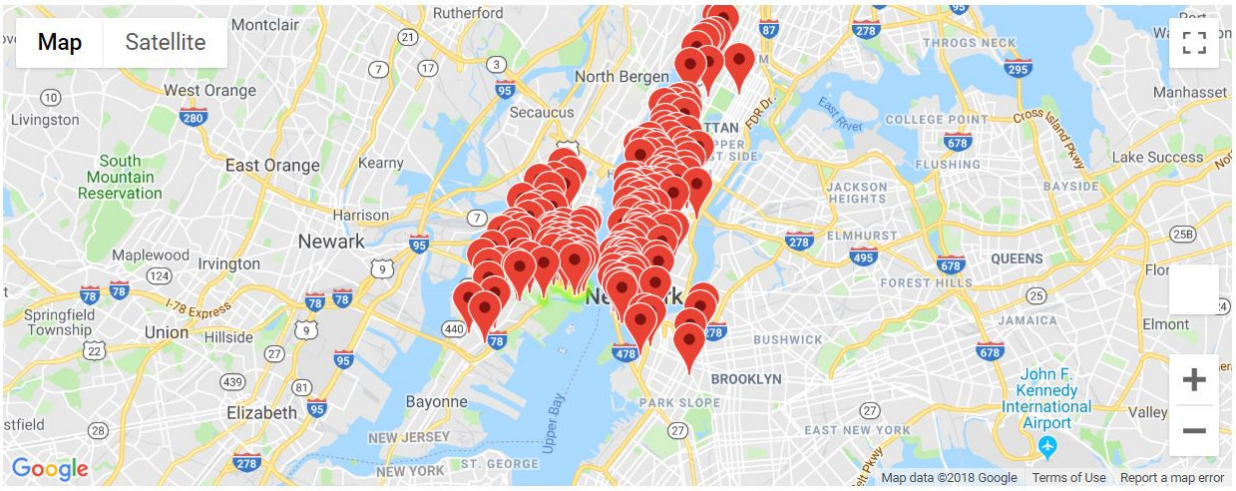
Inference: Users who use Citibike for transit to office due to New York's growing traffic use the bikes from one location to another.

To understand more about the Subscribers we try to understand the most famous routes taken by them

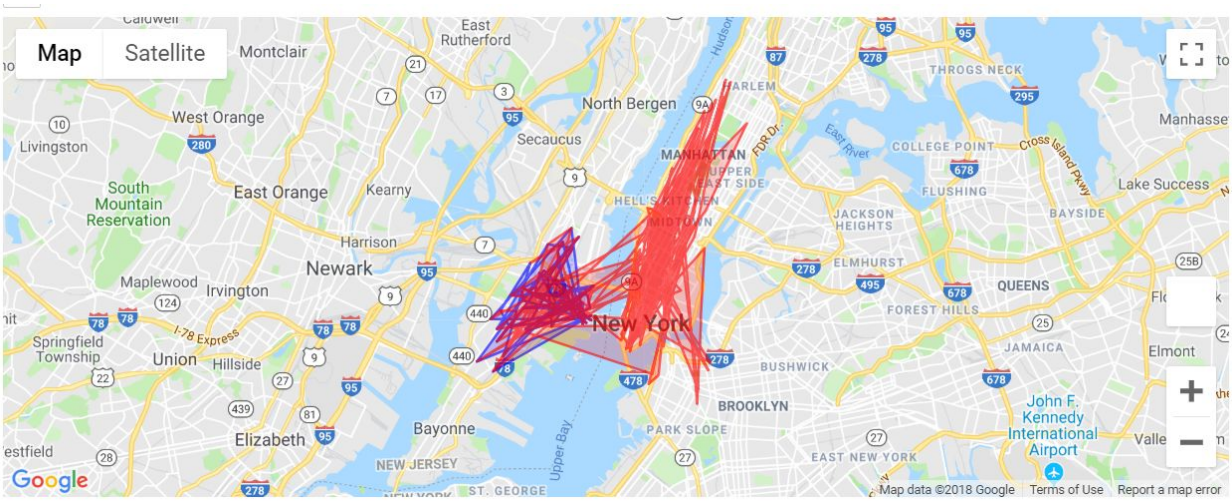


To check the pickup stations of all users to have an overall understanding of the locations of Citibike operations.

Drop stations are outside the boundary of the pickup locations



To understand more clearly we have drawn a polygon boundary where red is pickup locations and blue being drop locations



Final Verdict

CitiBike is spending extra capital in bringing the bikes from these new drop stations to the pickup stations, So CitiBike should invest in creating pickup stations where these new drop stations are present to get more profits. Since there are many drop off locations outside of New Jersey (where all of the pickup stations are present), setting up more stations in the Manhattan area will help in increasing Citibike's revenue and customer base and indirectly decrease in cost of transportation of bikes back to New Jersey.