Project Report
CitiBike Data Analysis

Fnu Karan (kx361)
Rudhika Kawar (rrk324)
Raghav Bansal (rb3526)

May 19, 2017

Instructor: Professor Rodriguez

Abstract

CitiBike is the largest bike sharing program in the Unites States, which was launched in New York City in May 2013. CitiBike data is one of the most prominent transportation system in New York City and it is used by several customers every single day.

This project examines data for years 2015 and 2016. Here we analyze the major factors behind the high ridership and how the ridership has improved over the time in these two years. We also analyze the frequency of ridership on stations, or to be precise, the traffic on the stations and then choosing the top pairs of stations. One of the greatest challenges with the Citi Bike system is expanding its customer base as well balancing the number of the bicycles that are available at the station where there is a high demand. It has become an integral part of the New York City public transportation system. We have analyzed the data with the help pf technologies like R, SparkR, PySpark and SparkML. In this report, we have addressed the steps that we performed to analyzez the data including challenges that we faced while doing that.

Acknowledgement

We owe a debt of deepest gratitude to our professor Mr. Juan C Rodriguez, Department of Computer Science for his constant guidance, support, motivation and encouragement throughout the period of this work carried out. His readiness for consultation at all times, his educative comments, his concern and assistance even with practical things have been invaluable. Mr. Juan C Rodriguez has been a great source of inspiration for us and we thank him from the bottom of our hearts. We are grateful to him for providing us with the necessary opportunities and for the completion of our project. We also thank the other staff members for their invaluable help and guidance.

Not to forget, we also owe our Teaching Assistants, Arjun Sehgal and Vishesh Kakarala, for their continuous support and helping us whenever required.

–
Fnu Karan
Rudhika Kawar
Raghav Bansal

Contents

Chapter 1

Introduction

A major feature of public bike sharing is that it allows the riders to ride the rented bicycles from one bike station to the other bike station. The urban bike sharing systems in the cities like Washington, DC, Chicago, Denver, Chicago let users to access the downtown business districts, commercial hubs by the bicycle for the daily commuters and the users. City Bike is one of the largest bike system in the world.

City Bike riders had taken more than 13.6 million trips, and more than 120,000 riders had taken the annual membership by paying $95 to become the annual members. As compared to the other bike sharing programs City Bike operates at a dense concentrated core that intertwines with the transit. Whenever there are some serious constraints with the mass transit than Citi Bike complements and supplements existing in the New York City.

Figure 1.1: CitiBike

Chapter 2

Methodology

2.1 Data Capture

Data was obtained from the CitiBike website, *https://s3.amazonaws.com/tripdata/index.html*. The Data gives the overview of where do the Citi Bikers ride? When do they ride? How far do they go?

Citi Bike has published the open source data to the people to explore further and analyze their data. The whole data has been provided according to the NYCBS Data Use Policy. The data includes the below fields:

- 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 used has been processed to remove the trips that are taken by the staff as they service and inspect the system, trips that are taken to/from any of our "test" stations (which were using more in June and July 2013), and any trips that were below 60 seconds in the length (potentially false starts or users trying re-dock a bike to ensure its secure).

We also requested weather data from NOAA (National Centers for Environmental Information). We have used the data for the year 2015 and 2016. The data was cleaned with the help of SparkR. It included the fields like STATION, STATION_NAME, DATE, PRECIPITATION, SNWD, SNOW, MAX and the MIN temperature.

We also used a data for the number of holidays for the year 2015 and 2016.

2.2 Tech Stack

- Amazon Web Services
- SparkR
- PySpark
- SparkML

- Python Matplot Lib
- RShiny
- Python Pandas

2.3 Data Cleaning

As mentioned earlier, data was taken for the years 2015 and 2016. Data cleaning involved several steps:

1. **Merging the data:** The data was obtained in multiple files, and since they had the same columns, it was pretty simple to merge them. The merging was done using SparkR.
2. **Changing column names:** All column names has certain spaces in them, which could potentially create problems for us in the future when we would have analyzed the data. Just removing them would not make any sense, so they were replaced by "_".
3. **Changing date formats:** Since there were multiple files in the data set that we obtained, we found that there are three different formats that are used for representing start date and stop date. Again, while implementing, we could have faced problems if the dates were not converted to a particular format.
4. **Removing NA values:** NA values are critical since they can affect the analysis of our data, depending upon the number of NAs that are present in the data set. In our case, NAs were present only in the column, birth_year, and the number of NAs were pretty significant. So, in order to remove them, we chose to replace these values with the average of other values rather than just removing those rows.
5. **Changing gender values:** The gender column in data contained the values as 0,1,2 and we replaced them with respective values. (Zero=unknown; 1=male; 2=female)
6. All the holidays were marked as one and non-holidays were marked as zero. It was having the fields date and the occasion describing the occasion for the holiday.

2.4 Data Analytics

Since we have obtained the complete data in section 2.3, we can go ahead and start performing some analytics on the data. Some of the major analytics that we performed are:

1. Initial loading of data and loading libraries
2. Number of rides per day for the duration of two years
3. Total trip duration per day for the duration of two years
4. Change of subscribers over months for the duration of two years
5. Change of customers over months for the duration of two years
6. Number of rides per month per hour
7. Number of rides per day of the week per month
8. Dashboard to observe when and where people ride the most. This also included the most popular stations, which could be considered as the stations/areas that are generating maximum revenue for the CitiBike.
9. A time series showing the number of rides for each month in both the years, 2015 and 2016.
10. A machine learning model, using SparkML which predicts the number of rides on a particular day considering the weather parameters like average temperature, minimum temperature, maximum temperature, snow and precipitation. Other than these parameters, we also included the holiday data as well, that could impact the number of rides on a day.

    We have studied the effect of the maximum and minimum temperature on the trips being taken. The data has been requested from the NOAA (National Oceanic and Atmospheric Administration). The model was prepared using the parameters: "TAVG", "TMIN", "TMAX", "SNOW", "PRCP".

Chapter 3

Results and Discussion

3.1 Number of rides per day for the duration of two years

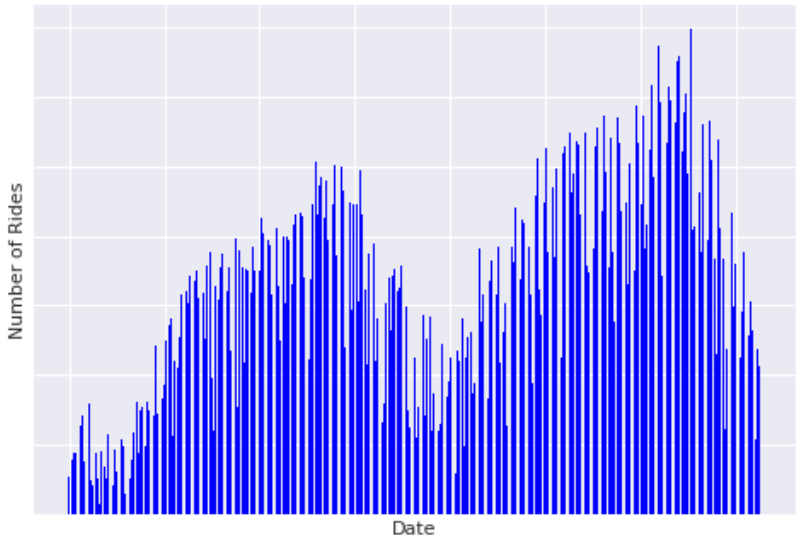

Figure 3.1: Number of rides per day for the duration of two years

As shown in the above figure, the rides from 2015 to 2016 increase significantly. The ride count goes down, during the months of November, December, January and February for both the years.

3.2 Total trip duration per day for the duration of two years

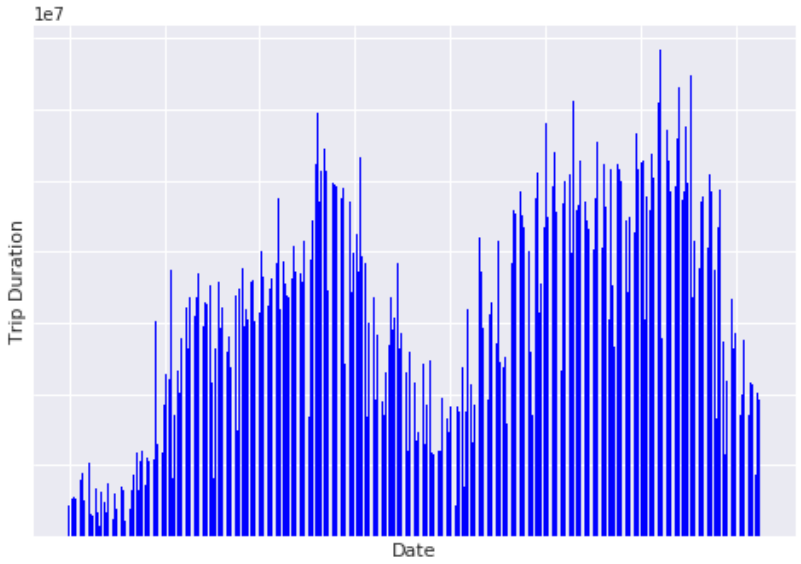

Figure 3.2: Total trip duration per day for the duration of two years

This figure also follows the same kind of pattern as the previous one as well. But, here we talk about the total duration of the rides for each day over the years.

3.3 Change of subscribers over months for the duration of two years

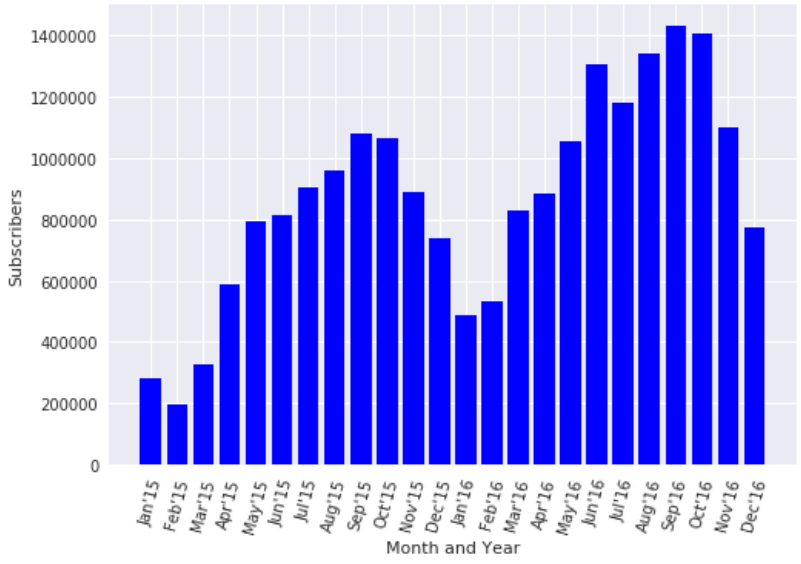

Figure 3.3: Change of subscribers over months for the duration of two years

The interesting thing to note here is that, the subscriber count reaches to a higher level in 2016, which is resonating with the previous results that we obtained. It can be inferred that the number of subscribers begin to drop during the winters as riding on Citi Bike becomes difficult as result extreme weather conditions in the winters.

3.4 Change of customers over months for the duration of two years

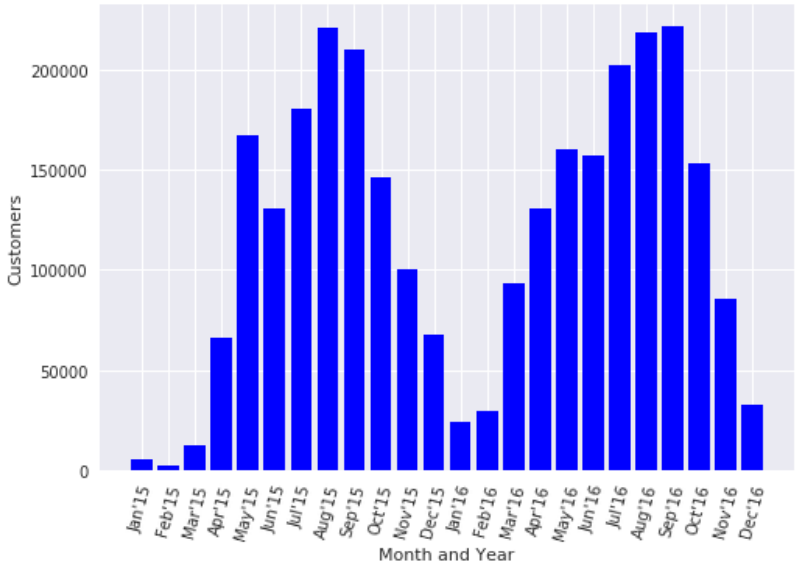

Figure 3.4: Change of customers over months for the duration of two years

This is a plot for the customers and here we can see that the regular customer follows the same trend in both the years. The rides for respective month here remains the same in both years. Concluding from the above 4 observations, we can come to a conclusion that the subscribers are the only ones that actually affect the ridership rather than that of the regular customer.

3.5 Number of rides per month per hour

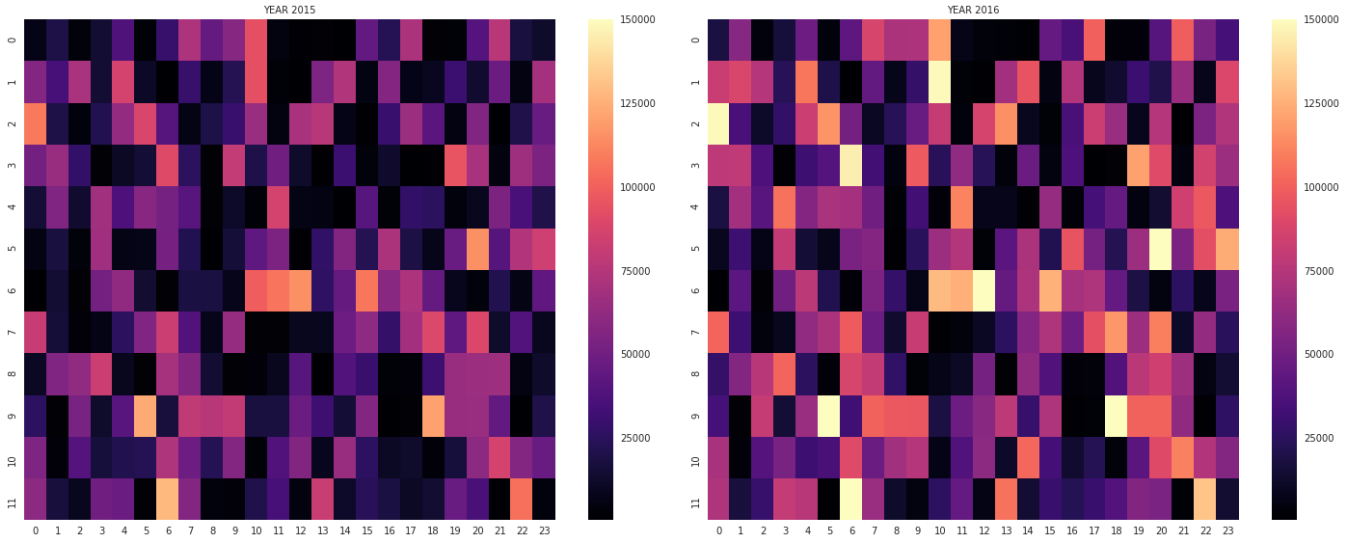

Figure 3.5: Number of rides per month per hour

Here, the numbers on X-axis denote the time or the hour of the day (0-23) and on Y-axis represent the month of the year. (0-January, 11-December).

3.6 Number of rides per day of the week per month

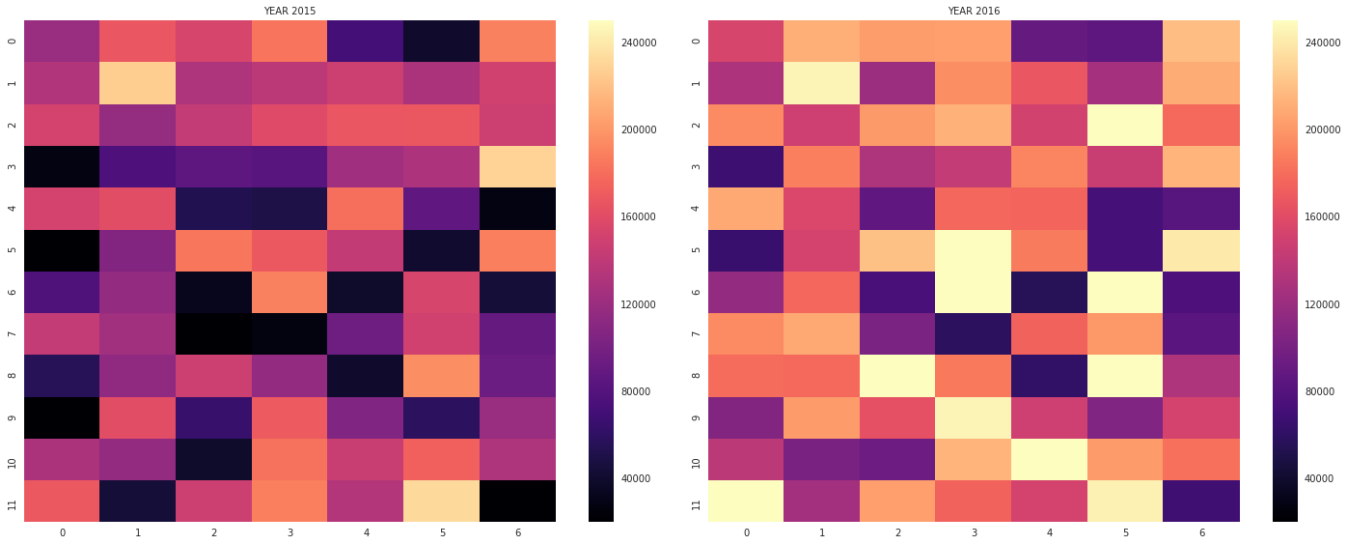

Figure 3.6: Number of rides per day of the week per month

Here, the numbers on X-axis represent the day of the week, zero=Monday and 6=Sunday, and Y-axis represent the month, same as the previous one.

3.7 RShiny Dasboard

Dashboard to observe when and where people ride the most. This also included the most popular stations, which could be considered as the stations/areas that are generating maximum revenue for the CitiBike.

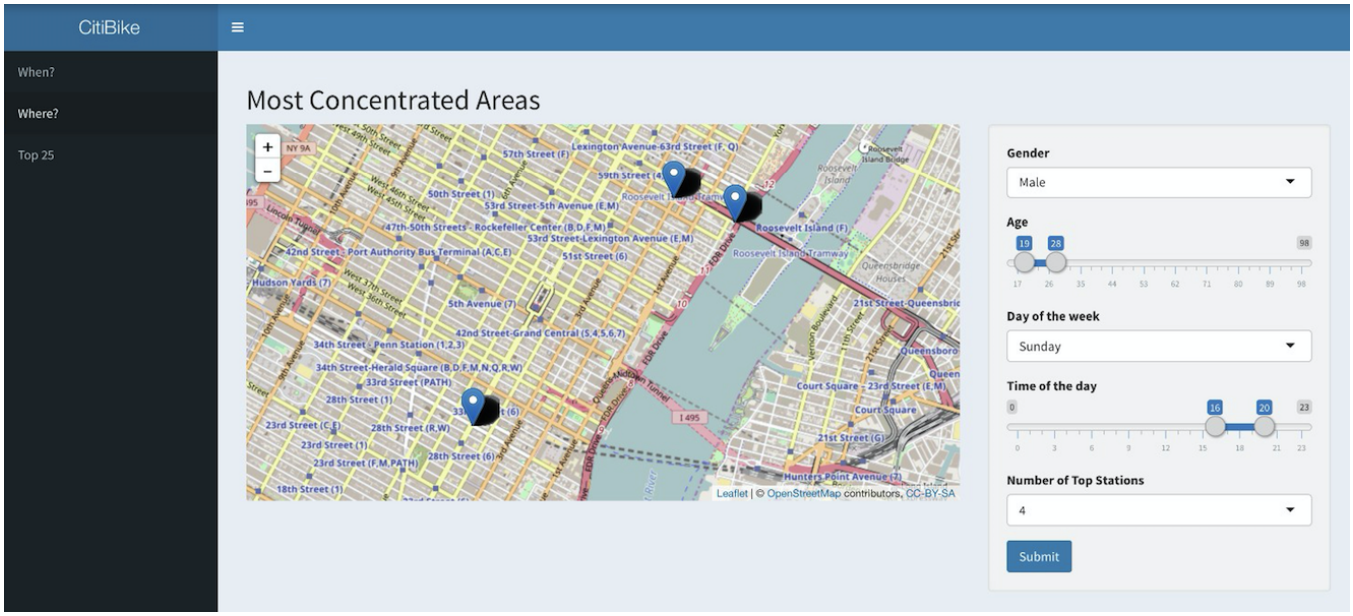

Figure 3.7: RShiny:Top stations from where people take rides

Figure 3.7 shows the most popular stations among the riders, and they cam be filtered using several parameters like gender, age, day of the week and time of the day. The results can also be filtered out for the number of most popular stations to be shown..

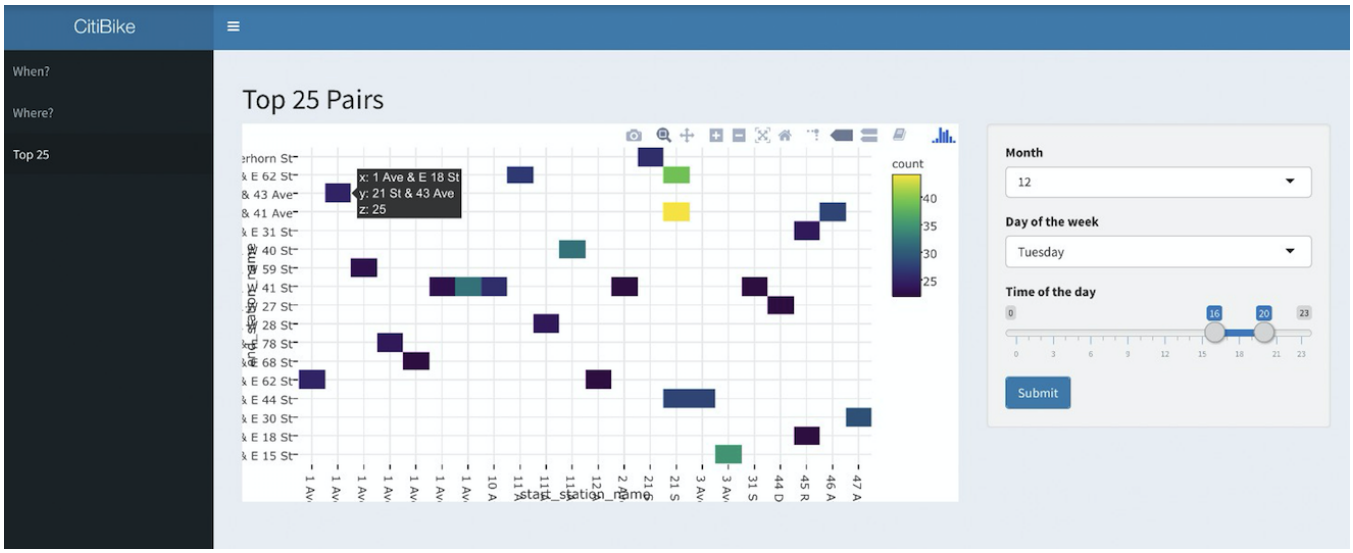

Figure 3.8: RShiny:Top 25 pair of stations generating maximum revenue for CitiBike

Figure 3.8 shows the top 25 pairs of stations that generate the maximum revenue for CitiBike. The results can be filtered using the parameters like month, day of the week and time of the day.

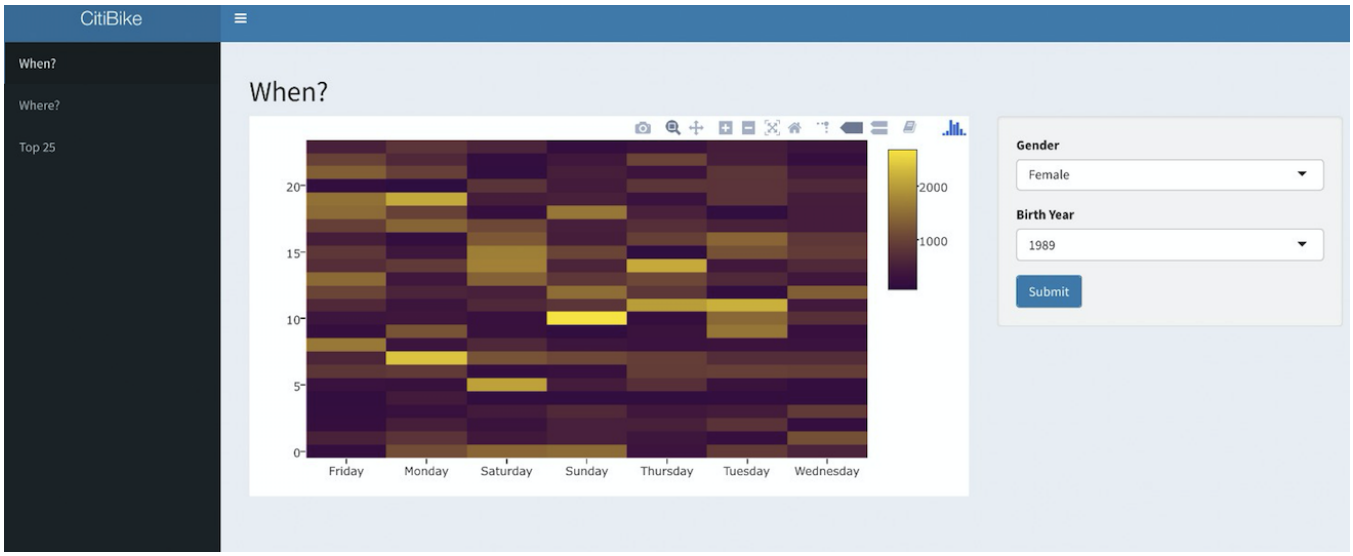

Figure 3.9: RShiny:When can people ride based on gender and birth year

Figure 3.9 shows the most busy hours of the day of a typical weekday and the these results can be filtered using gender and year of birth.

3.8 A time series showing the number of rides for each month in both the years, 2015 and 2016.

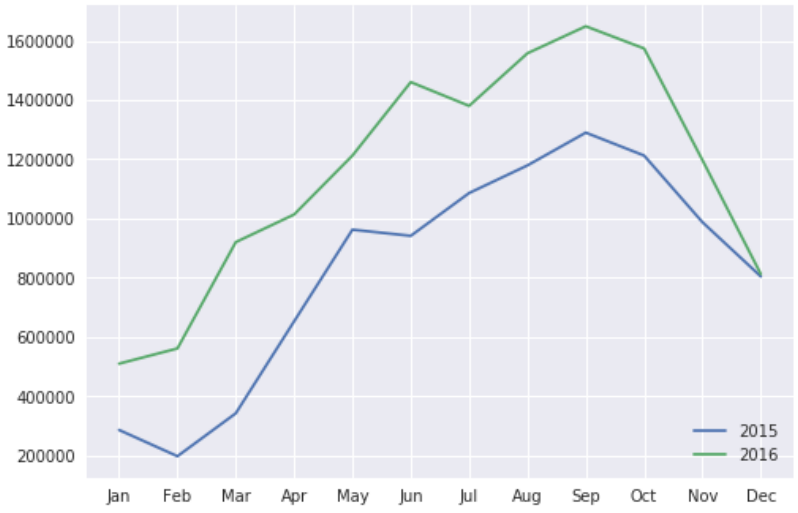

Figure 3.10: Time Series

Figure 3.10 shows the time series for the years 2015 and 2016. Showing the number of rides for each month of these years. As it is clearly evident from the graph, that CitiBIke has definitely grown over the years, and the significant improvement in the ridership is in the months of June to October. But during the other months, there is definitely an increase, but it is not significant. By looking at these results here, we chose to go ahead the predict the number of rides that can be there on a day.

3.9 ML Model

A machine learning model, using SparkML which predicts the number of rides on a particular day considering the weather parameters like average temperature, minimum temperature, maximum temperature, snow and precipitation. Other than these parameters, we also included the holiday data as well, that could impact the number of rides on a day. The results are shown in figure 3.11

We have studied the effect of the maximum and minimum temperature on the trips being taken. The data has been requested from the NOAA (National Oceanic and Atmospheric Administration).

The above four plots show us the counts of trips made when there is a certain maximum, minimum and average temperature. The plot is showing us the rides with the Maximum Temperature. We see a rise in rides when the temperature is around 70 and then we have another rise when the temperature is around 80-85.

Upon seeing the minimum temperature, of the day we see that there is an increase in the nuber of rides, where the min. temp is around 62-70 degrees. This would typically indicate a pleasant/mildly hot day.

We also tried to observe a measure of TAVG or Average Temperature. This was basically to measure what the entire day was like. This is being taken as an average measure to see whether it was a cold, hot or pleasant day in general.

The results of the ML Model are shown in the table are shown in the Table 3.1.

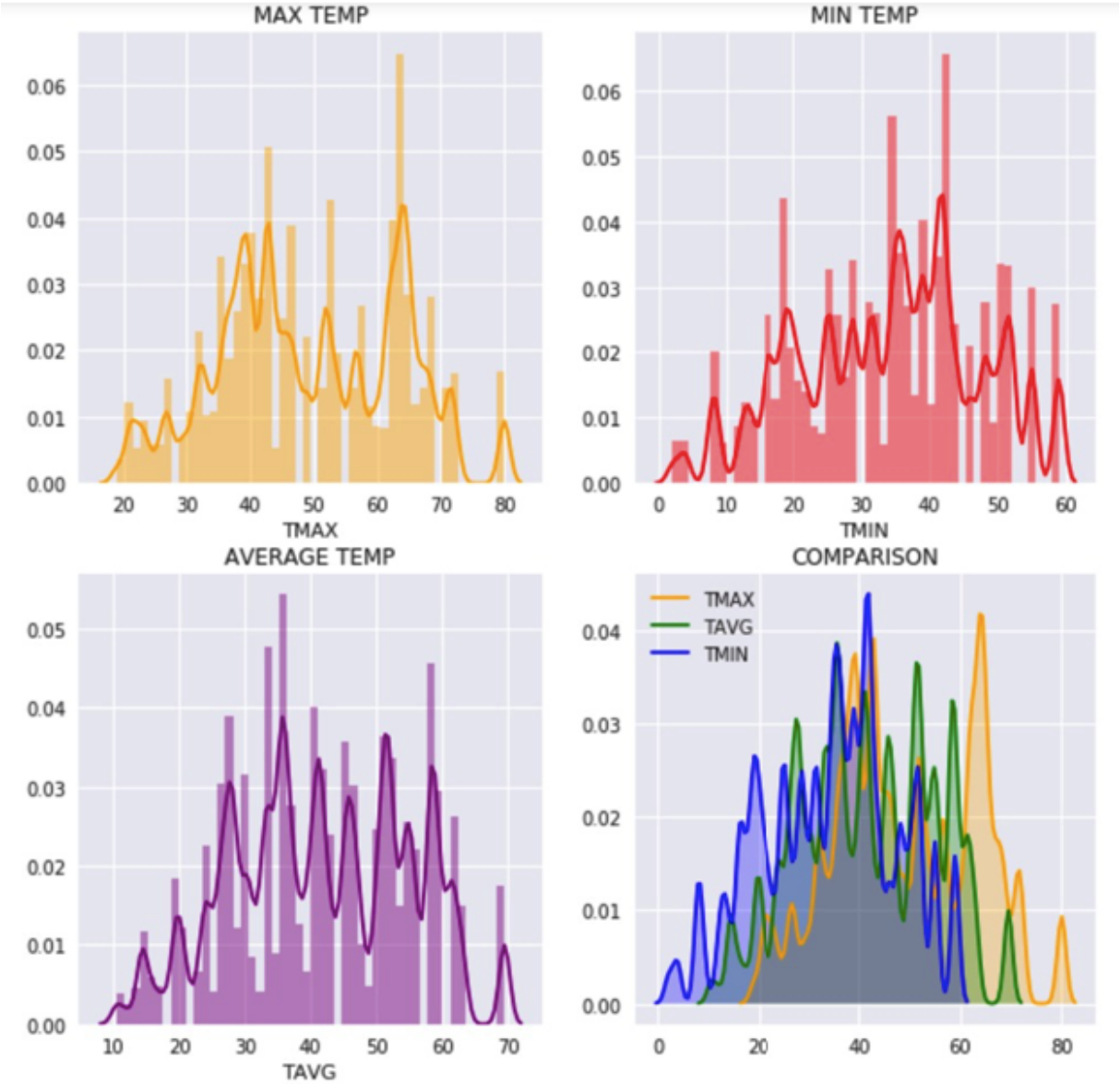

Figure 3.11: Effect of Maximum and Minimum Temperature on Trips

Actual Rides	Predicted Rides
33895	30869.51
16787	16310.11
21475	26949.60
29795	29904.25
37677	35613.07
9338	13642.56
5160	6108.01
32294	34279.91
30313	29568.02
6441	8301.75

Table 3.1: Results for the ML Model

Chapter 4

Code and References

4.1 Code

The code is present in the folders 'pyspark' and 'r'.

4.2 References

- https://www.citibikenyc.com/system-data
- http://www.noaa.gov/
- http://timeanddate.com
- http://spark.apache.org/docs/latest/sparkr.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.ml.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.mllib.html
- http://stackoverflow.com/