IST 652

Team 11

Pranali Shenvi
pshenvi@syr.edu

Srushti Samant
ssamant@syr.edu

Mikhail Pinto
mpinto01@syr.edu

Final Project Report

Analysis of NYC Citi Bike data

Submission Date: 12/07/2022

Total Pages: 13

Index:

Introduction:

- NYC is the most populated city in the USA with over 8.8 million people
- In 2013, a shared bicycle system known as Citi Bike has been available
- The benefits include reducing dependence on automobiles and encouraging public health through exercise
- The system has been expanding each year, with increases in the number of bicycles available
- The usage data provides a wealth of information which can be mined
- With such intelligence, the company would be better positioned to determine what actions might optimize its revenue stream

Data Gathering:

Making sure we are using the appropriate and proper amount of data was crucial because our project uses real-time generated data. To be sure we had a compiled dataset, we have taken a few steps. First, we took data files for the months of February through September in 2022. We got a total of 4.5M rows of this combined data for these months. After that, we sampled the combined data and prepared a dataset of 1,123,046 rows of data which is approximately 1.1M rows of data.

Goals:

Our goal is to explore the Citi Bike data and find insights to monitor the demand of Citi Bikes and analyze the trends. Some of the questions that we would like to answer are:

1. What are the most probable locations to take Citi Bike from?
2. What is the daily, monthly demand for Citi Bikes?
3. How is the Citi Bike trend on Weekends and Weekdays?
4. What is the average trip duration of Citi Bike rides?

Dataset description and attributes:

Sr. No.	Column Name	Description
1	Index	Index number
2	ride_id	Ride ID number
3	rideable_type	Type of bike (Electric, Classic, Docked)
4	started_at	Time and date the ride started
5	ended_at	Time and date the ride ended
6	start_station_name	Start station name from where the ride started
7	start_station_id	Start station ID from where the ride started
8	end_station_name	End station name from where the ride ended
9	end_station_id	End station name from where the ride ended
10	start_lat	The latitude of the bike from when the ride started
11	start_lng	The longitude of the bike from where the ride started
12	end_lat	The latitude of the bike from when the ride ended
13	end_lng	The longitude of the bike from where the ride ended
14	member_casual	Member ride or Casual ride

Official Citi Bike website: https://ride.citibikenyc.com/system-data

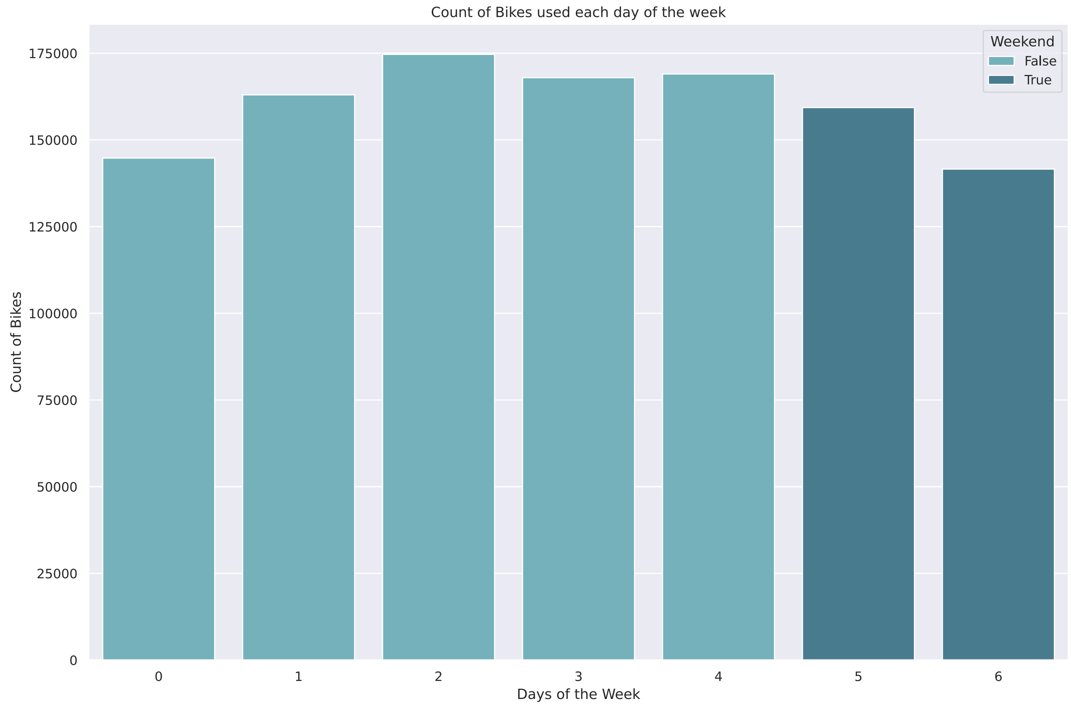
Data ranges from February 2022 to September 2022

The dataset consists of 1,123,046 rows and 14 columns

Data Cleaning and Transformation:

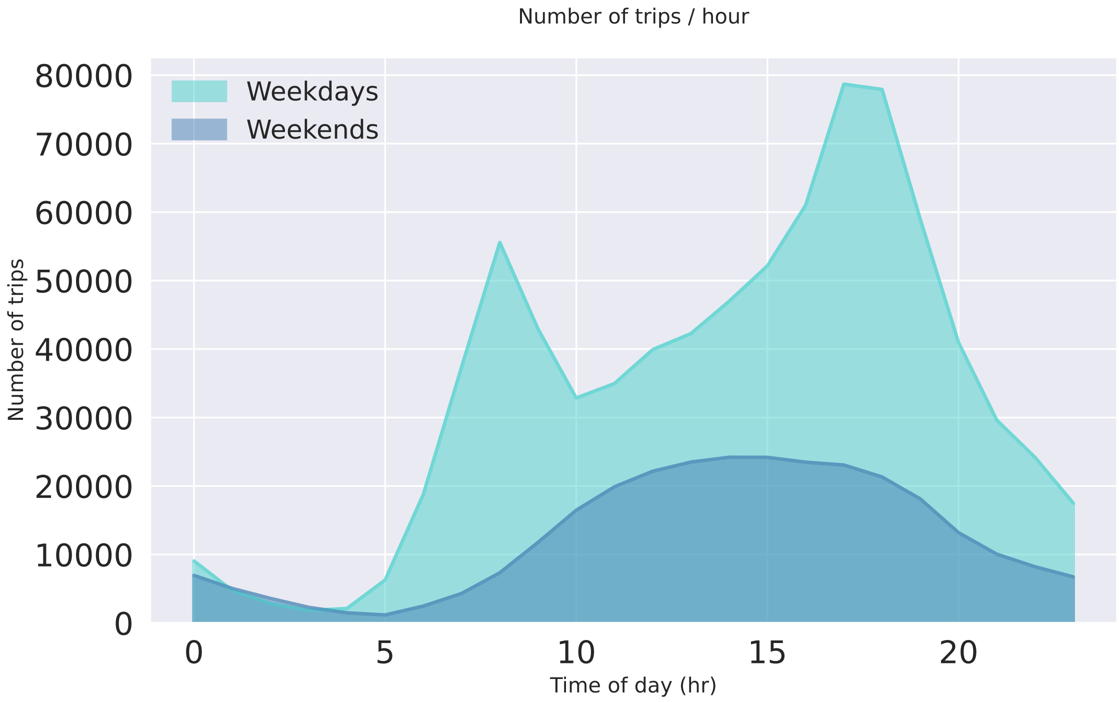
- After the data gathering stage, we pre-processed the data and performed data cleaning for the smooth analysis of the results
- We removed the null values. These values belonged to four columns - end_station_name, end_station_id, end_lat and end_lng. From this, we can say that the Citi Bike rides were not completed for these rides or probably there were issues in registering the end time for these rides. We removed a total of 2258 rows were removed
- Duplicate records from the data have been removed
- We have created four more columns using the time and date column which show the year, day, month, and quarter the ride was started. We were also able to classify the days into weekends and weekdays for our analysis

Exploratory Data Analysis:



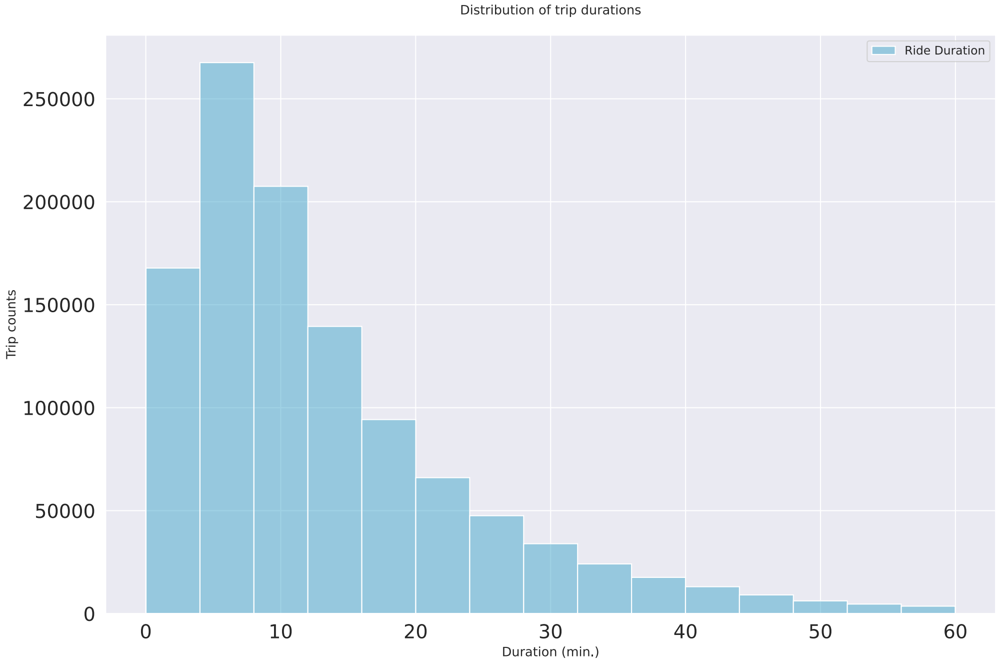
Count of Citi Bikes used each day of the week

In the above graph we have plotted days in the form of 0-6 where 0 stands for Monday and 5,6 stand for Saturday and Sunday respectively. We can see that there is not much difference between the count of bikes for weekends and weekdays. Our analysis is that the ones who use Citi Bikes on the weekends are travelling alone, the ones who go out as a family, will not be using Citi Bikes for their travel. On the other hand, the ones who use Citi Bikes on weekdays are the ones who are going to their workplace or office commute. This can further be seen in another way. On Monday, the count of Citi Bikes used are comparatively less than the other days. After a weekend, it is unlikely that people will commute to their workplace using bikes, they will travel by private automobiles or other modes of transport.



Number of trips per hour of the day

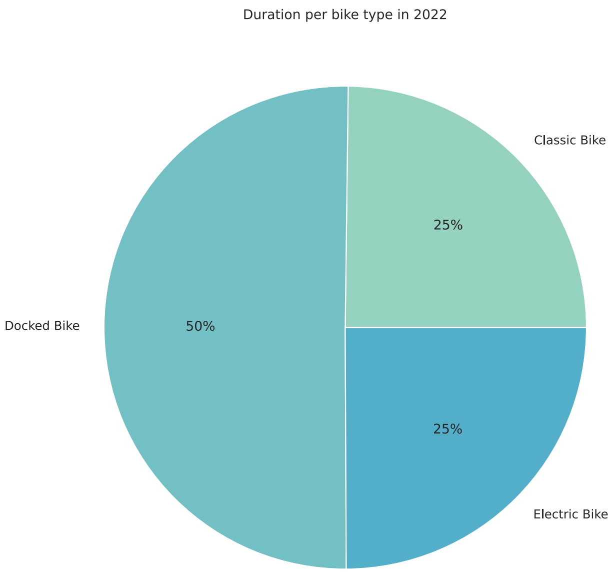
We have plotted the graph of number of trips on weekends and weekdays on different hours of the day. We can see that the number of trips on weekdays are more compared to weekends. Also, it can be seen that the graph peaks on time duration of 8 am and 6 pm for weekdays. This means that on weekdays this time is the peak time for people to use the Citi Bikes. This can mean than these are the people travelling to their workplace and leaving from there after the day. For weekends, on the other hand, we can see that there is no exact peak value available. The maximum number of trips is between the time frame of 12 pm to 6 pm. This can mean, that on weekends people tend to go out for recreational purpose during this timeframe.



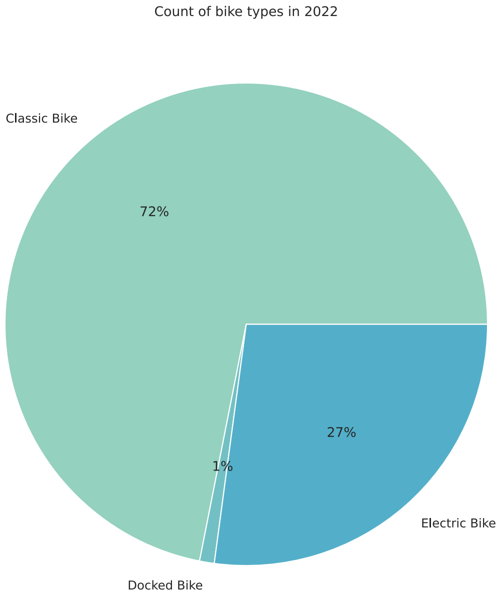
Distribution of the trip durations with respect to the trip counts

We can see that the graph above is for the number of trips compared to the duration of the trip in minutes. The highest number of rides have a trip duration of about 2.5 – 7.5 minutes. The highest count of rides is about 260k. It is a right skewed graph.

We can say that most of the rides are for short duration and the count of ride decreases as the time duration increases.

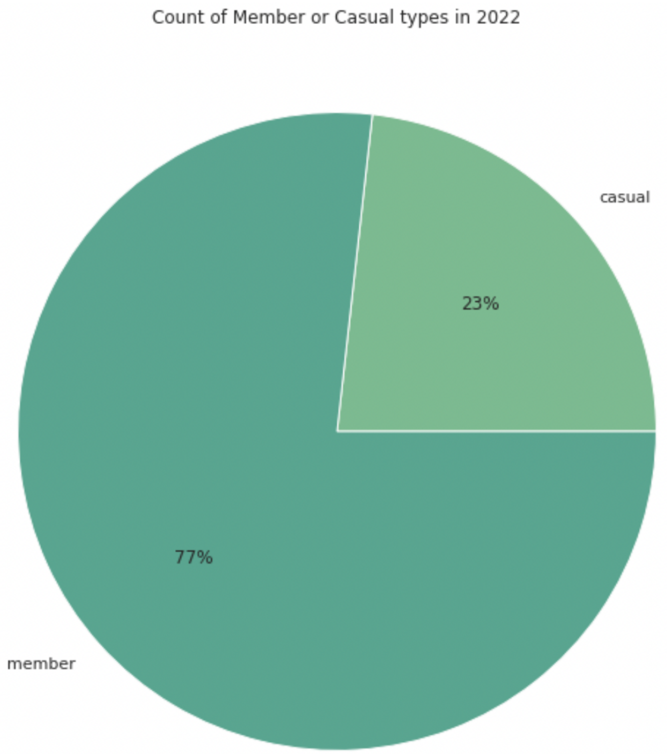


Percentage of duration of ride per bike types in 2022



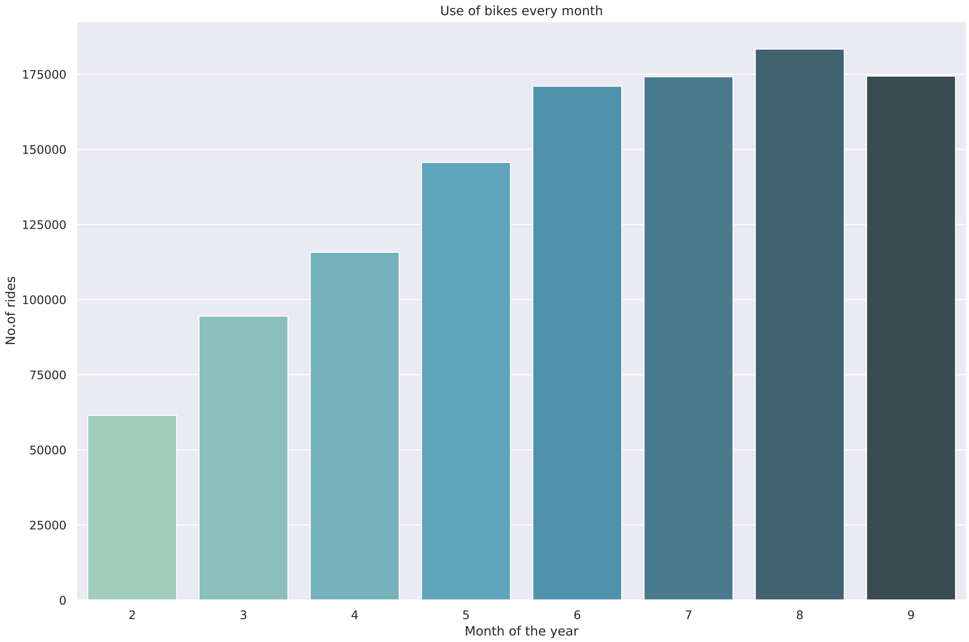
Percentage of count of bike types in 2022

The above two graphs complement each other. We can see that the Docked bikes have the highest percentage of trip duration which is about 50% followed by the Classic and Electric bikes which are 25% each. Out of all the rides, 72% of the bikes are classic bikes. The least number of bikes used are Docked bikes which is around 1%. We can see that the Docked bikes are less in count as it is very expensive to maintain the docks needed for these bikes. But we can also see that these Docked bikes have the highest percentage of duration of time they are used. We can conclude that if the company adds a greater number of Docked bikes, they will definitely be used more as the highest time duration of rides have been utilized by the Docked bikes.



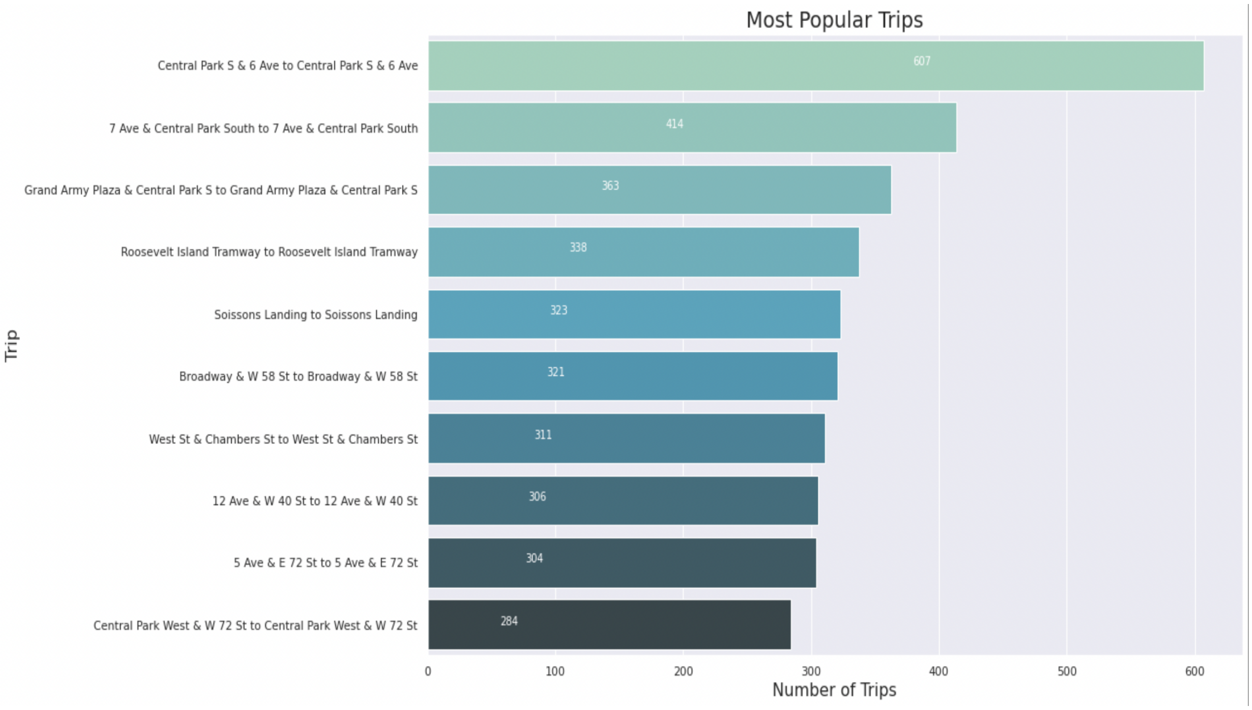
Percentage of the count of Member or Casual types in 2022

From the above graph we can see that the greater number of percentages of bike riders are Members compared to Casual type.



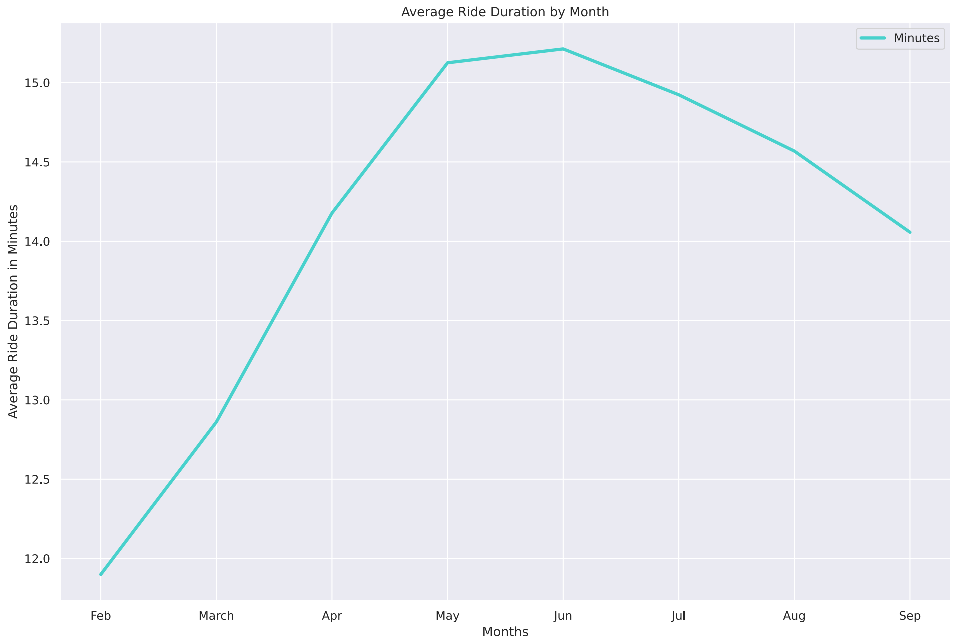
Count of bikes used each month in the year 2022

From the above graph we can see that the maximum count of the rides can be seen in the month of August and the minimum rides can be seen in the month of February. The months where it is colder or snows the most has few rides counts compared to the hotter months. We have used the months column that we created from the time and date column. The maximum number of rides is around 188k.



The 10 most popular trip locations and the number of trips

We can see that the 10 most popular rides are displayed. These maximum number of trips are around 600 for the most popular trip. We can observe that these trips are round trips, which is that the start and end locations of these trips is the same. We can say that these are recreational trips where people start and end at the same place. Also, we can see that these locations are around Central Park which means that many people come around this area for bike ride exercising.



Average ride duration by month in 2022

We can see the average ride duration by month in the above graph. We can see that the maximum duration in minutes is 15 minutes, and it is in the month of May and June. Whereas, the months February and September have low duration rides.

During the cleaning of the data, we also added a filter to remove those rides which are above 3 hours in duration. We have removed them as it is ideally not possible to continuously have a ride for that long and we kept the threshold as 3 hours. We considered the others as outliers and removed them.

Insights:

- There is not much difference in the count of rides on weekends and weekdays
- The bikes from stations where the number of bikes used are less, can be removed and added to the stations which are busy

- This will make sure there are enough bikes at the busy stations and also prevent extra supply of bikes
- In the months of winter, the number of trips is less compared to the summer months
- The average duration of trips is around 15 minutes

Future Work:

- Use Google API to track the location of the stations using latitude and longitude to get a visual on the map, it will be easier to get to know the locations and make better analysis
- We can also calculate the distance using the latitudes and find and predict the ride distances in the future, again, we will need to implement Google API for the same as we cannot use the straight distance as it is of not much use
- Combine the dataset with data related to weather to find out more information
- Adding data of different years to dive deeper

Conclusion:

- Right placement of the number bikes in the correct location will be very helpful to take care of the high demand of the bikes
- Member subscription fees if any can be manipulated to get more subscribers
- The type of bikes used can be changed according to the demand