Radhika Balasubramaniam

Rice University Data Analytics Bootcamp

TableAu- Citibike Analysis

Introduction 3

ETL Process 3

Analysis 3

Growth Trends 3

Ridership Growth Trends 3

Short Term Vs Subscriber Trends 3

Ridership Trends and Patterns 4

Monthly Analysis 4

Weekday vs Weekend Ridership Analysis 4

Peak hour Trends and Patterns 4

Peak Hour Trends 4

Peak Hour analysis for Summer and Winter 4

Top and Bottom stations Report 4

Station Analysis 4

Ridership by Demographics 4

Ridership by Age 4

Ridership by Gender 4

Bike Analysis 5

Average Miles by Bike 5

Bikes Most likely due for Repair or Inspection 5

Station Map 5

# Introduction

[New York Citi Bike](https://en.wikipedia.org/wiki/Citi_Bike) Program, is the largest bike sharing program in the United States. Since 2013, the Citi Bike Program has implemented a robust infrastructure for collecting data on the program's utilization. Through the team's efforts, each month bike data is collected, organized, and made public on the [Citi Bike Data](https://www.citibikenyc.com/system-data) webpage. However, while the data has been regularly updated, the team has yet to implement a dashboard or sophisticated reporting process. City officials have several questions on the program, and they want to build a set of data reports to provide the answers.

The following are some questions they wish to tackle.

* How many trips have been recorded total during the chosen period?
* By what percentage has total ridership grown?
* How has the proportion of short-term customers and annual subscribers changed?
* What are the peak hours in which bikes are used during summer months?
* What are the peak hours in which bikes are used during winter months?
* Today, what are the top 10 stations in the city for starting a journey? (Based on data, why do you hypothesize these are the top locations?)
* Today, what are the top 10 stations in the city for ending a journey? (Based on data, why?)
* Today, what are the bottom 10 stations in the city for starting a journey? (Based on data, why?)
* Today, what are the bottom 10 stations in the city for ending a journey (Based on data, why?)
* Today, what is the gender breakdown of active participants (Male v. Female)?
* How effective has gender outreach been in increasing female ridership over the timespan?
* How does the average trip duration change by age?
* What is the average distance in miles that a bike is ridden?
* Which bikes (by ID) are most likely due for repair or inspection in the timespan?
* How variable is the utilization by bike ID?

City officials would also like to see a dynamic map that shows how each station's popularity changes over time (by month and year). Again, with zip code data overlaid on the map.

# ETL Process

The data was downloaded from the [Citi Bike Data](https://www.citibikenyc.com/system-data" \t "_blank) webpage. The data was unzipped and using a python script, the data was cleaned and stored as single csv file. The records from 2016 until 2020 was used for the analysis.

# Analysis

For analysis and building the dashboard, over 3M records were analyzed. The data was analyzed from various viewpoints. A story board with 8 dashboards was created in tableau. The analysis is described below.

## Growth Trends

The first focus was to analyze the general growth trends for the Citi bike program. The average growth and comparison chart of short-term vs Subscriber chart was created as a part of analysis.

### Ridership Growth Trends

It was identified that the total ridership grew over the years. The percentage of growth compared to previous year went down each year. But the overall trend showed a growth for the past 5 years.

### Short Term Vs Subscriber Trends

Rider subscription has grown over the last 5 years. The short-term riders also grew in small proportions. One interesting observation: In 2020 due to COVID-19 situation the rider subscription went down by almost 40%. But the short-term customers almost tripled during 2020.

## Ridership Trends and Patterns

The next focus was to analyze the general ridership trends and patterns for the Citi bike program. The monthly analysis graphs, by year and comparison chart of Weekend vs Weekday chart was created as a part of analysis.

### Monthly Analysis

The Ridership monthly analysis graph shows the ridership typically increases from March and stays high during spring, summer, and fall.

The ridership is highest for the months of August and September and October.

The trends remained constant over the last five years. However, In 2020, due to COVID-19 situation, the ridership fell during April and May compared to previous years.

### Weekday vs Weekend Ridership Analysis

The graph shows the ridership is at least 20% higher on the weekdays compared to weekends.

## Peak hour Trends and Patterns

The next focus was to analyze the peak hour trends and patterns for the Citi bike program. The peak hour trend graphs, by year and comparison chart of Summer vs Winter peak hour charts was created as a part of analysis.

### Peak Hour Trends

The ridership remained high between 7 AM and 9 AM and it peaked at 8:00 AM in the morning and between 5:00 PM and 7:00 PM in the evening. The trends remained the same when further drilled down through the years.

### Peak Hour analysis for Summer and Winter

The peak hour riders remained high between 7:00 AM and 8:00 AM in the morning for both the summer and Winter months. However, the number of riders fell short by 50% during the winter.

## Top and Bottom stations Report

The next focus was to identify the top and bottom stations for the Citi bike program. The top 10 popular start and end stations as well as bottom 10 start and end stations were created as a part of analysis.

### Station Analysis

Grove Saint Path station has highest ridership. Both the start and end point are higher for the station.

The top 10 starting point stations has on an average total rides between 150K and 300K whereas the bottom starting point stations has on an average the total rides lesser than 5K.

The top 10 end stations also have on average total rides between 5K and 150K, whereas on the bottom it is less than 2 rides.

## Ridership by Demographics

The next focus was to analyze the demographic influence for the Citi bike program. Ridership by age is compared with Average trip times for different age group. Also gender based analysis chart were created as a part of analysis.

### Ridership by Age

Riders between the age group 20-30 are highest user group and they are consistent across the years. They are followed by the riders in the age group 30-40.

### Ridership by Gender

Males riders are using bikes (almost 2.5 times) more compared to Female riders. The average ride times are same for both males and female.

The promotions to improve female riders seems to have little influence as the growth is not significant.

## Bike Analysis

The next focus was to analyze the bike usage statistics for the Citi bike program. Average bike usage was calculated. Also identified the bikes that may require repair or Inspection.

### Average Miles by Bike

Average trip duration ranges from 400 Miles to 1400 Miles.

### Bikes Most likely due for Repair or Inspection

Based on the distance travelled in the last quarter, if above 5000 Miles they are due for inspection or repair

## Station Map

The last part of the analysis was to identify the popular stations and the average rides over the years for the Citi bike program. A dynamic map that shows how each station's popularity changes over time (by month and year). Again, with zip code data overlaid on the map was generated.