

# Citibike Trip Analysis for Jan 2020

Group 12 (Theertha Pillai, Nikhil Phatale, Ketan Kshirsagar)



## Table of Contents

1. Purpose
2. Background
3. Data Analysis
  - 3.1. Demographic Analysis
  - 3.2. Time Series Analysis
  - 3.3. Geographic Analysis
  - 3.4. Year-Over-Year Analysis
4. Conclusion
5. References

## 1. Purpose

This report analyses City Bike's data from New York city, for the month of January 2020. We have attempted to analyse the data from various perspectives such as Demography, Time-Series, Geography, etc. To do so, we have created visualizations by using different types of plots such as Leaflets, Network Diagrams, Line Graph, Bar Chart, Donut Graph, Scatter Plot and Pi Chart. Some of these visualizations are dynamic, helping the reader interact with data and fetch valuable insights from the same. Additionally, we have also provided our inferences and suggestions based on our comprehension about the data.

## 2. Background

Citi Bike is the USA's largest and privately operated public bicycle sharing program serving the New York City boroughs of the Bronx, Brooklyn, Manhattan, and Queens, as well as Jersey City, New Jersey, and Hoboken, New Jersey. The dataset used here for analyzing has been retrieved from Citi Bike's website which contains real time data for the bicycle service.

From this data, we are trying to inspect the data from different viewpoints and discover patterns and inferences from it. This report has touched upon various aspects of data and answers questions like; Who are the customers of Citi Bike?, How does factors such as Time, Weather and Day affect the sales? Which geographic location is more profitable than others?, How has business been doing for the past three years?, etc.

Moreover, we have taken data from Jan 2020 to 2022 to understand the company's business during pre-pandemic, pandemic and post-pandemic situation.

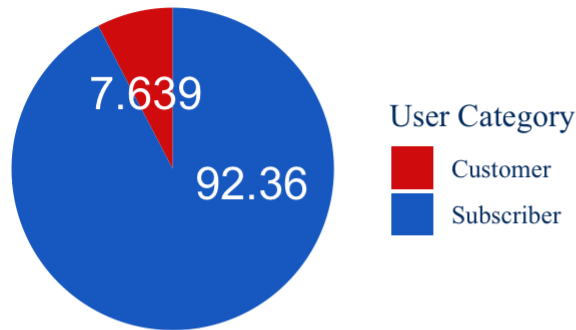
## 3. Data Analysis

The Data Analysis is done in four sections i.e. Demographic, Time-Series, Geographic and Year-Over-Year Analysis

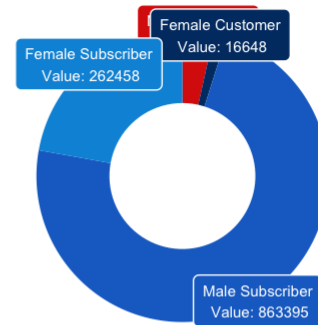
### 3.1. Demographic Analysis

To Analyse the demographical trends in data for Jan 2020, we have plotted the following graphs. The Pi chart shows consumer segments based on User Type i.e. Customers and Subscribers. Customers are one time users who can rent a Citi Bike from a rental location without a membership. Subscribers, as the name suggests, holds membership which can be annual, monthly or casual. The donut chart provides us a gender-based segregation of customers and subscribers. The stacked-bar graph provides an overview of customer based on their age-groups and gender.

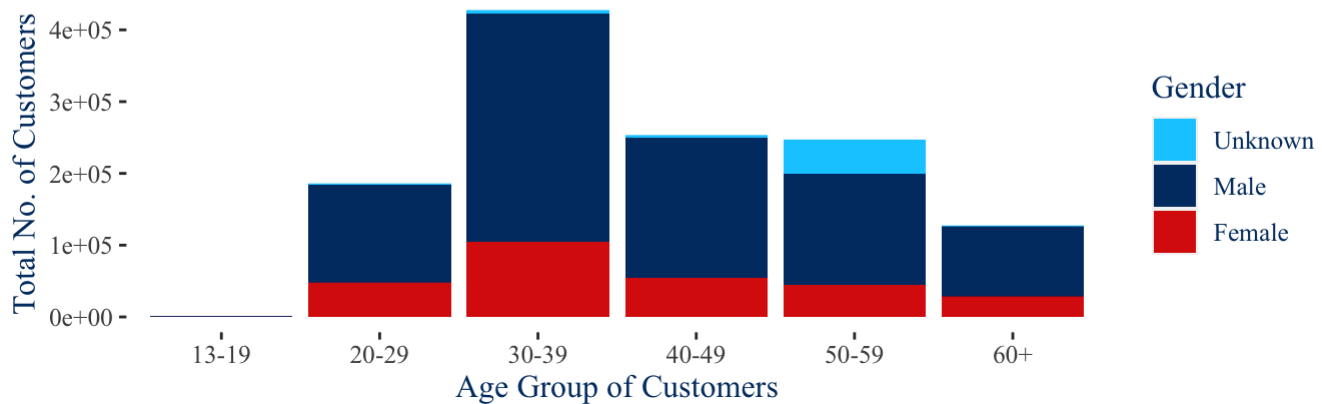
**Fig 1: Percentage Distribution of Consumer Segments in Jan 2020**



**Fig 2: Gender Classified User Segments in Jan 2020**

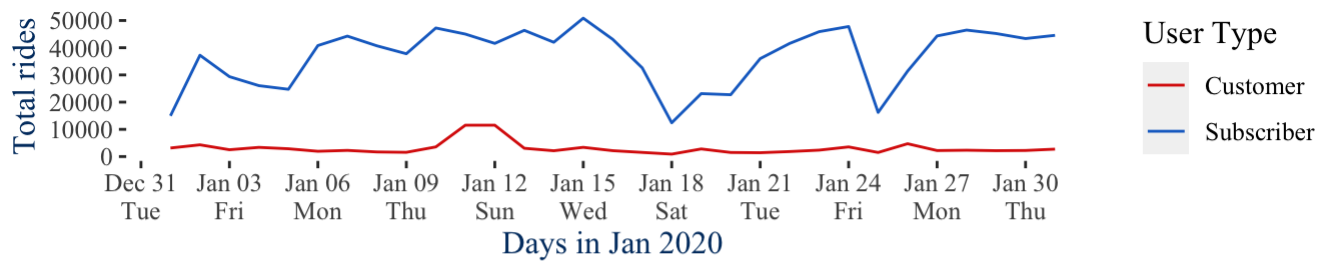
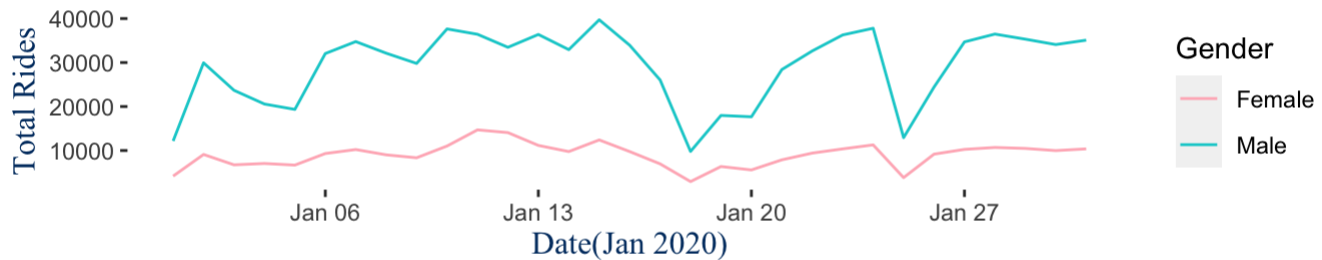
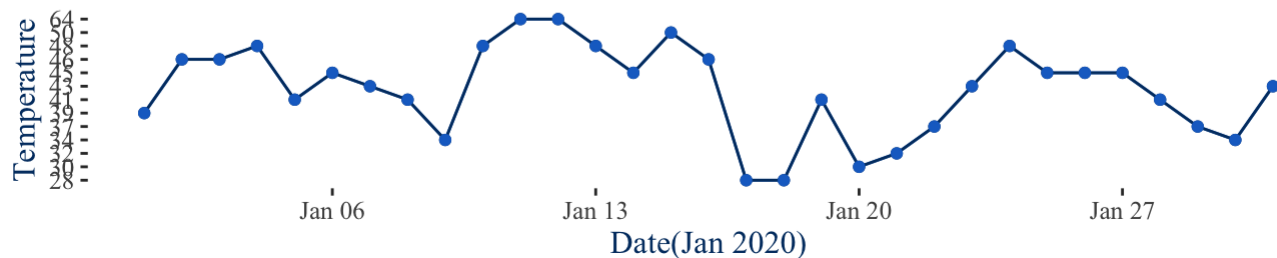


**Fig 3: Customer Age-Group Distribution**



The above visualizations clearly states that a major chunk of Citi Bike's Revenue is generated through subscriptions. Data provided on Citi Bike's website suggest revenue of approximately \$0.97M for the month of Jan 2020 just through subscriptions. It can also be inferred that, 75% of the total users are male subscribers. R&D can be done and appropriate strategies can be established to promote Citi Bike's to Female Customers. Looking at the Bar plot, we can see that the users below the age of 20 are negligible. Citi Bikes can potentially look into the opportunity to invest into rental bikes for kids and young adults.

The next two line graphs provide an overview of number of bikes used per day and categorize it based on user type and gender. The third graph was created using external data to find the correlation between bikes used and temperature on that particular day

**Fig 4: Day-wise Distribution of Customers****Fig 5: No. of Male riders vs Female riders****Fig 6: Temperature in NYC in the month of Jan 2020**

From the first graph it can be concluded that there is no significant co-relation between the day of the week and bikes used by subscribers. However, there is a considerable amount of increase in the number of bikes used by customers on weekends. Therefore, for the fraction of people who only uses Citi Bikes on weekends and do not have a yearly membership, a weekend pass would be a good idea to implement. The second line graph shows that females use less bikes as compared to men but the overall usage trend for both genders remain the same. The number of bikes used per day also varied with temperature. The third graph implies that whenever the temperature is low, bike usage for that day goes down as well.

### 3.2. Time-Series Analysis

The Time series graphs help us analyse data patterns using data collected over a specific time interval. The first graph helps us understand which hour of the day is most preferred by customers for renting bicycles. The second graph is an interactive graph that shows the average time period for which the bike was used. This graph can be viewed as either a Scatter Plot or a Bar Graph.

Fig 7: Count of users in each hour of the day

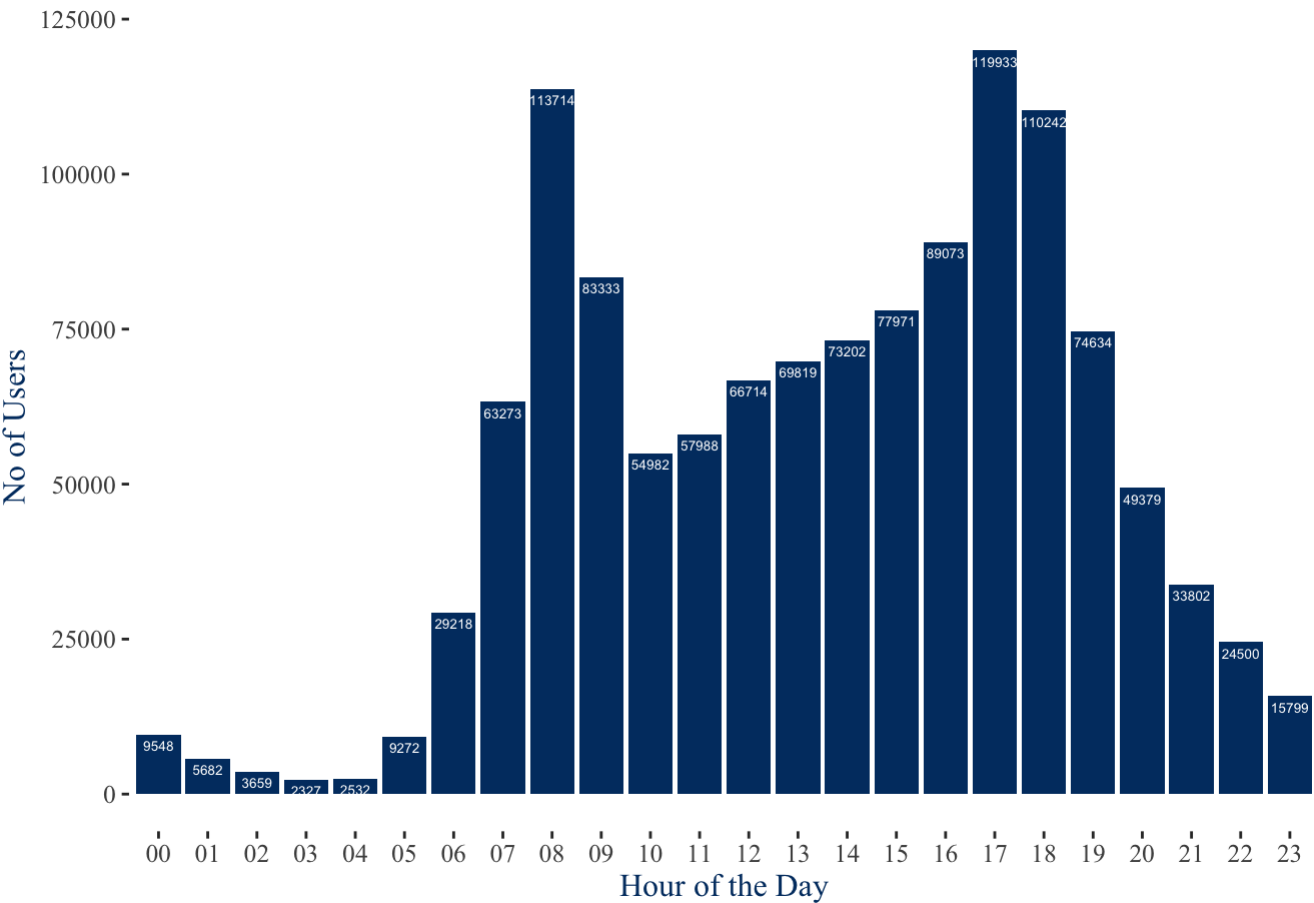
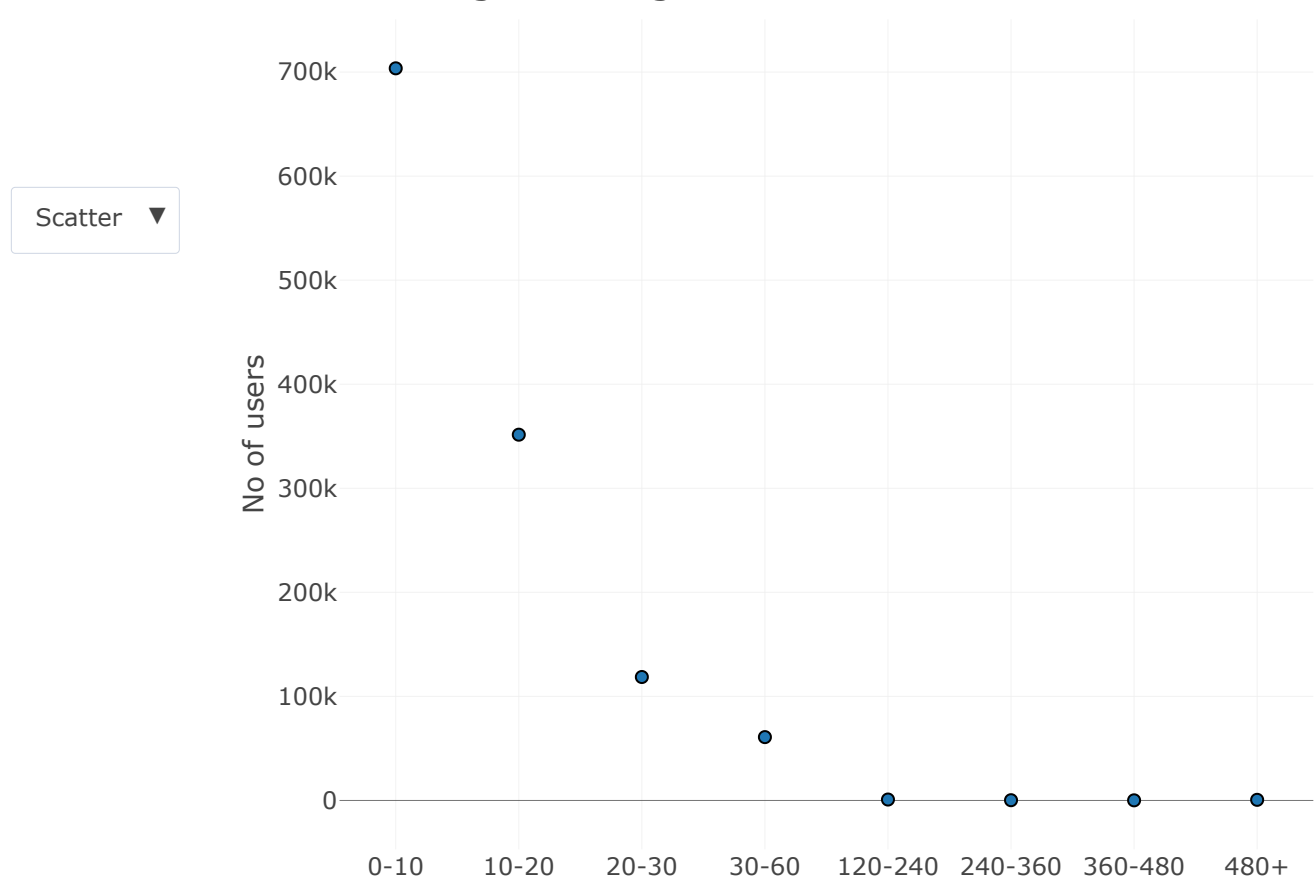


Fig 8: Average no of time



We can see that the first graph shows a major spike in the number of bikes used during the start and end of office hours i.e. 8 a.m. to 6 p.m on weekdays. Solutions can be provided to boost the usage of bikes at other times of the weekday without affecting the current peak usage times. Majority of people use bikes for 0-30 mins which will lead to more number of bikes being frequently available for customers. More schemes can be made by partnering with the government to boost tourism during the office hours which can generate more traffic during the office hours.

### 3.3. Geographic Analysis

Geographic Analysis helps find Geographic patterns using data. The below given leaflets provide a geographic representation of number of bikes picked and dropped per station along with its location. The bike locations have been color coded with color radiation to reflect the total no. of bikes picked/dropped from that station.

Click on the graph to get the details of station and no. of bikes picked

Fig 9: No. of Bikes Picked Per Station.

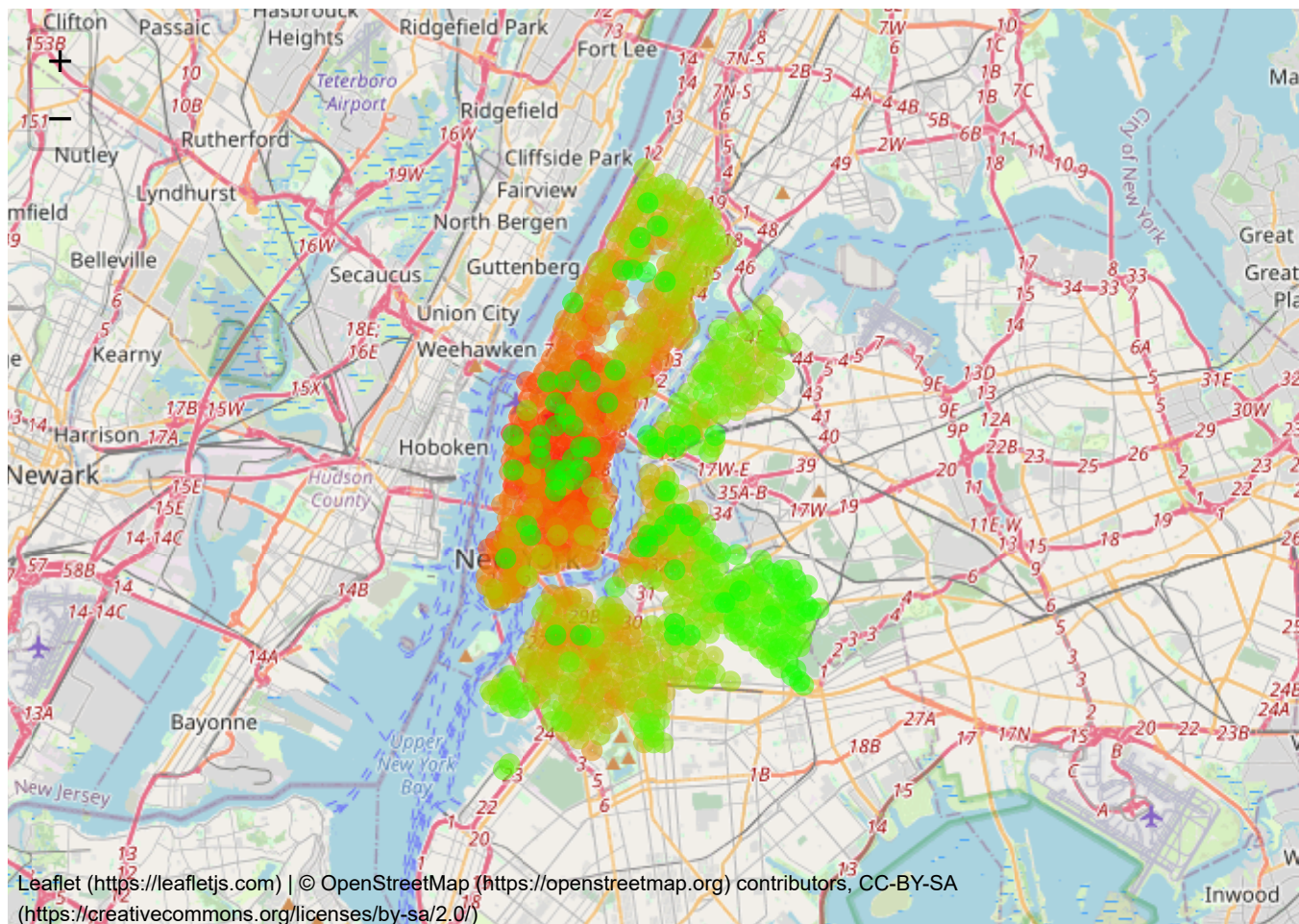
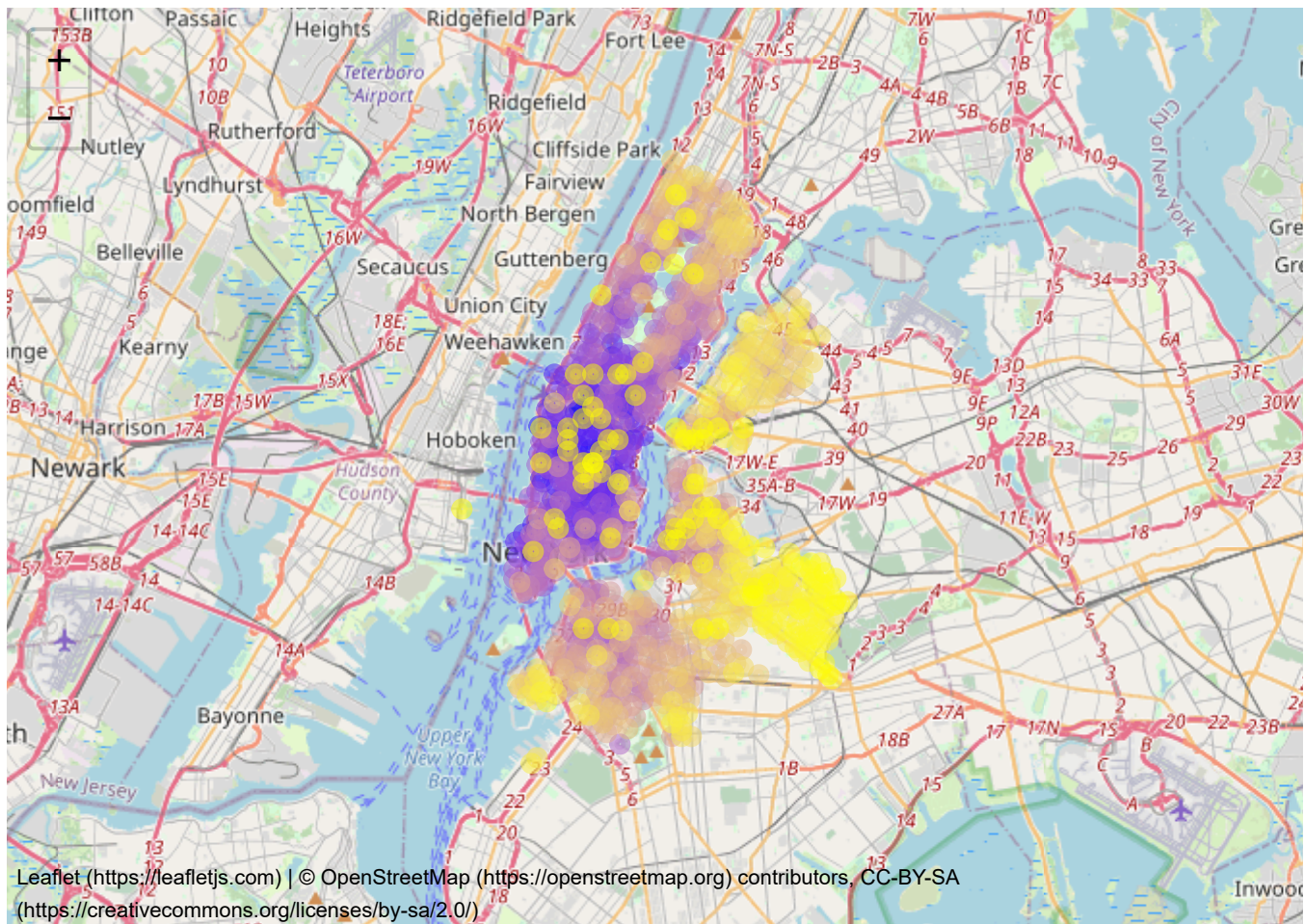


Fig 10: No. of Bikes Dropped Per Station.



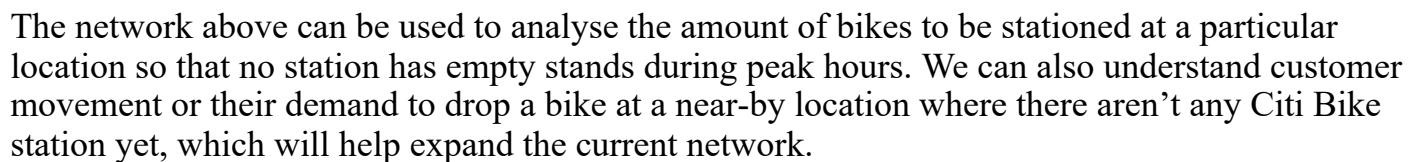


As we can infer from the above leaflets, Manhattan area has the highest number of pickups and drops on most of its stations. Research can be done to capitalize this demand by understanding user activity. Some possible solutions can be to establish more stations within close proximity of the existing ones or increase the number of racks per station

We have used a network graph to display the frequently used routes to understand the flow of traffic. This will help Citi Bike innovate new ideas to improve services in these areas.

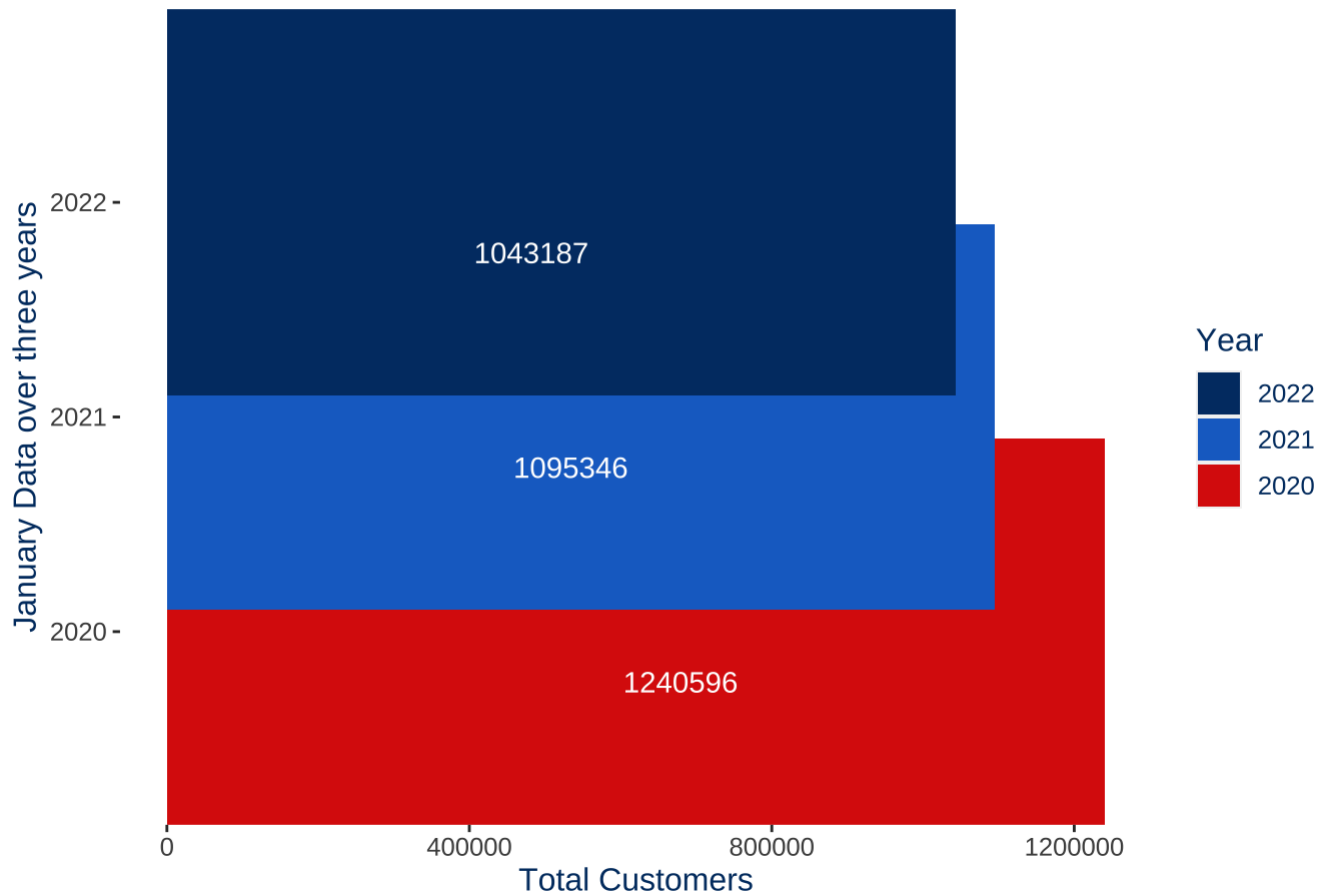
Hover over and drag a node(stations) to highlight its most connected routes.

Fig 11: Route Tracking and Density



Using additional datasets for Jan 2021 and Jan 2022, we have plotted a grouped bar chart to compare the total number of bikes used for the same month across three years.



**Fig 12: No. of Bike Users in the Month of January over 3 Years**

The graph shows an overall decrease in the number of Citi Bikes used in January 2022 over the previous years. This might be a result of the Covid-19 pandemic or the Climate Change over these years. The trend can be expected to change this year as the pandemic has subsided.

## 4. Conclusion

The Citi Bike data examined and visualized in the above sections has helped us gain insights on the general trends and areas of improvements based on the inferences made after examining those trends. One of the major area of improvement can be understand why Citi Bikes are not being used by women - is the bike design one of the reason behind it? Another area of improvement is to introduce Citi Bikes for kids and young adults. Citi Bike can also look at different ideas to boost use of bikes for longer duration during office hours. Introduction of a weekend pass is another scheme that can be proposed to boost use of bikes on weekends. Increasing the number of docks at a station and introducing new stations can be considered as a viable option based on the geographical analysis conducted above. Lastly the year-on-year comparison shows a dip in the total no. of users in the recent years. Covid-19 could be one of the reasons but, detailed examination should be done to understand the root-cause as a preventive business measure.

## 5. References

[https://en.wikipedia.org/wiki/Citi\\_Bike](https://en.wikipedia.org/wiki/Citi_Bike) ([https://en.wikipedia.org/wiki/Citi\\_Bike](https://en.wikipedia.org/wiki/Citi_Bike))  
<https://ride.citibikenyc.com/system-data> (<https://ride.citibikenyc.com/system-data>)  
<https://www.youtube.com/watch?v=hPTBZelmAh4> (<https://www.youtube.com/watch?v=hPTBZelmAh4>)  
<https://statisticsglobe.com/change-font-size-of-ggplot2-plot-in-r-axis-text-main-title-legend>  
(<https://statisticsglobe.com/change-font-size-of-ggplot2-plot-in-r-axis-text-main-title-legend>)  
<https://community.rstudio.com/t/dplyr-way-s-and-base-r-way-s-of-creating-age-group-from-age/89226> (<https://community.rstudio.com/t/dplyr-way-s-and-base-r-way-s-of-creating-age-group-from-age/89226>)  
<https://www.youtube.com/watch?v=dx3khWsUO1Y> (<https://www.youtube.com/watch?v=dx3khWsUO1Y>)  
<https://stackoverflow.com/questions/50132459/how-to-add-title-to-a-networkd3-visualisation-when-saving-as-a-web-page> (<https://stackoverflow.com/questions/50132459/how-to-add-title-to-a-networkd3-visualisation-when-saving-as-a-web-page>)