DSGA 1007 Capstone Project Report: Analysis of Citibike Activity in 2019 and 2020

Group 7: *Isidora Filipovic, Jenna Ellis, Kristin Mullaney, Elizabeth Wheeler, Eugenia Fomitcheva*

Introduction

While in pursuit of a topic, we decided it would be relevant and of interest to New Yorkers like ourselves to explore mobility trends and patterns through the usage of Citibikes in recent years. In particular, we chose to focus on similarities and differences between 2019 and 2020 as we anticipated to see shifts in transit patterns due to Covid-19 and an increase in usage of Citibikes as their popularity grew. In this report, we will briefly describe the dataset and consequently propose some relevant questions for Citibike which we will answer through analysis performed on their users, bikes, and stations.

The Dataset

Citibike readily provides access to its system data which is uploaded on a monthly basis and includes several interesting features about rides taken by Citibike users. These details include pertinent information about the demographic of the user while remaining anonymous, as well as timestamped information about the trip taken and where the bike was picked up and dropped off. The Citibike data has also already been pre-processed to exclude staff service trips during which they inspect the system and trips with less than one minute duration as these could indicate false starts or bike re-docking. For our analysis we additionally chose to eliminate any rides from riders over the age of 100 as we view these as outliers and likely to be user input mistakes, as well as any trips with a duration of over 10 hours.

Analysis

Understanding Citibike's riders

We broke down our inquiries about Citibike into two primary groups, questions about Citibike users themselves and questions about bike activity. We will begin by describing some of the interesting findings with regards to user demographics, namely age, gender, and user type as these are the three features of anonymized personal information that were provided in our dataset.

In order to determine who Citibike riders are, and how the makeup of riders had changed over the two year time period, we first examined the rides by gender. As evidenced in **Figure 1**, males have consistently been the majority group of Citibike customers and have ranged anywhere between a high of 75% of riders in January 2019 to a low of 55% of riders in May 2020. When the male population dipped in May 2020, both the female and unknown groups saw an increase as a percent of overall riders with females accounting for 31% of riders and the unknown category rising to 14%.

Citibike classifies all riders as one of two user types: subscribers, who pay a subscription for membership, and customers, or "casual users", who pay per ride. Much like with gender, the user type breakdown between subscribers and customers remained fairly consistent, with subscribers outnumbering casual users every month. **Figure 2** illustrates the change in user type composition over time. Interestingly, between March and May of 2020, the proportion of rides taken by Subscribers fell precipitously while the proportion of rides taken by casual users climbed. One possible explanation for this observation is that many subscribers pre-pandemic used Citibike to

commute to work. As remote working became the norm during the first few months of the pandemic, the commuting subscribers no longer used Citibike as regularly.

In order to gain insight into the typical profile of casual riders and subscribers, we analyzed the gender breakdown of both user groups. Notably, about 45% of casual riders have their gender listed as unknown as compared to less than 2% for subscribers. It may be that data capture for this information differs for the two groups, resulting in incomplete information in the casual rider subgroup. Filtering out unknown values (gender = 0), we plotted the proportions of casual riders and subscribers for men (**Figure 3**) and women (**Figure 4**). There seems to be a skew towards men as subscribers. Men consistently account for a larger chunk of the subscriber rides than they do casual rides. Women seem to be the opposite, consistently accounting for a larger chunk of the casual rides as compared to subscriber rides. Both casual and subscription riders followed a similar, likely seasonal pattern of seeing fewer female riders around January, a phenomenon that held in both 2019 and 2020.

Lastly, we investigated riders by age, determining that over the 2019-2020 time period, the average age of riders was 39 years old. **Figure 5** displays the changes in average age by month. During both 2019 and 2020, the average age of riders attained a minimum in August and rose to a maximum in January. Between February and April of 2020, the average age of Citibike riders increased unlike the same period in 2019 which saw a decrease in average age. One possible explanation for this difference could be that older individuals were more cautious at the onset of Covid, electing to ride bikes outside rather than risk exposure on the subway and other closed-space public transport systems.

We then considered the distribution of ages by month and found that all months over the two-year period reflected a similar distribution (**Figure 6** depicting age distribution in July 2020 is provided as an example) with a local maximum around age 30 and a large spike at about the age of 50. Such a distinct jump in Citibike popularity for this age group seemed unlikely, leading us to hypothesize that it could be due to data collection or processing methods. For example, if the user chooses not to select a birth year upon creating an account, the system may convert their age to a set value. The Unix epoch (the start of the Unix clock) is January 1, 1970, however that time is set for Greenwich, England and is subject to timezones. Therefore, the start of the Unix clock for anyone west of England will be December 31, 1969. It is possible that somewhere in Citibike's data pipeline, birth years were converted into Unix time values with missing values treated as 0, which would be a valid input and converted to 12/31/1969. This theory is supported by our investigation into the age distribution by customer type. The proportion of casual riders with a birth year of 1969 is consistently above 40% each month. Since casual riders are far more likely to be missing demographic information, including birth year, seeing a particularly high amount of casual riders born in 1969 is unsurprising if missing values are being inadvertently converted to that year.

Hoping to gain more insight into age distribution, we grouped ages, in increments of five years, and examined the percentage of riders in each group by month (**Figure 8**). From the data, it is evident that younger riders tend to account for a larger percentage of ridership. This can be observed in the rainbow effect of the graph, with younger age groups represented in red and older in blue. The two

groups that seem to be outliers are the 15-19 and 50-54 age groups, with the 50-54 age group accounting for a far larger percentage of ridership than its adjacent age groups combined. This echoes the overrepresentation of individuals born in 1969 previously discussed. However, here we see that the 50-54 age group also exhibited an anomalous increase in ridership between January and May of 2020. In this time, this group went from 11% to almost 19% of overall riders. By comparison, the 55-59 age group decreased from 6% to just over 4% of riders during this same time period. The 15-19 age group, conversely, seems to account for a smaller percent of ridership as compared to the next youngest age groups. This could be in part due to Citibike's minimum age being 16, although another reasonable explanation might be that riders aged 15-19 are less likely to be subscribers. As subscribers vastly outnumber customers in terms of rides per month, this may explain the low percentage contribution of the 15-19 age group. It is also worth noting that both the 15-19 and 20-24 year-olds exhibit more seasonal variation than other groups with a large uptick in ridership in the summer months.

In order to better understand the relationship between age group and user type, we determined the breakdown of user type in each group, as shown in **Figure 9**. We see that the proportion of subscription rides decreases steadily from the 25-34 age group onward, while casual rides increase in frequency up to the 45-54 age group. Subscriber rides outnumber customer rides in every age group with the notable exception of age groups 15-24 and 45-54. This observation is consistent with the previous finding that riders with a listed birth year of 1969 account for at least 40% of all customer rides across all 24 months and supports the theory that riders aged 15-19 are less likely to be subscribers than riders in other age groups.

Understanding Citibike trips and transit patterns

With a clearer understanding of the makeup of Citibike riders, we turned our attention to ride volume, station popularity and changes to common routes and traffic patterns from a pre-pandemic NYC to December 2020. Despite the confounding factor of the pandemic, the line graph in **Figure 10** suggests a seasonality to ride volume with a decrease in trips during colder months and a gradual increase with warmer weather. **Figure 11** provides a clearer picture of the difference in ride volume between the two years. The number of trips taken between March and July 2020 was below the volume of the same months in 2019. By August, however, and for the rest of 2020, ride volume increased and surpassed the number of rides from the previous year. We see that the dip in rides taken corresponds to the beginning of the Covid-era when most New Yorkers faced a lot of uncertainties. Conversely, the significant increase in activity following April, we speculate, may be attributed to individuals choosing outdoor methods of transport, growing Citibike popularity, and the arrival of spring and summer months.

In order to gain insight into where people are traveling to and from, we determined which Citibike stations were most popular. The 2019 results were not all that interesting; Pershing Square North held the top spot every month, likely the result of proximity to Grand Central, making the location a convenient starting point for midtown workers and travelers alike. However, in April 2020 we again observed a shift -- 1 Ave & E 68 St and then 12 Ave & W 40 St become the most popular starting stations. Such a change may be the result of businesses adopting remote working environments, as riders no longer had to travel to midtown office locations and possibly relied on Citibike for more

recreational purposes, as the jump in average trip duration in April of 2020 displayed in **Figure 12** would suggest. In fact, our analysis shows that while some of the top routes in 2019 were commuter routes, the most popular trips in 2020 were roundtrips at a variety of locations across the city. The top two trips in particular were in Prospect Park and Central Park, where residents likely took to biking as an activity in and of itself.

After establishing top-activity stations, we delved into the type of traffic those stations saw. For example, some spots, like Pershing Square North, saw a healthy balance of bikes docking and leaving the station throughout the day. A few others didn't maintain such balance. For example, the station at 8 Ave & W 31 St saw a lot of bikes leaving with fewer docking in the mornings and more bikes leaving in the evenings (**Figure 13**). Meanwhile, West St & Chambers St saw the opposite when in 2019, more bikes were arriving in the morning than leaving. In 2020 however, we see the imbalance at West St & Chambers St levels off as fewer people were coming into their Wall St. offices (**Figure 14**). Nonetheless, with certain stations seeing these traffic patterns, Citibike must maintain a reasonable distribution of bikes to prevent overcrowding or empty stations, or could alternatively consider expanding hub locations to accommodate riders at peak travel times.

We decided to further investigate bike traffic in order to better understand rider movement throughout the day. Theorizing that bike traffic differed on weekdays versus weekends, we determined that the distribution of rides across hours of the day on weekdays is bimodal, as bike traffic seems to be heaviest between the hours of 8-9am, presumably the result of riders commuting to work, and between the hours of 5-6pm as they return home. This is distinctly different from the more normally distributed bike traffic during the weekend. While this pattern is consistent across months until April 2020, in April 2020 and May 2020, the distribution of weekdays loses its bimodal appearance and converges towards the normal distribution of the weekend (**Figure 15)**. As the shift in weekday distribution coincides with the start of the pandemic, a reasonable explanation for this change is again the adoption of remote work. When analyzing the hourly distribution of rides by user type rather than day of the week, a similar pattern appears with subscribers distributed bimodally while the time stamp for casual riders appears normal, suggesting that most weekday riders are subscribers riding to and from work, while weekend riders are more often casual users. Indeed, the shift observed in April 2020 for the weekday versus weekend distributions is again present in the distributions by user type with the subscriber distribution losing its bimodality in April and May 2020 and appearing to converge to the distribution of casual users.

Finally, we compared average bike usage in 2019 and 2020 in terms of trips and hours ridden (**Figure 16**). We found that while the average number of monthly trips in 2020 remained largely below what we observed in 2019, the average monthly usage of bikes in hours increased in May 2020. This interesting result suggests that Citibikers took less trips in 2020 but spent a decent amount more time per trip in 2020 versus 2019. We previously hypothesized that biking became more of a recreational activity than necessarily a mode of transport to and from the office, and this data further supports this. As we can see, 2020 turned out to be an impactful and dynamic-shifting year for Citibike with prevalent trends that may be worth investigating in 2021 as an expansion of this project and case for further exploration.

Appendix

Figure 1

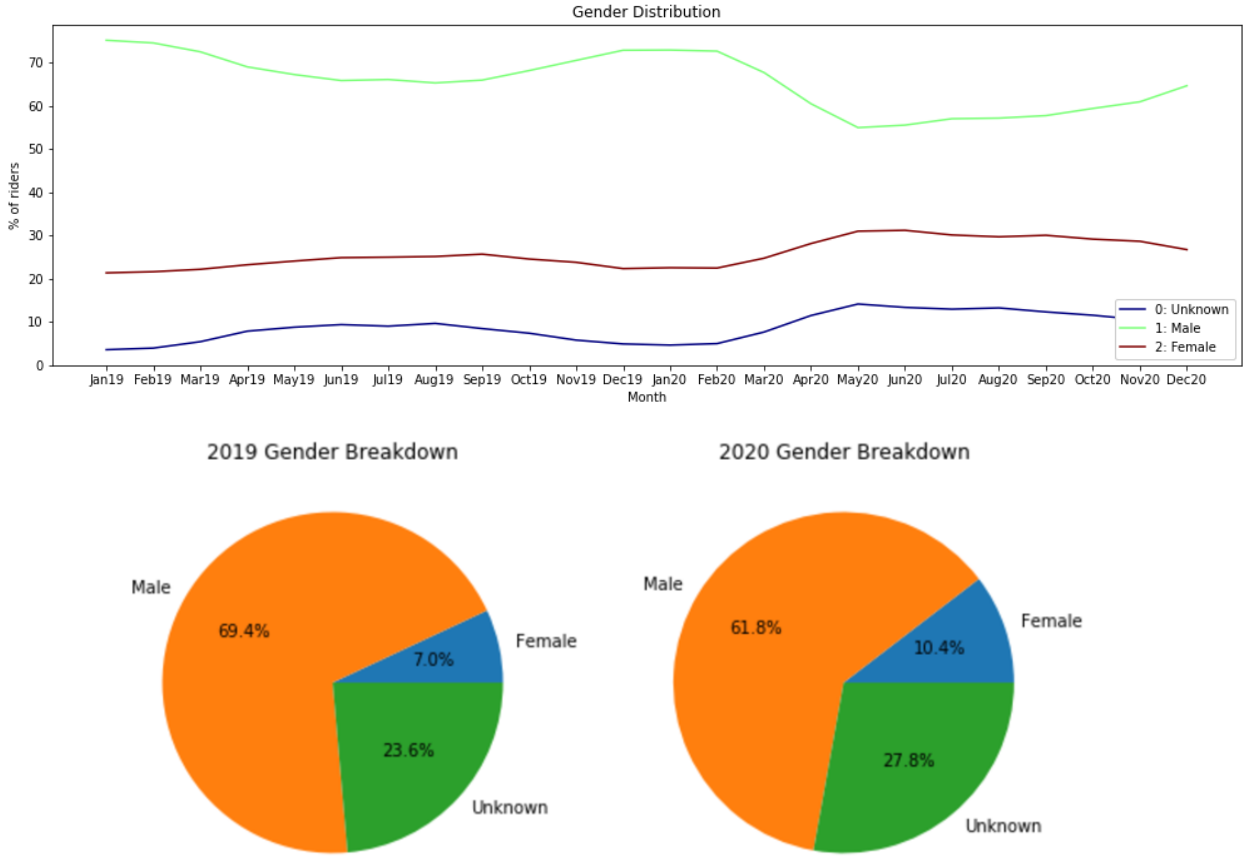

Figure 2

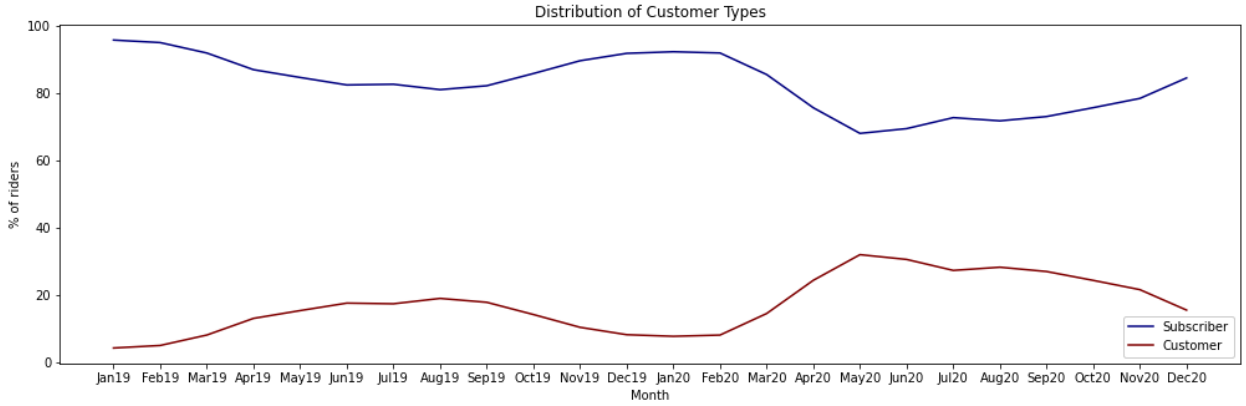

Figure 3

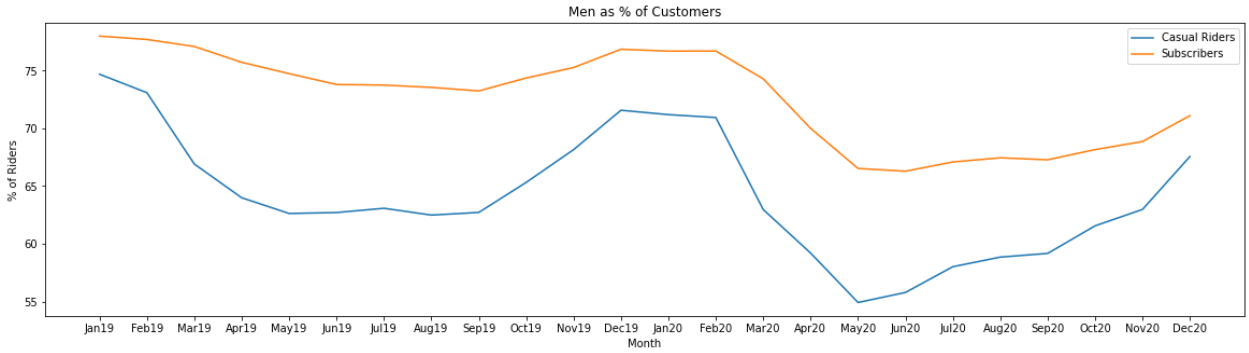

Figure 4

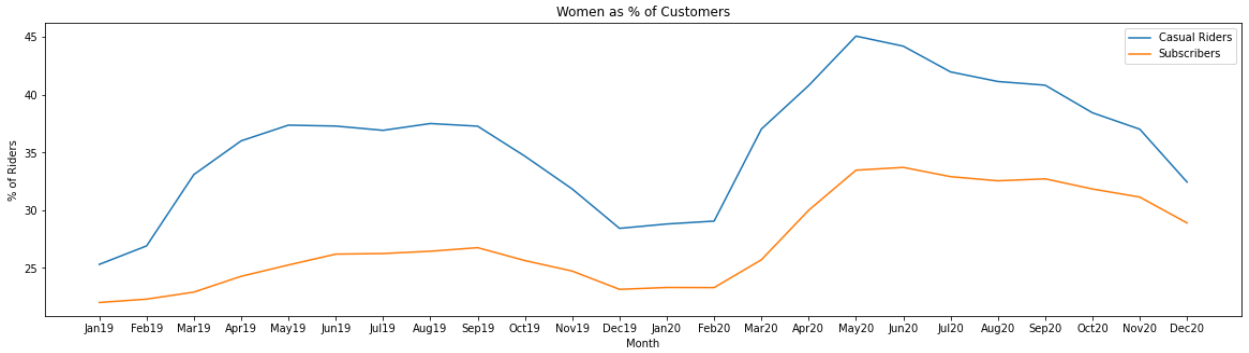

Figure 5

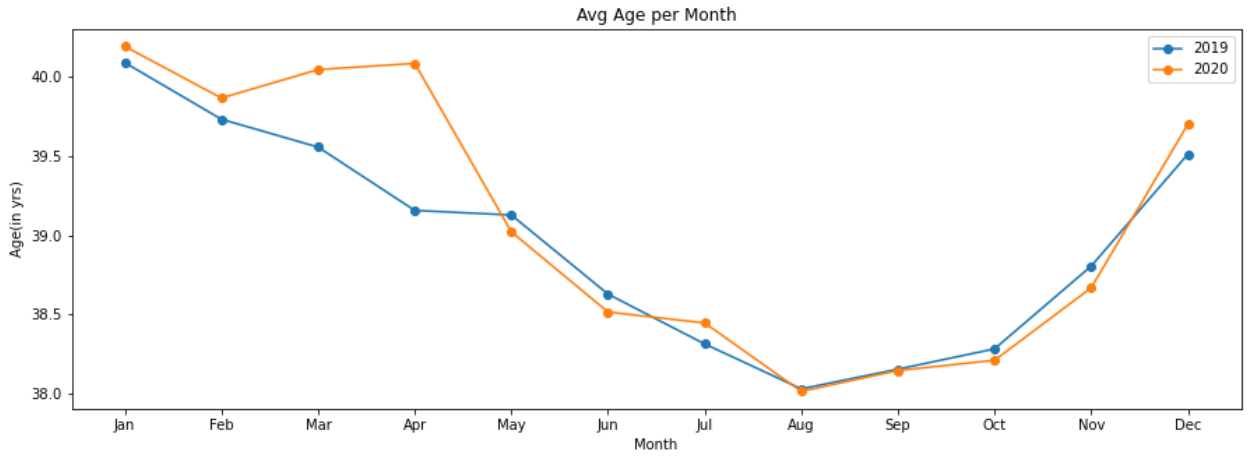

Figure 6

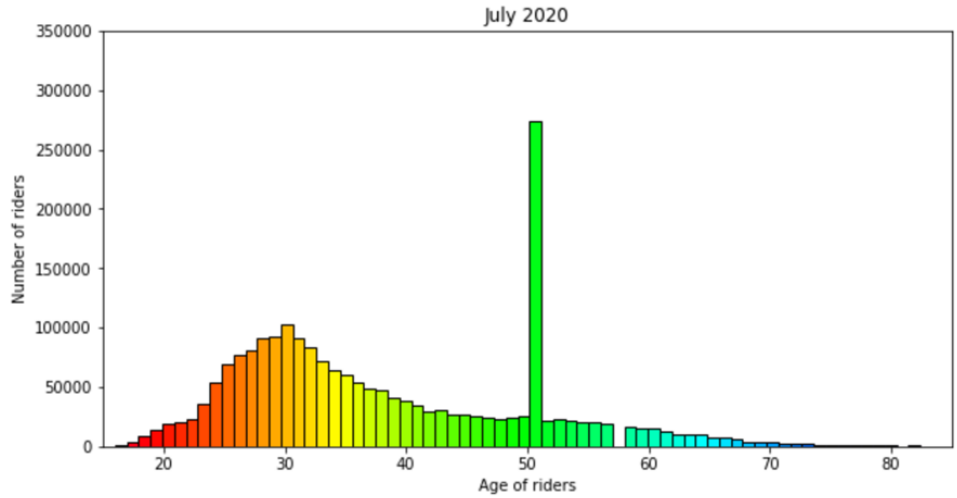

Figure 7

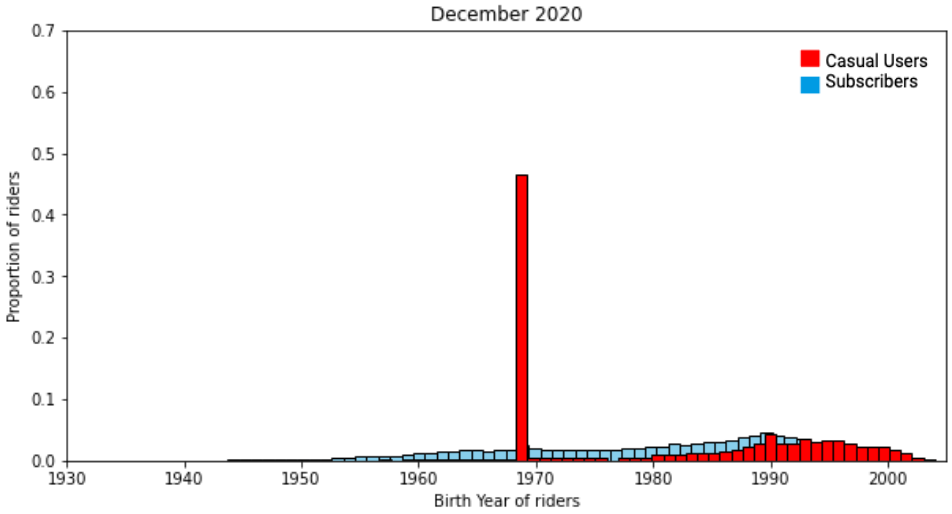

Figure 8

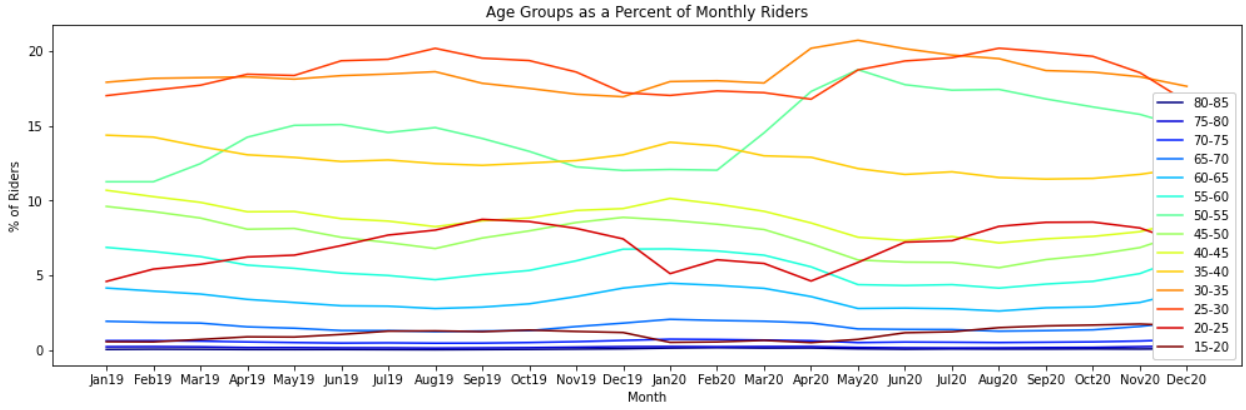

Figure 9

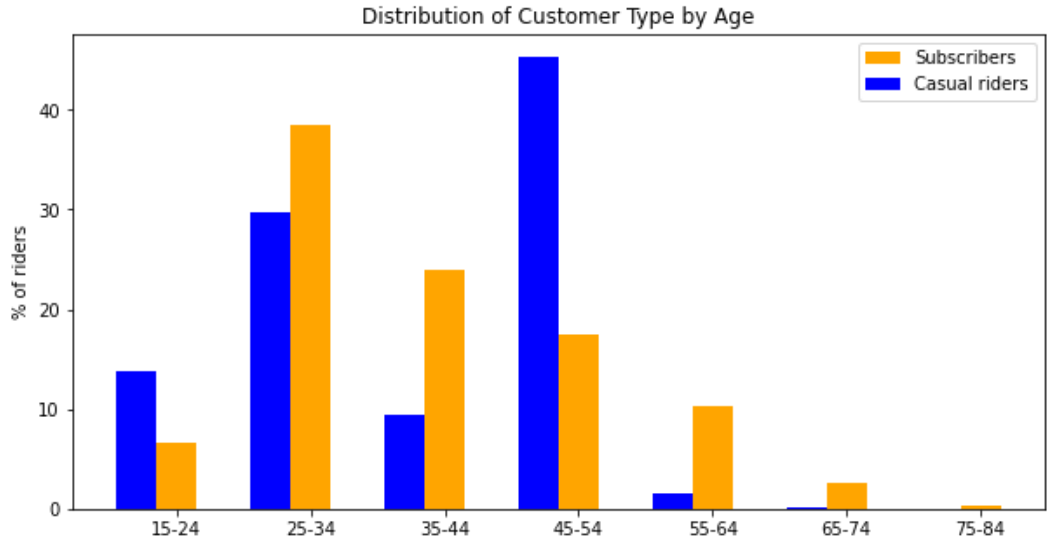

Figure 10

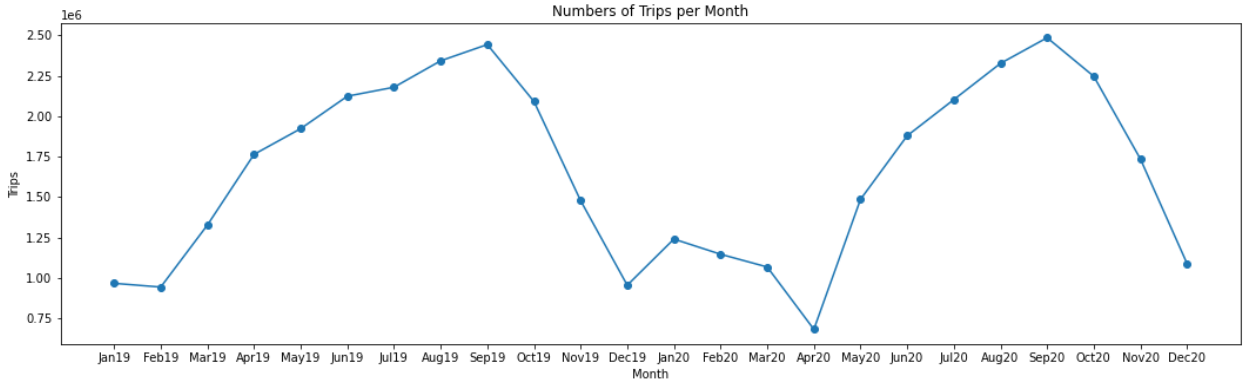

Figure 11

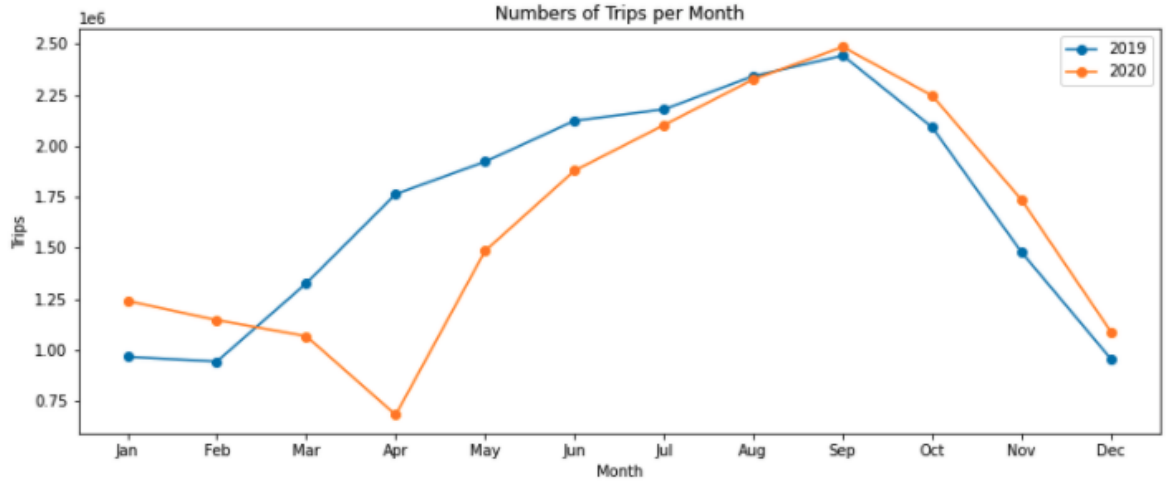

Figure 12

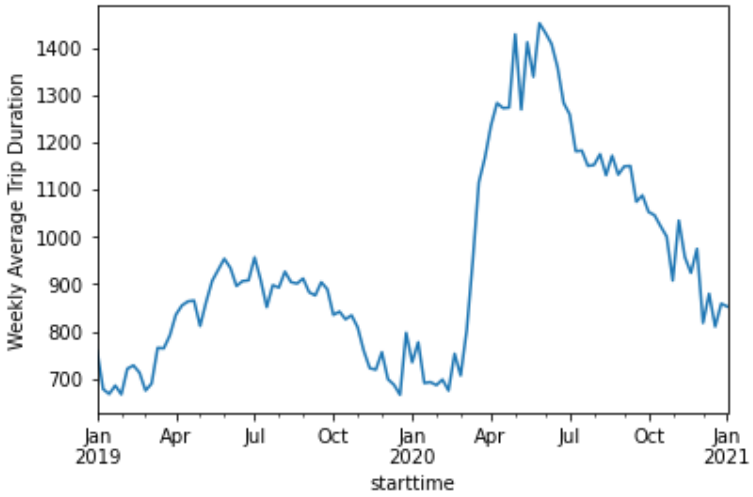

Figure 13

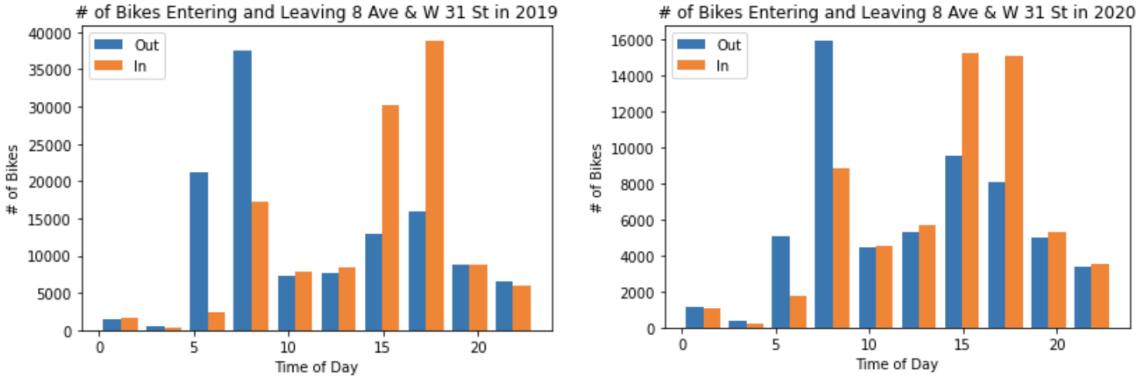

Figure 14

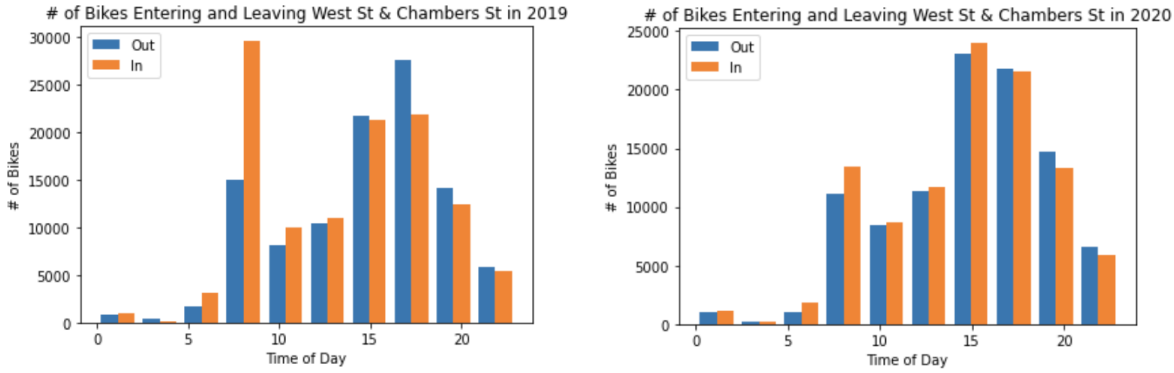

Figure 15

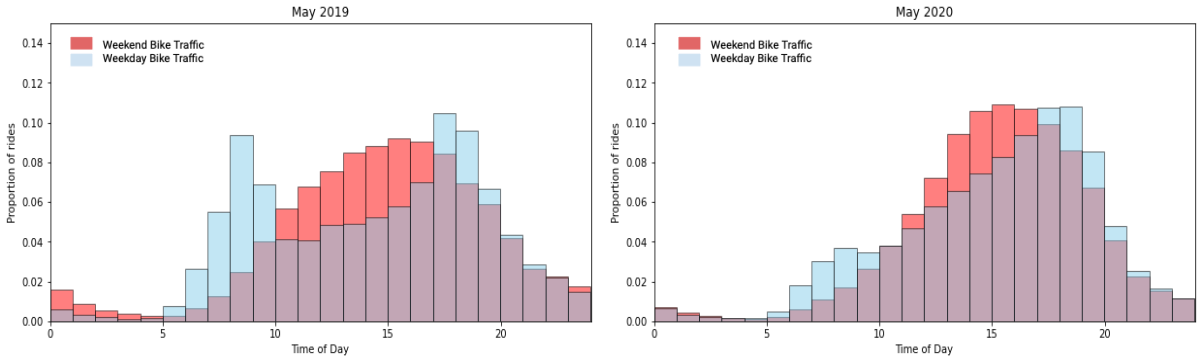

Figure 16