A picture containing bicycle, outdoor, blue, parked

Description automatically generated

Impacts of COVID-19

on

Citi Bike

User Profiles

For

Syracuse University

Professor Corey Jackson

24 March 2021

Patrick Bush (pabush@syr.edu)

Farahin Choudhury (fchoudhu@syr.edu)

Destiny McDaniel ([dmcdanie@syr.edu](mailto:dmcdanie@syr.edu))

Clayton Monroe ([cbmonroe@syr.edu](mailto:cbmonroe@syr.edu))

A Leigh Sausville (alsausvi@syr.edu)

### Table of Contents

1. [Project Abstract](#_avc5bsiwrje3)
2. [Contribution Statement](#_57ua8j8t3uqe)
3. [Introduction](#_da5yv2jg2rx1)
4. [The Dataset](#_se9xcxmvakhj)
5. [Data](#_bkvyqbe6mg90) Conversion and Preparation
6. Data [Analysis](#_n4xlewhnel30) Methods
7. Analysis [Summary](#_2g7vp4tc6ijw)
8. Statistical Assessments
9. [Conclusions](#_tb4hsi8a2abh)
10. Recommendations
11. [References](#_uzhu5z6y25mv)
12. [Appendix](#_hn5qmg7z1mfu)

### 

### PROJECT ABSTRACT

The Director of Citi Bike Data Analysis Division requested the Data Analytics Team 2 to conduct a “quick look” analysis examining the impacts of COVID on user behavior. The goal of this report is to collect and analyze Citi Bike user data and describe changes occurring after the onset of government restrictions prompted by health and safety factors related to the pandemic. The intent is to understand how changes in user behavior patterns could support decisions related to future company operating models. The analysis will focus on user demographic profiles before and after COVID restrictions.

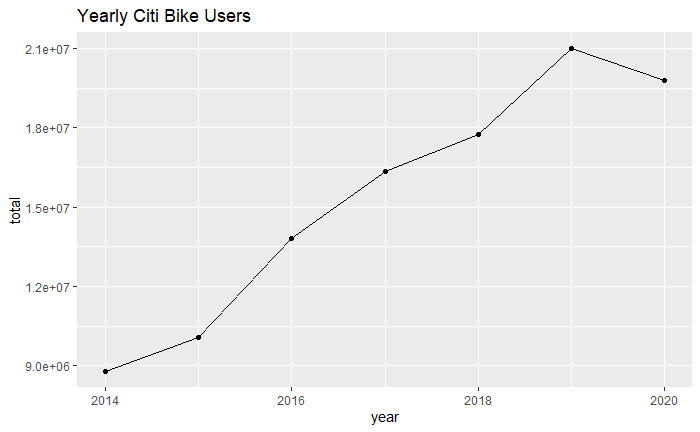
### CONTRIBUTION STATEMENT

* 1. **Farahin Choudhury - Data Scientist**
  2. **Clayton Monroe - Data Analyst**
  3. **Patrick Bush - Technical Writer**
  4. **Destiny McDaniel - Graphics Analyst**
  5. **Leigh Sausville - Research Analyst**

**INTRODUCTION**

Bike sharing programs are increasingly popular across major cities in the United States. They provide low-cost transportation alternatives within densely populated areas. Citi Bike is New York City’s bike sharing system, sponsored by the Department of Transportation, and owned by Lyft. The bike sharing service is available 24 hours a day and currently offers 1,166 active stations.[[1]](#footnote-1) Operating a bike sharing service requires an understanding of the user base and utilization of data analytics to identify trends in user behavior, predict future requirements and optimize service delivery to sustain a feasible and competitive operations model. Bikes are in the docking stations spread out across New York City.[[2]](#footnote-2) Users can either purchase annual memberships or buy short-term day passes. Citi Bike has experienced annual average growth rates of 18% between 2014 and 2020. The steady growth in customer demand has been supported by the growth in the number of bikes in the fleet and docking stations for pick up and return. **Figure 1** provides a snapshot of the yearly growth in Citi Bike users.

The COVID pandemic hit the United States in late 2019. Citi Bike was not immune to feeling the impacts on ridership. This resulted in a significant lag in ridership of its users; while there was an increase in seasonal traffic between winter and spring, the overall growth was lower compared to prior years.



**Figure 1 Snapshot of Citi bike User Growth 2014-2020**

The data analysis team collected data for user profiles and compared categories of the data before and after COVID restrictions. The team analyzed data from May through July for 2019 and 2020. The following questions were addressed in the analysis:

* + - How did overall ridership change?
    - Did trip durations change between 2019 and 2020?
    - Did customer membership change?
    - Were there changes in the peak time of day for customer trips?
    - Did the top locations change?

**THE DATASET**

The team used data provided in Citi Bike Monthly operating reports[[3]](#footnote-3). The team used this data to understand the context of Citi Bike user data. The raw data was imported into R Studio for analysis.[[4]](#footnote-4) The second data set was much more diverse and richer with specific user characteristic information. The data was obtained from Citi Bike web pages and was provided in CSV format.[[5]](#footnote-5) **Appendix A** provides a snapshot of the data set. The data categories are included in **Table 1**.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 3-day pass user; Subscriber = Annual Member) | Gender (Zero=unknown; 1=male; 2=female) | Year of Birth |  |

**Table 1 Data Categories in Citi Bike Data Set**

Several issues were identified by the team when exploring the data sets including:

* Extremely large files – one month file contains over 1 million entries.
* Data in certain categories was inconsistent.
* With repeating data values (pertaining to birth year and gender), it is hard to tell if it's the same individual making the trip. For example, if there are multiple records in the dataset for a male born in 1990, there's no way to know if it's the same person for all those records.
* Station names and locations expanded and changed between certain months.
* Date of birth years entered incorrectly.
* Gender identification inconsistent and missing input
* Station names not organized by borough.

**DATA CONVERSION AND PREPARATION**

The team extracted the data files from Citi Bike’s website for preparation and consolidation. The files were group together by their columns in R Studio, producing consolidated and clean files per month per year. **Appendix B** provides an example of the Citi Bike data in R Studio. The team combined monthly data sets for May, June and July into a comprehensive three-month data set for 2019 and 2020. The data for those two years were combined into a single comprehensive data set specific for this analysis.

p ten most high traffic areas and concentrate our analysis on those on

### DATA ANALYSIS METHODS

An examination of the monthly trends in users between 2014 and 2020 was completed. The team also looked at monthly comparisons across those years to assess the trends over time. A regression analysis was completed to assess the variables on user ridership. With the strategic profile of the data in place the team examined specific characteristics of user data between the selected months for 2019 and 2020. A t-test was used to compare the number of user types with trip duration. Ggplot2 was used to plot changes in most popular docking stations. Other visualizations helped describe the results and formulate recommendations.

### ANALYSIS SUMMARY

**Appendix C** provides an overview of Citi Bike user growth between the years 2014-2020[[6]](#footnote-6). Key insights include:

* Citi Bike users have historically exhibited changes in user totals due to seasonal trends
* The expected “spring bounce back” in 2020 displayed a lag in growth compared to prior months in spite of increases in number of bikes and docking stations.
* By the middle part of the summer (August) in 2020 Citi Bike user totals returned to comparative levels observed in 2019 and are consistent with previous years.

**Appendix D** provides a more detailed look at monthly comparisons for Citi Bike users for the month of May and July between the years 2014 and 2020.

**Figure 2** is a visualization that compares user trip duration between May and July for 2019 and 2020. The data indicates that Citi Bike users took longer trips in 2020 after government restrictions were enacted as a result of COVID-19. On average users in 2019 took trips with a duration of 17.6 minutes. In 2020 the average duration trip between May and July was 28 minutes. The data indicates Citi Bike users were taking trips roughly 37% longer during the months when COVID-19 restrictions were in place.

Chart, bar chart

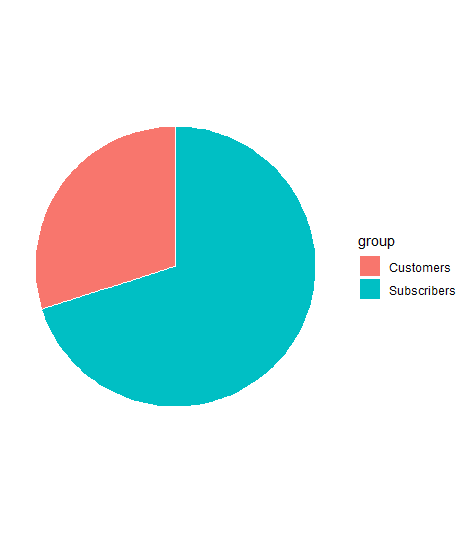
Description automatically generated

**Figure 2 Citi Bike User trip duration comparison May-July 2019-2020**

The key findings from **Figure 2** include:

* An increase in trip durations in 2020 could help offset the impact to total revenue resulting from a decrease in total riders.
* The increase in trip durations may have impacts on mean time of failure for each bike and could increase the requirements for regular maintenance cycles, acquisition of bike replacements and the resources needed for field support and service calls.

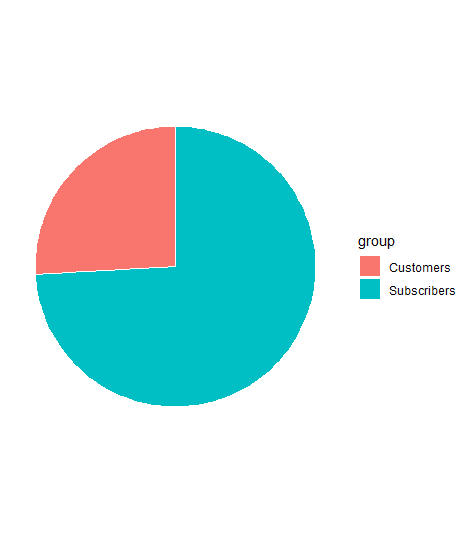
**Figure 3** shows the comparison between user types in 2019 and 2020. As expected, the government-imposed COVID-19 restrictions resulted in a decrease in the number of Citi Bike subscriber memberships. While there were more users in the “customer” category in 2020, the “subscriber” category remained in good standing despite experiencing a decrease. This is intuitive because the restrictions on work and travel would be expected to change the demand for Citi Bike services on a normal predictable pattern.



2020

70.1%

29.9%



2019

74.1%

25.9%

**Figure 3 Citi Bike User type comparison May-July 2019-2020**

This change in user behavior has strategic implications for the future growth of Citi Bike overall. If the external variables (such as government restrictions) remain, casual users may continue to grow and complicate traditional predictable patterns of number of users. Even if the restrictions ease up, the behavior patterns of users may not automatically revert back to what they were prior to the restrictions. Ridership could be impacted by permanent changes in tourism and less commuter traffic due to working remotely. Some implications include:

* Number of bikes required in the fleet
* Number of available docking stations
* Field support and customer service
* Locations and time of peak usage

**Figure 4** provides a comparison of the peak time of day when Citi Bike users are starting their rides, for both 2019 and 2020.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Figure 4 Citi Bike trip start time of day**

Results depicted in **Figure 4** are intuitive for the normal Citi Bike operating model in 2019. Peak periods correspond to rush hour in the city when many Citi Bike customers are using the service to commute to and from work. Usage surges during morning commute times and stay relatively constant during the afternoon. A second surge is observed beginning around 4 pm when users are getting off work, tourists are moving about the city, residents are running errands, going to dinner, or pursuing a variety of entertainment options.[[7]](#footnote-7) In 2020 changes in user trip start times are evident. Rather than peaks and valley’s corresponding to rush hour there is a steady climb in riders throughout the data starting at 6 am and peaking around 6 pm. These changes have implications for Citi Bike service delivery approaches to handle customer requests for assistance and respond to service calls throughout the city.

**Figure 5** provide the map visualizations of the result of this analysis using ggplot. The initial result indicate that the top ten most popular stations changed in 2020 showing a more concentrated footprint in lower Manhattan, Midtown and south of Central Park. In 2020 the top location centered arounds parks and attractions compared to 2019 when the concentrated around secondary transportation hubs such as Grand Central Station and Madison Square Garden.

2020

2019

Map

Description automatically generated

Map

Description automatically generated

**Figure 5 Top Ten Citi Bike Stations 2019 and 2020**

### STATISTICAL ASSESSMENTS

**Figure 6** provides the results of a regression analysis performed on the total user data. The regression was done using monthly Citi Bike summary data to examine the relationship between number of bikes available, number of docking stations available and data on number of users. Number of users was the dependent variable. The Adjusted R square value of 46% indicate that the dependent variables (Bikes, Docking Stations and Date) explain about half of the impact on changes to the independent variable. Over 50% of the changes in overall user population is impacted by external variables not related to number of bikes, docking stations or date. To develop more accurate predictions and optimize operating models Citi Bike needs additional data on variables impacting user behavior, such as weather, percentage of people working from home and trends in tourists visiting the city.

While number of bikes appears to be statistically significant based on the p value, further data is required to make substantive decisions on the variables that impact user behavior. Date was determined to not be statistically significant which was inconsistent with the overall yearly user profile. The Group believes weather is a more accurate variable that should be included in further analysis. The intuitive conclusion is that expected number of users lagged expected growth and the change cannot be explained by number of bikes, docking stations or date. The impact on total number of users was the result from other external variables. Because Citi Bike user data tends to be nonlinear regression analysis might not be the best means of developing a predictive model. Citi Bike data analysts will need to investigate the use of an exponential data model to better reflect the non-linear nature of Citi Bike clients.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.018e+06 4.414e+06 -1.590 0.1158

Date 5.525e-03 3.323e-03 1.663 0.1003

Bikes 1.334e+02 4.793e+01 2.784 0.0067 \*\*

Stations -1.933e+03 1.016e+03 -1.903 0.0606 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

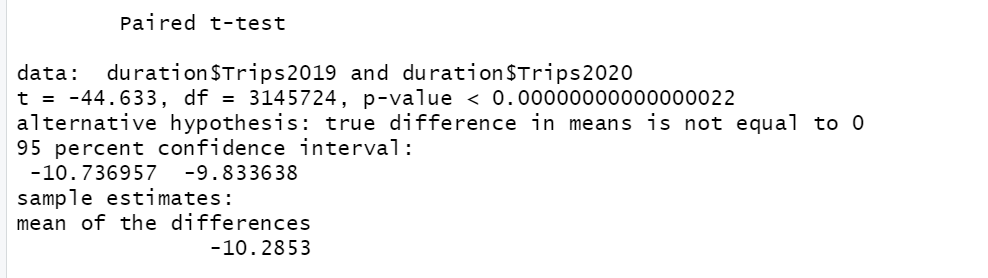
Residual standard error: 422800 on 80 degrees of freedom

Multiple R-squared: 0.4835, **Adjusted R-squared: 0.4641**

F-statistic: 24.96 on 3 and 80 DF, p-value: 1.69e-11

**Figure 6 Regression Analysis for total Citi Bike users between 2014-2020**

### The Team also conducted a two-sided paired t test to compare average trip duration between 2019 and 2020. Figure 7 provides the results of the t test.



**Figure 7 Two-Sided Paired t Test**

The t test was conducted to compare the means for trip duration in 2019 and 2020. With the null hypothesis being equal to no change in the population from 2019 to 2020 (Citi Bike riders), in terms of mean trip duration, we found there was in fact a change from 2019 to 2020. The results provided a p value of well below .05 indicating with 95% confidence the population had evolved.

### With such a sharp increase in trip duration between 2019 and 2020, we were quite confident that ridership in 2019 was different in 2020 (if not in the users themselves, in the users’ behavior) and this test provided confirmation that this shift was not the result of natural variation.

### CONCLUSIONS

* + Data Analysis indicates COVID restrictions impacted Citi Bike user profiles.
  + Total users lagged expected growth levels based on yearly trends.
  + By summer 2020 user rates recovered to expected seasonal levels.
  + User profiles changed in 2020.
    - Subscriber customers declined between 2019 and 2020.
    - Casual riders increased between 2019 and 2020.
    - Trip durations increased between 2019 and 2020.
    - Primary docking stations shifted away from traditional secondary transportation hubs recorded in 2019 to focus more around parks and attractions in 2020.
  + The Citi Bike performance appears to be stable and growing, however:
    - User base is subject to seasonal trends probably based more on weather than date.
    - The user base appears to be dynamic in response to external conditions and government regulations that impact citizens’ personal and professional environments and routines.
    - Projecting outyear user totals must account for external variables (currently undefined) beyond number of bikes available.
    - While the customer behavior changed the operations model remains solid, the user profile changes re-enforce the need for Citi Bike to increase the use of real time data analytics to anticipate and respond to shifts in user behavior that impact where and when bikes and bike docking stations need to be available.

**RECOMMENDATIONS**

Group C collected the data, ran analytics, and based on the results developed some visuals to help explain the results. The results of the analysis led us to highlight four top recommendations for Citi Bike:

**Go all in on data analytics** - to help describe patterns and trends, develop predictive models, and optimize operating approaches. Define data standards for collection and reporting to make analysis go more smoothly and allow results to be consistent. Invest in data scientists to work across all corporate departments. Data should be treated as a corporate asset.

**Link to Financial Forecasting -** develop a stronger link between the data analytics team and the program controls department to better predict revenues and develop out year operating budgets based on analysis of variables and external conditions.

**Integrate data analysis into acquisition plans** - that focus on defining how and when to purchase new bikes, how to balance operations and maintenance plans and implementation – to optimize where and when bikes are needed.

**Develop an expertise with data analysis techniques** - that enrich the service delivery performance consistent with a broad and diverse user base. Data analytics can help predict which bike docking stations are used and when - and also help define where growth in new bike docking stations will be most effective.

### REFERENCES

*An Introduction to Data Science*, Jeffrey S Saltz, Jeffrey M Stanton, SAGE Publishing, 2018.

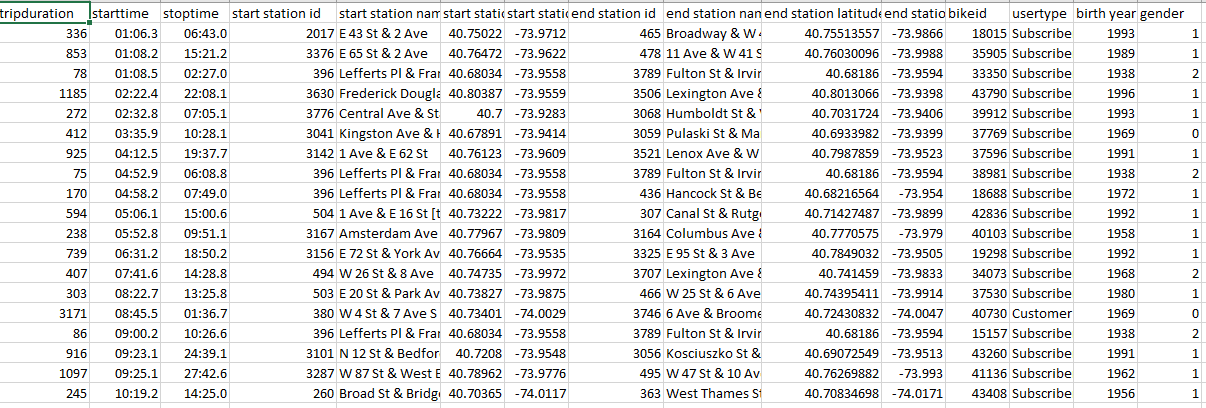
Data Science for Business, What You Need to Know About Data Mining and Data-Analytic Thinking, Foster Provost and Tom Fawcett, O’Reilly Publishing, second edition 2013

Practical Statistics for Data Scientists, 50+ Essential Concepts Using R and Python, Peter Bruce, Andrew Bruce and Peter Gedeck, O’Reilly Publishing, second edition 2020

R in a Nutshell, A Desktop Quick Reference, Joseph Adler, O”reilly Publishing, second release 2015

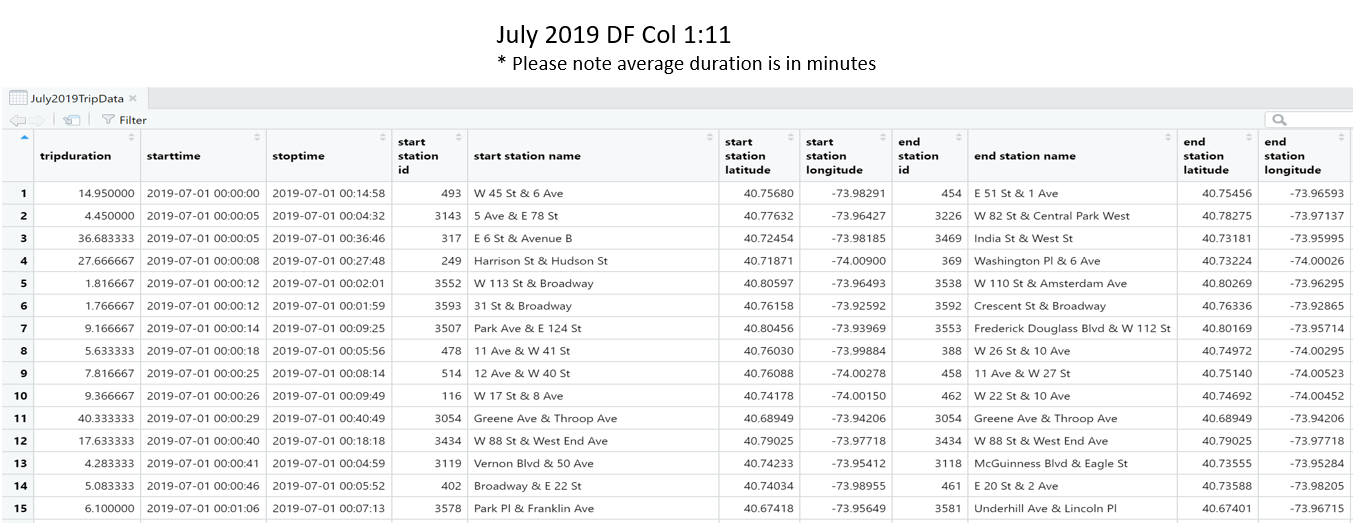
### 

### Appendix A



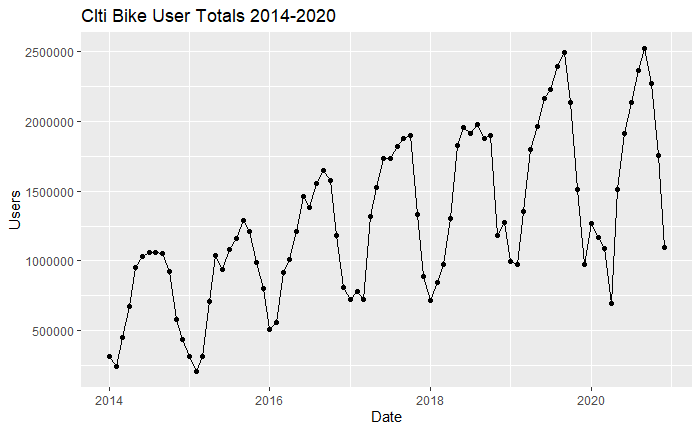
**Example of raw data supporting Citi Bike COVID-19 Impact Study**

**Appendix B**

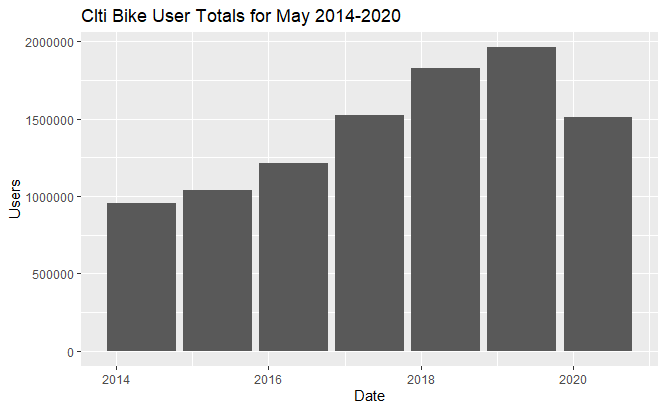


**Snapshot of Citi Bike Data imported into RStudio for analysis**

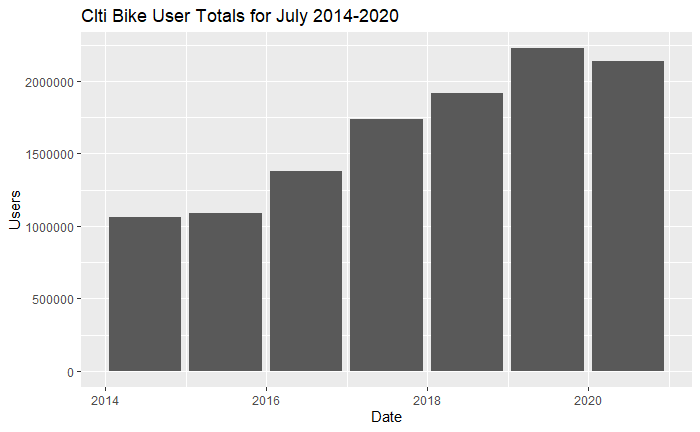
**Appendix C**

**Monthly user totals for Citi Bike 2014-2020 - Seasonal Variation**

**Appendix D**



**Total Citi Bike users for May between 2014-2020**



**Citi Bike Users for the Month of July 2014-2020**

1. <https://www.citibikenyc.com/how-it-works>, Citi Bike web page, “Experience NYC in a whole new way” [↑](#footnote-ref-1)
2. Citi Bike services Manhattan, Brooklyn, Queens, the Bronx and Jersey City [↑](#footnote-ref-2)
3. <https://www.citibikenyc.com/system-data/operating-reports>, Citi Bike Monthly Operating Reports. [↑](#footnote-ref-3)
4. Appendix A provides a PDF file of the R Markdown with further details on the code used and additional visualizations on Citi Bike user trends [↑](#footnote-ref-4)
5. Citi Bike home page, <https://www.citibikenyc.com/> and Citi Bike System Data <https://www.citibikenyc.com/system-data>. Citi Bike provides monthly “tripdata” in zipped excel files [↑](#footnote-ref-5)
6. Data for total Citi Bike Users in 2020 only goes through August. [↑](#footnote-ref-6)
7. This analysis is consistent with previous Citi Bike assessments highlighting peak hour statistics between Oct 2018 and March 2019. https://wbrueske.github.io/Citi-Bike-Analysis [↑](#footnote-ref-7)