Bike Sharing Usage During COVID-19

Yuchuan Huang
huan1531@umn.edu

1. Introduction

Bike sharing has become a popular transportation method in many places around the world. According to a survey by NACTO[1], bike sharing ridership in the US reached 50 million trips in 2019 (see Figure 1). However, as the COVID-19 started in the year of 2020, it is undetermined how the pandemic affects the bike sharing business. A first intuition is that the pandemic could both decrease or increase the bike sharing usage. On one hand, as many businesses shut down and people started to work from home, there would be fewer trips. On the other hand, compared with other public transportation like buses and light-rails, bike sharing has larger social distance and better air quality, which is crucial for preventing COVID-19. Hence it might also gain some growth.

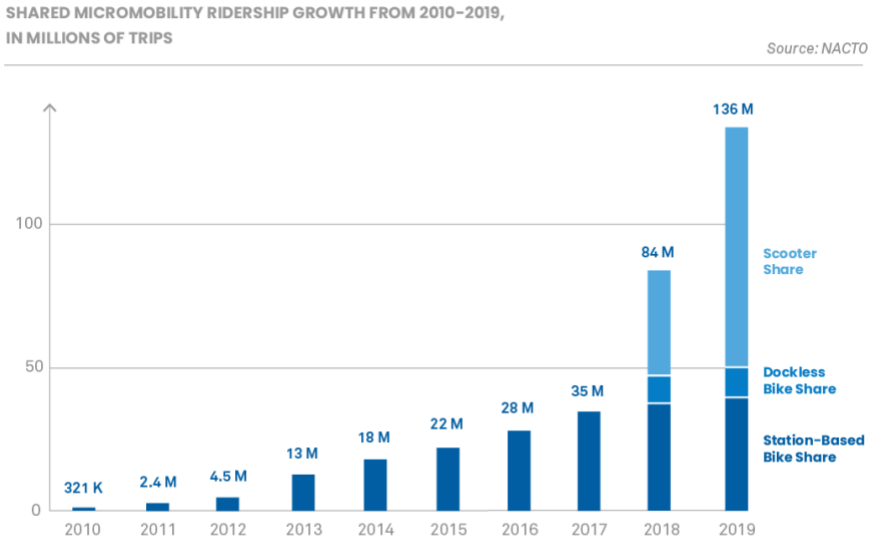

Figure 1. Shared micromobility ridership growth 2010-2019 (by NACTO)

Understanding the effect of pandemic on bike sharing has multifold benefits:

1. For bike sharing operators, they can better understand why and how their businesses grow/shrink. In this way, they can adjust their future business operations (for example, deploy more or fewer bikes) according to the spreading of COVID-19.
2. For governments, they can better understand how the public adopts bike sharing during pandemic. Therefore, the Transportation department can decide whether to increase or decrease the investment on bike sharing; the Health department can decide whether bike sharing is an important part of pandemic control.
3. For general customers, they can know whether bike sharing is still a reliable transportation during the pandemic and then make their trip choices.

[1] National Association of City Transportation Officials. https://nacto.org/shared-micromobility-2019/

In this project, we use the city of Minneapolis and its bike sharing operator, NiceRide[2], as a case to study how the pandemic affects the bike sharing business. This project assumes a major audience as the NiceRide company itself and focuses on the usage changing during COVID-19.

The rest of this report is organized as follows. Section 2 describes the data we find available to this topic. Based on the available data, Section 3 formally proposes the business questions we need to answer in this project. Section 4 gives an overview of the project's workflow, which is composed of three main components, i.e., data acquisition, data preparation and data analysis. Section 5 describes the methodology in these components respectively. Section 6 shows and discusses the results we find in this project. Finally, Section 7 concludes the technical lessons learned and challenges faced in this project.

2. Data Availability

NiceRide has been publishing its trip data[3] since 2010. Each trip record consists of the information like:

- Start/End datetime
- Start/End station: ID, name and location (latitude and longitude)
- User type: either single trip customer or annual/monthly subscriber
- Other miscellaneous information

In this project, we use the data from 2018 to 2021, which is two years before COVID-19 (2018 and 2019) and two years after (2020 and 2021). Please note that NiceRide only operates from April to November each year.

As for the dataset of COVID-19, the ideal data would be the daily update of new cases in Minneapolis. However, that data is not publicly available. Therefore, we use the daily new cases in Minnesota to show how we develop the method. The data consists of the new positive cases per day since March 5, 2020.

3. Business Questions

In this project, we try to understand the bike sharing usage change at different granularity. Specifically, we pose questions from three levels: global, station and trip

3.1 Global Level

1. How does the annual usage change before and after COVID-19?
2. Correlation analysis: does the NiceRide daily usage change with COVID-19 cases?

At the global level, we first want to understand how the amount of trips changes before and after COVID-19. Then we want to figure out if there is a significant correlation between the number of new cases and the number of bike sharing trips per day. Answering these two questions will give us a bird's-eye view of bike sharing usage.

3.2 Station Level

1. Which station's usage decreases most as a start station (Top 10)?
2. Which station's usage decreases most as an end station (Top 10)?
3. Which station's usage increases most as a start station (Top 10)?
4. Which station's usage increases most as an end station (Top 10)?

[2] https://niceridemn.com/

[3] https://niceridemn.com/system-data

At the station level, we want to find out how the usage changes at each station, so we find the stations with usage decrease/increase most. Please note that each station it can be a start or an end station of a trip, so we discuss the start and end stations separately. Answering these questions will give us a sense that in which areas bike sharing get more or less popular during COVID-19.

3.3 Trip Level

1. Which trip's usage decreases most (Top 10)?
2. Which trip's usage increases most (Top 10)?

At the trip level, we want to find which trips, i.e., pairs of start and end stations, decrease/increase most. Answering these questions will show that for what trips people tend to use or not use bike sharing during COVID-19.

4. Project Plan

Figure 2 gives an overview of the project process, as well as a flow of data transferring throughout the project. We first acquire the online data by downloading the NiceRide trip data from its S3 bucket and crawling the COVID-19 data from the website. We follow the "ELT" manner for data preparation: we import raw data into Spark directly after acquisition and then perform the data transferring against the Spark DataFrames. Then we conduct data analysis, answer the above business questions and visualize the results. All components in the gray box are conducted in the Databricks Notebook environment. All functionalities in the orange box are supported by Apache Spark.

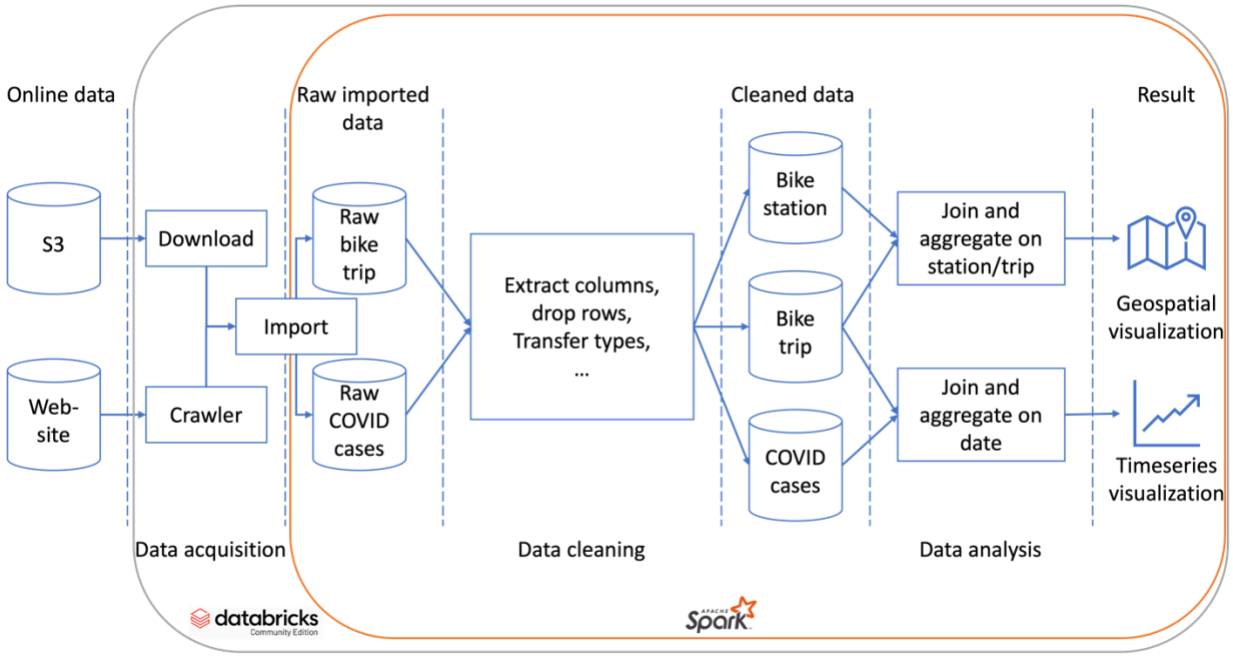

Figure 2. Project Overview

5. Methodology

In this section, we show the methodology we used in this project.

5.1 Data Acquisition

NiceRide Data. NiceRide publishes its monthly trip data on a public S3 bucket[4]. To get such data, we write a shell script using "wget" to download all data from 2018 to 2021.

COVID-19 Data. The Minnesota Department of Health updates the daily new cases on its website[5], as a table, and there is no downloading port for that data. To get such data, we write a web crawler in Python and use the "read_html" function from Pandas to perform the crawling.

5.2 Data Cleaning: ELT

NiceRide Data. The NiceRide data is in CSV files after downloading and unzipping. Before we load the data, we use the command line tool "head" to take a quick glance at the raw data and so we can specify the schema explicitly. Since the data in 2018 and 2019 are in different schemas as they are in 2020 and 2021, we cannot load all of them into a single DataFrame. Therefore, we load data and create DataFrame for each year.

The data cleaning target for NiceRide data is to have four bike trip tables, one for each year with following schema:

(trip_id, start_time, start_station_id, end_time, end_station_id, usertype)

And a station table with following schema:

(station_id, station_name, latitude, longitude)

The *start_station_id* and *end_station_id* in bike trip tables are foreign keys referencing the *station_id* of the station table.

There are three main data cleaning tasks for the NiceRide data:

1. Union data of four years. The first cleaning we do is to extract the desired common columns of the four years and then union the data into a single DataFrame, so the further transfer can be applied on all data. For simplicity, we drop all the rows that contain Null value. We also generate a uuid for each trip.

2. Extract station information. Station information is important to answer the station and trip level question, so we must check its quality. Unfortunately, station information has quality issues like: a) Same station name has different IDs in different years; b) Same station name has different latitudes and longitudes in different years. To clean up and unify the station information across the four years, we select the distinct station names as the full set of stations. For station ID, we drop the old ID and assign a new ID for each station. For latitude and longitude, we average the various values in the raw data.

3. Transfer usertype column. The usertype field has different values in different years: it is "customer" and "subscriber" in 2018 and 2019, but "casual" and "member" in 2020 and 2021. To unify the usertype across the four years, we use a user defined function (UDF) to transfer the column. The UDF returns 1 for "subscriber" or "member" and returns 0 for "customer" or "casual".

At last, we split the trip data into four years. We save the four trip DataFrames and the station DataFrame as tables in Databricks (using the function df.write.saveAsTable). This caches the DataFrames and saves time for later analysis.

[4] https://s3.amazonaws.com/niceride-data/index.html

[5] https://www.health.state.mn.us/diseases/coronavirus/situation.html

COVID-19 Data. The COVID-19 data is in good quality in general. The only issue is that we load the data to Spark from Pandas DataFrame so the Date column is loaded as string since it is not in a common format. We use the "to_date(col, format)" function in Spark SQL to transfer it to the date type. In the end, we have COVID-19 data in following schema:

(date, new_case_count)

5.3 Data Analysis

We use Spark SQL to answer all the business questions. Then we use Plotly in Python to visualize the results. In this section, we briefly describe the idea in writing each query. For specific Spark SQL queries, please refer to the Notebook.

How does the annual usage change before and after COVID-19? We count the number of trips for each year. We also count for casual trips and member trips respectively.

Does the NiceRide daily usage change with COVID-19 cases? We first count the number of trips for each day and then join on the date with the COVID-19 data.

Which station's usage decreases most as a start station (Top 10)? We count the number of trips starting from each station in 2019 (before COVID-19) and 2020 (after COVID-19). Then we rank the stations based on the difference in the two years and pick the top 10. The idea is the same for all station level questions.

Which trip's usage decreases most (Top 10)? We count the number of trips for each start-end station pair in 2019 and 2020. Then we rank the station pairs based on the difference in the two years and pick the top 10. The idea is the same for all trip level questions.

6. Results and Visualization

In this section, we visualize and discuss the result for each question. We combine the discussion for the station and level quesitons as they lead us to the same conclusion.

6.1 Global level

How does the annual usage change before and after COVID-19?

Figure 3. Annual NiceRide trips 2018-2021

Figure 3 shows the annual NiceRide trips from 2018 to 2021. Even before the COVID-19, the trips decreased significantly from 2018 to 2019. A main reason is the popularity of shared scooters in 2019. However, the decrease only happened in the casual trips. The member trips actually increased a little bit. Whereas, from 2018/2019 to 2020/2021, the number of member trips decreased significantly, due to the reason of COVID-19.

Correlation analysis: does the NiceRide daily usage change with COVID-19 cases?

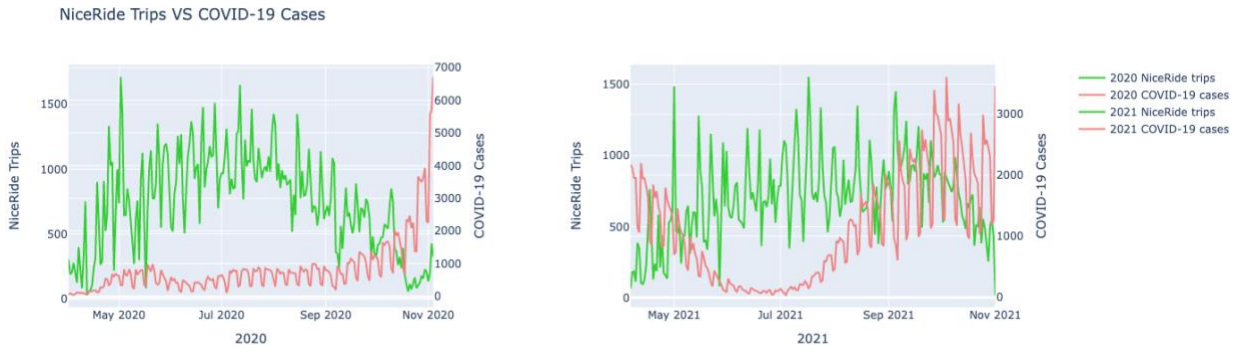

Figure 4. Daily trips and cases 2019 and 2020

Figure 4 shows the daily trips and cases during April to November in 2020 and 2021. In 2020, the number of COVID-19 cases were stable until October and went to a peak in November. The NiceRide trips were roughly stable for most of the time and then significantly went down in November, but it could also because of the weather condition. In 2021, the COVID-19 had a valley around July, while the NiceRide trips did not have significant change at that time. Generally speaking, we do not observe a significant correlation between the trips and the cases.

6.2 Station and trip level
Start, end stations and trips decrease most.

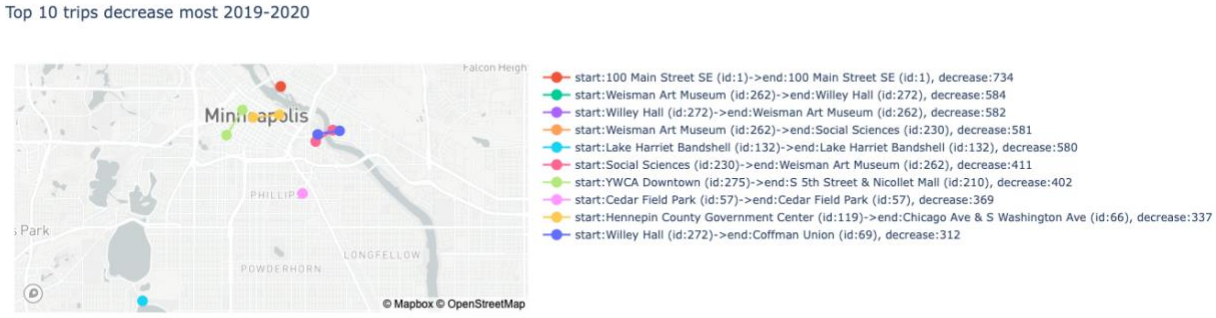

Figure 5. Top 10 trips decrease most 2019-2020

Figure 5 shows the trips that decreased most from 2019 to 2020. Figures for start and end stations are in the Appendex (Figure 7 and 8). Most of the decrease happened in the downtown area and in the University campus. This shows that the work/study-from-home policy is indeed a major reason for the decrease of NiceRide trips.

Start, end stations and trips increase most.

Figure 6. Top 10 trips increase most 2019-2020

Figure 6 shows the trips that increased most from 2019 to 2020. Figures for start and end stations are in the Appendex (Figure 9 and 10). Most of the increase happened in the park area, for example the Lake of the Isles and Bed Maka Ska. Another worth-noticing thing is that all the increasing trips are same origin-destination trips, which means that these trips are for entertainment, instead of commuting. This means that, during the COVID-19, some people do prefer riding a public bike in an outdoor place as an entertainment.

6.3 Conclusion

To sum up our findings in this project, COVID-19 has both negative and positive effects on NiceRide business. On the negative side, it does decrease many member trips and a lot of commuting trips in the downtown and university area. On the positive side, more people tend to use NiceRide as an entertainment or exercise method during the COVID-19. Therefore, for NiceRide, it could decrease their deployment in the downtown and university area and move more bikes to the park area.

7. Lessons and Challenges

The most important lesson I learned in this project is that data science is not only about analysis. I spent most of my time doing stuffs outside analysis: downloading, cleaning and transfering. Another take of this lesson is that, your data analysis work heavly relies on your cleaning quality. If you didn't spend enough time on cleaning the data, you will spend much more time in the analysis step. On the contrary, if your data is well cleaned, your analysis will be much easier. Therefore, it is always a good investment to clean the data before analyzing it.

Another lesson I leanred is that "ELT" does seem like a better manner than "ETL". I used to have the raw data in Pandas DataFrames and did the cleaning and transfering with Pandas APIs. Some APIs are hard to use, and I had to write long code. Later I imported the raw data into Spark directly and did the cleaning and transfering with Spark SQL. It is a much better experience since Spark is more powerful, and SQL is more concise and readable.

A challenge I faced is the efficiency problem of using the community edition of Databricks. I used to wait for about 5 minutes for each analysis result. Because of the lazy execution of Spark, each analysis runs from the cleaning step. Later I cached the clean data as tables in Databricks. The caching still takes time, but the analysis is much faster.

Appendex

Figure 7. Top 10 start stations decrease most 2019-2020

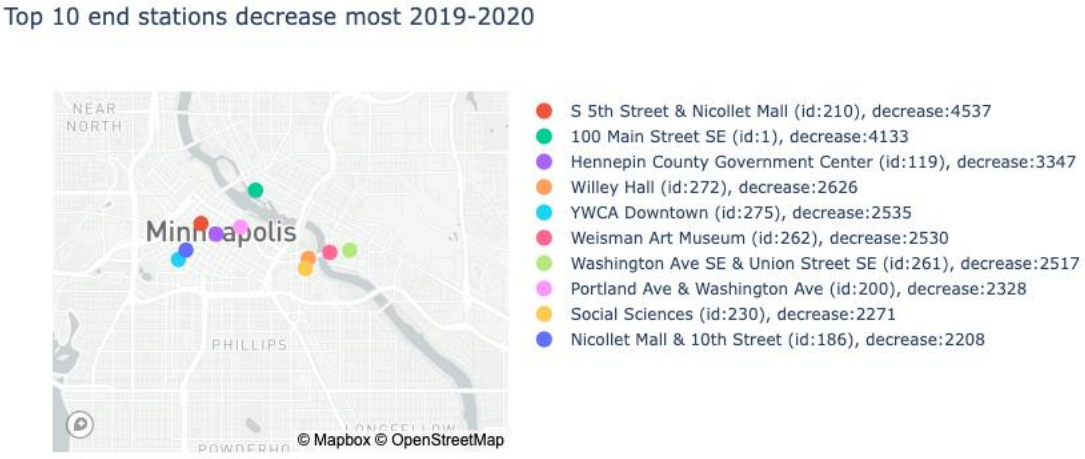

Figure 8. Top 10 end stations decrease most 2019-2020

Figure 9. Top 10 start stations increase most 2019-2020

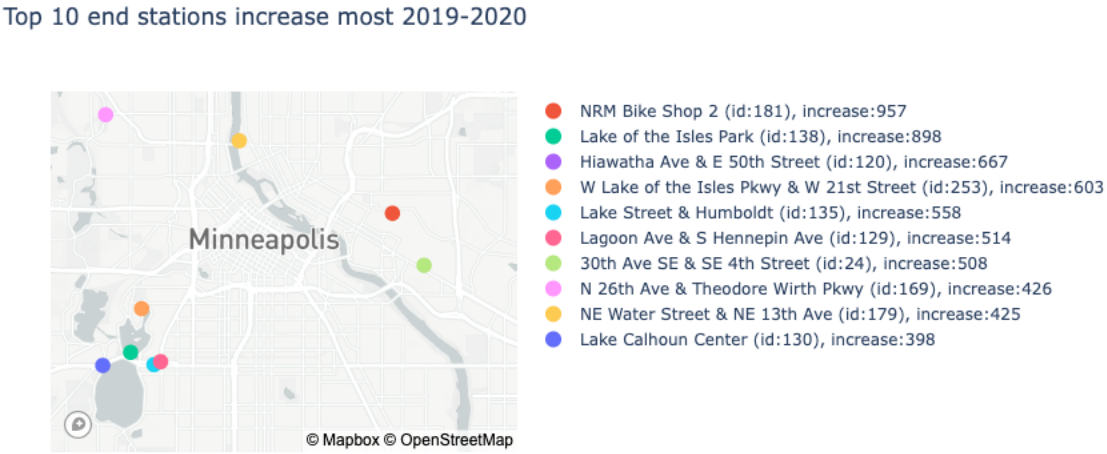

Figure 10. Top 10 end stations increase most 2019-2020