**A glance on the Bike-Share Company: *Cyclistic***

*By: Bruno Trevisan*

Brazil  
2022

Topics

[1. The Business Task 3](#_Toc101538772)

[1.1. Context 4](#_Toc101538773)

[1.2. What is to be done 4](#_Toc101538774)

[2. Data Used and Metadata 4](#_Toc101538775)

[2.1. Data Used 4](#_Toc101538776)

[2.2. Metadata 4](#_Toc101538777)

[3. Data Cleaning and Manipulation 4](#_Toc101538778)

[3.1. Tools Used 4](#_Toc101538779)

[3.2. Discovering the data 4](#_Toc101538780)

[4. Data Viz and Key Findings 4](#_Toc101538781)

[5. Recommendations 4](#_Toc101538782)

[5.1. Membership Option 4](#_Toc101538783)

[5.2. Digital Media 4](#_Toc101538784)

[6. Future Projects 4](#_Toc101538785)

# The Business Task

## Context

In 2016, *Cyclistic* launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

## What is to Be Done

This company now has the necessity of a better understanding of their entrepreneurship. For that reason, I downloaded, cleaned, and analyzed their data to see what kind of tendencies their users have.

There are three main questions to be answered:

* How do Casual riders (single-ride and full-day users) and members (annual memberships) use the bikes differently.
* Why would casual riders buy *Cyclistic* annual memberships?
* How can *Cyclistic* use digital media to influence casual riders to become members?

# Data Used and Metadata

## Data Used

The data used is available in the following link:

*<https://divvy-tripdata.s3.amazonaws.com/index.html>*

Every month start new data is uploaded from the previous month from the Bikeshare service. I didn’t collect the data, and I’m not part of that company, all I did was the cleaning and analyze part. This work is under this [*License*](https://ride.divvybikes.com/data-license-agreement).

This is the way, after downloaded, the data was organized inside the folder:

Text

Description automatically generated with low confidence

*It is important to say that each file was downloaded individually. As I wanted to make an analysis month by month and along the year, I selected that specific files.*

## Metadata

* The data comes in the .csv format
* The variables are: **ride\_id, rideable\_type, started\_at, ended\_at, start\_station, end\_station, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual.**

*Where:*

**ride\_id**: Character type variable. It Is a code generated individually for each bike rented.

**rideable\_type:** Character type variable. There are three types of bikes: Docked Bike, Classic Bike and Electric Bike.

**started\_at:** Date-Time object. Date and time that bike was rented.

**ended\_at:** Date-Time object. Date and time that bike was delivered.

**start\_station:** Character type variable. Street or avenue the bike was rented.

**end\_station:** Character type variable. Street or avenue the bike was delivered.

**start\_lat:** Exact latitude coordinates the bike was rented.

**start\_lng:** Exact longitude coordinates the bike was rented.

**end\_lat:** Exact latitude coordinates the bike was delivered.

**end\_lng:** Exact longitude coordinates the bike was delivered.

**member\_casual:** Whether the person is casual or an annual member cyclist.

# Data Cleaning and Manipulation

## Tools Used

I used two different tools to work with this part:

* RStudio
* Power Bi

What I’m going to share is my thought process and how did I approach them with the tools I’ve chosen. And they are listed in the sequence I thought at that specific moment.

## Discovering the data

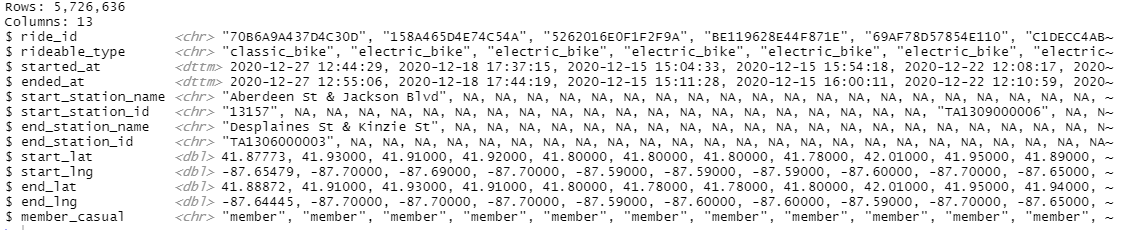
To load the data inside RStudio, I wanted a solution that was easy and would allow me to combine all the csv files into only one dataset. Since they have the same variables and structure I was able to use the following code:

*data = dir(pattern = "\*.csv", full.names = T) %>% map\_df(read\_csv)*

As you only can see the rainbow if you see the storm first, I wanted to see the kind of storm I had, and what “devastations” it could bring to my analysis if not anticipated.

This code below allows me to see how my data is organized and some useful information, like the type of the data and *NA* values.

*data %>%   
 glimpse()*



This is the result after the code was executed.

It helps me to understand what kind of structure and values I have. First the number of rows and columns it has. It also shows the variables and its format. What I could see also, was some *NA* values, which might not be good for some analysis. What I know is that these *NAs* appear in *cases like:*

* Incorrect typing
* Field not filled
* Some calculation with an impossible answer, like a division by 0.

Even I can’t see any other variable showing *NA* values, that does not mean there isn’t other variables with *NAs,* because I cannot see the entire dataset. To be sure that I have only variables with this behalf, I decided to filter based on a specific criteria.

data %>%

filter(!complete.cases(.)) %>%

View(.)

This is the code used, and here is the result:

Graphical user interface, application, table, Excel

Description automatically generated

With the data filtered, I am now sure that all the data I’m seeing have at least one variable with a *NA* in it.

Based on the information I had, I could not affirm what was the case for these *NAs* to appear, but since I need all the variables to be complete for the analyze I wanted to do, I decided to drop those rows with *NAs*. It is important to note that this type of view, shows how much rows this temporary dataset has, so, dropping all that rows will affect the analysis, even though I don’t like it, it will help for a complete analysis.

After dropping these *NA* values, I wanted to add a variable that represented the difference between the ride start and end time. Here is how I approached it:

*data\_cleaning\_dates$duration <- data\_cleaning\_dates$ended\_at - data\_cleaning\_dates$started\_at*

Since the data class was already in the right format (*POSIXct* and *POSIXt)* which is the common class for date and time variables in R I was able to do that.

But as soon as I did it and checked the results, I realize there was a problem, some of the results end up being negative. The reason for that is because some of the *ended\_at* variable, were apparently swapped with the *started\_at,* that could have happened because of a system error input or typing. Even with this “error”, the value would be the same but positive, so instead of just dropping these values, I decided to swap back these cases where my new *“duration”* variable was negative. This is how I did it:

*for (i in seq(1:nrow(data\_cleaning\_dates))) {*

*if(data\_cleaning\_dates$ended\_at[i] < data\_cleaning\_dates$started\_at[i]){*

*temp <- data\_cleaning\_dates$ended\_at[i]*

*data\_cleaning\_dates$ended\_at[i] <- data\_cleaning\_dates$started\_at[i]*

*data\_cleaning\_dates$started\_at[i] <- temp*

*}*

*}*

There might be a better way to do this, but that’s how I solved this problem. After that I realized that there were some results with a duration of 0 seconds, 1 second, 2 seconds, and so on. I had some assumptions about why this happened, like:

* System error at the moment the bike was rented.
* The rider just changed their mind after renting the bike, and just returned it to the bike station

But since I’m not sure about it, I decided to filter the rides duration under 5 minutes (300 seconds), that might be wrong to do this, and it is all about my interpretation, but what came to my mind is that in a short distance no one would rent a bike, and I considered to be like they have changed opinion or had some issue during the ride.

To filter these rows out, first I had to transform it into a numeric format, since they were in a *“secs”* format, and I needed just the numbers itself. The filtering of these values was made inside Power Bi but I changed the number format in R.

*data\_cleaning$duration <- as.numeric(data\_cleaning$duration)*

And then renaming the column name.

*data\_cleaning <- data\_cleaning %>%   
rename(duration\_seconds = duration)*

Satisfied with the dataset I got, I saved it in a csv format so I could take it inside Power Bi for some other manipulations and analyzing.

This is how I did it:

*write.csv (data\_cleaning\_filtered, "divvy-tripdata-cleaned.csv", row.names = FALSE, sep = ",")*

So, after running this whole code, you should get a new .csv file with no *NA* values and the modifications I wrote here.

Now with the dataset inside Power Bi, for analysis purposes, I decided to create two tables:

* A calendar table
* A time table

If I was to create new columns to the original dataset which is almost 5 Mi rows long, it would be 5 Mi more rows for each column that I’d create, so for that and other reasons it’s much better to make a new *table* with unique values for this kind of data.

This is how I approached it.



This is the Dax function I used to create the Calendar table.

Table

Description automatically generated

And that is the table after some additional information I wanted that I would be using later in the analysis step. Each Column was created also using Dax.

The timetable was a little more complicated.

Table

Description automatically generated

First, I created my “base” column which was all the minutes we have in a day (0 to 1439, completing 1440 minutes). The second column is the result of a division of the first column by 1440, which gives me some fractional numbers, and then I simply transformed it into *Date which gave me the “Time to the Minute”* column. After that to create the *bucket* columns I just used the Divide (integer) option in Power Query and put 30 and 60 respectively, that would be my base columns for the *interval* columns which would be the actually categorized time values. Then to create that *interval* columns I used the *Custom Column* option, and inputted the following function:

*[30\_bucket] \* 30 / 1440 = 30\_minute\_interval*

*[Hourly\_bucket] \* 60 / 1440 = Hourly\_minute\_interval*

*And that gave me the 2 columns I wanted for the analysis. Here are the 3 tables together*

Graphical user interface, text

Description automatically generated

After that, doing the data modeling is my next step.

Graphical user interface

Description automatically generated

With that I can start my analysis.

# Data Viz and Key Findings

So, after a few explorations on the data, I found some interesting results:

Graph 1: Weekly Quantity fluctuation Casual vs Member Riders.

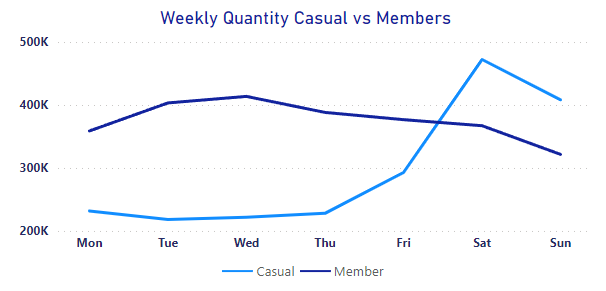


Figure 1: Casual vs Members Ride Quantity Along the Week

As expected, casual riders are more likely to go after bikes to rent in the weekends, probably for a recreational reason. The members on the other hand, are more “stable” along the week, since they probably use the bike as a transport and commute to work. They have slightly drop on weekends since they are not working, but still, most of them might use it for fun.

**Graph 2: Annual Comparison Between Member and Casual Riders**

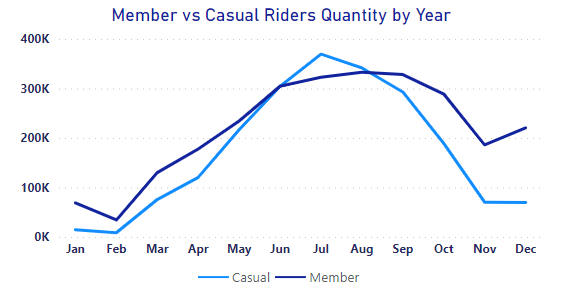


Figure 2

Here we can see a tendency for a higher demand starting in April and going through September, which also makes sense, because this time of the year is when Chicago has the highest temperatures and the lowest wind speed in average making the ride more appreciable (*source:* [*https://www.weather-us.com/en/illinois-usa/chicago-climate?c,mm,mb,km*](https://www.weather-us.com/en/illinois-usa/chicago-climate?c,mm,mb,km)*).*

Okay! But… What about along the day? What might be expected? Let’s see!

**Graph 3: Distribution of Rides along the Day.**

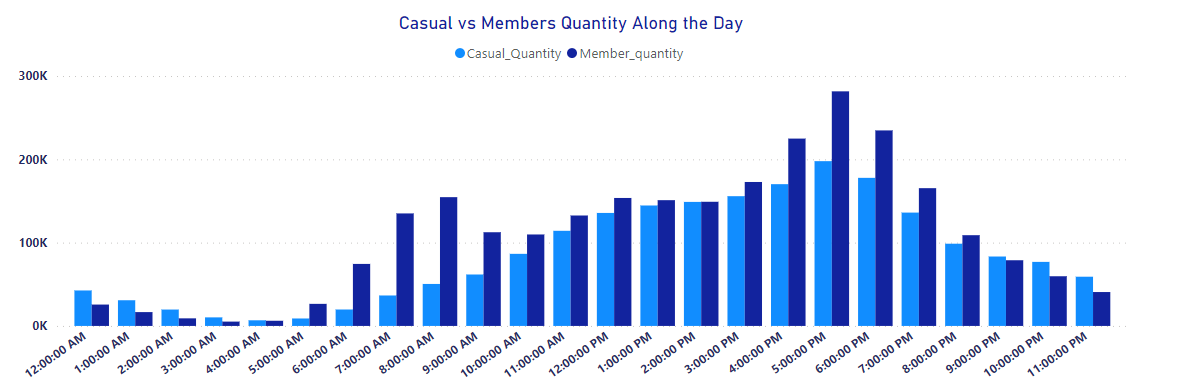


Figure 3 Daily Fluctuation of Riders

It looks like the members are more likely to use the service when returning from home when getting out from the work then using it going to work. But equally, the casual riders have a slightly higher demand at the same hour (between 4 and 6 PM). They might use as a commute to home as well. But members tend to use the bikes in the morning more compared to casual riders.

**Graph 4: Rideable type Usage in Quantity**

Chart, bar chart

Description automatically generated

Figure 4 Bike Types Comparison

Here we can see that even though the members and casual riders use the classic bike more, casual riders use the docked type much more. Which with further exploration and data might be a key difference.

**Graph 5: Difference in Minutes of Ride duration.**

Chart, bar chart

Description automatically generated

Figure 5 Average Rides Duration.

On this graph, I think is the major difference between them. Casual riders because of their tendency to use the bikes for leisure, instead of using it for commuting to work for example. Here I found some outliers values like a ride duration of 38 *days*, although, testing the average, even with these extreme values the impact in the average value was around 10% for both Casual AND Members maintaining the proportionality. The cause of these extreme values is unknown since I didn’t have enough information on the business as a whole but one thing that crossed my mind was that some bikers might have problems when returning the bike to the station, whether for a system problem or else, or some of them can just let the bike anywhere and live, until someone take it back to the station.

Chart, line chart

Description automatically generated

Figure 6

Confirming the Figure 5 along the week we can see the average duration is kind of stable

# Recommendations

## Membership Option

The major difference between Casual and Member riders seems to be the duration of the ride and days of the week usage as we can see on Figure 5, since they use the service for different purposes. Members tends to use it along the week, and Casual riders use the bikes more on weekends, but when using it, Casual riders tend to ride for a longer time then the Members.   
  
Also, there might be a tendency of growing demand between March and July, a good explanation to that is that between these months the temperatures are higher compared to other months and lower wind speed as well.

With all that said, I can see that instead of having just one kind of membership, *Cyclistic* could have another option that makes it worth for casual riders that use their service for a longer duration but for less days in the week. Could be a cheaper version of the annual membership which favors the usage duration instead of the quantity, because those who pays a full-day pass or single-ride pass, can’t see any advantage of paying a membership since they won’t enjoy it at its fullness.

## Digital Media

Although it might be a little bit of a cliché, but one way *Cyclistic* could use digital media is by warning the population about pollution by CO2 emissions and stress that affect life quality in a negative way, and both are increasing as the time goes which is true, and with that the benefits of using the bikes to exercise more. (Source: https://www.bbc.com/news/science-environment-59148520 and https://sitn.hms.harvard.edu/flash/2020/a-stressful-new-decade-the-latest-information-on-how-stress-shapes-our-minds-and-bodies/ ).

# Future Projects

When talking about business decisions and analysis, there is a whole world to explore and Key Performance Indicators that can become handy. Thinking about it I’ve come with some ideas that would need further data collection to be able to accomplish in order to go even deeper in the analysis, which are:

* Data about user, like a unique ID for that specific rider and maybe their age as well would be of a good fit for analyzing about how some a specific group of them uses the service, and with that be able to redirect marketing strategies based on that, making them more efficient.
* A better description about the bike used, if it was a cargo bike, an inclusive one (for people with some disability), etc. In parallel with the above topic, it would help to make a more detailed analysis about the riders.