Cyclistic Case Study

Ian Morris

2022-03-25

# Introduction

Hello, my name is Ian. This is a case study I’ve completed for the Google Data Analytics Career Program. It is for a fictional company, Cyclistic, and uses real world data from [Divvy](https://divvybikes.com/) under this [license](https://ride.divvybikes.com/data-license-agreement). This is the first case study I’ve completed and it’s been an invaluable learning experience. Unfortunately I did run into some blocks that I was unable to overcome under the circumstances. Specifically these were in cases when I needed more information about the data and usage of bikes, but was unable to get help from classmates or mentors, or find the answers online. There is also definitely more that I would like to do and explore with this data or the creation of case studies in general, and I hope to build upon this foundation in the future. Thank you for your interest in my work.

## Background

Cyclistic is a bike-share program in the Chicago area, featuring more than 5,800 bicycles and 600 docking stations. There are three methods of purchasing the usage of a bike: single-ride passes, full-day passes, and annual memberships. Single-ride pass and full-day pass users are collectively referred to as “casual riders.” The company’s finance analysts have determined that annual members are significantly more profitable than casual riders. The marketing director believes that maximizing the number of annual members will be the key to future growth and that there is a strong possibility of converting casual riders into annual members. They’re interested in analyzing Cyclistic’s historical bike trip data to identify trends.

## Task

Determine any key differences in bike usage between the two major types of customers, annual members and casual riders, allowing Cyclistic to understand why casual riders would purchase an annual membership. Following that, Cyclistic can then determine the best ways to encourage casual riders to become annual members, increasing growth.

## Data Preparation

The data is the previous 12 months of Cyclistic bike trip data, from March 2021 to February 2022. It is initially organized into 12 separate excel spreadsheets, 1 for each month. The data for each ride includes a unique ride id, bike type, ride start and end date-time, ride start and end station name and id where applicable, ride start and end latitude and longitude, and membership type. Having looked through some of the spreadsheets, it is apparent that there are some potential issues with the data, such as missing values in the start and end station name and id columns.

All personal information was removed prior to accessing the data to follow data privacy guidelines. The data was stored in a cloud service for ease of access. No further security steps were taken as this is a publicly accessible data set with no personal information. Backups of the original data were stored in a separate folder.

## Data Processing

As I began to review the data, it became apparent that some information, including station data and end gps coordinates, for some rides is absent. There are also some unusual starting position gps coordinates that appear to be significantly far from the Chicago area. I chose to use R for this case study, as it offers all of the tools I’ll be needing to process and analyze the data, then compile it into a usable document for review by stakeholders.

### Preparing to work in R

I’ve loaded the tidyverse, janitor, lubridate, ggrepel, and hms libraries into R for working with the data set. Following that, I’ve loaded the spreadsheets from each month, and then combined them into a single data frame. I then reviewed the structure and values of the data to better understand the data and look for any obvious anomalies.

### Further investigation of data anomalies

Reviewing the data, there are some notable anomalies. There are 4,617 null (missing) values for the end latitude and longitude. There are also null values for the station name and id for many entries. There is also an anomalous maximum value for the starting latitude and longitude.

For the start and end station names and id’s, there are about 750,000 null values in each of those 4 columns, roughly 12% of the dataset. I’m interested if there’s any additional unifying characteristic for these, or if perhaps the station values can be added. I’ve reached out to my “colleagues” (classmates and mentors) for assistance to see if they know why these values are missing. My initial hypothesis was that my assumption that all bikes had to be docked or undocked from a station was incorrect, but after further investigation, I found that bikes are required to be docked at the end of a ride. This leads me to the hypothesis that the station names were simply not recorded for these entries. I then checked to see if there was any relationship between gps coordinates and stations already listed in the dataset, but there doesn’t appear to be. I checked if membership type might have any relationship, and found it was a fairly even split of 45% casual riders and 55% members. I checked if the bike type was related, and found that all missing start and almost all missing end stations were for the electric bike type. Later, after calculating new fields, I checked if there was a correlation with the time period. Looking at months first, I found that the null values correlate roughly with the number of rides, so if there are more rides, then there are more missing station values. Going forward I chose to keep these records in the dataset, as they still contain relevant information for analysis.

With the missing end location gps coordinates, I also reached out for assistance to see if there was a known explanation for these, but unfortunately I haven’t heard back yet as of writing this. I looked into these entries further, and I found that many of them have ride lengths in excess of 24 hours, which seems to be an unlikely amount of time to be riding a bike. These entries do not seem to be reliable sources of information, and I have proceeded to remove them. It might make sense to adjust that cutoff, making it 12 hours instead of 24, or perhaps an even lower number. I will discuss this further below.

The anomalous start location that is far outside the other values appears to be from a test station. It is only one value, and can be safely removed without affecting the results of the analysis. This should be done, given that it appears to be a test case outside of the data we’re interested in.

The results of this initial processing were saved as a new data frame before proceeding to preserve the original data.

### Creating new columns

To proceed with the analysis, there are some additional values I’m interested in that need to be calculated. These are the start and end times for each ride, created from the started at and ended at columns that have date and time information. There is the ride length, which is the time difference between the start and end times. There is the change in latitude, change in longitude and change total fields, created by finding the difference between the start and end gps coordinates and finding the distance between the two points. There is the starting and ending day of the week, found from the date-time column. There is the start month, which is the month the ride began in, and in most cases the month the ride ended in.

### Checking newly calculated values

I wanted to check the newly created columns, ride length in particular, to see if there was anything unusual. I found that there are about 1,300 rides longer than 24 hours, and 4,000 rides that are longer than 12 hours. I can’t say for certain the context these rides occurred in, but they are a relatively small portion of the overall data. These values seem very unrealistic. I also see that most of the missing end location gps coordinates are present for these unusual ride length entries. For entries with values over 24 hours in ride length, 2,721 of the 4,038 entries have the missing coordinates mentioned earlier, which is one of the reasons I chose to remove those entries.

Reviewing the Cyclistic’s (Divvy’s) information informed me that rides typically max out at about 3 hours, which is the longest ride you can take without incurring a fee, and that is available with the day pass, that ends at 24 hours. Members can have unlimited 45 minute rides, and single rides are 30 minutes long before incurring a fee. I have chosen to retain the other entries with ride lengths over 24 hours, as while those values seem very unrealistic, and likely values past the 12 hour mark are as well, there is no clear cut-off as to when a ride becomes unrealistic. Ideally I would contact colleagues or management for further information at this point to see if they know more or have a preference about these types of values.

I also found that there are some ride lengths with values less than 0. These are not realistic, and appear to occur for rides with an end time 1 second before the start time. These will not provide any insightful information and have been removed.

## Analyze

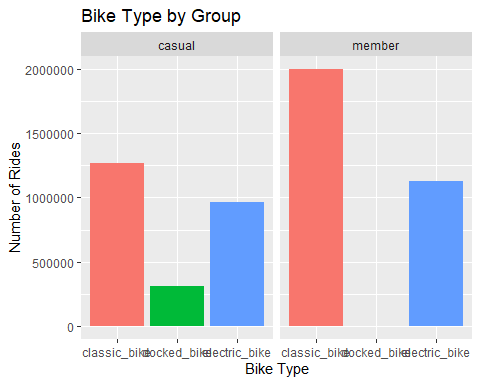
I found several trends in the data. Most rides and most casual rides occur in the summer. There are still some casual riders in the winter, but they’re a very small portion of the total rides. Year round, most casual rides occur in the afternoon. Casual riders prefer the weekends, while members appear to have a slight preference for weekdays. Casual rides tend to be longer than member rides, more than twice as long on average, and with a median length of about 16 minutes compared to 9.5 minutes.

Casual riders have a definite preference for a few specific station locations, namely Streeter Dr & Grand Ave, Millennium Park, Michigan Ave & Oak Street, Theater on the Lake and Shedd Aquarium. Members do not share these station preferences.

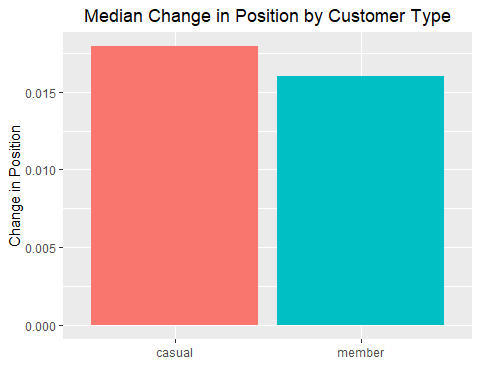
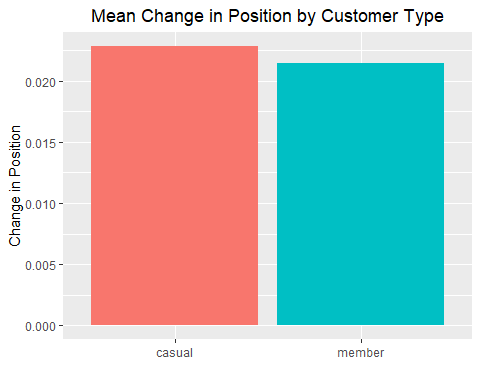
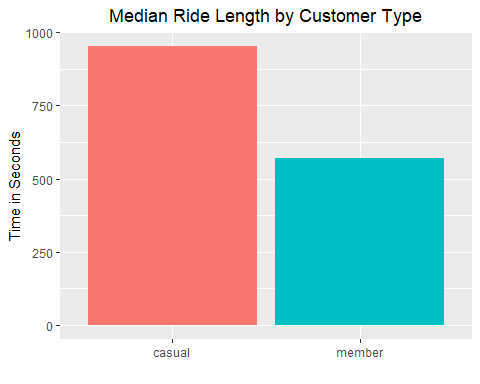
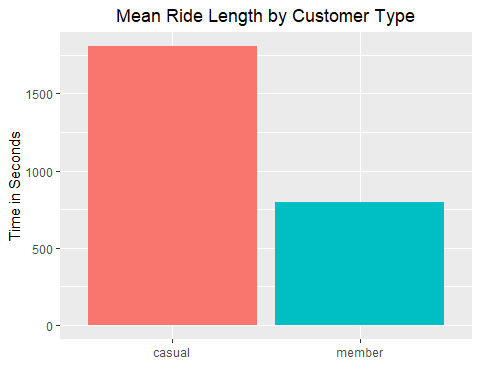
## Share

### The What

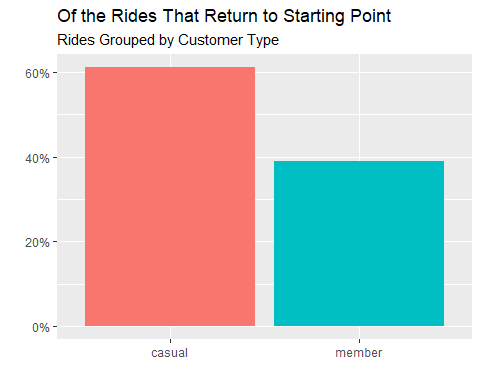
The first main question I asked about the difference in bike usage between rider types was: What? What does each type prefer to do with the bikes? First I checked what bikes each group prefers. Only casual riders use docked bikes. Beyond that, casual riders show a slight preference for classic bikes, while members show a strong preference for classic bikes.



Following that, I checked the ride length and change in position values for both groups. With ride lengths, the average and median of casual riders is significantly higher, but the values vary much more than with members. With the change in position, which is the change in gps coordinates from start to end of a ride, there is a slightly greater change with casual riders than members, but it is much less pronounced than ride length. So there seems to be a general trend of casual rides being longer in time, but not going much farther in distance.

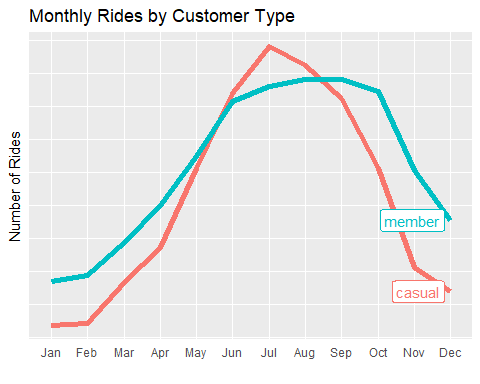


I also found that many rides begin and end at the same location. I checked this by looking for rides where there was no change in the gps position. We can see that roughly 60% of this group were casual riders and 40% were members.

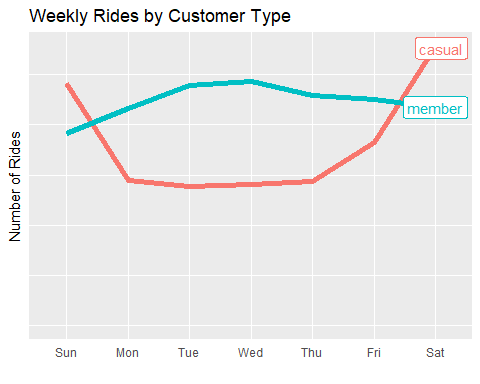


### The When

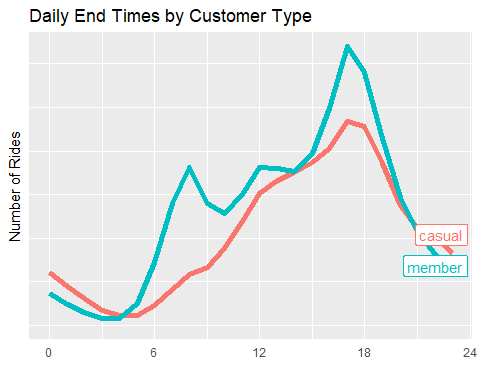
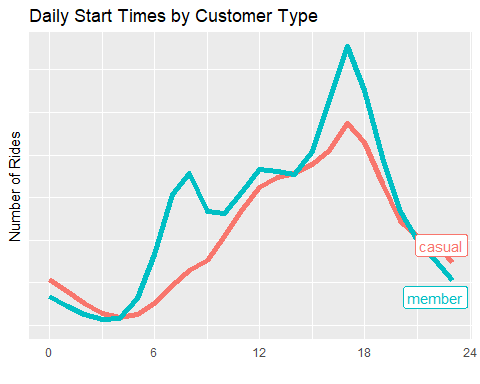
The next main question I asked was: When? When does each type prefer to use the bikes? I looked at monthly, weekly, and daily preferences. On the monthly chart, we can see that most rides occur in the warmer summer months. Casual rides in particular occur mostly from May to October. Both groups decrease during the colder winter months, but casual rides decrease much more significantly than member rides during that time.



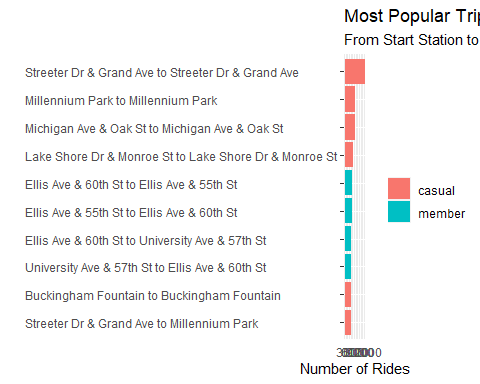
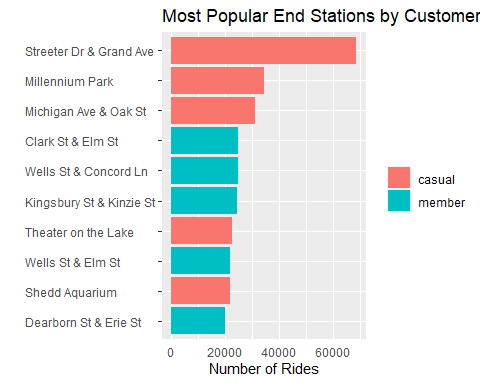
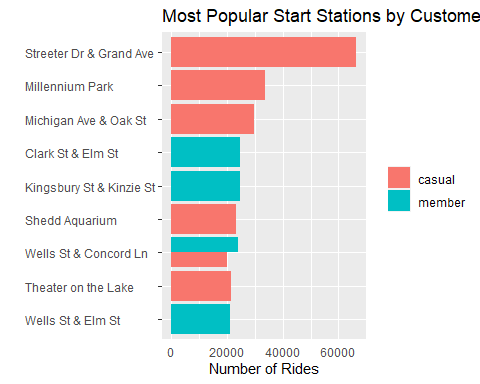
Following the monthly trends we have the weekly trends. Here we see a clear preference by casual riders for the weekend, with higher numbers on Friday, Saturday, and Sunday, and lower numbers during the other weekdays. The trend is much more even for member rides, with roughly the same numbers each day with a slight preference for weekdays.



The daily preferences have some broad trends as well. Both groups have peak rides in the afternoon and evening, but members also have many rides in the morning. We can also see there isn’t a substantial difference in how many rides are beginning or ending in a given time period.



### The Where

The last main question I asked on was: Where? Where does each type prefer to use the bikes? I looked at the top stations for each group. Both have different preferences here. The five most popular stations for casual riders are Streeter Dr & Grand Ave, Millennium Park, Michigan Ave & Oak Street, Theater on the Lake and Shedd Aquarium. Streeter Dr & Grand Ave in particular is significantly more popular than other stations. While there are some stations that are more popular with members, the difference is not nearly as large. We can also see a preference when looking at specific start and end stations for each trip. We can see that the four most popular trips are with casual members, and they all begin and end at the same station. With members, their four most popular trips are paired. It appears that they tend to go to and from these locations, but I’m unable to prove that without being able to identify individual riders. I do believe that is most likely the case. 

It should also be noted that this does not reflect all of the data, as roughly 750,000 start and end station values were not recorded for electric bikes, which is roughly 12% of the overall data.

## Act

My recommendations are to focus on the when first. Members use Cyclistic bikes more consistently through the year, week, and day. From our data we also know when most casual rides occur as well, how most occur during warmer summer months, on weekends, and in the afternoon and evenings.

Second, we also know that members prefer classic bikes compared to ebikes, and that casual rides have a much lower preference for classic bikes while also being the only users of docked bikes.

Third, we know that some of our stations are significantly more popular with casual riders, and that member rides don’t have as much of a focus on specific stations. We also know that casual riders will tend to return to their start location more, and that there isn’t a very pronounced difference in the change in position with each group.

With all of this in mind, we should be able to begin answering the next question in our marketing program: Why would casual riders buy Cyclistic annual memberships?

### Additional Data

Having explored the available data thoroughly, there are some questions that additional data could help in solving. For example, we don’t know exactly how long a path riders have traveled. Perhaps it is different between the two groups. If there was user id data to track usage on a per user basis, that would be very useful, but we would definitely need to be careful with that information and keep data privacy in mind, which might preclude our ability to use it.