Capstone Project: Bike-Share

Stephen Banks

2023-10-29

Cyclistic Bike-Share Analysis

Introduction:

The stakeholders at Cyclistic would like to know how casual and member riders use Cyclistic bikes differently, with the aim to convert more casual riders into members.

Before getting into the data, here are my assumptions:

- Members ride the bikes more frequently than casual riders, and they mostly like use the bikes primarily Mon-Fri to commute to work.
  - I also assume that they will be using their membership to utilize the bikes on weekends, although they may not use them as much as during the regular work week.

  - Because members are more likely commuters I would expect to see a higher use out of the E-bikes as long as they are available. I assume this would be the case because most people don't want to arrive at work sweaty and have to change or shower.
- Casual riders most likely use the bikes more on weekends and less frequently during the work week.
  - I would also expect that the casual riders are more likely to use more of a variety of the bikes offered.

  - Assuming the casual riders are more likely to use the bikes on the weekend, then I would assume that the duration of the bike usage one weekends would be higher than the members who are most likely commuters.

  - I also expect that casual riders would use the bikes for longer periods of time because they aren't trying to get somewhere by an expected time and they are using the bikes more for pleasure.
- Overall, I expect there to be some seasonality because this bike company is based in Chicago, IL and it is difficult to ride bikes in the snow. In addition to the snow which makes it difficult to ride, the cold weather doesn't make it any easier.
  - I would expect to see higher rides during the colder seasons for members vs. casual riders just because the members have already paid for their bike usage.

Loading Packages

```r
library(tidyverse)
library(dplyr)
library(readr)
library(lubridate)
library(ggplot2)
```

Aggregating the Monthly CSV Files Into One Dataframe

```r
combined_csv <- list.files(pattern = "\\.csv$", full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
```

Data Cleaning Time

```r
all_data <- na.omit(combined_csv)
all_data <- select(all_data, -c(start_lat, start_lng, end_lat, end_lng, ride_id))
all_data_clean <- subset(all_data,
                         !start_station_name == "HQ QR"&
                          !difftime(ended_at, started_at) <0)
```

Data Manipulation

```r
all_data_clean <- all_data_clean %>%
  mutate(start_station_end_station = paste(start_station_name, end_station_name,
                                           sep = " to "),
         start_id_end_id = paste(start_station_id, end_station_id, sep = " to "),
         ride_duration = difftime(ended_at, started_at),
         ride_duration_minutes = as.numeric((ride_duration), units = "mins"),
         date = as.Date(started_at),
         month = format(date, format = "%b"),
         day = format(as.Date(date), "%d"),
         year = format(as.Date(date), "%Y"),
         day_of_week = format(as.Date(date), "%A"),
         weekday = wday(started_at, label = TRUE))

all_data_clean$day_of_week <- ordered(all_data_clean$day_of_week,
                                      levels = c("Sunday", "Monday", "Tuesday",
                                                 "Wednesday", "Thursday",
                                                 "Friday", "Saturday"))
all_data_clean$month <- ordered(all_data_clean$month,
                                levels = c("Jan", "Feb", "Mar", "Apr", "May",
                                           "Jun", "Jul","Aug", "Sep", "Oct",
                                           "Nov", "Dec"))
```

Reviewing the Reorganized Data

```r
summary(all_data_clean$ride_duration_minutes)
summary(all_data_clean$rideable_type)

aggregate(all_data_clean$ride_duration_minutes ~ all_data_clean$member_casual,
          FUN = summary)
aggregate(all_data_clean$ride_duration_minutes ~ all_data_clean$member_casual +
             all_data_clean$day_of_week, FUN = summary)
```

Further Reviewing Data & Adding Tables for Charts

```r
result <- all_data_clean %>%
  group_by(member_casual) %>%
  summarize(count = sum(ride_duration_minutes > 2*60))
result
```

```
## # A tibble: 2 x 2
##   member_casual count
##   <chr>         <int>
## 1 casual        51522
## 2 member         3522
```

```r
total_duration_count <- all_data_clean %>%
  group_by(member_casual) %>%
  summarize(total_count = n())
total_duration_count
```

```
## # A tibble: 2 x 2
##   member_casual total_count
##   <chr>               <int>
## 1 casual            2048335
## 2 member            2539851
```

```r
outlier_percentage <- left_join(result, total_duration_count, by = "member_casual")
outlier_percentage <- outlier_percentage %>%
  mutate(percentage = (count / total_count) * 100)

outlier_percentage
```

```
## # A tibble: 2 x 4
##   member_casual count total_count percentage
##   <chr>         <int>       <int>      <dbl>
## 1 casual        51522     2048335       2.52
## 2 member         3522     2539851       0.139
```

```r
total_ride_count <- all_data_clean %>%
  group_by(member_casual, rideable_type) %>%
  summarize(ride_count = n())

total_ride_count_by_member <- total_ride_count %>%
  group_by(member_casual) %>%
  summarize(total_ride_count = sum(ride_count))

bike_use_percentage <- total_ride_count %>%
  left_join(total_ride_count_by_member, by = "member_casual") %>%
  mutate(bike_percentage = ((ride_count / total_ride_count) * 100),
         bike_percentage = round(bike_percentage, 2))
```

Plotting the Data

Number of Rides by Member Type & Day of the Week
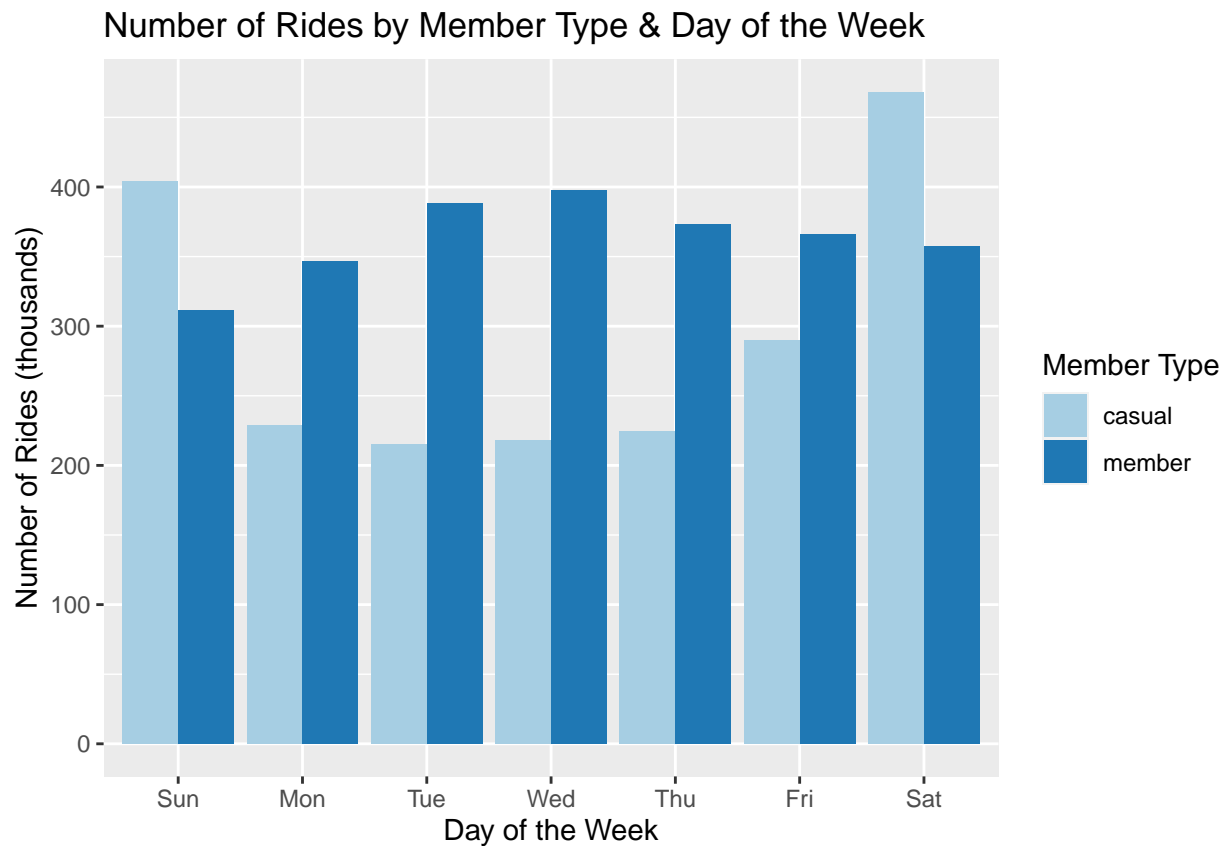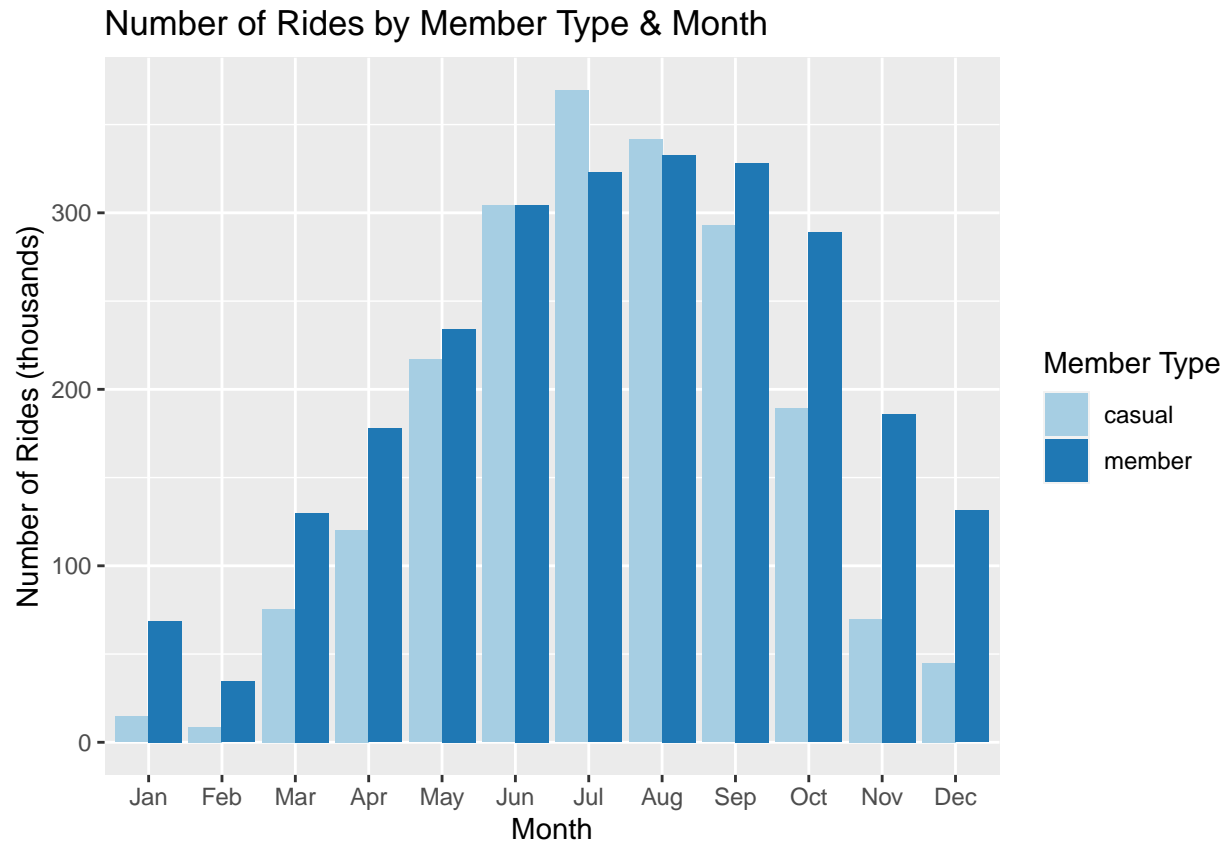

Results/Interpretation

As expected, the member users ride more often from Mon-Fri, while casual users ride more often on the weekend. There are still a steady amount of rides from Mon-Fri for casual riders, so I definitely think there are some opportunities to convert casual users into members. Ideally, I'd like to view and analyze more individualized user accounts in order to identify and target specific users for promotions.

Median Ride Duration by Member Type & Day of the Week
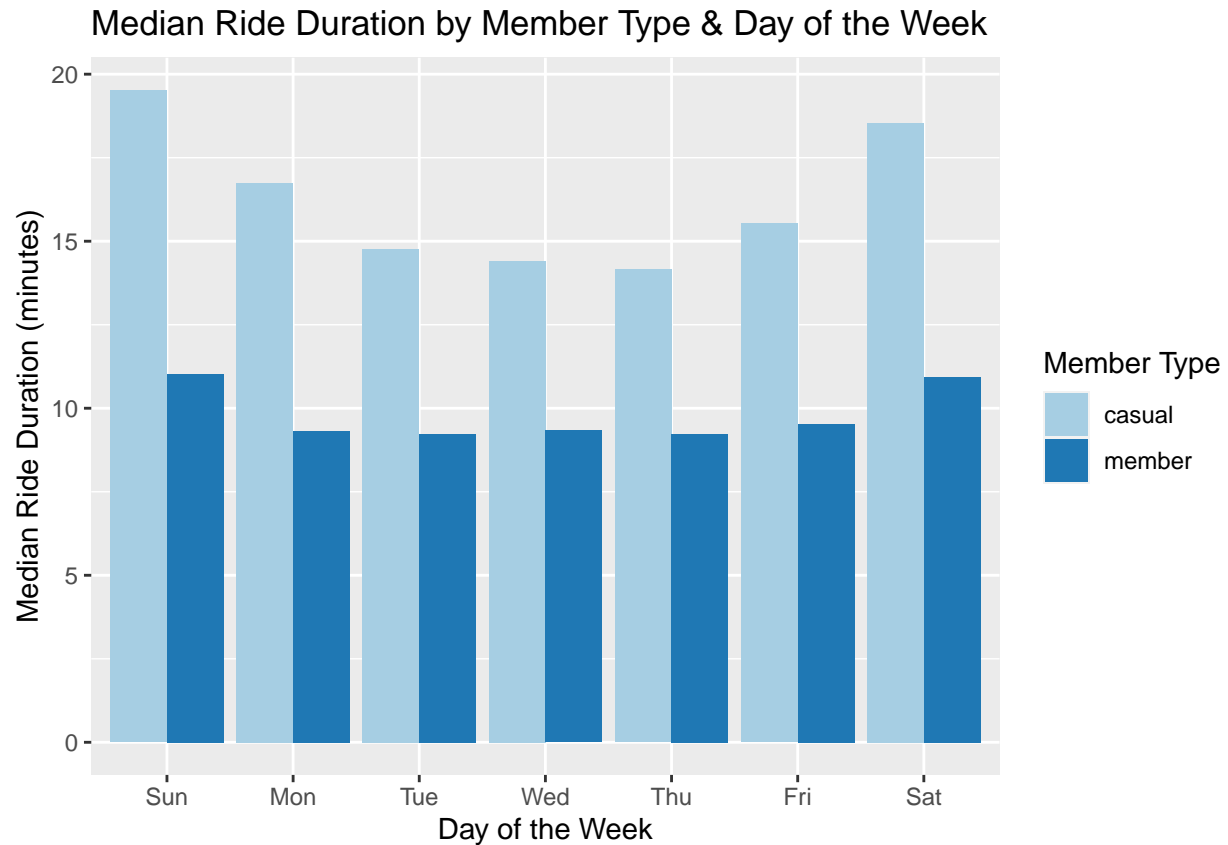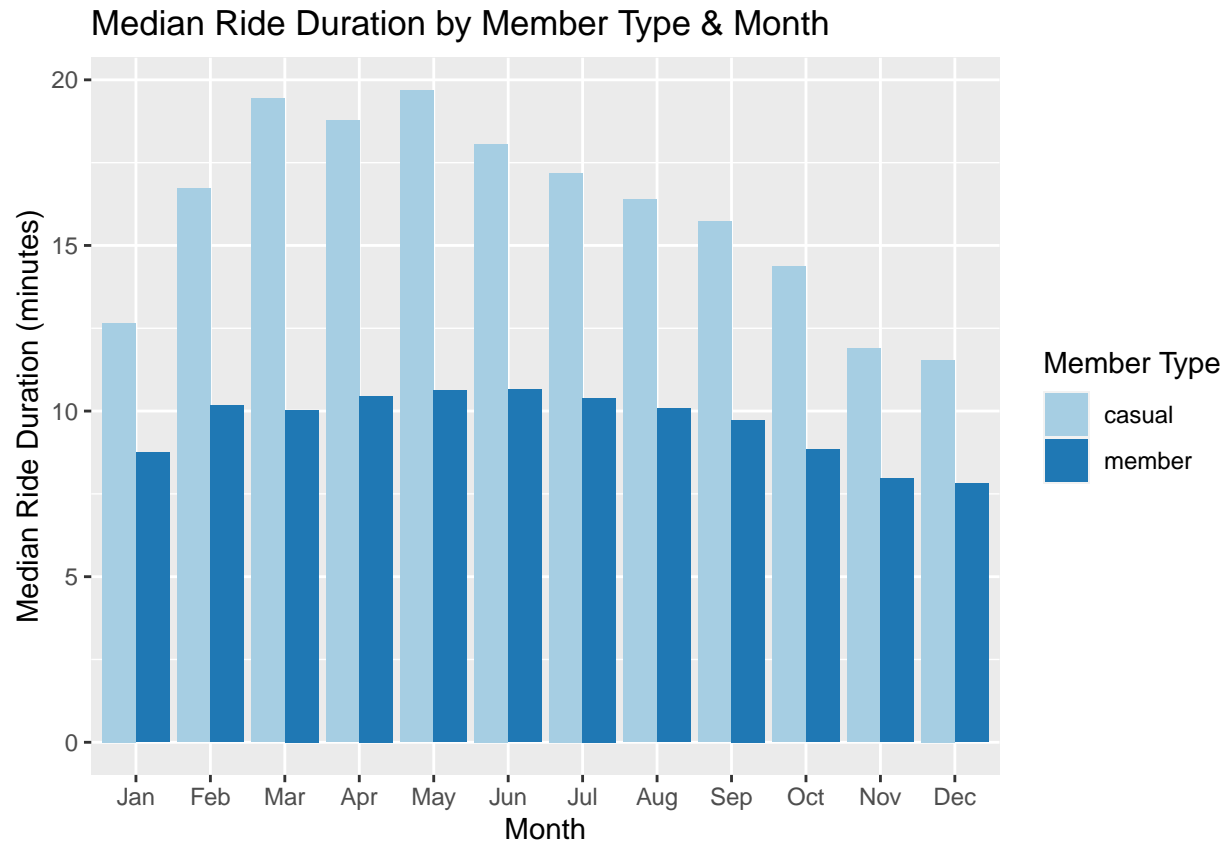

Results/Interpretation

As expected, casual riders use the bikes longer than member users do, especially on weekends. I also noticed that casual users have an inverted bell curve when it comes to duration and day of the week, with Thursdays having the lowest duration. With the assumptions that members are most likely commuters, seeing a steady duration from Mon-Fri was expected, and seeing higher duration rides on the weekend is also expected since they most likely aren't expected to be at a place by a set time.

* I used median for duration instead of mean (avg) because there were some pretty significant outliers, and median would provide a more accurate view for analysis.

Number of Rides by Member Type & Month
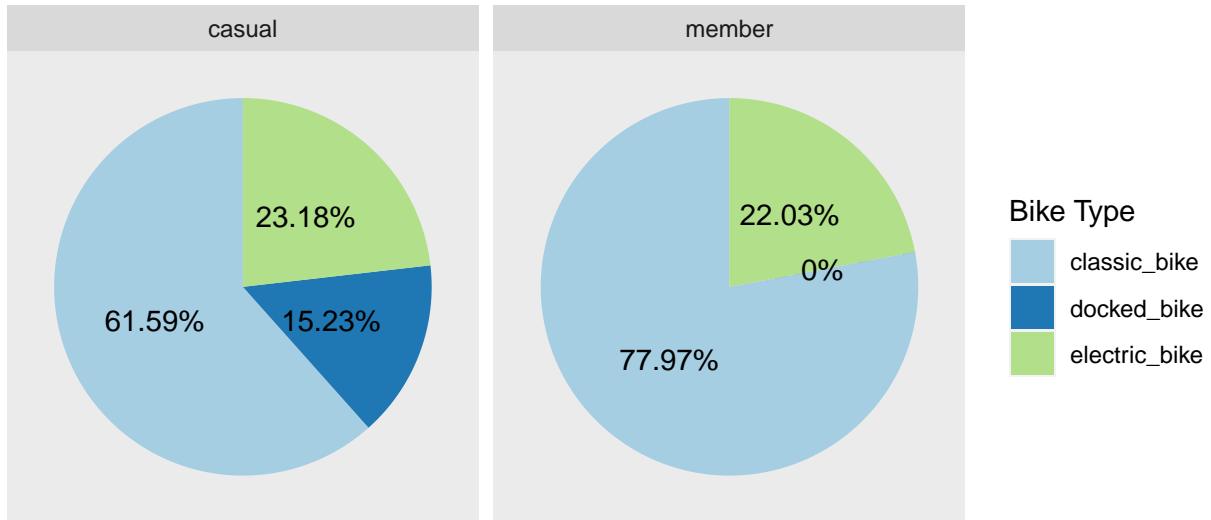
Results/Interpretation

Here we see that the colder months, Jan-Mar and Oct-Dec, have lower rides for both member types. As expected, while member rides are still lower during the colder months, the member rides are significantly higher than casual rides during those colder seasons. With the peak season of rides for casual riders between May-Sep and being very close to the number of member rides during that period, I think there is an opportunity to convert casual riders to members soley based on the "busy" season of bike usage for casual riders.

Median Ride Duration by Member Type & Month

Results/Interpretation

While it is unsurprising that the member ride duration was relatively consistent throughout each month, it was surprising that there was ascending ride duration from Jan-Mar for casual riders which then descended towards the end of the year. July had the highest number of rides for casual riders out of all the months, but it had the fifth highest ride duration for the year. If I were to look into this even further, I would expect to see that weather overall, or by area, played a part in the ride duration.

Percentage of Rides by Bike & Member Type

Results/Interpretation

As expected, casual riders use more of a variety of the bikes that are offered. For members, however, I expected to see a larger use of the e-bikes because member users are more likely to be commuters. Could this be because there aren't as many e-bikes readily available? Is there a higher tier membership that affects e-bike usage? These are questions that I would want to answer before my final presentation to Cyclistic's stakeholders. It also looks like member users don't use docked bikes at all, so I'd want to look into why that is also.

Conclusion:

The goal of the upcoming marketing campaign is to convert casual users into member users because the finance team has determined that it is more profitable and is the key to Cyclistic's future growth. In my experience with bike rentals, they are based on time or distance usage. Based on the data provided and analyzed, casual riders have significantly higher ride duration than member riders, so Cyclistic should utilize this information in their marketing campaign based on the month and user. Since the median ride duration is longer each month compared to the current member users, Cyclistic could show that, over the course of a year or month, a membership would be more beneficial to casual riders. Including different tiers, or the entire bike selection could also be beneficial when trying to convert casual riders to members. In addition to that, Cyclistic could offer seasonal promotions to improve their ridership during the slower seasons of the year, Jan-Mar and Oct-Dec.

References

The data I used came from Divvy, and I used all of the 2021 csv files **here**. **Here** is the license for this data as well.