Capstone Report

Spencer Eiserman

2024-10-06

### Introduction

As directed by Lily Moreno, I used our Cyclistic bike-share data from Q3 of 2023 through Q2 of 2024 to analyze the differences between our members and casual riders. To do this, I needed the rider type, day of the week, year and month, bike type, and the length of the trip. Other types of data could be collected and would be useful in understanding both casual riders and members. For example, a column describing the purpose of a ride in categories like leisure, commute, errands, and miscellaneous. We could also see what percent of each group has a car, uses public transit, or uses other ride-share services. However, with this dataset, we get a glimpse into how each group behaves.

### 

### Data Source

This fictional case study uses data collected by the Divvy bike-sharing company in Chicago. I downloaded 12 files from [this link](https://divvy-tripdata.s3.amazonaws.com/index.html), and then read all 12 files into a single tibble of 5,734,381 rows. Divvy has data on their bikes dating back to 2013, but, for the last few years, the data are structured in monthly files. I used July of 2023 to June of 2024 in order to sync up with the start of the third quarter to the end of the second quarter. This results in the most recent data that contains a full fiscal year with complete quarters, which will also highlight the seasonality of rides.

### 

### Data Cleaning/Manipulation

First, I filtered out any rows that were missing the **start\_station** and/or **end\_station** information. These rides were often abnormal in some respect. Bikes are occasionally lost, damaged, or stolen, and these could account for a portion of the trips in question. Without additional data on these rides, it cannot be known for sure. In total, this filter removed 1,459,927 rows, or about 25% of my raw trip data.

In order to compare groups, I created the **trip\_length** column by calculating the difference between the **started\_at** and **ended\_at** column and dividing by 60 to convert to minutes. Next, I created a **year\_month** column by extracting the year and month from the **ended\_at** column. Then, I converted **year\_month** to a date for the added benefit when making visualizations. For a later analysis, I extracted the day of the week (e.g. Monday, Tuesday, etc) from the **ended\_at** column.

At this point, I reduced the number of columns in my data view by selecting only **trip\_length**, **year\_month2**, and **rider\_type**. In this chunk of code, I also took out any data that had a **trip\_length** of less than 0 minutes or over 24 hours. Negative time could mean there was an error in the automatic reporting, or that bikes can be manually clocked in by Cyclistic maintenance crews after being offline. Trips over 24 hours are presumably day-pass riders who are returning bikes late, however, only 230 of the 4.27 million rows were affected.

### 

### Analysis

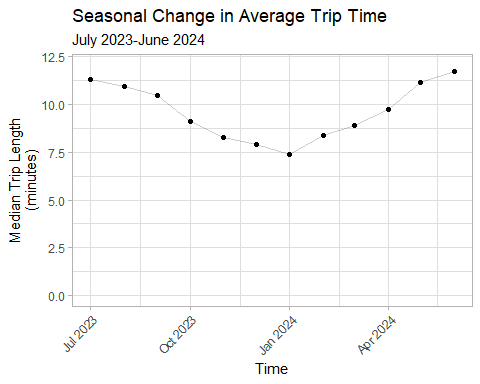
One aspect of our business is the seasonal change in demand for our services. First, I grouped the data by month. I used the **trip\_length** column to find the average and median monthly trip lengths and then got a trip count for each month. Figures 1 and 2 represent the seasonal change in demand that we experience from all of our customers.

data\_sum <- df\_filter2 %>%  
 group\_by(year\_month2) %>%  
 summarize(avg\_trip\_length = mean(trip\_length),  
 med\_trip\_length = median(trip\_length),  
 trip\_count = n()) %>%   
 ungroup()

**Figure 1**

ggplot(data\_sum, aes(x=year\_month2, y=med\_trip\_length)) +  
 geom\_point() +  
 geom\_line(alpha=.2) +  
 scale\_y\_continuous(limits=c(0,12)) +  
 labs(title="Seasonal Change in Average Trip Time",  
 subtitle = "July 2023-June 2024",  
 x = "Time",  
 y = "Median Trip Length\n(minutes)") +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle = 45,  
 vjust = 1,  
 hjust = 1)

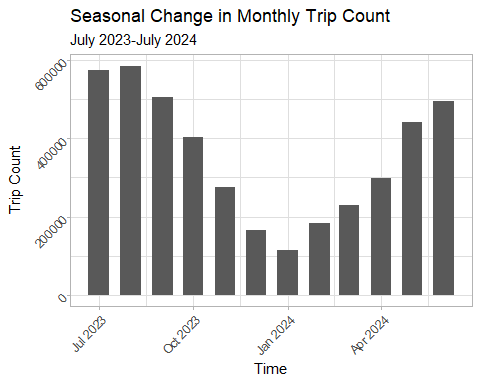
)



*A note here that for the sake of accuracy, the median trip length column was used instead of the average. The average was frequently close to or higher than the third quartile of the data. There is a good possibility that the day passes are skewing our data averages higher than is representative since a majority of our trips are 30 minutes or less. These could be reviewed separately for more insights, but this data does not denote a day pass as a variable.*

**Figure 2**

ggplot(data\_sum, aes(x=year\_month2,y=trip\_count)) +  
 geom\_col(width = 20) +  
 labs(title="Seasonal Change in Monthly Trip Count",  
 subtitle="July 2023-July 2024",  
 x="Time",  
 y="Trip Count") +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle =45, vjust = 1, hjust = 1),  
 axis.text.y = element\_text(angle = 45)  
 )



##### 

##### Members vs. Casual Riders

Now that we understand the general trend of our business as a baseline, we can analyze differences between members and casual riders. First, I grouped by the month and rider type, and then I got summary data like the first and third quartile, median trip length, trip count, and the percentage of the monthly total for each rider type.

mem\_cas <- df\_filter2 %>%  
 group\_by(year\_month2, rider\_type) %>%  
 summarize(first\_quart = quantile(trip\_length, 0.25),  
 med\_trip\_length = median(trip\_length),  
 avg\_trip\_length = mean(trip\_length),  
 third\_quart = quantile(trip\_length, 0.75),  
 trip\_count = n()) %>%   
 ungroup() %>%   
 group\_by(year\_month2) %>%   
 mutate(monthly\_total = sum(trip\_count)) %>%   
 mutate(percent = (trip\_count/monthly\_total)\* 100)

## `summarise()` has grouped output by 'year\_month2'. You can override using the  
## `.groups` argument.

Figure 3 shows the seasonal change in median trip length broken down by rider type. Note that the casual riders’ median trip length falls much more during the winter when compared to the members. There is also a higher degree of variability in the trips casual riders take as seen by how much larger their interquartile range is when compared with members.

**Figure 3**

ggplot(mem\_cas, aes(x=year\_month2, y=med\_trip\_length, color = rider\_type)) +  
 geom\_crossbar(aes(ymin = mem\_cas$first\_quart, ymax = mem\_cas$third\_quart)) +  
 scale\_y\_continuous(limits=c(0,30)) +  
 labs(title="Member vs. Casual Rider Trip Lengths",  
 subtitle = "July 2023-June 2024",  
 x = "Time",  
 y = "Median Trip Length\n(minutes)",   
 color = "Rider Type") +  
 facet\_wrap(~rider\_type) +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle = 45,  
 vjust = 1,  
 hjust = 1),  
 legend.position = "none"  
 )

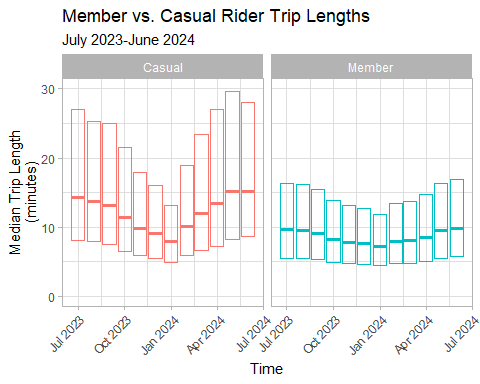
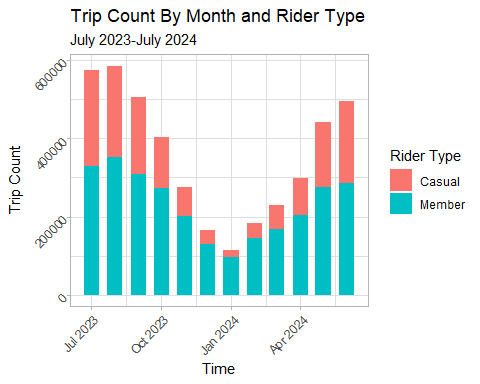


Figure 4 shows how the monthly trip count is split between members and casual riders. Even though both groups experience a seasonal dip in demand, members' trip counts decrease less in the winter compared to casual riders.

**Figure 4**

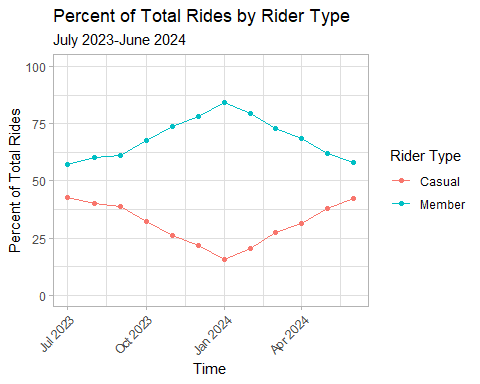
ggplot(mem\_cas, aes(x=year\_month2,y=trip\_count, fill = rider\_type)) +  
 geom\_col(width = 20) +  
 labs(title="Trip Count By Month and Rider Type",  
 subtitle="July 2023-July 2024",  
 x="Time",  
 y="Trip Count",  
 fill = "Rider Type") +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle =45, vjust = 1, hjust = 1),  
 axis.text.y = element\_text(angle = 45)  
 )



In Figure 5, I wanted to accentuate this difference, so I plotted the monthly percent of total rides by rider type. The members’ share of rides goes up during the winter. So far, we have only seen rates go down in the winter. This insight helps us know what part of the year we might be more likely to reach our target customer base, the casual rider. Not only do casual riders’ trips decrease during the winter, but because they decrease at a higher rate than members, their share of the monthly rides also goes down.

**Figure 5**

ggplot(mem\_cas, aes(x = year\_month2, y = percent, color = rider\_type)) +  
 geom\_point() +  
 geom\_line() +  
 ylim(0,100) +  
 labs(title="Percent of Total Rides by Rider Type",  
 x="Time",  
 y="Percent of Total Rides",  
 subtitle="July 2023-June 2024",  
 color = "Rider Type") +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)  
 )



##### 

##### Comparing Groups by Bike Type

The next variable I chose to consider was the **bike\_type**. First, I grouped by the month, rider type, and type of bike. Then, I removed outliers in the data labeled “docked\_bike”. It only showed up in two months of the casual riders’ data and was a very small percent that did not affect the trends whether included or excluded. I calculated a trip count for each item I grouped by, and then ungrouped. Regrouping without the rideable type allowed me to calculate the percent of trips taken on classic bikes vs. electric bikes broken down by month and rider type.

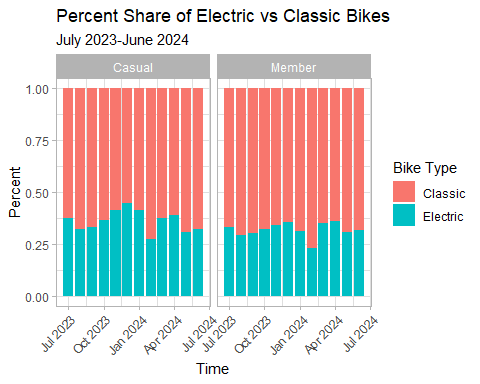
df\_ana <- df\_proc %>%   
 filter(between(trip\_length, 0, 1440)) %>%  
 select(rider\_type, bike\_type, year\_month2, trip\_length) %>%  
 group\_by(year\_month2, rider\_type, bike\_type) %>%  
 filter(bike\_type!="docked\_bike") %>%   
 summarize(trips=n()) %>%  
 ungroup() %>%   
 group\_by(rider\_type, year\_month2) %>%   
 mutate(totalbygroup = sum(trips)) %>%   
 mutate(percent = trips/totalbygroup)

## `summarise()` has grouped output by 'year\_month2', 'rider\_type'. You can  
## override using the `.groups` argument.

Notice that casual riders and members behave similarly when choosing the kind of bike they want throughout the year. Even though there is no clear or consistent trend as with seasonal demand, both groups behave almost identically.

**Figure 6**

ggplot(df\_ana, aes(x=year\_month2,y=percent,fill=bike\_type)) +  
 geom\_col() +  
 facet\_wrap(~rider\_type) +  
 labs(title="Percent Share of Electric vs Classic Bikes",  
 subtitle="July 2023-June 2024",  
 x="Time",  
 y="Percent",  
 fill="Bike Type") +  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle=45, hjust=1)  
 )



##### 

##### Comparing Groups of Day of the Week

Lastly, I considered that members and casual riders may not ride at the same rates depending on the day of the week. First, I selected the rider type, weekday, and trip length. I grouped by weekday and rider type, then I calculated the median trip length and trip count.

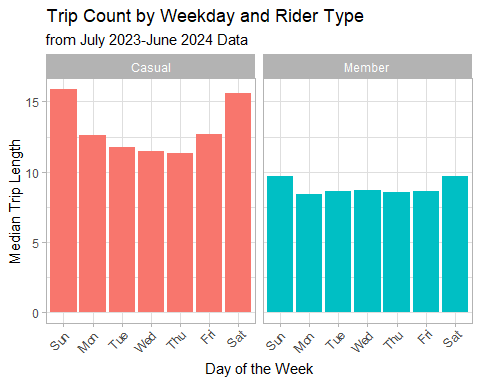
df\_ana2 <- df\_proc %>%  
 select(rider\_type, weekday, trip\_length) %>%   
 group\_by(weekday, rider\_type) %>%  
 summarise(trip\_count = n(),  
 med\_trip\_length = median(trip\_length))

## `summarise()` has grouped output by 'weekday'. You can override using the  
## `.groups` argument.

The data about median trip length by weekday and rider type confirms the stability of member behavior throughout the week and shows that casual riders take longer trips on weekends.

**Figure 7**

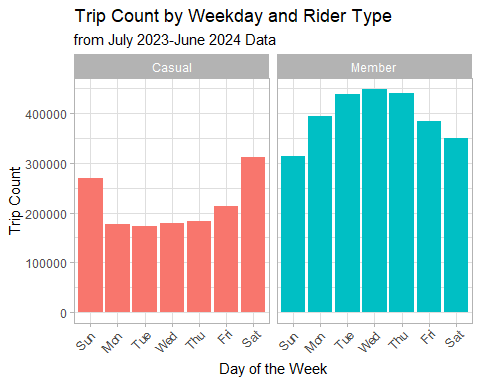
ggplot(df\_ana2, aes(x=weekday, y=med\_trip\_length, fill=rider\_type)) +  
 geom\_col() +  
 facet\_wrap(~rider\_type) +  
 labs(title="Trip Count by Weekday and Rider Type",  
 subtitle="from July 2023-June 2024 Data",  
 x="Day of the Week",  
 y="Median Trip Length") +  
 theme\_light() +  
 theme(axis.text.x=element\_text(angle=45,  
 hjust=1.2,  
 vjust=1.2),  
 legend.position= "none"  
 )



The data on trip count by weekday and rider type shows that members and casual riders behave oppositely; members ride more often during the week while casual riders show an drastic increase in trips on Saturdays and Sundays only.

**Figure 8**

ggplot(df\_ana2, aes(x=weekday, y=trip\_count, fill=rider\_type)) +  
 geom\_col() +  
 facet\_wrap(~rider\_type) +  
 labs(title="Trip Count by Weekday and Rider Type",  
 subtitle="from July 2023-June 2024 Data",  
 x="Day of the Week",  
 y="Trip Count") +  
 theme\_light() +  
 theme(axis.text.x=element\_text(angle=45,  
 hjust=1.2,  
 vjust=1.2),  
 legend.position= "none"  
 )



Although the data do not explicitly tell us whether each ride was taken for commute or for leisure, the trends we observe allow us to make evidence-based inferences. Commuting happens for most people during the week and less often on weekends. The fact that members take more trips during the week, coupled with the fact that they behave more consistently throughout the year than casual riders, is evidence to suggest that members are more likely to use Cyclistic bikes for commute. This consistency implies some level of necessity for a portion of our members. Casual riders take many more trips on Saturdays and Sundays than they do during the week, and they tend to be much more variable in their need for Cyclistic’s services depending on the time of year. This variability implies that casual riders are more likely to ride for convenience or leisure than out of necessity.

### 

### Recommendations

With the dataset I had available to me, I analyzed the differences between Cyclistic members and casual riders. There were some notable patterns in the data that led me to three recommendations:

1. The marketing team should advertise to casual riders from early May through August.
2. The marketing team should advertise more on Fridays-Sundays than the weekdays.
3. Create a new membership option for casual riders who bike more frequently in the summer.

Though the data were limited, one thing was clear: members ride more consistently throughout the year than casual riders. In Figure 4, though the members’ trip count goes down during the winter by 72.7%, casual riders’ total trips go down by 92.7% in comparison. In Figure 3, we can clearly see a much higher degree of variability in casual riders than in members. Furthermore, Figure 7 shows how consistently members ride throughout the week.

In order to reach more of our target market, any advertising campaign should be done at a time of year when the casual riders’ share of the trip count increases to its highest point. In July and August, the casual riders take about 40% of the total trips, but from April to May, there is a bump in casual riders’ share of the trip count. I recommend starting a marketing campaign targeted at casual riders in May and running it through August.

Casual riders clearly differ from members when it comes to the most popular days of the week to ride. They take many more trips on weekends and slightly more on Fridays too. I would recommend advertising more on Fridays-Sundays when casual riders are more likely to take a ride.

My last recommendation would be to expand the options for memberships to capitalize on our casual riders needs. There are a few options for doing this, the primary option involves creating a monthly membership tier, so that riders who need short-term transportation or want to ride regularly for only part of the year, would have a more suitable choice. Another option is creating a seasonal pass, not dissimilar from amusement park summer passes. Of course, some of our annual members will choose this alternate option. More data could be collected to analyze this risk, because the current data does not denote whether each ride was taken for leisure, errands, or commute. We can surmise from these data that some portion of our members use bikes year-round for commuting and errands, so we will always have that customer base for the winter. The income potential from casual riders becoming monthly/seasonal members most likely out-weighs the deficit from some annual members becoming monthly/seasonal members. We should take this risk in order to exploit the yearly demand curve for larger margins when demand is rising to its peak in July and August.

### 

### Resources

Here is the Capstone project packet with the setup and context for this fictional case study. (<https://www.coursera.org/api/rest/v1/asset/download/pdf/1XKhm37HS9iPXHfAIEBaRQ?pageStart=&pageEnd=>)

This link shows where I downloaded my data from. I used 202307-divvy-tripdata.zip through 202406-divvy-tripdata.zip. (<https://divvy-tripdata.s3.amazonaws.com/index.html>)