Google Data Analytics Capstone Project:

Cyclistic Case Study by: Nathan Park

Hello, I will be analyzing bikeshare data from Cyclisitic in order to find conclusions and make recommendations. I will be using 6 steps: ***Ask, Prepare, Analyze, Share,*** and ***Act.*** This is a way to break down the analytics process into a clear structure. I will be using R and Tableau to analyze and visualize the data.

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Stakeholders

Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to
commute to work each day.
Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns
and initiatives to promote the bike-share program. These may include email, social media, and other channels.
Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and
reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy
learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic
achieve them.

Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the
recommended marketing program.

About the company

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. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.
Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

Cyclistic is trying to increase the number of annual members by transitioning casual riders into annual members in order to maximize the growth of the company.
Three questions will guide the future marketing program:
1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?
Lily Moreno has assigned us the first question for me to answer. Our business task is to analyze how annual members and casual riders use Cyclistic bikes differently in order

to gain insight on how to effectively market to casual riders in order to increase annual membership.

Prepare

The data source that we are using is historical bike trip data from Cyclistic. The trip data can be found here. Out of all the historical data, we will select the previous 12 months, which would be October 2021 to September 2022. The data is formatted in .csv files. The data has been licensed and made available by Motivate International Inc. by this license. (Disclaimer: because Cyclistic is a fictional company, the data is under a different name, but for the purposes of this case study, the data is appropriate) The data is from a public source, however data-privacy issues prohibit using riders' personally identifiable information. Because of this, I won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Process

The Cyclistic historical trip data contains millions of records. Due to the amount of data, tools like Google Sheets or Microsoft Excel would not be able to process the data efficiently. Because of this, I chose to use Rstudio to clean and analyze the data. Rstudio can process, analyze and visualize large amounts of data.

First off, I installed and loaded packages that would allow me to use functions to clean and organize the data.

```
# setting up packages for analysis
install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")
install.packages("dplyr"))
install.packages("readr")
install.packages("janitor")
install.packages("skimr")
# loading packages
library("tidyverse")
library("lubridate")
library("ggplot2")
library("dplyr")
```

```
library("readr")
library("janitor")
library("skimr")
```

Next, I imported the trip data into Rstudio so that I can begin processing it. I imported the previous 12 months worth of trip data using the **read_csv()** function.

```
# importing the data

oct_2021_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202110-divvy-tripdata.csv")
nov_2021_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202111-divvy-tripdata.csv")
dec_2021_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202112-divvy-tripdata.csv")
jan_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202201-divvy-tripdata.csv")
feb_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202202-divvy-tripdata.csv")
mar_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202203-divvy-tripdata.csv")
apr_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202204-divvy-tripdata.csv")
may_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202205-divvy-tripdata.csv")
jun_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202206-divvy-tripdata.csv")
jul_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202207-divvy-tripdata.csv")
aug_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202208-divvy-tripdata.csv")
sep_2022_trips <- read_csv("C:/Users/natha/OneDrive/Desktop/Bikeshare
Data/202209-divvy-publictripdata.csv")
```

I used the **colnames()** function to see if there are any discrepancies in variable names amongst the 12 months of data.

```
# checking that column names are the same
colnames(oct_2021_trips)
colnames(nov_2021_trips)
```

```
colnames(dec_2021_trips)
colnames(jan_2022_trips)
colnames(feb_2022_trips)
colnames(mar_2022_trips)
colnames(apr_2022_trips)
colnames(may_2022_trips)
colnames(jun_2022_trips)
colnames(jul_2022_trips)
colnames(aug_2022_trips)
colnames(sep_2022_trips)
```

Since the variable names are all the same, I combined all the month datasets into one dataframe using the **bind_rows()** function. This dataframe will be called all_trips. This made it easier for me to analyze all the data together without having to call all 12 datasets every time.

```
# binding all the month data frames into one

all_trips <-
bind_rows(oct_2021_trips,nov_2021_trips,dec_2021_trips,jan_2022_trips,feb_2022_trip
s,

mar_2022_trips,apr_2022_trips,may_2022_trips,jun_2022_trips,jul_2022_trips,
                        aug_2022_trips,sep_2022_trips)
```

In order to keep the data concise, I removed unneeded variables from the dataframe. This would clear up the data and keep only the variables that we are going to analyze.

```
# removing unneeded columns
all_trips <- all_trips %>%
        select(-c(start_lat, start_lng,end_lat,end_lng))
```

I also changed variable names using the **rename()** function to ones that make more sense and are easier to recognize.

```
# updating column names
all_trips <- all_trips %>%

rename(bike_type=rideable_type,start_time=started_at,end_time=ended_at,
                                        user_type=member_casual)
```

After that, I added year, month, and day columns so that we can have more variables to analyze by date. This was done using the **format(as.Date))** function. That allows us to get a more detailed look on membership types based on date. This would allow us to aggregate by year, month or day.

```
# adding month,day, and year columns
all_trips$date<-as.Date(all_trips$start_time)
all_trips$month<-format(as.Date(all_trips$date),"%m")
all_trips$day<-format(as.Date(all_trips$date),"%d")
all_trips$year<-format(as.Date(all_trips$date),"%Y")
all_trips$day_of_week<-format(as.Date(all_trips$date),"%A")
```

I also added the ride_length column in order to create a variable to see how long each ride is by using **difftime()**. This would provide another variable that we can analyze.

```
# add a ride_length calculation since there is no trip duration column
all_trips$ride_length<-difftime(all_trips$end_time,all_trips$start_time)
```

Next, I used the **str()** function to check the formatting of the columns. The ride_length column should be numeric in order to run calculations on it. I used the **typeof()** function in order to check the formatting of the ride_length column. The **as.numeric()** function changes the format to numeric and I used **is.numeric()** to make sure that ride_length is numeric.

```
# check structure of the columns
str(all_trips)
#convert ride_length to numeric in order to run calculations
typeof(all_trips$ride_length)
all_trips$ride_length<-as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

Finally, there are some trips that have data that doesn't make sense. Cyclistic performed quality checks on a certain number of bikes. So we must remove the trips that started at "HQ QR." There are also some trips with negative ride_lengths, so we must remove them as well. I also used the **na.omit()** function to get rid of NA values. In doing this, the data frame that we use is now all_trips_v3.

```
# take out bad data
```

```
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" |
all_trips$ride_length<0),]
# remove all na values
all_trips_v3<-na.omit(all_trips_v2)
```

Analyze

To start analyzing the data, I used the **summary()** function to find the mean, median, minimum and maximum value of the all_trips_v3 data frame. This gives us more insight on the data and we can start to see where the data is trending.

```
# summary of ride_length
summary(all_trips_v3$ride_length)
```

In order to address the business task, I needed to find summary data and compare the data based on the usertype, annual vs casual riders. Annual members are under the member column and casual riders are under the casual column. I compared them using the **aggregate()** function.

```
# compare members vs casual riders
aggregate(all_trips_v3$ride_length~all_trips_v3$user_type,FUN=mean)
aggregate(all_trips_v3$ride_length~all_trips_v3$user_type,FUN=median)
aggregate(all_trips_v3$ride_length~all_trips_v3$user_type,FUN=max)
aggregate(all_trips_v3$ride_length~all_trips_v3$user_type,FUN=min)
```

From the data, we can see that casual riders had longer ride lengths overall than annual members did. In order to explore the data further, I used the **aggregate()** function again to analyze the difference in ride length between annual and casual riders per day of the week. This would give me a better idea of how the ride lengths differ throughout the week.

```
# compare average ride lengths
aggregate(all_trips_v3$ride_length~all_trips_v3$user_type+all_trips_v3$day_of_week,F
UN=mean)
```

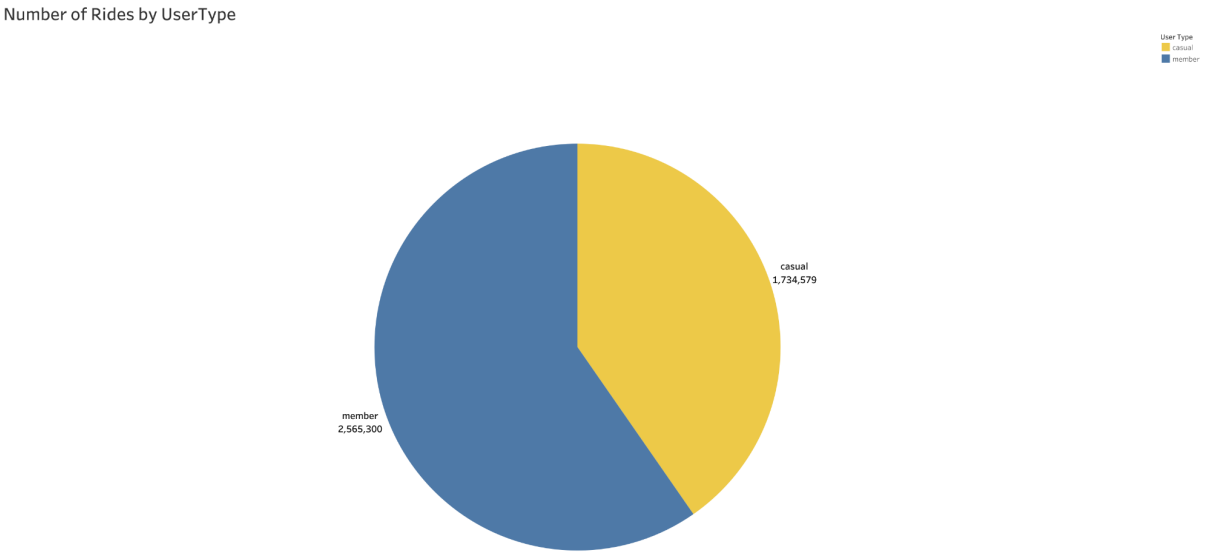
Finally, I wanted to see how the number of rides differ between annual and casual riders per day. I also wanted to see what the average duration of each ride was. This can give us more insight on the difference in behavior between annual and casual riders.

```
#analyze number of rides and average duration per day
all_trips_v3%>%
                mutate(weekday=wday(start_time,label=TRUE))%>%
                    group_by(user_type,weekday)%>%

summarise(number_of_rides=n(),average_duration=mean(ride_length))%>%
                                    arrange(user_type,weekday)
```

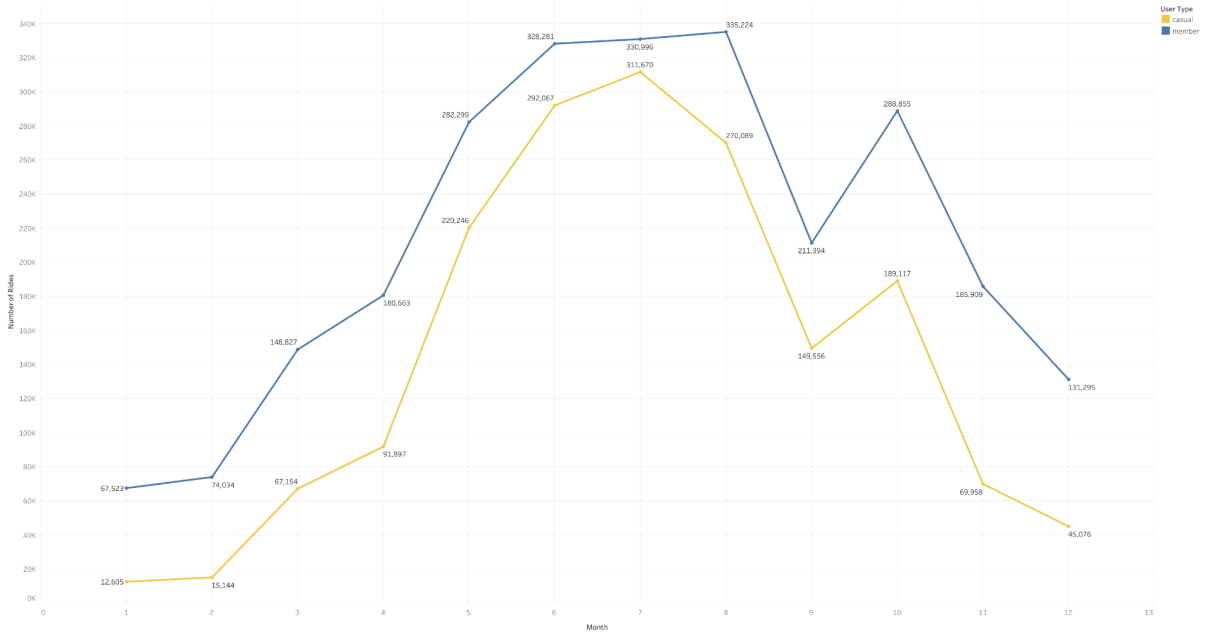
Share

In order to create the most effective visualizations, I chose to use Tableau as a tool to create and showcase them. To do this, I needed to export the data from Rstudio. I used the **write.csv()** function to export the data. Once exported, I uploaded the data to Tableau Public and started creating visualizations.



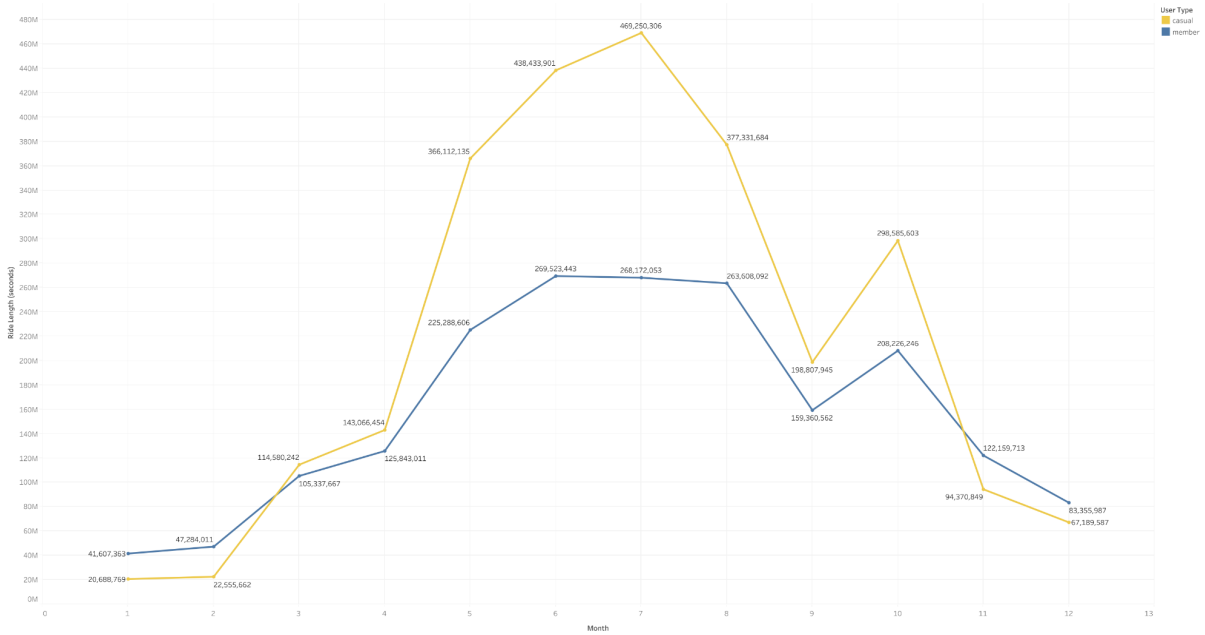
This first visualization shows us the total number of rides. Annual members have a much larger total number of rides.

Number of Bike Rides by Month

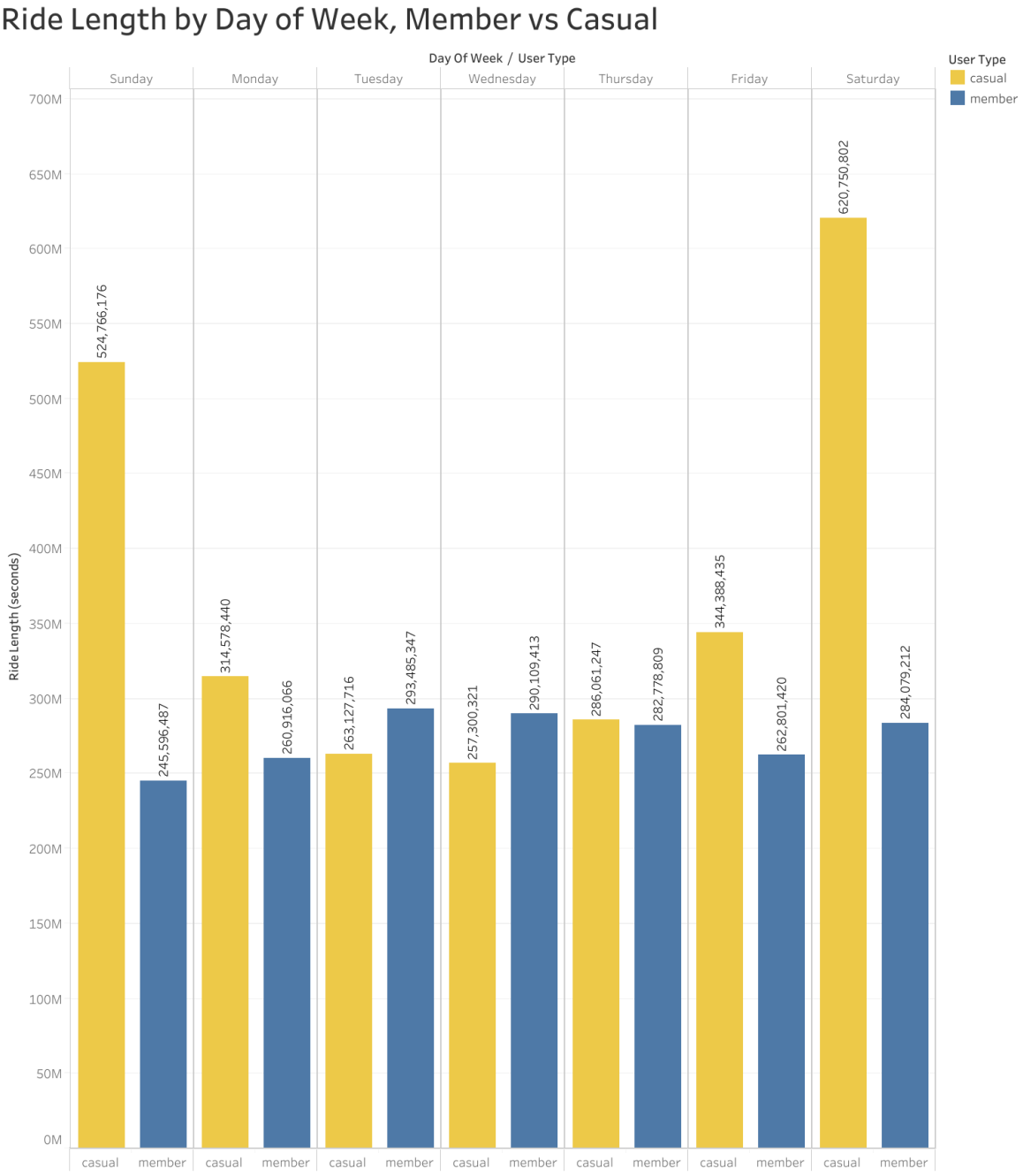


This visualization shows the number of bike trips taken per month by comparing usertype. It allows us to see how the number of trips differ month to month. The line graph shows us that the number of bike rides peak for both casual riders and annual members from June to August. There seems to be a spike in bike use during summer months.

Length of Rides by Member Type per Month

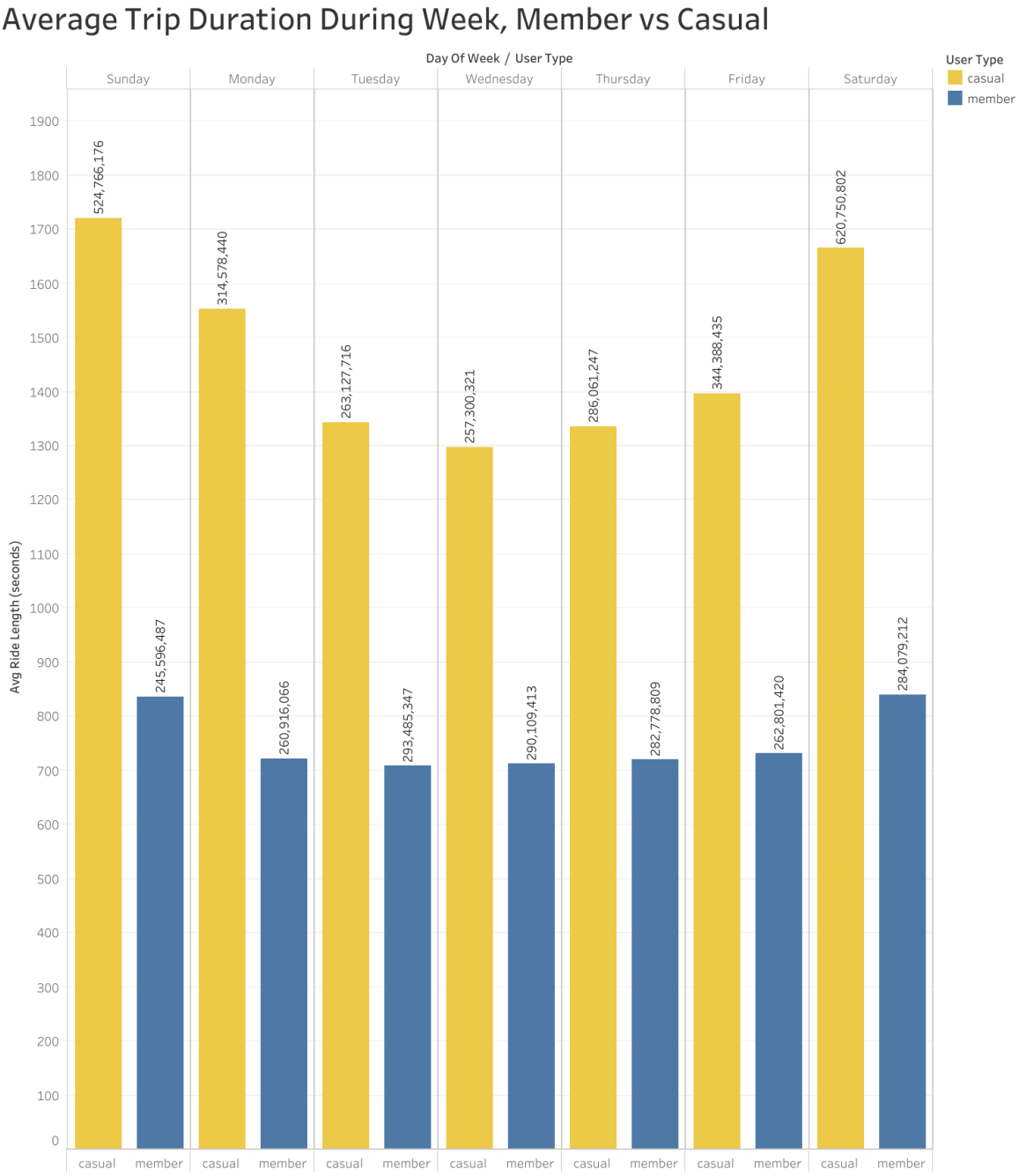


This graph shows the length of rides by usertype over months. The previous visualization showed that annual members take more rides than casual riders. However, this graph reveals that the length of rides of casual riders far exceed the length of rides of annual members. Again, ride lengths seem to peak during the summer months. However, there is an additional peak during October.

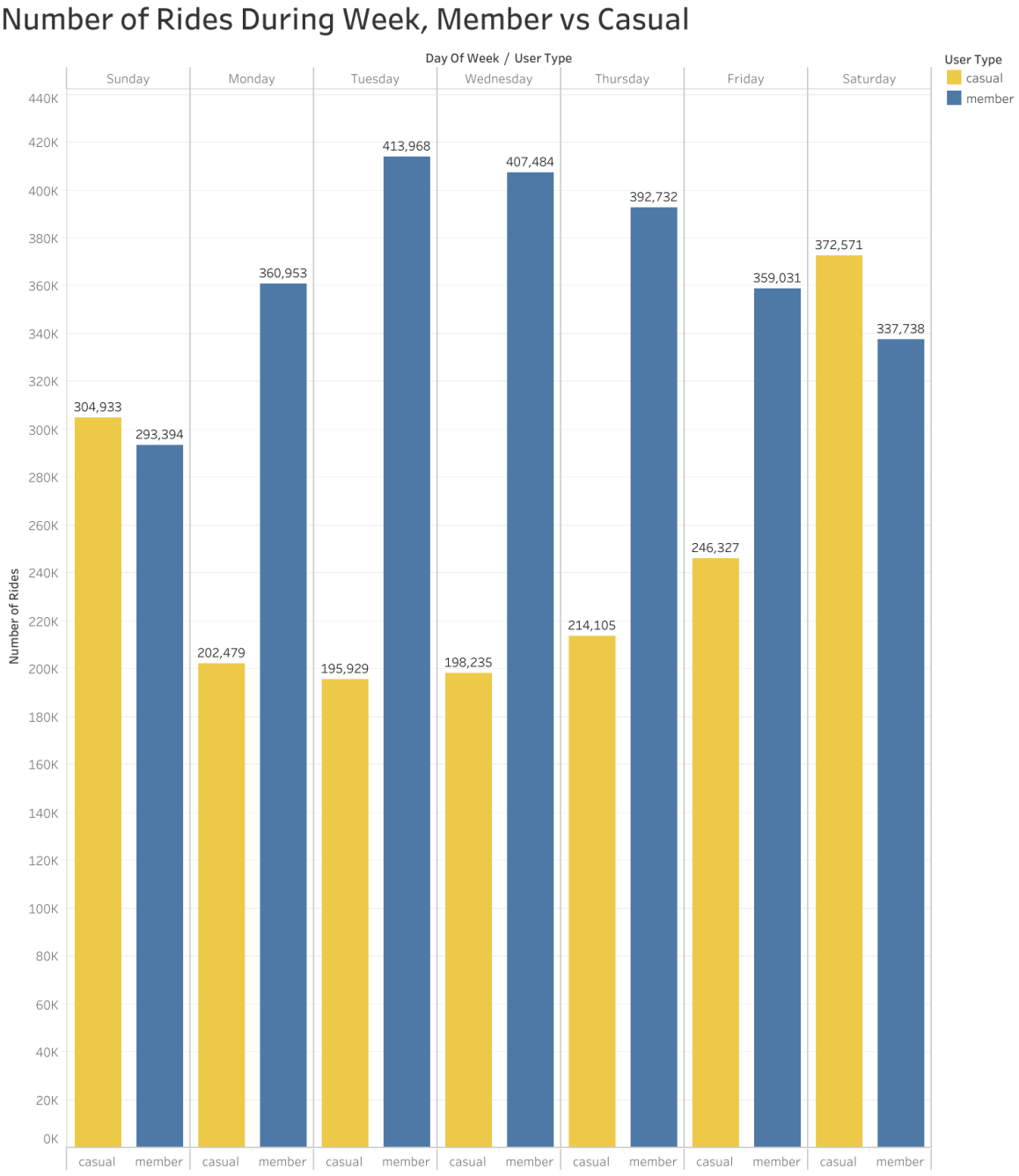


This barchart shows that the ride lengths of casual riders peak during the weekend. However,

the ride lengths of annual members stay relatively consistent throughout the week.



On average, casual riders seem to take longer rides than annual members. The average peaks during the weekend for both casual riders and annual members.



From this bar graph, it is revealed that annual members actually take more rides than casual riders do. Casual riders do have more rides on the weekends, but overall, annual members take more bike rides.

Act

From the data that I analyzed and visualized, I was able to identify and understand some important differences between casual riders and annual members.

- Annual members take more rides in total than casual riders do, almost double in fact. I believe this is because annual members are more inclined to take more bike rides because they are paying for a membership.

- When we look at the difference in user type by month, we see that although annual members have more total number of rides, the length of rides that casual riders take are much longer. Both ride length and number of riders peak during summer months of June to August with another spike in October. This increase in rides and ride length could be indicative of the effect of temperature and climate of the warmer months on bike rides. Warmer temperature leads to people being more willing to go outside and exercise. Also, children who are still in school are more likely to use bikes in their summer breaks. Rainy and snowy weather could account for the lower numbers in the winter months. I hypothesize that the additional spike in rides in October is an outlier and could be the result of warmer weather in October than usual. However, more data and analysis would need to be done to form a definitive conclusion.

- Finally, when we look at the difference in user type by day of week, we see that casual riders have longer total rides and longer rides on average than annual members. However, annual members have consistent ride lengths throughout the week while casual riders peak on the weekends. Annual members take more rides throughout the week, however, casual riders tend to have more rides on the weekends. I believe this is because annual members use bikes consistently, indicating they use bikes for commuting or for daily exercise. Because casual riders take bike rides less frequently but at the same time take much longer rides, this reveals that casual riders use bikes for temporary leisure, transportation, or recreation.

Here are some marketing strategies that I recommend based on the data.

1. Add more forms of membership with Cyclisitic. Based on the data, people who pay for memberships are more likely to use more bikes. For people who are uncomfortable with committing to a full year of membership, there should be monthly and weekend only payment options. These should be modeled after gym memberships, with annual memberships being a one time yearly fee, but

cheapest overall. Then monthly memberships would be next cheapest and then weekend only memberships. However, since monthly and weekend only memberships have more frequent payments, the payments won't seem to be too much. Weekend and monthly members who are satisfied with Cyclistic would then transition over to annual memberships over time. This strategy would allow the company to keep the pricing flexibility that attracts customers and at the same time increase profitability by increasing the number of members.

2. Cyclistic should have promotions and advertisements in the beginning of and the months leading up to summer. Summer is when users are most active so this season should be heavily prioritized to bring in more members. A promotional discount in memberships would attract many people who are planning to bike in summer anyways, and bring in new customers who were unsure. These discounts should be advertised near schools. Students would have much more free time in the summers, and that free time aligns with the peak number of bike rides.
3. The company should add another promotional discount in October where there is an additional peak in rides. This would allow summer members to be retained for the winter months while at the same time taking advantage of the increase in riders in October. This allows Cyclists to address two things, the increase in engagement in October, and allows the company to mitigate the loss of riders in the winter months.

I appreciate you for taking the time to read my case study.