Cyclistic Data Analysis Case Study

Introduction

Cyclistic is a fictional bike-share company that launched in 2016. These 5,824 company bicycles are geotracked and locked into a network of 692 stations in Chicago. These bikes can be unlocked from any station and returned to another at any time; In this case study, I am a theoretical junior data analyst uncovering insights from the first Quartiles of 2019 and 2020 to promote consumer growth in the future years.

I. Asking the Questions(s)

The primary objectives to guide the future marketing is as follows:

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

Stakeholders:

- Lily Moreno: Director of marketing and my theoretical manager.
- Cyclistic marketing analytics team: Data analysts responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
- Cyclistic executive team: Decides to approve or decline recommended marketing program.

II. Preparation

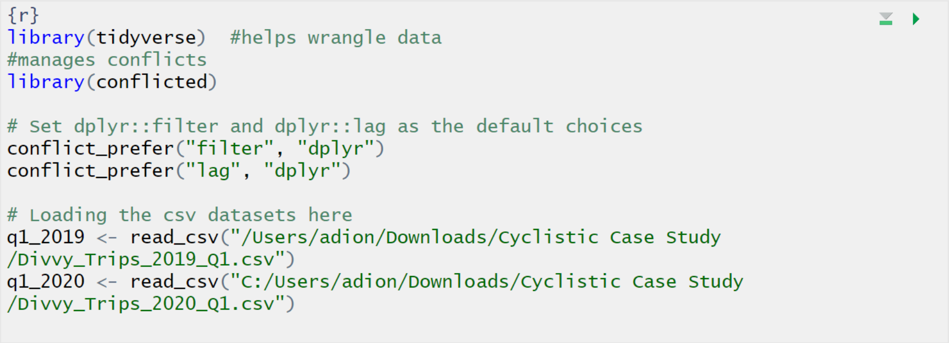
Data Source and Location: The data has been made available by Motivate International Inc. under this data license agreement. The dataset is hosted on Amazon S3 and can be accessed through the Divvy Trip Data portal.This is public data that allows the exploration of consumer types using this service but prohibits the use of riders' personally identifiable information. This privacy-protection prohibits connecting pass purchases to credit card numbers to determine if casual riders in the area live in the Cyclistic service area or if they have purchased multiple single passes.

Data Organization: The data is organized in monthly CSV files from the years 2014-2020; for this analysis, the dataset focused on consumer insights from 2019-2020. The dataset used has been analyzed for errors, repetition, anomalies, and inconsistencies to ensure accuracy and reliability.

III. Process

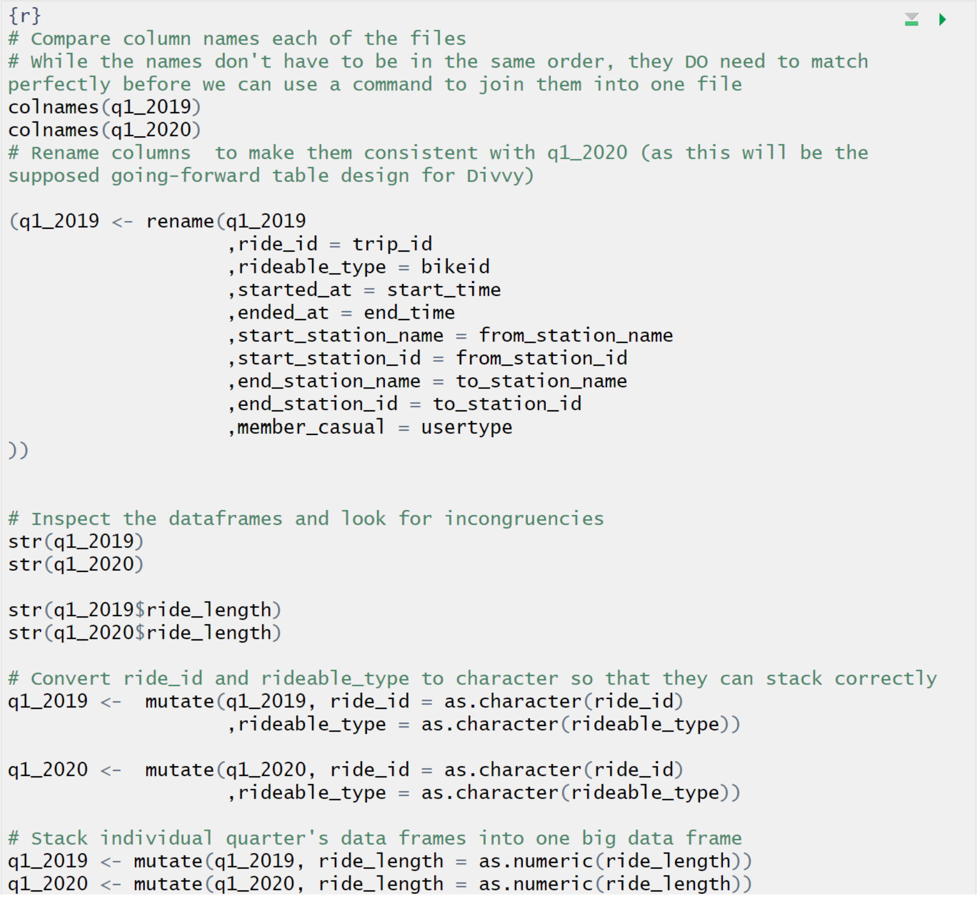
The primary tool used for data analysis in this study is R as it is ideal for its abundant libraries and capabilities to handle, analyze, and visualize data. The secondary tool used for data visualization is Tableau due to its multitude and sophisticated representations to better spotlight insights.

Data Loading R Code



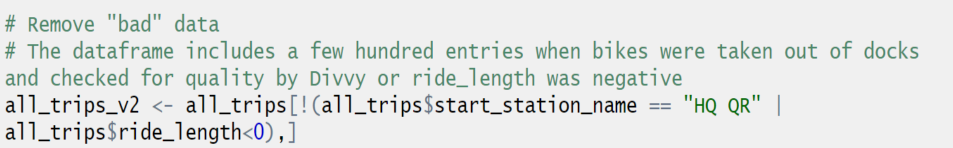
```
{r}
library(tidyverse)  #helps wrangle data
#manages conflicts
library(conflicted)

# Set dplyr::filter and dplyr::lag as the default choices
conflict_prefer("filter", "dplyr")
conflict_prefer("lag", "dplyr")

# Loading the csv datasets here
q1_2019 <- read_csv("/Users/adion/Downloads/Cyclistic Case Study
/Divvy_Trips_2019_Q1.csv")
q1_2020 <- read_csv("C:/Users/adion/Downloads/Cyclistic Case Study
/Divvy_Trips_2020_Q1.csv")
```

Cleaning the Data and Re-Formatting



```
{r}
# Compare column names each of the files
# While the names don't have to be in the same order, they DO need to match
perfectly before we can use a command to join them into one file
colnames(q1_2019)
colnames(q1_2020)
# Rename columns  to make them consistent with q1_2020 (as this will be the
supposed going-forward table design for Divvy)

(q1_2019 <- rename(q1_2019
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype
))


# Inspect the dataframes and look for incongruencies
str(q1_2019)
str(q1_2020)

str(q1_2019$ride_length)
str(q1_2020$ride_length)

# Convert ride_id and rideable_type to character so that they can stack correctly
q1_2019 <-  mutate(q1_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

q1_2020 <-  mutate(q1_2020, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

# Stack individual quarter's data frames into one big data frame
q1_2019 <- mutate(q1_2019, ride_length = as.numeric(ride_length))
q1_2020 <- mutate(q1_2020, ride_length = as.numeric(ride_length))
```

```
all_trips <- bind_rows(q1_2019, q1_2020)

# Remove lat, long, birthyear, and gender fields as this data was dropped beginning
in 2020
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, birthyear, gender,
"tripduration"))

colnames(all_trips)  #List of column names
nrow(all_trips)  #How many rows are in data frame?
dim(all_trips)  #Dimensions of the data frame?
head(all_trips)  #See the first 6 rows of data frame.  Also tail(all_trips)
str(all_trips)  #See list of columns and data types (numeric, character, etc)
summary(all_trips)  #Statistical summary of data. Mainly for numerics

all_trips <-  all_trips %>%
  mutate(member_casual = recode(member_casual
                                ,"Subscriber" = "member"
                                ,"Customer" = "casual"))

# Check to make sure the proper number of observations were reassigned
table(all_trips$member_casual)
all_trips$date <- as.Date(all_trips$started_at) #The default format is yyyy-mm-dd
all_trips$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")

# Add a "ride_length" calculation to all_trips (in seconds)
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)

# Inspect the structure of the columns
str(all_trips)

# Convert "ride_length" from Factor to numeric so we can run calculations on the
data
is.factor(all_trips$ride_length)
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

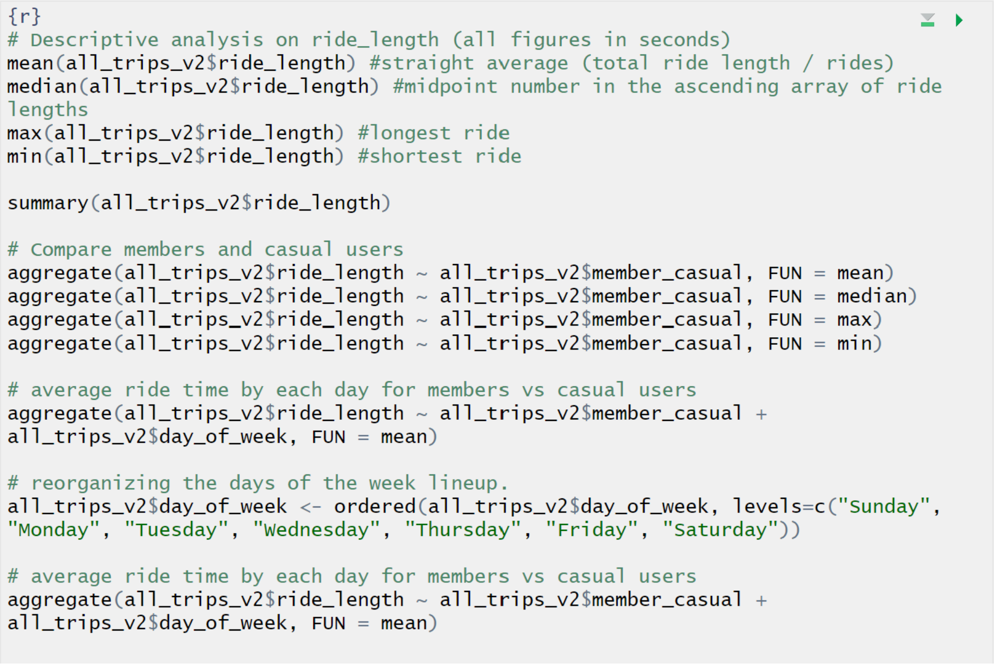


```
# Remove "bad" data
# The dataframe includes a few hundred entries when bikes were taken out of docks
and checked for quality by Divvy or ride_length was negative
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" |
all_trips$ride_length<0),]
```

IV. Analysis

Exploratory Data Analysis (EDA)

This analysis involves calculating summary statistics, exploring the relationships between different variables and comparing to identify patterns and trends in consumer behavior.



```
{r}
# Descriptive analysis on ride_length (all figures in seconds)
mean(all_trips_v2$ride_length) #straight average (total ride length / rides)
median(all_trips_v2$ride_length) #midpoint number in the ascending array of ride
lengths
max(all_trips_v2$ride_length) #longest ride
min(all_trips_v2$ride_length) #shortest ride

summary(all_trips_v2$ride_length)

# Compare members and casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)

# average ride time by each day for members vs casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)

# reorganizing the days of the week lineup.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday",
"Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

# average ride time by each day for members vs casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)
```

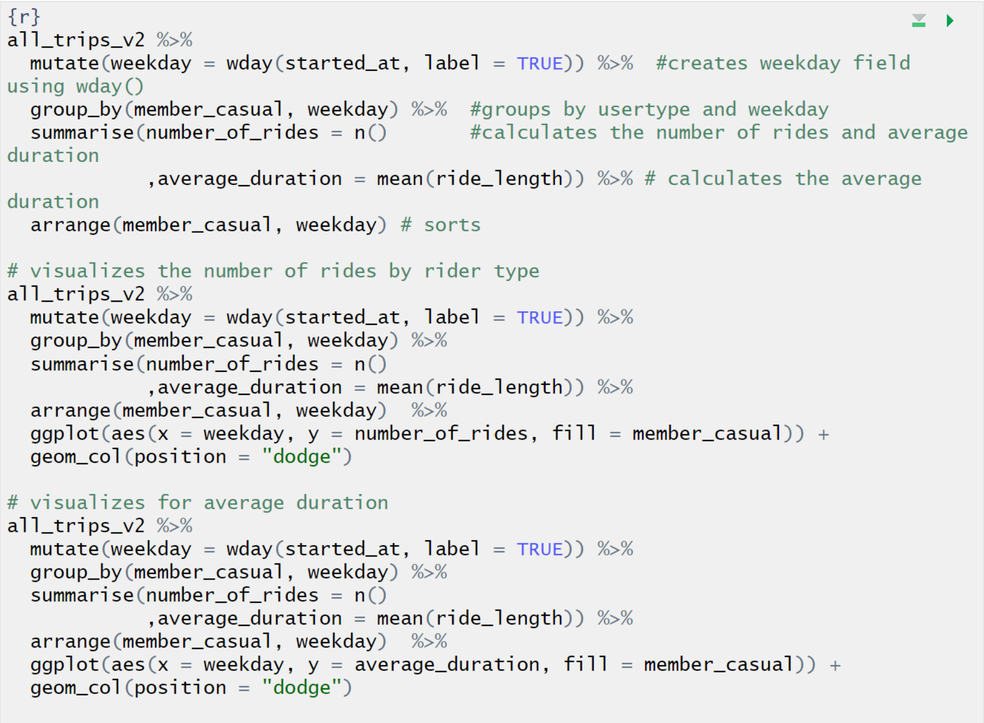
Interpretation of Results

Most Cyclistic users used this service from Tuesdays to Thursdays during 2019 and 2020 in the first three months, where Sundays and Saturdays had the least users. There were more members who use this service than casual customers overall; but the average duration of riders was higher for casual riders than members, where the average ride duration was highest on Thursday for this group. The mean ride length for Cyclistic users is 1,189 seconds or 19.82 minutes, where the minimum is 1 second while the maximum ride_length is 10,628,422 seconds or 177,40.367 minutes.

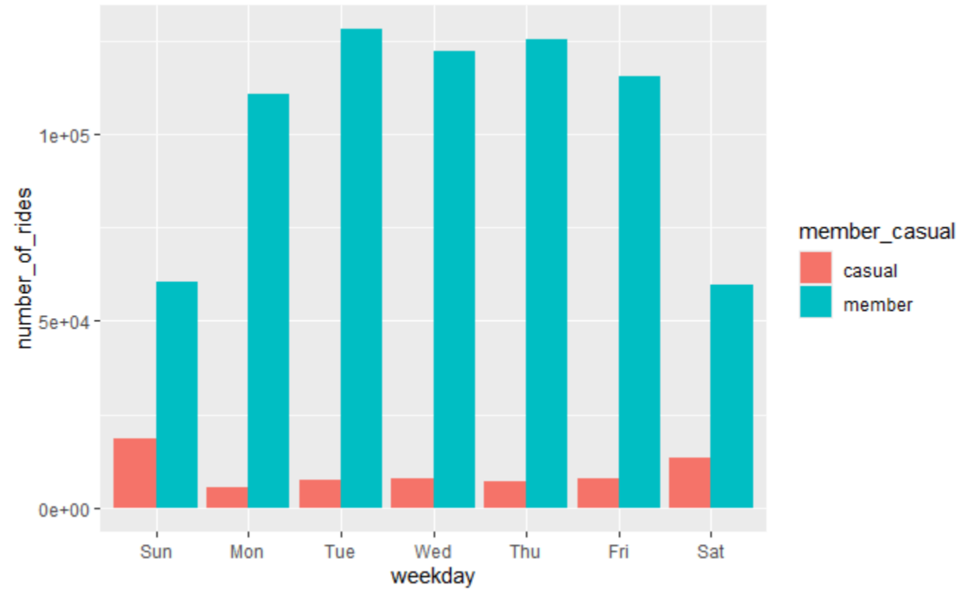
Having other datasets from later in the year would definitely streamline comparison bases. For instance, to put into perspective that this dataset is collected from the early months of 2019 and 2020 (winter season), other assumptions are left unanswered: such as, are these consumer behaviors atypical during the cold? Are users less likely to use bike-share overall, or more? Despite these incongruences, I am able to analyze overall consumer trends with the data presented such as that bike-share use is more frequent from mid-week to later in the week, and the least frequent during the weekend/beginning of the week.

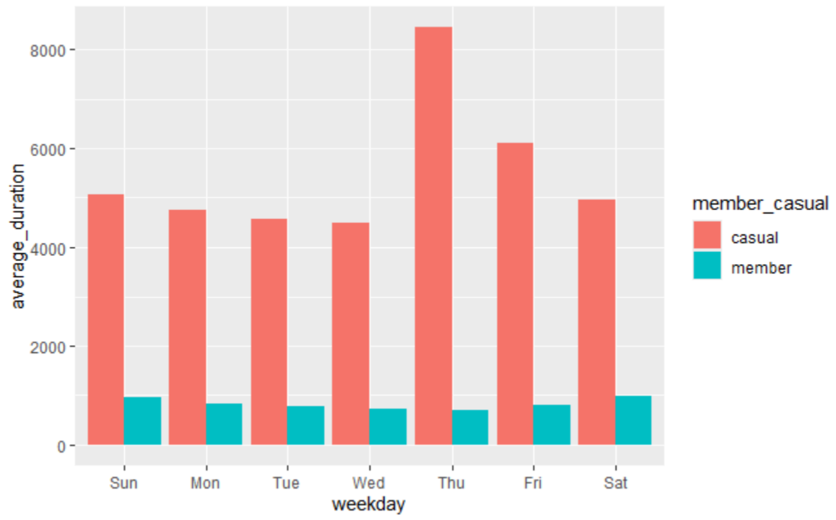
V. Sharing

Visualization

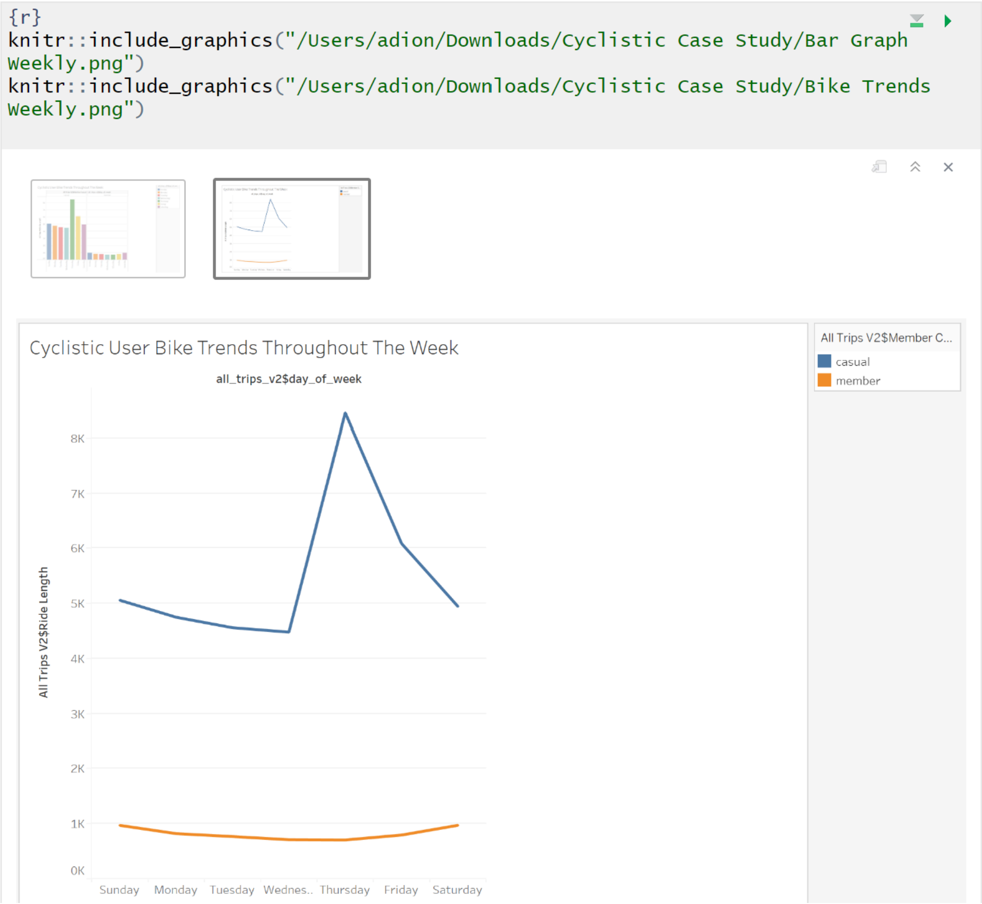


```
{r}
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field 
using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()       #calculates the number of rides and average 
duration
            ,average_duration = mean(ride_length)) %>% # calculates the average 
duration
  arrange(member_casual, weekday) # sorts

# visualizes the number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")

# visualizes for average duration
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

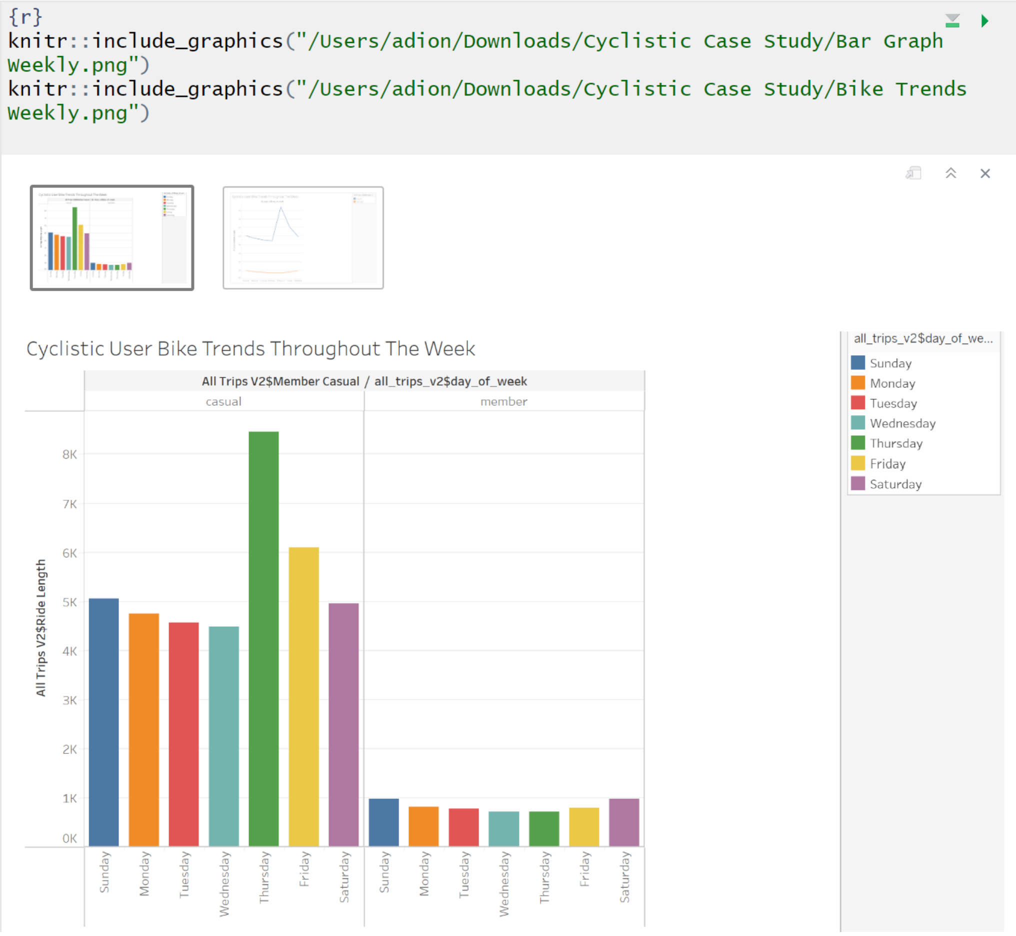




Further visualization using Tableau:



```
{r}
knitr::include_graphics("/Users/adion/Downloads/Cyclistic Case Study/Bar Graph
Weekly.png")
knitr::include_graphics("/Users/adion/Downloads/Cyclistic Case Study/Bike Trends
Weekly.png")
```



```
{r}
knitr::include_graphics("/Users/adion/Downloads/Cyclistic Case Study/Bar Graph
Weekly.png")
knitr::include_graphics("/Users/adion/Downloads/Cyclistic Case Study/Bike Trends
Weekly.png")
```

Interpretation of the Visualizations

The bar graphs presented indicate consistent bike-share usage among rider members throughout the week compared to their casual rider counterparts. We see that the bike-share duration is highest on Thursdays for casual riders and throughout Saturdays in the Tableau bar graph; this spike is evidenced also in the contrast of the line graph. The number of riders overall is highest among the rider members subscribed to the service as seen in the first two figures.

VI. Acting

Conclusion and Recommendations

Consumer-geared Marketing Campaigns:

- **Weekend Promotions:** Since casual riders are more active during weekends, Cyclistic should focus on weekend promotions or discounts to encourage casual riders to transition into members.
- **Longer Ride Packages:** Offer special packages or memberships that cater to longer ride durations, as this aligns with the usage patterns of casual riders.

Digital Media Strategies:

- **Social Media Engagement:** Utilize social media platforms to promote the benefits of membership, highlighting convenience, cost savings, and additional perks. Focus on targeting users during mid-week to weekends when casual usage is higher. Partnerships with fitness influencers would also take advantage of the market for wellness.
- **Email Campaigns:** Personalized email campaigns can be developed based on rider behavior, encouraging casual riders to become members by showcasing the potential savings and added value of membership.

Service Enhancements:

- **Loyalty Programs:** Introduce loyalty programs that reward frequent casual riders with discounts or points that can be redeemed towards an annual membership, or even a "free ride" promo after a certain benchmark of rides on days not similar to the peak in riders.
- **Improved User Experience:** Enhance the user experience on the Cyclistic app and website to make the transition from casual rider to member as seamless as possible. Consider offering a trial period where casual riders can experience the benefits of membership.
- **Personalized Rider-Wellness Goals:** Integrate and launch wellness features such as a mileage or minute or ride tracker based on the rider's fitness/biking goals.

Data-Driven Decision Making:

- **Continuous Monitoring:** Regularly monitor usage patterns and update marketing strategies based on evolving trends. For instance, seasonality might affect rider behavior, and marketing efforts should be adjusted accordingly.
- **Surveys and Feedback:** Implement rider surveys to gather feedback on what features or incentives would most encourage casual riders to become members, and use this data to refine marketing tactics and service offerings.

It is evident that there are distinct patterns between annual members and casual riders. Members tend to use the service consistently throughout the week, while casual riders show higher usage toward the end of the week, particularly on Thursdays and Saturdays. Casual riders generally have longer ride durations compared to members; this indicates that while members rely on the service for regular, perhaps shorter commutes, casual riders might use it more for leisure or longer, occasional trips.