**Case Study 1**

**How Does a Bike-Share Navigate Speedy Success?**

**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 visualization.

**Ask**

1. How do annual members and casual riders use Cyclistic bikes differently?

The business task is to understand how annual members and casual riders use Cyclistic bikes differently.

1. Why would casual riders buy Cyclistic annual memberships?

The business task is to understand how to make annual memberships appeal more to casual riders.

1. How can Cyclistic use digital media to influence casual riders to become members?

The business task is to understand casual riders' demographic and habits so that we can target them better with advertising.

**Prepare**

We will be using Cyclistic’s historical trip data to analyze and identify trends, starting from Jan 2022 to Dec 2022 (most recent 12 months available).

All files have been extracted into separate folders for each month and sorted chronologically.

Naming conventions have also been improved to reduce confusion for other team members.

Old name : 202205-divvy-tripdata  
New name : 2022-05

The Data has been made available by Motivate international Inc.

**Process**

I chose to use Rstudio to process the data since it seemed way too much for spreadsheets.

These are the steps I took:

1. Installing and loading the necessary packages for the task.

#Installing the necessary packages

install.packages('tidyverse')

install.packages('janitor')

install.packages('lubridate')

#Loading the necessary packages

library('tidyverse')

library('janitor')

library('lubridate')

1. The next step is to import all the data that was provided and assign it new names so that it’s much easier to use.

#Importing data and assigning it simplified names

Jan2022 <- read\_csv("divvy-tripdata/2022-01/2022-01.csv")

Feb2022 <- read\_csv("divvy-tripdata/2022-02/2022-02.csv")

Mar2022 <- read\_csv("divvy-tripdata/2022-03/2022-03.csv")

Apr2022 <- read\_csv("divvy-tripdata/2022-04/2022-04.csv")

May2022 <- read\_csv("divvy-tripdata/2022-05/2022-05.csv")

June2022 <- read\_csv("divvy-tripdata/2022-06/2022-06.csv")

July2022 <- read\_csv("divvy-tripdata/2022-07/2022-07.csv")

Aug2022 <- read\_csv("divvy-tripdata/2022-08/2022-08.csv")

Sep2022 <- read\_csv("divvy-tripdata/2022-09/2022-09.csv")

Oct2022 <- read\_csv("divvy-tripdata/2022-10/2022-10.csv")

Nov2022 <- read\_csv("divvy-tripdata/2022-11/2022-11.csv")

Dec2022 <- read\_csv("divvy-tripdata/2022-12/2022-12.csv")

1. Next, I wanted to merge all these data sets into one data frame but first I must confirm that the data sets are compatible for such a merge.

This entails any extra or missing columns, data types, and any other discrepancy.

I decided to use the str() function to perform this verification.

#Inspecting data structure to verify its format

str(Jan2022)

str(Feb2022)

str(Mar2022)

str(Apr2022)

str(May2022)

str(June2022)

str(July2022)

str(Aug2022)

str(Sep2022)

str(Oct2022)

str(Nov2022)

str(Dec2022)

Luckily, the data already matched in terms of type as well as columns so there was no need to transform or adjust anything prior to the merge.

1. Next we merge all datasets into one table that we can call tripdata\_df, and we do so using the bind\_rows function.

#Creating data frame of all the data

tripdata\_df <- data.frame()

#Merging all rows of our data under shared columns

tripdata\_df <- bind\_rows(Jan2022, Feb2022, Mar2022, Apr2022, May2022, June2022, July2022, Aug2022, Sep2022, Oct2022, Nov2022, Dec2022)

1. Now that we have our data frame, we can finally apply functions to our entire data.

We start with cleaning names and removing nulls.

#Cleaning and removing any spaces or parentheses

tripdata\_df <- clean\_names(tripdata\_df)

#Removing all rows and columns from data frame that are composed of NA values

remove\_empty(tripdata\_df, which = c())

1. Now that the data is ready to be used, we must look back to our business task and figure out if any additional data manipulation is required.  
   We conclude that it would be useful to add new columns to our data frame that include “day of the week”, “start hour”, “month”, and “trip duration”.

#Extracting the time in "hour" format from a column inside the data frame

tripdata\_df$starting\_hour <- format(as.POSIXct(tripdata\_df$started\_at), '%H')

#Extracting the date from a column inside the data frame

tripdata\_df$month <- format(as.Date(tripdata\_df$started\_at), '%m')

#Calculates the time difference between two columns within a data frame and displays it in desired unit

tripdata\_df$trip\_duration <- difftime(tripdata\_df$ended\_at, tripdata\_df$started\_at, units = 'sec')

1. Our final step is to clean the data one last time and put everything we got in a new data frame.This data frame excludes any trip duration of 0 or less seconds.  
   We also add the “day of the week” here after creating “cleaned\_df”

#Creating a new data frame that excludes any row with trip duration of 0 seconds or less

Cleaned\_df <- tripdata\_df[!(tripdata\_df$trip\_duration <= 0),]

#Extracting the day from the date column

day\_of\_the\_week <- wday(tripdata\_df$started\_at, label = T, abbr = T)

**Analyze & Share**

It’s very important to always think about the business task at every stage of the analysis.

The questions asked were:

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?

To answer these questions, we must first understand how annual members and casual riders use Cyclistic bikes differently.

We will be using both visual and descriptive analysis to understand the relationships that exist within our data.

For our visual analysis:

#Graphing number of rides by member type, then saving it

#Using scipen to remove scientific notation, and position = dodge to make sure member and casual don't stack on top of each other

options(scipen = 999)

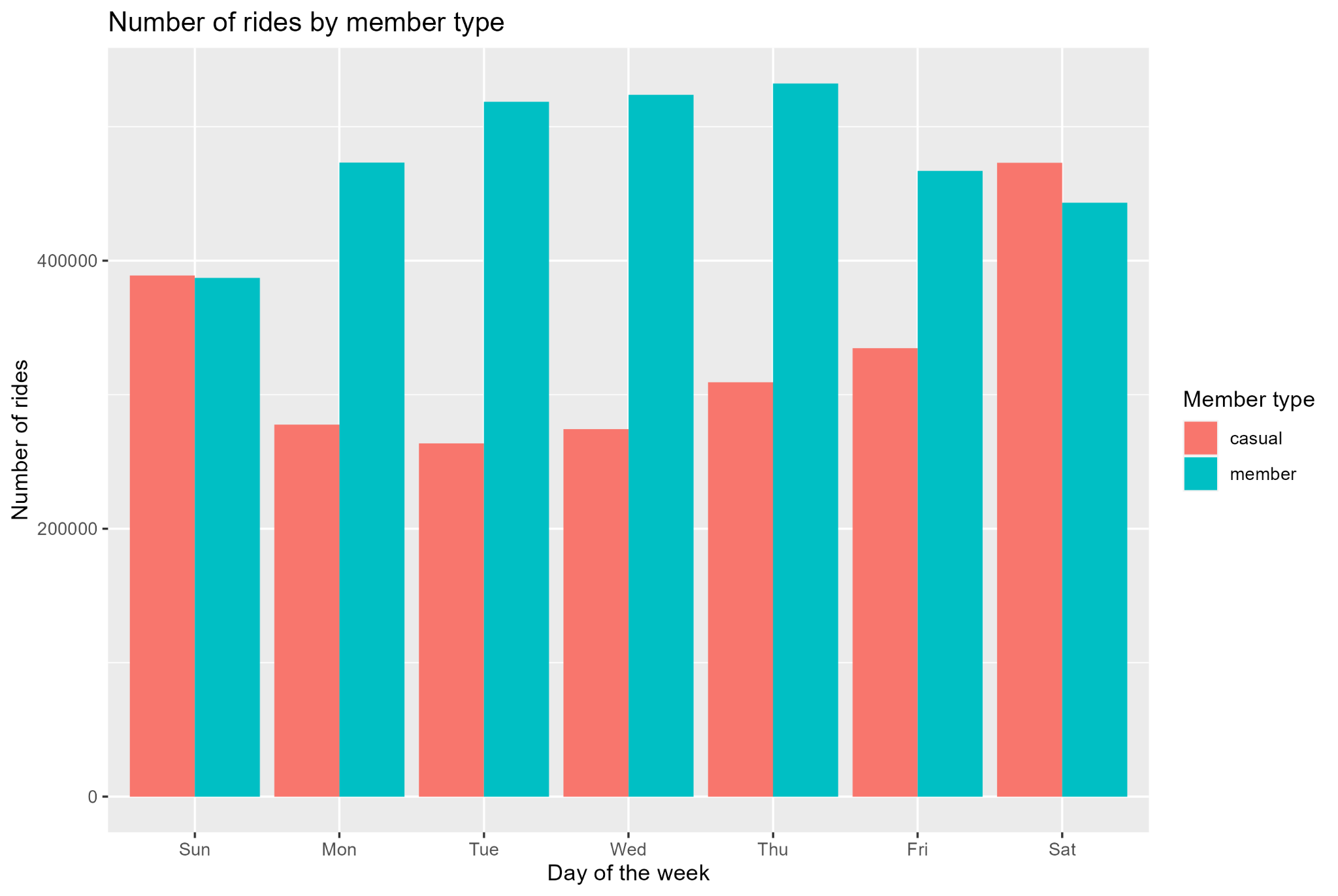
ggplot(data = Cleaned\_df) +

aes(x = day\_of\_the\_week, fill = member\_casual) +

geom\_bar(position = 'dodge') +

labs(x = 'Day of the week', y = 'Number of rides', fill = 'Member type', title = 'Number of rides by member type')

ggsave("number\_of\_rides\_by\_member\_type.png")



During the week, members tend to use Cyclistic bikes almost twice as much as casuals; however on the weekend both members and casuals use Cyclistic bikes at the same rate.

This can indicate that the members are using these bikes to commute to work or school during the weekdays; however on the weekends their use of these bikes becomes the same as the casuals.

#Graphing number of rides per month, then saving it

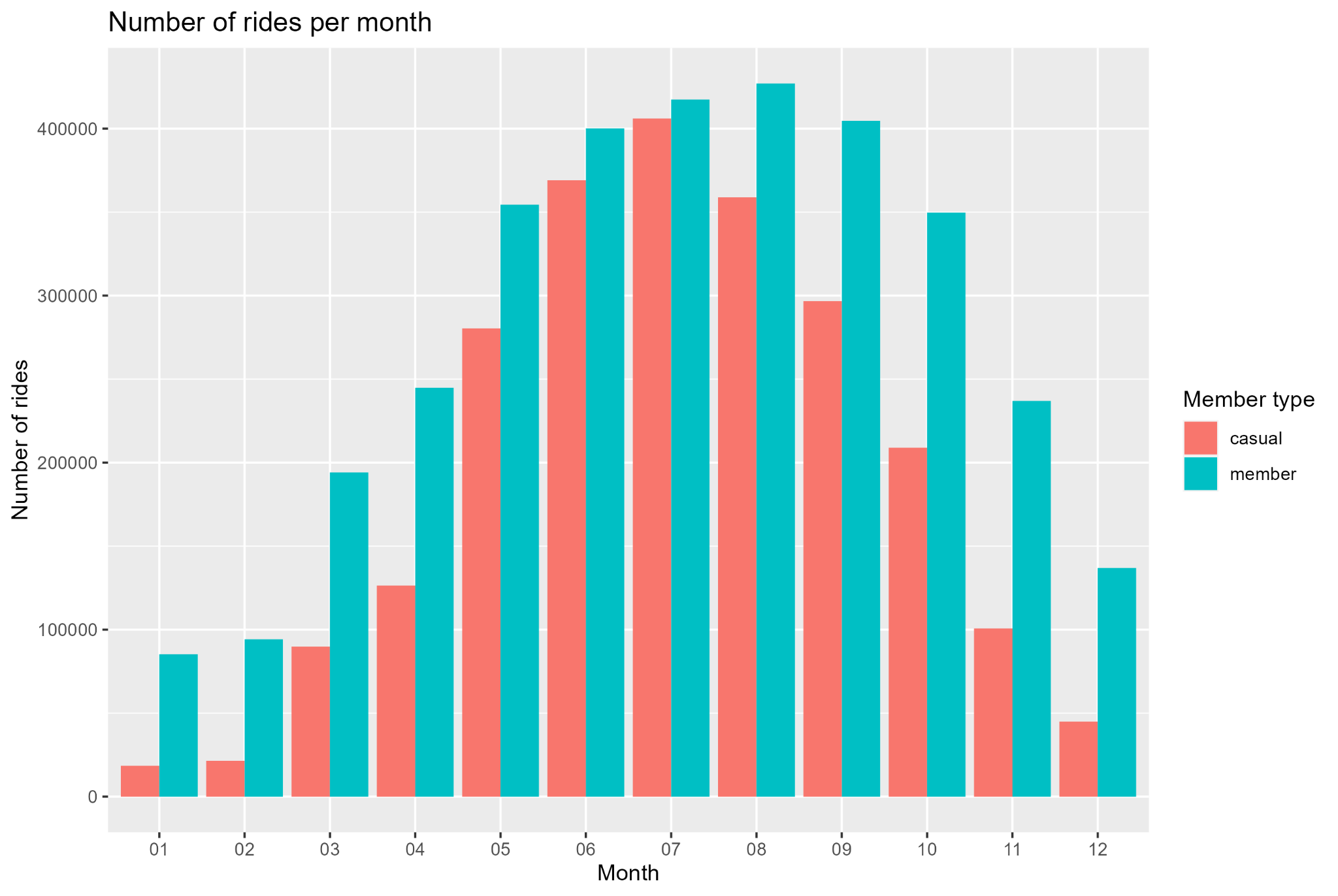
ggplot(data = Cleaned\_df)+

aes(x = month, fill = member\_casual)+

geom\_bar(position = 'dodge')+

labs(x = 'Month', y = 'Number of rides', fill = 'Member type', title = 'Number of rides per month')

ggsave("number\_of\_rides\_per\_month.png")



After observing this graph of the number of rides per month for each type of member, we can clearly see that the use of bikes is greatly less for the colder months such as January, February, March, April, November, and December. This is especially true for casual members. In the warmer months during the year such as June and July, it’s clear that both members and casuals are using bikes at a similar rate. This relationship is proof that biking is a seasonal activity; however members are still using bikes throughout the whole year more consistently than casuals. This indicates the value in converting casuals into members.

#Graphing the hourly use of bikes throughout the week, then saving it

ggplot(data = Cleaned\_df)+

aes(x = starting\_hour, fill = member\_casual)+

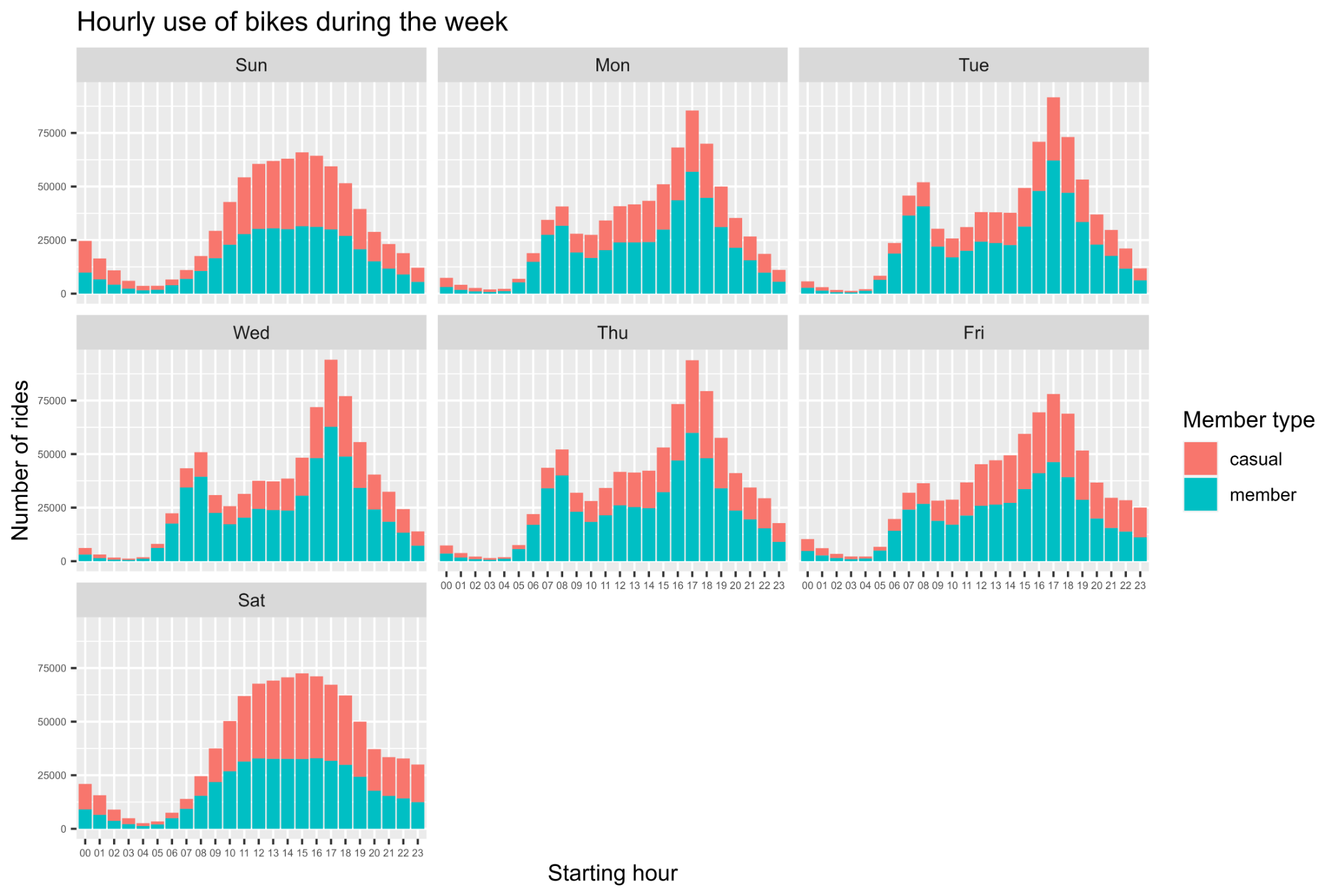
facet\_wrap(day\_of\_the\_week)+

geom\_bar()+

labs(x = 'Starting hour', y = 'Number of rides', fill = 'Member type', title = 'Hourly use of bikes during the week')+

theme(axis.text = element\_text(size = 5))

ggsave("Hourly\_use\_of\_bikes\_during\_the\_week.png", dpi = 1000)



After observing the series of graphs that were generated by the facet\_wrap function of the hourly use of bikes on all seven days of the week. We can clearly see that both members and casuals are actively using the bikes around the same time each day for a similar amount of rides. This observation is especially true on the weekdays; however the casuals are using the bikes much more frequently during the weekends than the members at almost double the rate.

The reason for this can be that casuals are using the bikes for leisure purposes while members use it for transportation mostly.

For our descriptive analysis:

#Using aggregate to compare trip duration to member type on each day of the week

aggregate(Cleaned\_df$trip\_duration ~ Cleaned\_df$member\_casual + day\_of\_the\_week, FUN = mean)

Cleaned\_df$member\_casual day\_of\_the\_week Cleaned\_df$trip\_duration

1 casual Sun 2043.6343 secs

2 member Sun 841.9355 secs

3 casual Mon 1751.3805 secs

4 member Mon 736.2531 secs

5 casual Tue 1549.5189 secs

6 member Tue 727.8171 secs

7 casual Wed 1485.1319 secs

8 member Wed 726.3364 secs

9 casual Thu 1533.0238 secs

10 member Thu 737.6191 secs

11 casual Fri 1682.8110 secs

12 member Fri 751.8978 secs

13 casual Sat 1957.0725 secs

14 member Sat 848.4573 secs

When observing the results of our aggregate function, it’s clear that the trip duration for casuals is much higher than it is for members. This is true for every day of the week.

Longer trips are often for leisure purposes which further confirms our hypothesis from before.

#counting the number of times each station used then sorting it to find out the most popular stations per member type

count(filter(Cleaned\_df, member\_casual =='member'), start\_station\_name, sort = T)

# A tibble: 1,497 × 2

start\_station\_name n

*<chr>* *<int>*

1 NA 485645

2 Kingsbury St & Kinzie St 24936

3 Clark St & Elm St 22037

4 Wells St & Concord Ln 21298

5 University Ave & 57th St 19949

6 Clinton St & Washington Blvd 19827

7 Ellis Ave & 60th St 19501

8 Loomis St & Lexington St 19128

9 Wells St & Elm St 18986

10 Clinton St & Madison St 18931

count(filter(Cleaned\_df, member\_casual =='member'), end\_station\_name, sort = T)

# A tibble: 1,483 × 2

end\_station\_name n

*<chr>* *<int>*

1 NA 483758

2 Kingsbury St & Kinzie St 24636

3 Clark St & Elm St 22364

4 Wells St & Concord Ln 21914

5 University Ave & 57th St 20531

6 Clinton St & Washington Blvd 20529

7 Clinton St & Madison St 19725

8 Ellis Ave & 60th St 19197

9 Loomis St & Lexington St 18903

10 Wells St & Elm St 18721

count(filter(Cleaned\_df, member\_casual =='casual'), start\_station\_name, sort = T)

# A tibble: 1,599 × 2

start\_station\_name n

*<chr>* *<int>*

1 NA 347393

2 Streeter Dr & Grand Ave 58083

3 DuSable Lake Shore Dr & Monroe St 31860

4 Millennium Park 25524

5 Michigan Ave & Oak St 25264

6 DuSable Lake Shore Dr & North Blvd 23655

7 Shedd Aquarium 20265

8 Theater on the Lake 18450

9 Wells St & Concord Ln 16215

10 Dusable Harbor 14101

count(filter(Cleaned\_df, member\_casual =='casual'), end\_station\_name, sort = T)

# A tibble: 1,617 × 2

end\_station\_name n

*<chr>* *<int>*

1 NA 408778

2 Streeter Dr & Grand Ave 59870

3 DuSable Lake Shore Dr & Monroe St 29602

4 Millennium Park 26675

5 Michigan Ave & Oak St 26449

6 DuSable Lake Shore Dr & North Blvd 26143

7 Theater on the Lake 19417

8 Shedd Aquarium 18631

9 Wells St & Concord Ln 15506

10 Clark St & Armitage Ave 13849

When observing the results of our count functions, it’s clear that “Kingsbury St & Kinzie St” is the most popular start and end station for members. While “Streeter Dr & Grand Ave” is the most popular start and end station for casuals.

**Act**

One final time, let’s refer back to our business task questions.

1. How do annual members and casual riders use Cyclistic bikes differently?

The business task is to understand how annual members and casual riders use Cyclistic bikes differently. Based on our analysis, we learned that casuals are mainly composed of tourists or individuals that love to spend weekends sightseeing and enjoying the outdoors. We also have learned that annual members are mainly composed of working adults or students that use the bikes as a form of transportation while sometimes also enjoying sightseeing and the outdoors on the weekend.

1. Why would casual riders buy Cyclistic annual memberships?

The business task is to understand how to make annual memberships appeal more to casual riders. Based on our analysis, we learned that casuals enjoy using our services since it seems that the majority of their use happens to be for leisure purposes. We also know it’s greatly beneficial for us to convert casual members to annual members so that even during colder months in the year, we can still expect their business. In conclusion, the best method to convert casual users to annual members is to offer them a special promotion. Casuals clearly enjoy the services provided but perhaps don’t think the commitment behind the membership is worth it in terms of saving money.

1. How can Cyclistic use digital media to influence casual riders to become members?

The business task is to understand casual riders' demographic and habits so that we can target them better with advertising. Based on our analysis it’s clear that casual riders mainly consist of tourists or individuals that love to spend weekends sightseeing and enjoying the outdoors. We also know that casual riders use our services the most in the months of June and July. We also know that they’re most active on the weekends and between the hours of 12:00 pm to 8:00pm.

Finally, we know their favorite stations in order of most popular to least popular. Advertising to our casual riders using these parameters should be very effective.