Google Data Analytics Certificate Capstone Project 1 - Cyclistic Bike Share

Divvy_Exercise_Full_Year_Analysis

This analysis is based on the Divvy case study "'Sophisticated, Clear, and Polished': Divvy #and Data Visualization" written by Kevin Hartman (found here: [link phase] (https://artscience.blog/home/divvy-dataviz-case-study (https://artscience.blog/home/divvy-dataviz-case-study)). The purpose of this script is to consolidate downloaded Divvy data into a single dataframe and then conduct simple analysis to help answer the key question: "In what ways do members and casual riders use Divvy bikes differently?"

#install packages

```
install.packages("tidyverse")
install.packages("janitor")
install.packages("lubridate")
install.packages("readxl")
install.packages("dplyr")
```

#install libraries

```
library(tidyverse)
library(janitor)
library(lubridate)
library(readxl)
library(ggplot2)
library(scales)
library(dplyr)
```

#read data from csv files

```
df1<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202103-divv
y-tripdata.csv")
df2<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202104-divv
y-tripdata.csv")
df3<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202105-divv
y-tripdata.csv")
df4<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202106-divv
y-tripdata.csv")
df5<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202107-divv
y-tripdata.csv")
df6<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202108-divv
y-tripdata.csv")
df7<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202109-divv
y-tripdata.csv")
df8<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202110-divv
y-tripdata.csv")
df9<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202111-divv
y-tripdata.csv")
df10<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202112-div
vy-tripdata.csv")
df11<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202201-div
vy-tripdata.csv")
df12<- read.csv("/Users/dean/Desktop/Google Data Analytics Course Certificate/202202-div
vy-tripdata.csv")
```

#Combine into 1 dataset

```
bike_rides <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
dim(bike_rides)
bike_rides <- janitor::remove_empty(bike_rides,which = c("cols"))
bike_rides <- janitor::remove_empty(bike_rides,which = c("rows"))
dim(bike_rides)
```

#Change Column Names for Start and End Date

```
bike_rides$start_date <- as.Date(bike_rides$started_at)
bike_rides$end_date <- as.Date(bike_rides$ended_at)
```

#Calculate Casual vs Member Riders

```
table(bike_rides$member_casual)
```

#Modify Start and End Date Format

```
bike_rides$date <- as.Date(bike_rides$started_at) #The default format is yyyy-mm-dd
bike_rides$month <- format(as.Date(bike_rides$date), "%m")
bike_rides$day <- format(as.Date(bike_rides$date), "%d")
bike_rides$year <- format(as.Date(bike_rides$date), "%Y")
bike_rides$day_of_week <- format(as.Date(bike_rides$date), "%A")

bike_rides$ride_length <- difftime(bike_rides$ended_at,bike_rides$started_at)
```

#Create new bike rides v2 variable with null data/QA removed

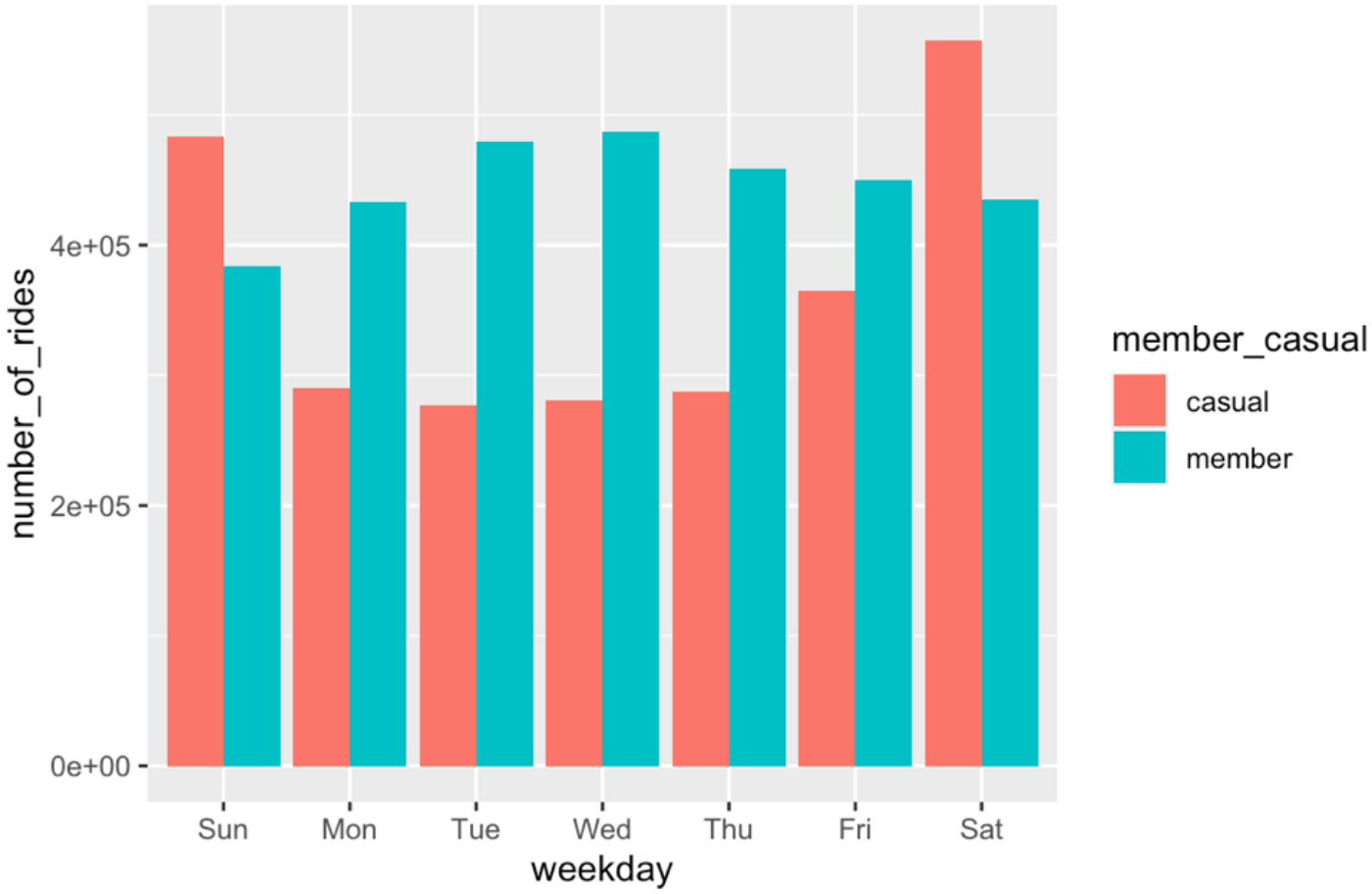
```
is.factor(bike_rides$ride_length)
bike_rides$ride_length <- as.numeric(as.character(bike_rides$ride_length))
is.numeric(bike_rides$ride_length)


bike_rides_v2 <- bike_rides[!(bike_rides$start_station_name == "HQ QR" | bike_rides$ride
_length<0),]
```

#Descriptive analysis on ride_length (all figures in seconds)

```
mean(bike_rides_v2$ride_length) #straight average (total ride length / rides)
median(bike_rides_v2$ride_length) #midpoint number in the ascending array of ride length
s
max(bike_rides_v2$ride_length) #longest ride
min(bike_rides_v2$ride_length) #shortest ride

aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual, FUN = mean)
aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual, FUN = median)
aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual, FUN = max)
aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual, FUN = min)
aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual + bike_rides_v2$day_of
_week, FUN = mean)

bike_rides_v2$day_of_week <- ordered(bike_rides_v2$day_of_week, levels=c("Sunday", "Mond
ay", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

aggregate(bike_rides_v2$ride_length ~ bike_rides_v2$member_casual + bike_rides_v2$day_of
_week, FUN = mean)

bike_rides_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wda
y()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                           #calculates the number of ri
des and average duration
            ,average_duration = mean(ride_length)) %>%      # calculates the average dur
ation
  arrange(member_casual, weekday)
```

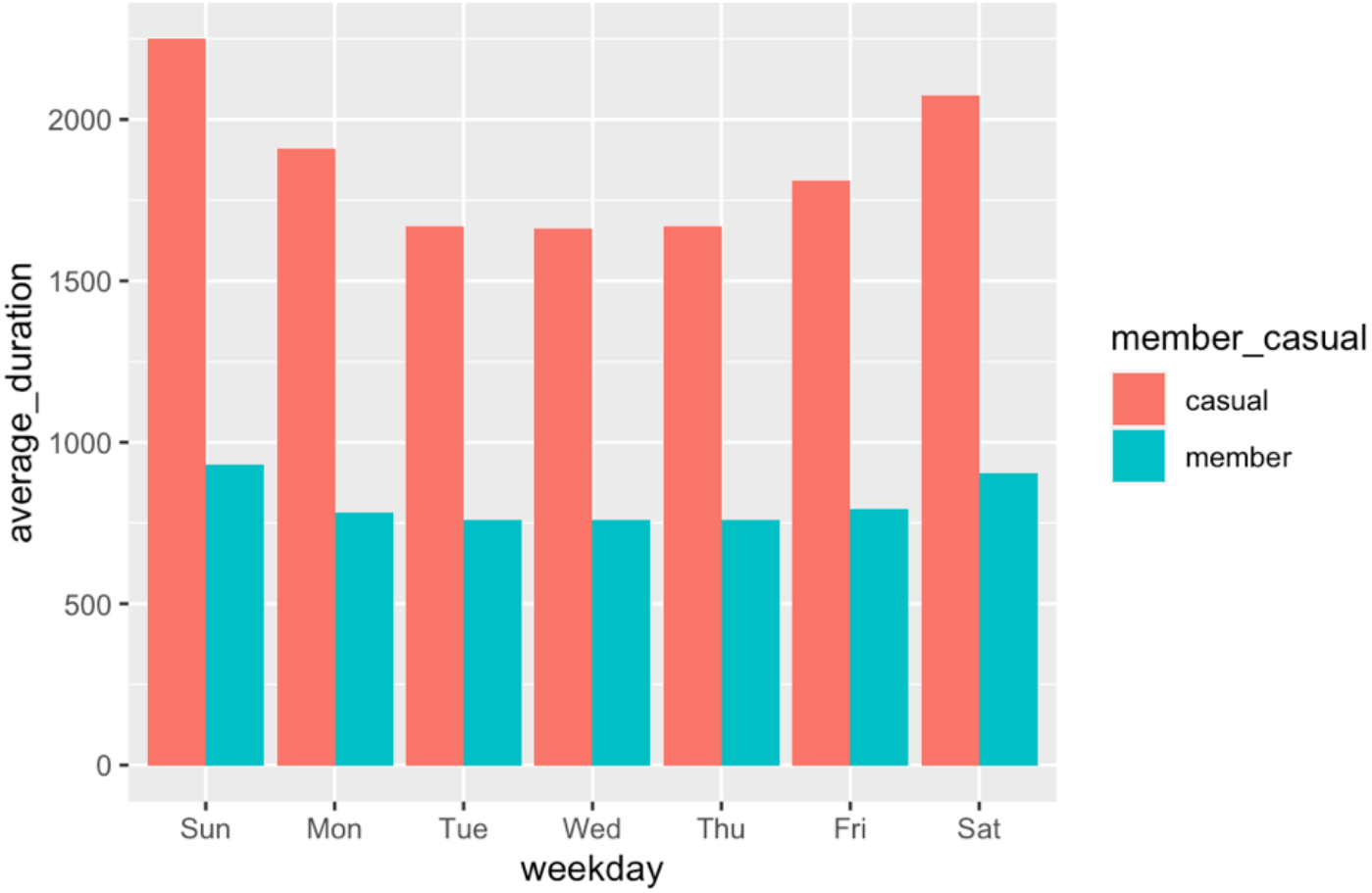
#Plot of bike usage between casual and paid riders during the week

```
plot1 <- bike_rides_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")
```



#Plot of bike ride length between casual and paid riders during the week

```
plot2 <- bike_rides_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")
```



Conclusion: I would preface that while we have pieces of data that suggest how casual and member cyclists use Cyclistic's services differently, we do not understand conclusively why the 2 groups behave differently based on the data we have available. For instance – we don't have any social-economic data for the casual and member riders, or why members chose to become paid subscribers in the first place.

However, given that we know casual riders utilize bike services more on the weekends and for longer durations – Cyclistic's management could target a marketing campaign that promotes more casual riders to use bicycles on the weekend which include:

- Offer a middle pricing tier subscription allowing for unlimited rides in the weekend, whereas the weekdays would be pay as you go.