Introduction

Welcome to this Data Analytics Case Study! This is the capstone project for the google data analytics professional certificate on coursera.

Scenario

In this project Cyclistic is 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.

Business task

Use the power of data analytics to turn more casual riders into annual members!

Cyclistic
Data Analytics Case Study

PHASE 1: ASK

Objectives:

1. **Identify the Business Task**
   Questions to be answered:
   - How do annual members and casual riders use Cyclistic bikes differently?
   - Why would casual riders buy Cyclistic annual membership?
   - How can Cyclists use digital media to influence casual riders to become members?
2. **Consider Key stakeholders**
   - The key stakeholders of this case study are the Cyclistic executive team, the Cyclistic marketing analytics team and Lily Moreno the director of marketing.
3. **Statement of the Business task**
   - Work with the analytics team to help create Cyclistic marketing strategy identifies the behavior and trend within the data, with the goal to convert casual riders into annual members. Create data insight through visualization recommendations Cyclistic executives will approve.

PHASE 2: PREPARE

Objectives:

Download data and store it appropriately

- The data is originally stored on Index of bucket "divvy-tripdata"
  The data is now stored both locally and on the github for data redundancy, and data is password protected for security measures.

Identify how it's organized.

- The data are in the .csv(comma separated values) file format. Each row represents a trip and each trip contains 12 columns of attributes. They are :trip's id, start and end time of each trip, the bike's type, the trip duration in seconds,start and end bike station ID and name, then user type(subscriber or not) and the position of the start and end point.

Sort and filter the data.

- The data is stored in 12 separate .csv files each containing more than 100 thousands rows of data, for the process phase ahead we need to import the data into Rstudio for sort and filter, because Excel and Spreadsheet Can't handle data this size efficiently.
- All the csv files will be combined into one single data set.

First we will load csv into Rstudio data set

```
> data1 <- read.csv("CSV/202101-divvy-tripdata.csv")
> data2 <- read.csv("CSV/202102-divvy-tripdata.csv")
> data3 <- read.csv("CSV/202103-divvy-tripdata.csv")
> data4 <- read.csv("CSV/202104-divvy-tripdata.csv")
> data5 <- read.csv("CSV/202105-divvy-tripdata.csv")
> data6 <- read.csv("CSV/202106-divvy-tripdata.csv")
> data7 <- read.csv("CSV/202107-divvy-tripdata.csv")
>
> data8 <- read.csv("CSV/202108-divvy-tripdata.csv")
> data9 <- read.csv("CSV/202109-divvy-tripdata.csv")
> data10 <- read.csv("CSV/202110-divvy-tripdata.csv")
> data11 <- read.csv("CSV/202111-divvy-tripdata.csv")
> data12 <- read.csv("CSV/202112-divvy-tripdata.csv")
```

Then we will run “colnames()” on each data set to inspect that all column names matches

```
colnames()
```

Then we could combine all 12 data set into one with

```
trip_data <-
bind_rows(data1,data2,data3,data4,data5,data6,data7,d
ata8,data9,data10,data11,data12)
```

Rstudio returns and the merge was successful, the trip_data now has 5595063 rows and 13 columns.

trip_data	5595063 obs. of 13 variables

And at this stage the data is prepared for further process. The last step in this phase is to remove duplicated data.

```
duplicated(trip_data)
```

Phase 4: process

Cyclistic
Data Analytics Case Study

Objectives:

1. **Check data for errors**
    - Clean empty data

      Upon inspection it's clear that some of the data was missing from the data set, the best strategy here is to remove them and work on the rest.

      ```
      na.omit(trip_data)
      ```

      And the Rstudio returned

```
[ reached 'max' / getOption("max.print") -- omitted 5590216 rows ]
```

2. **Choose right tool**
    - The R tool package we are using as follow:

```
library(tidyverse)
library(janitor)
library(skimr)
library(lubridate)
```

3. **Transform the data**
    - To make the analysis process more effective, we will perform data transformations.
        1. First we will need the day of week extracted from the data set.

```
##-- extract the day of week from the started_at
> trip_data$day <- format(as.Date(trip_data$started_at),"%A")
```

        2. We will add a column called “ride_length_s” and “ride_length_m” to represent each rides’ time length in seconds and minutes.

```
##-- calculate the time difference between the start and end time of
each ride in second
trip_data$ride_length_s <-
difftime(trip_data$ended_at,trip_data$started_at)
##-- divide second by 60 to calculate time in minutes
trip_data$ride_length_m <- trip_data$ride_length_s/60
```

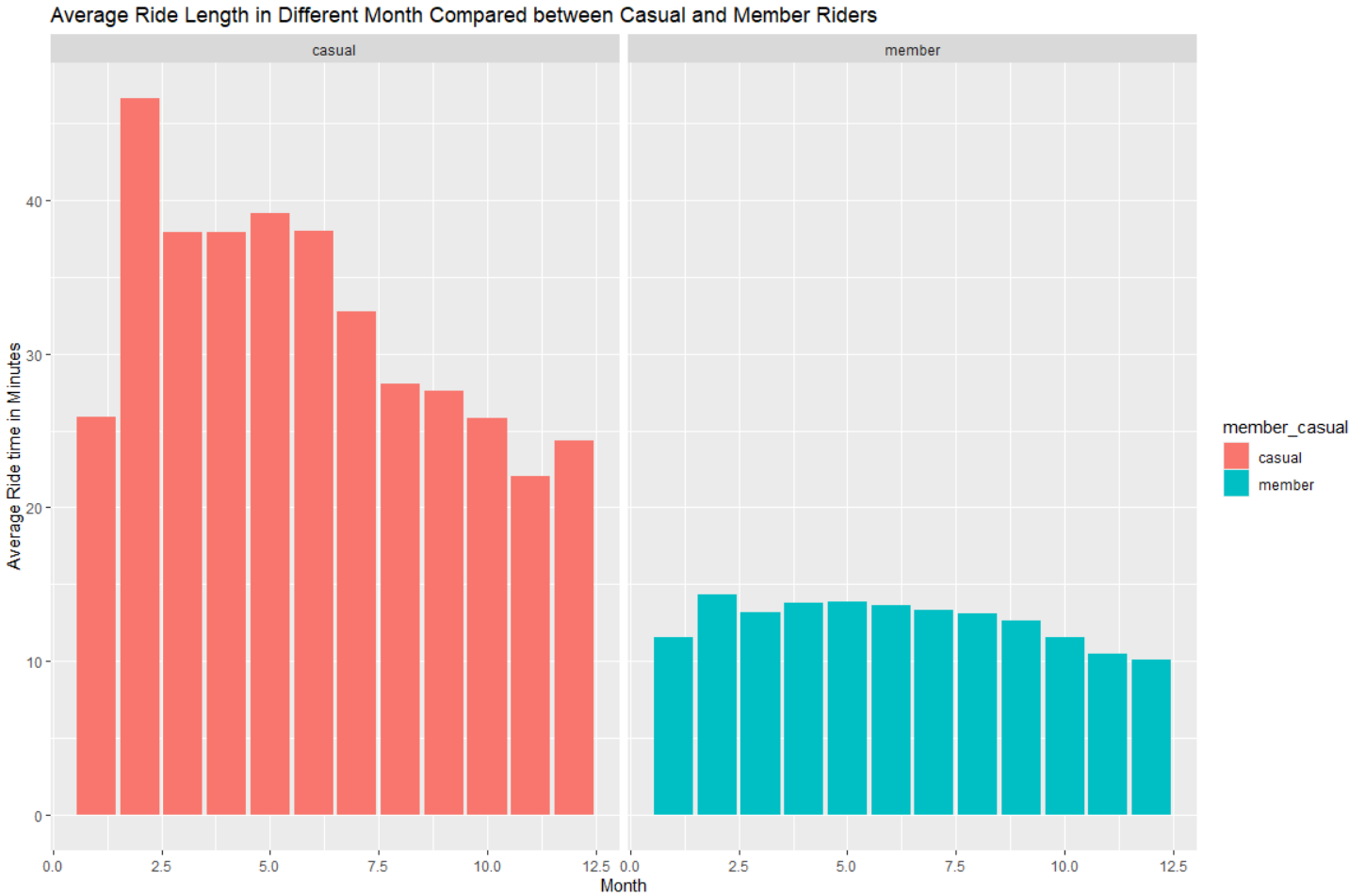
Phase 5: analysis & share

Cyclistic
Data Analytics Case Study

Analysis 1: how do casual riders and members ride differently?

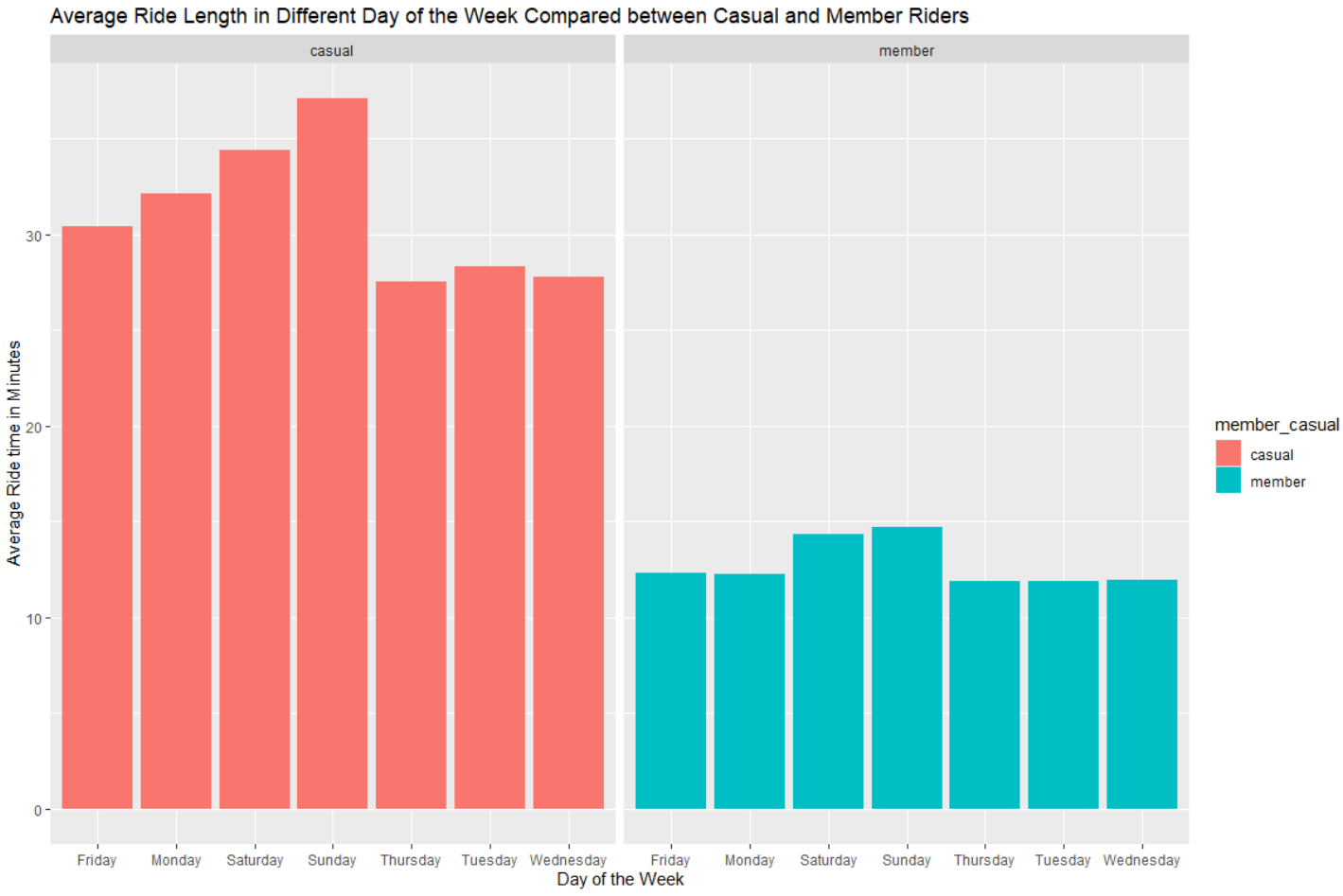
1. Different ride length

```
##-- plot the graph shows the relation between average ride time
##--length(minutes) vs Month For both casual and member riders

trip_data %>%
group_by(member_casual,month) %>%
summarise(num_of_ride = n(),average_length = mean(ride_length_m)) %>%
arrange(member_casual) %>%
ggplot(aes(x = month, y = average_length, group = member_casual)) +
geom_col(aes(fill = member_casual)) + facet_wrap(~member_casual) +
labs(title = "Average Ride Length in Different Month Compared between
Casual and Member Riders") +
ylab("Average Ride time in Minutes") +
xlab("Month")
```



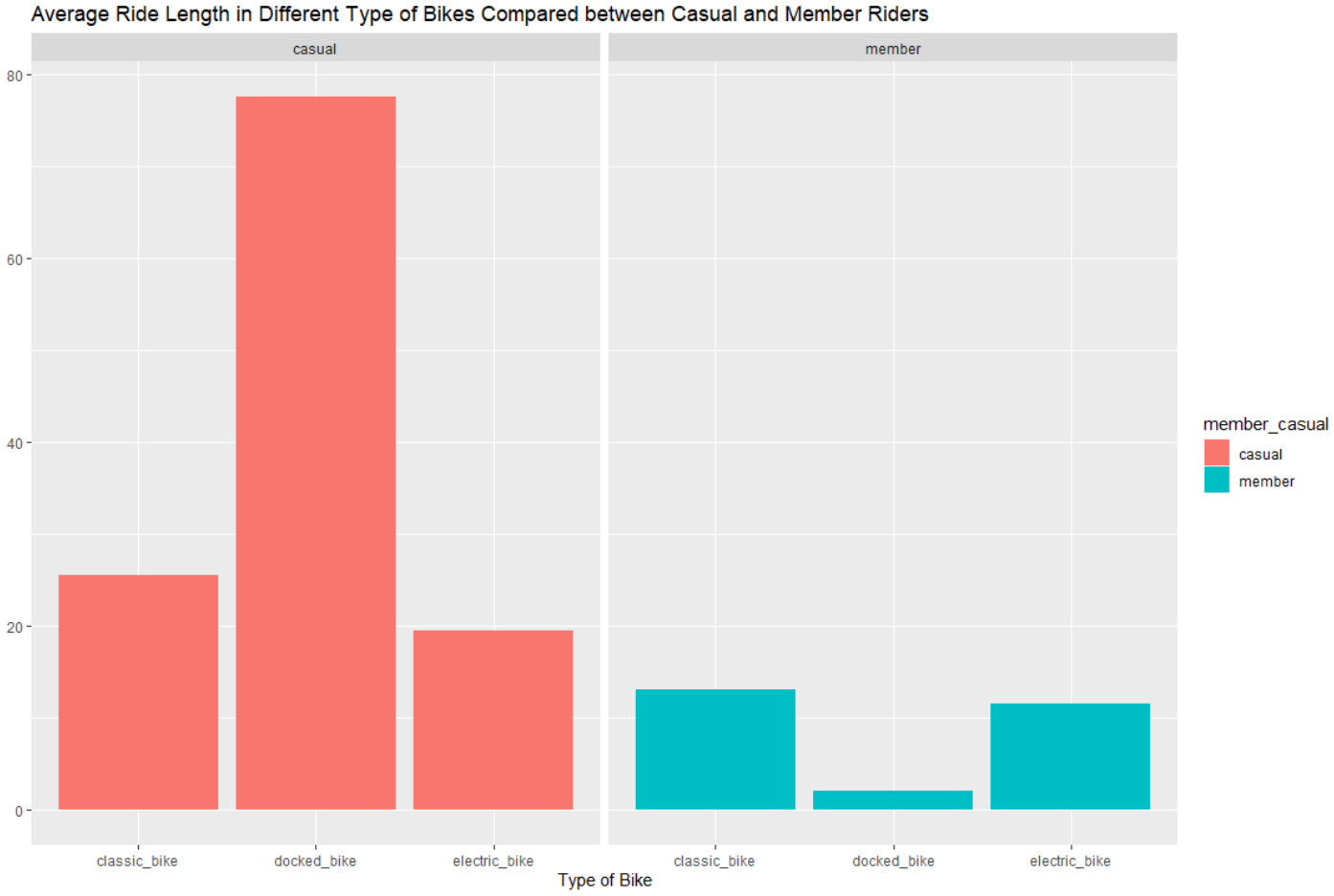
2. Different time of week

```
##-- Plot the graph between average ride time length and the day of the
##-- week for both Casual and Member riders

trip_data %>%
group_by(member_casual,day) %>%
summarise(num_of_ride = n(),average_length = mean(ride_length_m)) %>%
arrange(member_casual) %>%
ggplot(aes(x = day, y = average_length, group = member_casual)) +
geom_col(aes(fill = member_casual)) + facet_wrap(~member_casual) +
labs(title = "Average Ride Length in Different Day of the Week Compared
between Casual and Member Riders") +
ylab("Average Ride time in Minutes") +
xlab("Day of the Week")
```

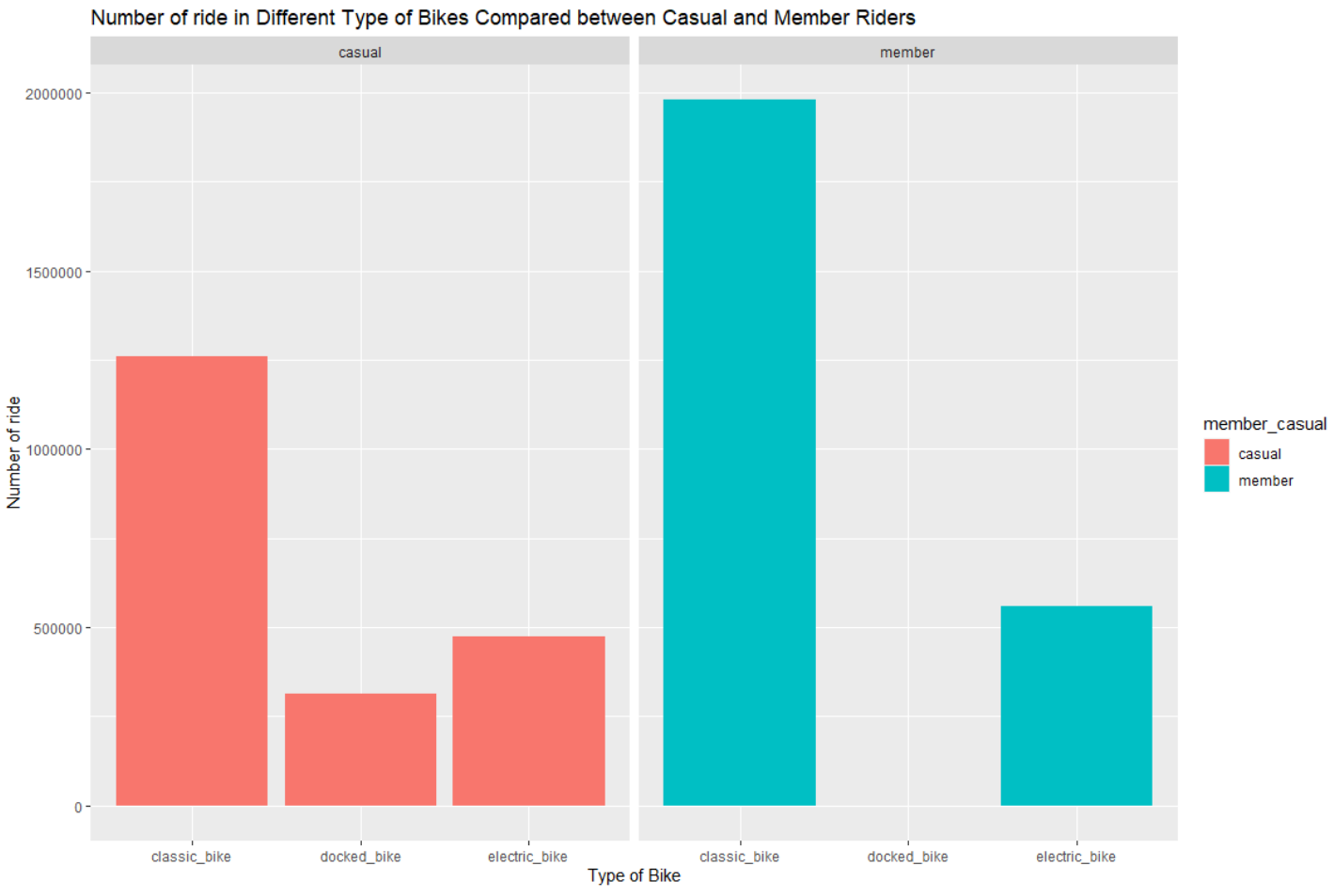


3. Different type of bike

```
##-- plot the graph show the relation between bike type and average ride
##-- time length for both casual and member riders

trip_data %>% group_by(member_casual,rideable_type) %>%
summarise(num_of_ride = n(),average_length = mean(ride_length_m)) %>%
arrange(member_casual) %>% ggplot(aes(x = rideable_type, y =
average_length, group = member_casual)) + geom_col(aes(fill =
member_casual)) + facet_wrap(~member_casual) + labs(title = "Average Ride
Length in Different Type of Bikes Compared between Casual and Member
Riders") + ylab("Average Ride time in Minutes") + xlab("Type of Bike")
```



```
##--plot the graph for number of rides in different type of bikes
compared ##--between Casual and member riders

trip_data %>%
group_by(member_casual,rideable_type) %>%
summarise(num_of_ride = n(),average_length = mean(ride_length_m))
%>% arrange(member_casual) %>%
ggplot(aes(x = rideable_type, y = num_of_ride, group =
member_casual)) + geom_col(aes(fill = member_casual)) +
facet_wrap(~member_casual) +
labs(title = "Number of ride in Different Type of Bikes Compared
between Casual and Member Riders") +
ylab("Number of ride") +
xlab("Type of Bike")
```

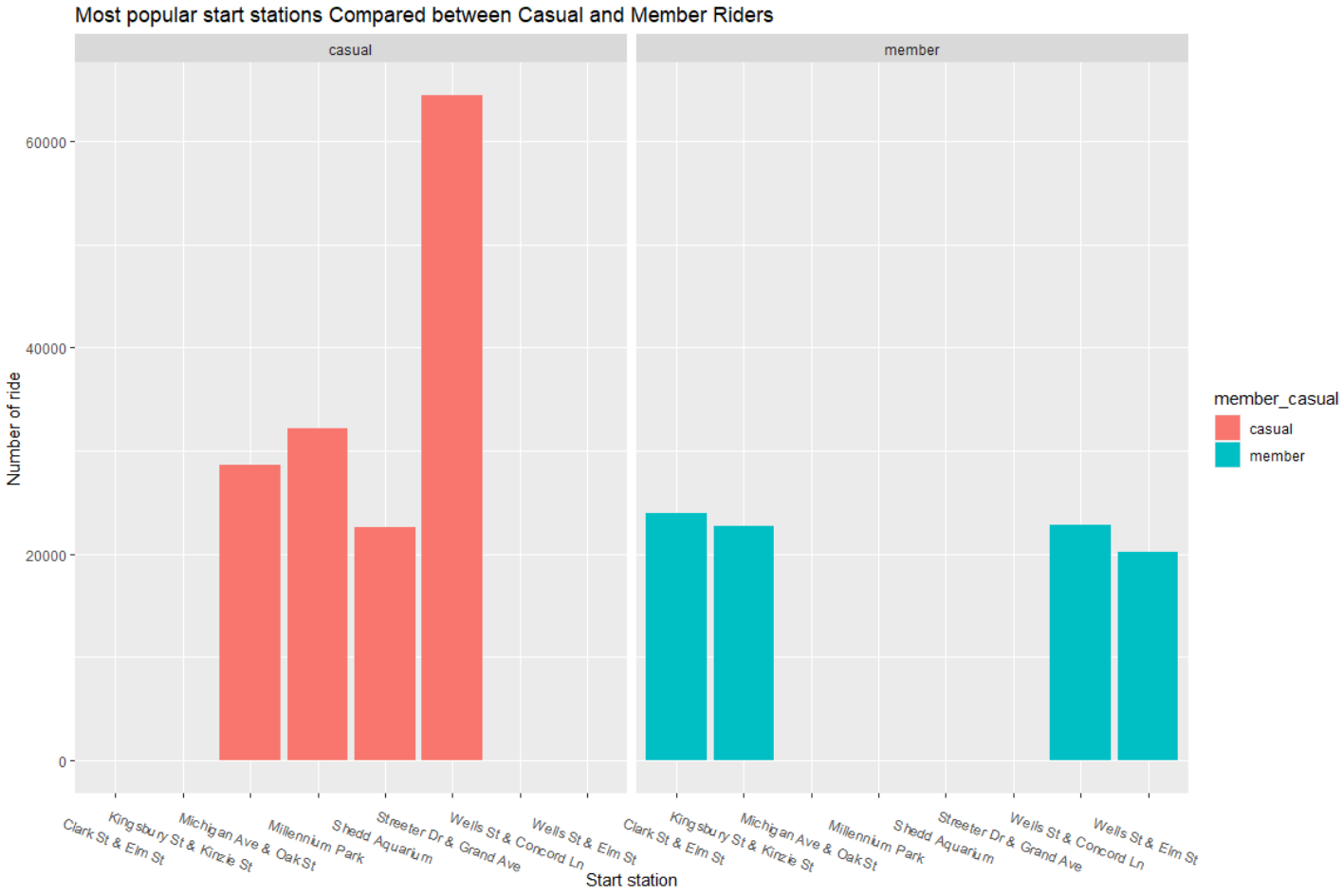


Cyclistic
Data Analytics Case Study

Analysis 2: where do riders start and end their ride?

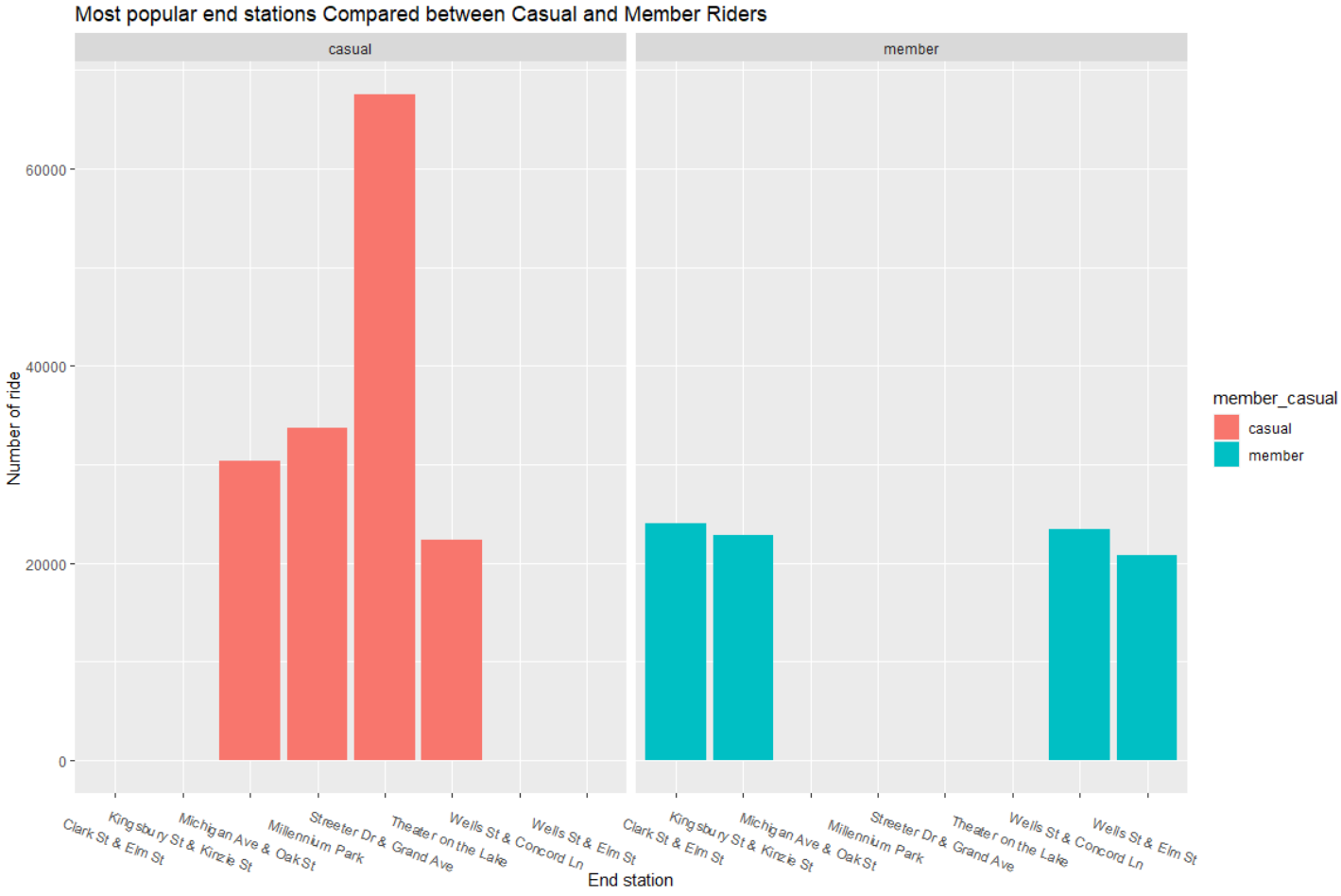
1. What is the most visited start station?

```
##-- plot graph for most visited start station
trip_data %>%
group_by(member_casual,start_station_name) %>%
summarise(num_of_ride = n(),average_length =
mean(ride_length_m),start_name = n()) %>%
arrange(desc(start_name))
%>%slice(1:4)
%>% ggplot(aes(x = start_station_name, y = num_of_ride, group =
member_casual)) +
geom_col(aes(fill = member_casual)) +
facet_wrap(~member_casual) +
labs(title = "Most popular start stations Compared between Casual and
Member Riders") + ylab("Number of ride") + xlab("Start station")+
theme(axis.text.x = element_text(angle = 340))
```



2. What is the most visited end station?

```
##-- Plot the graph for most visited end stations

trip_data %>% group_by(member_casual,end_station_name) %>%
summarise(num_of_ride = n(),average_length =
mean(ride_length_m),end_name = n()) %>%
arrange(desc(end_name))%>%
slice(1:4) %>%
ggplot(aes(x = end_station_name, y = num_of_ride, group =
member_casual)) + geom_col(aes(fill = member_casual)) +
facet_wrap(~member_casual) + labs(title = "Most popular end stations
Compared between Casual and Member Riders") +
ylab("Number of ride") +
xlab("End station")+
theme(axis.text.x = element_text(angle = 340))
```

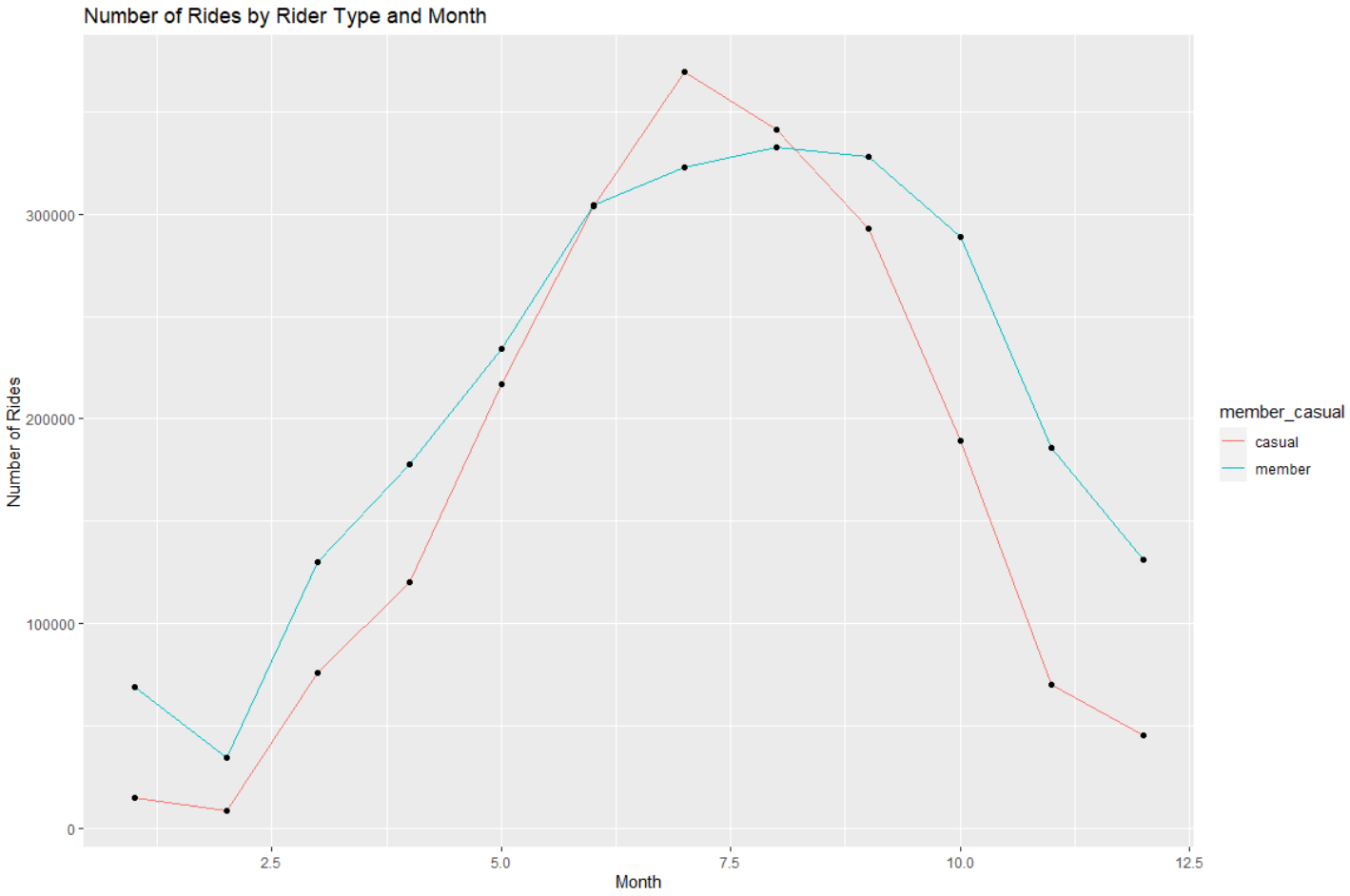


Cyclistic
Data Analytics Case Study

Analysis 3: How do causal riders ride at different times of year?

1. What is the busiest time of year for casual/ member riders?

```
##-- plot the graph the shows the relation between numbers of ride and the
##-- month around the year, grouped by member rider or causal rider

trip_data %>% group_by(member_casual,month) %>% summarise(num_of_ride =
n(),average_length = mean(ride_length_s)) %>% arrange(member_casual, month)
%>% ggplot(aes(x = month, y = num_of_ride, group = member_casual)) +
geom_line(aes(color = member_casual)) + geom_point() + labs(title = "Number
of Rides by Rider Type and Month")+ylab("Number of Rides") + xlab("Month")
+ scale_x_continuous(labels = comma)
```



Cyclistic
Data Analytics Case Study

Summary:

- The average ride time for casual riders are around 40 minutes and member riders average 15 minutes.

- The member users have a more consistent average ride time throughout the different days of the week, casual riders are less consistent in terms of rider time.

- The most visited start station for members is Wells St & Concord Ln, which is also the most visited end station for members

- The most popular start station for casual riders is also Well St & Concord Ln, the most popular end station is Theater on the Lake.

- Upon inspecting the “Number of Rides by Rider Type and Month” graph above we notice that both casual and member rides peak around mid-year June and July. and ride numbers have a trend to fall around September to February and rise between March and May.

- The casual rider spent more time on each ride compared to members, for all year long.

- The rental is peaked at weeknd for both member and casual riders.

PHASE 5: ACT

Three recommendation on your analysis:

1. From the summary and graph above we know that both casual and member rides peak around mid-year June and July. and ride numbers have a trend to fall around September to February and rise between March and May, this is probably due to the poor weather condition during the winter. consider moving some bikes into storage to decrease maintenance cost between September and February. As to increase annual membership, lower the membership price during the winter season to increase sales. And during summer month increase advertising efforts.

2. Promotional personnel at our most popular start station for casual riders. And create membership campaigns at most visited stations.

3. Provide registration bonus for casual riders, as they stack up their riding time offer them discount on membership fee.