Google Data Analytics Capstone – Case Study

Title – How Does a Bike-Share Navigate Speedy Success?

Author - Pyae Phyo Maung

Date - December 15, 2023

1. Introduction

This is Google Data Analytics Capstone – Case Study 1 (How Does a Bike-Share Navigate Speedy Success?) which can be found Google Data Analytics Capstone: Complete a Case Study course. The case study involves a bikeshare company's data of its customer's trip details over a 10-month period (January 2023 - October 2023). The data has been made available by Motivate International Inc. under this license.

2. Scenario

I am 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, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my recommendations, so they must be backed up with compelling data insights and professional data visualizations

3. Analyzing Data

In this analysis, the 6 phases of the Data Analysis process: Ask, Prepare, Process, Analyze, Share and Act are as follow.

3.1 Ask

The problem I am trying to solve is to answer one of the three questions that will guide the future marketing program – How do annual members and casual members use the bicycle bikes differently?

The main objective is to determine a way to build a profile for annual members and the best marketing strategies to turn casual bike riders into annual members. This will help the marketing team to increase annual members.

In this project, the key stakeholders are

- Lily Moreno: The director of marketing and my manager
- Cyclistic marketing analytics team
- Cyclistic executive team

The clear statement of the business task is 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.

3.2 Prepare

A total 10 CSV files of datasets starting from (**January-2023 to October-2023**) is chosen by my own. But Google also provided their own link with the same dataset with more years and station descriptions.

The combined size of all the 10 CSV files is very large. Data cleaning in spreadsheets will be time-consuming and slow compared to R. I am choosing R simply because I could do both data wrangling and analysis/ visualizations in the same platform.

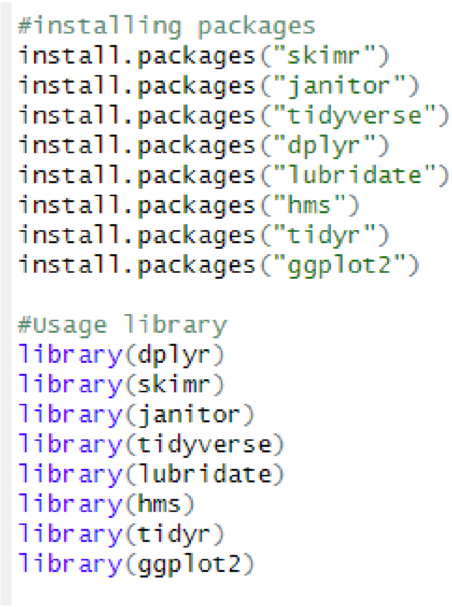
The data is located in R studio because I will use R studio for data analysis. The data is separated by month and its own csv file. It is not bias and cleaned but some information missing and have to delete in R studio during analysis. It's ROCCC because it's reliable, original, comprehensive, current and cited.

The company has their own license over the datasets. Besides that, the dataset doesn't have any personal information about the riders. All the files have consistent columns and each column has the correct type of data. It may have some key insights about the riders and their riding style. There is no problem with data except for some missing information that I have to clean it out.

3.3 Process

This step will prepare the data for analysis. All the csv files will be merged into one file to improve workflow.

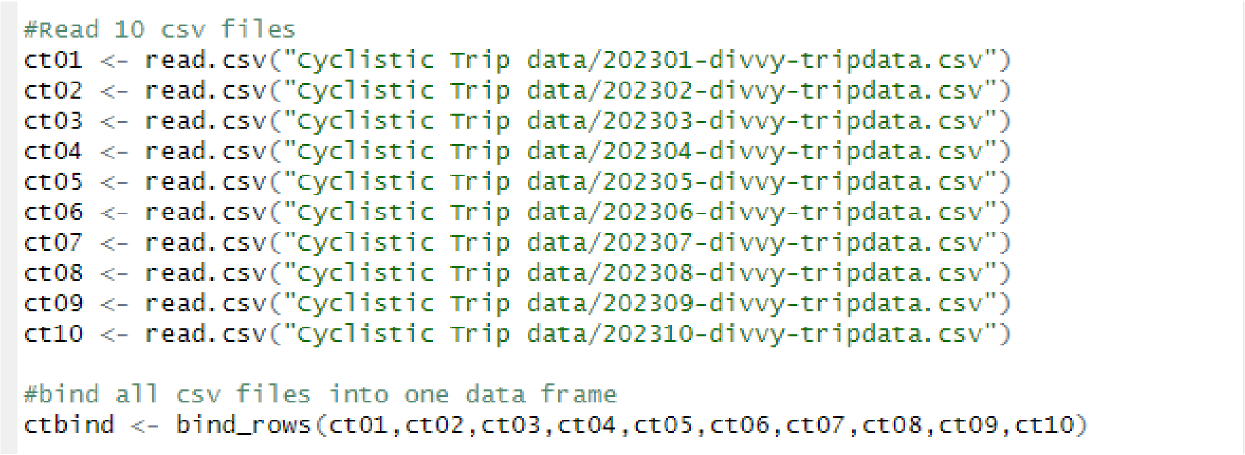
3.3.1 Install packages and load libraries



```
#installing packages
install.packages("skimr")
install.packages("janitor")
install.packages("tidyverse")
install.packages("dplyr")
install.packages("lubridate")
install.packages("hms")
install.packages("tidyr")
install.packages("ggplot2")

#Usage library
library(dplyr)
library(skimr)
library(janitor)
library(tidyverse)
library(lubridate)
library(hms)
library(tidyr)
library(ggplot2)
```

3.3.2 Read all csv files and combine into one data frame

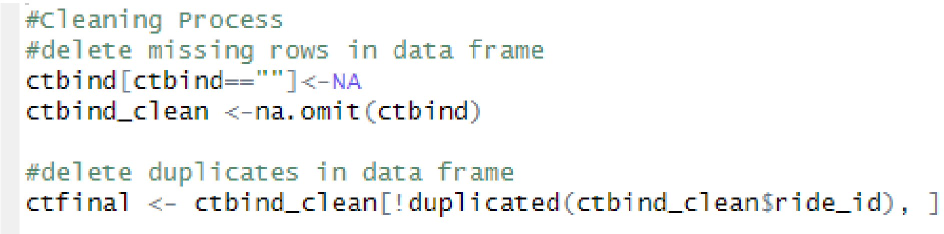
Read all csv files from exact location and bind into new dataset.



```
#Read 10 csv files
ct01 <- read.csv("Cyclistic Trip data/202301-divvy-tripdata.csv")
ct02 <- read.csv("Cyclistic Trip data/202302-divvy-tripdata.csv")
ct03 <- read.csv("Cyclistic Trip data/202303-divvy-tripdata.csv")
ct04 <- read.csv("Cyclistic Trip data/202304-divvy-tripdata.csv")
ct05 <- read.csv("Cyclistic Trip data/202305-divvy-tripdata.csv")
ct06 <- read.csv("Cyclistic Trip data/202306-divvy-tripdata.csv")
ct07 <- read.csv("Cyclistic Trip data/202307-divvy-tripdata.csv")
ct08 <- read.csv("Cyclistic Trip data/202308-divvy-tripdata.csv")
ct09 <- read.csv("Cyclistic Trip data/202309-divvy-tripdata.csv")
ct10 <- read.csv("Cyclistic Trip data/202310-divvy-tripdata.csv")

#bind all csv files into one data frame
ctbind <- bind_rows(ct01,ct02,ct03,ct04,ct05,ct06,ct07,ct08,ct09,ct10)
```

3.3.3 Data cleaning process

Delete rows with null values and removes duplicated rows from table.

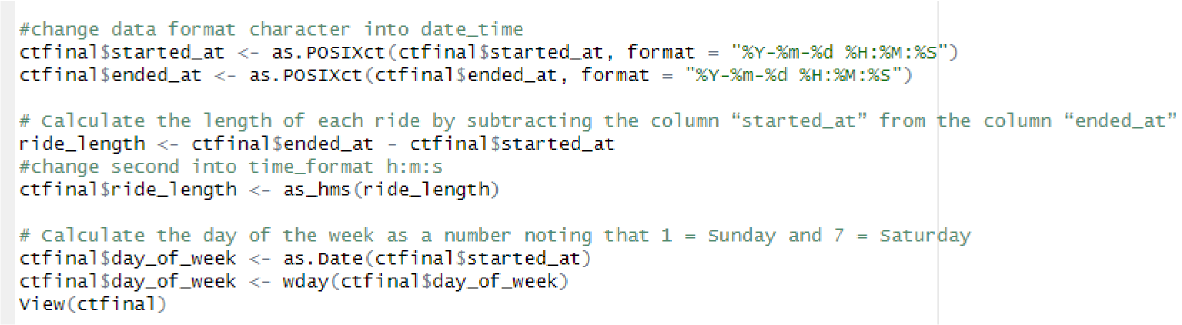


```
#Cleaning Process
#delete missing rows in data frame
ctbind[ctbind==""]<-NA
ctbind_clean <-na.omit(ctbind)

#delete duplicates in data frame
ctfinal <- ctbind_clean[!duplicated(ctbind_clean$ride_id), ]
```

3.3.4 Data Calculation (ride length, day of week)

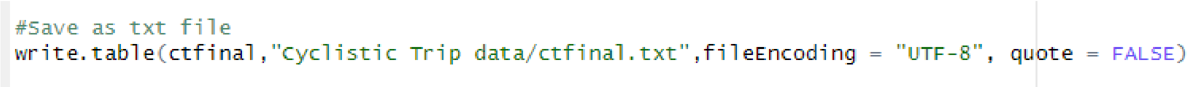
Calculate the length of each ride(**ride_length**) by subtracting the column “**started_at**” from the column “**ended_at**”.

The weekday will be useful to determine patterns of travels in the week.



```
#change data format character into date_time
ctfinal$started_at <- as.POSIXct(ctfinal$started_at, format = "%Y-%m-%d %H:%M:%S")
ctfinal$ended_at <- as.POSIXct(ctfinal$ended_at, format = "%Y-%m-%d %H:%M:%S")

# Calculate the length of each ride by subtracting the column “started_at” from the column “ended_at”
ride_length <- ctfinal$ended_at - ctfinal$started_at
#change second into time_format h:m:s
ctfinal$ride_length <- as_hms(ride_length)

# Calculate the day of the week as a number noting that 1 = Sunday and 7 = Saturday
ctfinal$day_of_week <- as.Date(ctfinal$started_at)
ctfinal$day_of_week <- wday(ctfinal$day_of_week)
View(ctfinal)
```

3.3.5 Saving the result as txt file



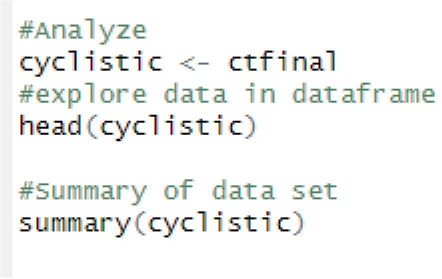
```
#Save as txt file
write.table(ctfinal,"Cyclistic Trip data/ctfinal.txt",fileEncoding = "UTF-8", quote = FALSE)
```

3.4 Analyze

The data frame is now ready for descriptive analysis that will help us building a profile for annual members and how they differ from casual riders.

3.4.1 Summary of dataset

To quick start, let's generate a summary of the dataset.

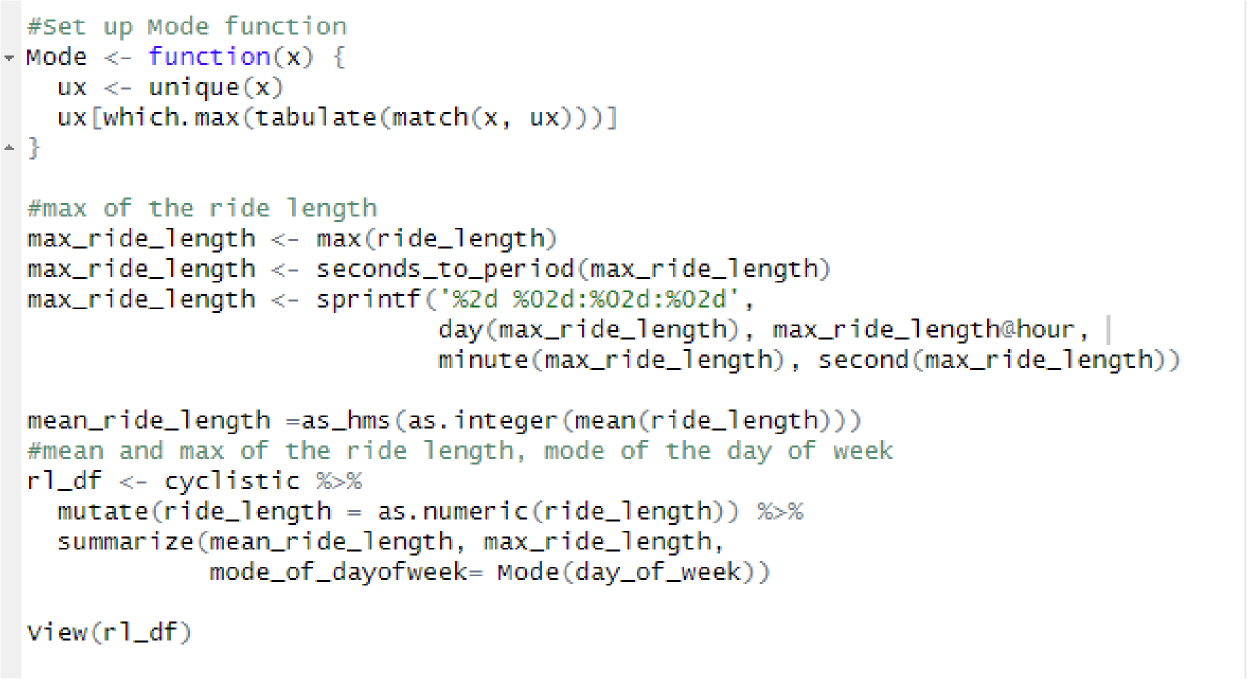


```
#Analyze
cyclistic <- ctfinal
#explore data in dataframe
head(cyclistic)

#Summary of data set
summary(cyclistic)
```

3.4.2 Some Calculations for data analysis

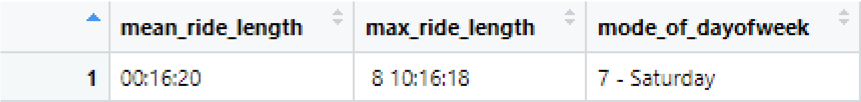
Run a few calculations in one file to get a better sense of the data layout.

- Calculate the mean of **ride_length**.
- Calculate the max **ride_length**.
- Calculate the mode of **day_of_week**.



```r
#Set up Mode function
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

#max of the ride length
max_ride_length <- max(ride_length)
max_ride_length <- seconds_to_period(max_ride_length)
max_ride_length <- sprintf('%2d %02d:%02d:%02d',
                           day(max_ride_length), max_ride_length@hour,
                           minute(max_ride_length), second(max_ride_length))

mean_ride_length =as_hms(as.integer(mean(ride_length)))
#mean and max of the ride length, mode of the day of week
rl_df <- cyclistic %>%
  mutate(ride_length = as.numeric(ride_length)) %>%
  summarize(mean_ride_length, max_ride_length,
            mode_of_dayofweek= Mode(day_of_week))

view(rl_df)
```

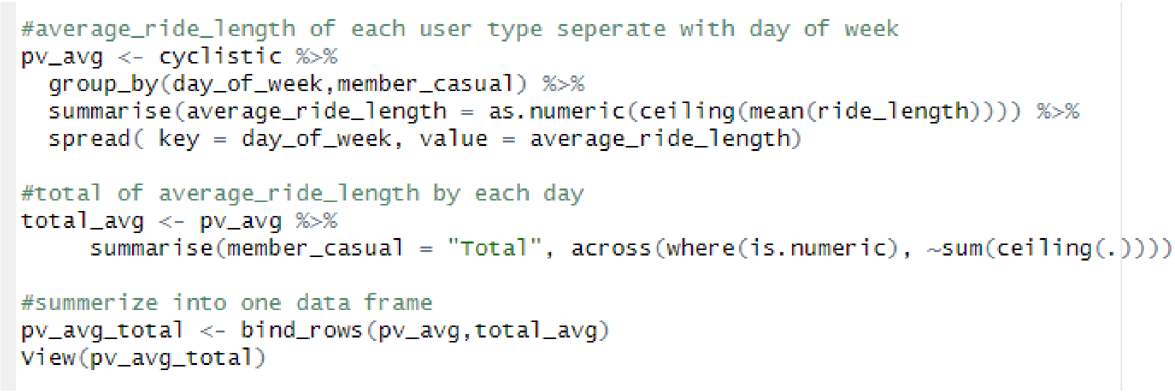
Result dataset.



	mean_ride_length	max_ride_length	mode_of_dayofweek
1	00:16:20	8 10:16:18	7 - Saturday

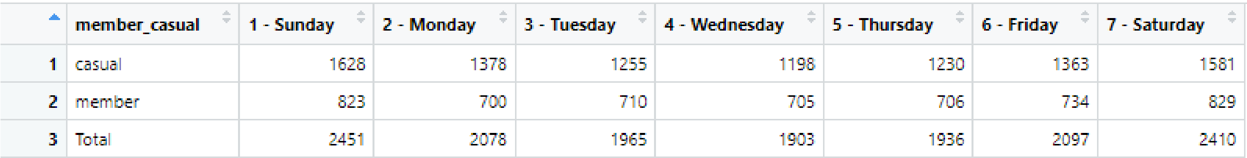
3.4.3 Create a pivot table to quickly calculate and visualize the data.

- Calculate the average **ride_length** for members and casual riders by **day_of_week.**



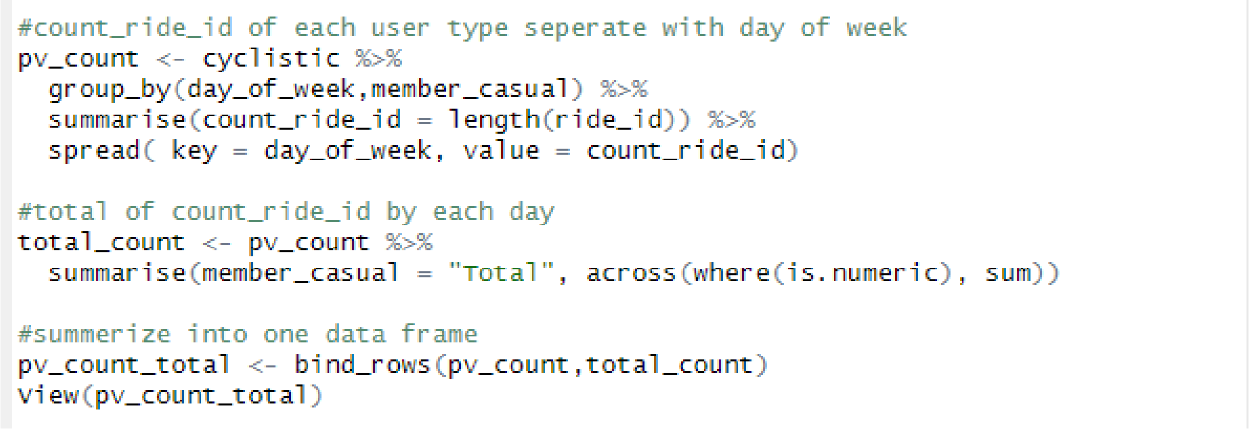
```r
#average_ride_length of each user type seperate with day of week
pv_avg <- cyclistic %>%
  group_by(day_of_week,member_casual) %>%
  summarise(average_ride_length = as.numeric(ceiling(mean(ride_length)))) %>%
  spread( key = day_of_week, value = average_ride_length)

#total of average_ride_length by each day
total_avg <- pv_avg %>%
     summarise(member_casual = "Total", across(where(is.numeric), ~sum(ceiling(.))))

#summerize into one data frame
pv_avg_total <- bind_rows(pv_avg,total_avg)
View(pv_avg_total)
```

Result pivot table.



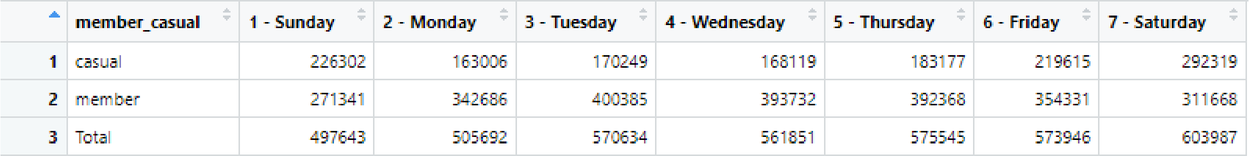
	member_casual	1 - Sunday	2 - Monday	3 - Tuesday	4 - Wednesday	5 - Thursday	6 - Friday	7 - Saturday
1	casual	1628	1378	1255	1198	1230	1363	1581
2	member	823	700	710	705	706	734	829
3	Total	2451	2078	1965	1903	1936	2097	2410

- Calculate the count **ride_id** for members and casual riders by **day_of_week**.



```
#count_ride_id of each user type seperate with day of week
pv_count <- cyclistic %>%
  group_by(day_of_week,member_casual) %>%
  summarise(count_ride_id = length(ride_id)) %>%
  spread( key = day_of_week, value = count_ride_id)

#total of count_ride_id by each day
total_count <- pv_count %>%
  summarise(member_casual = "Total", across(where(is.numeric), sum))

#summerize into one data frame
pv_count_total <- bind_rows(pv_count,total_count)
view(pv_count_total)
```

Result pivot table.



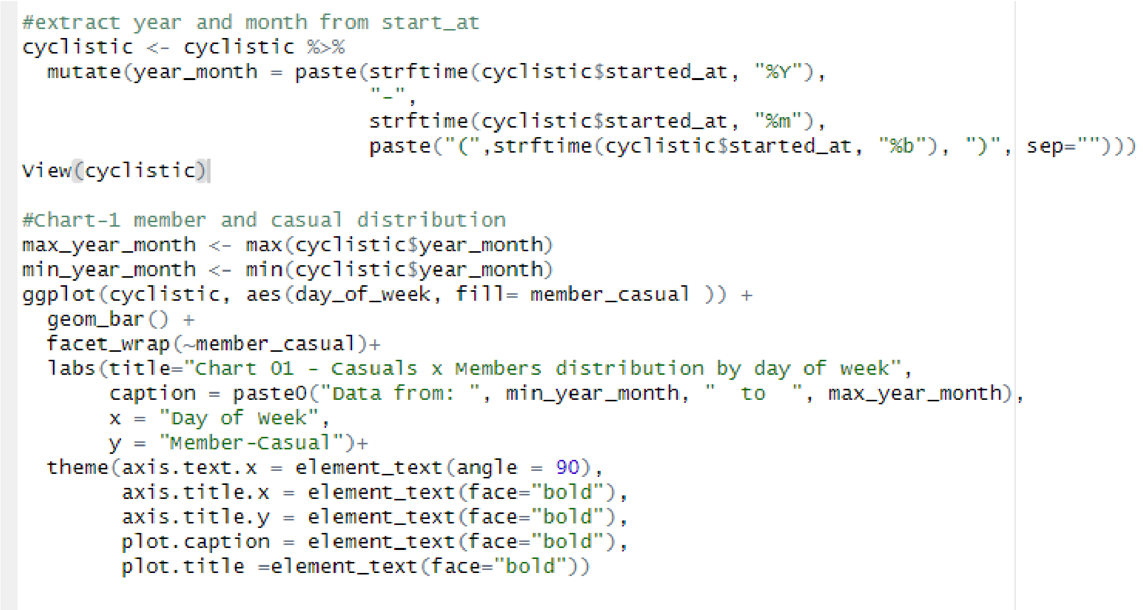
	member_casual	1 - Sunday	2 - Monday	3 - Tuesday	4 - Wednesday	5 - Thursday	6 - Friday	7 - Saturday
1	casual	226302	163006	170249	168119	183177	219615	292319
2	member	271341	342686	400385	393732	392368	354331	311668
3	Total	497643	505692	570634	561851	575545	573946	603987

3.5 Share

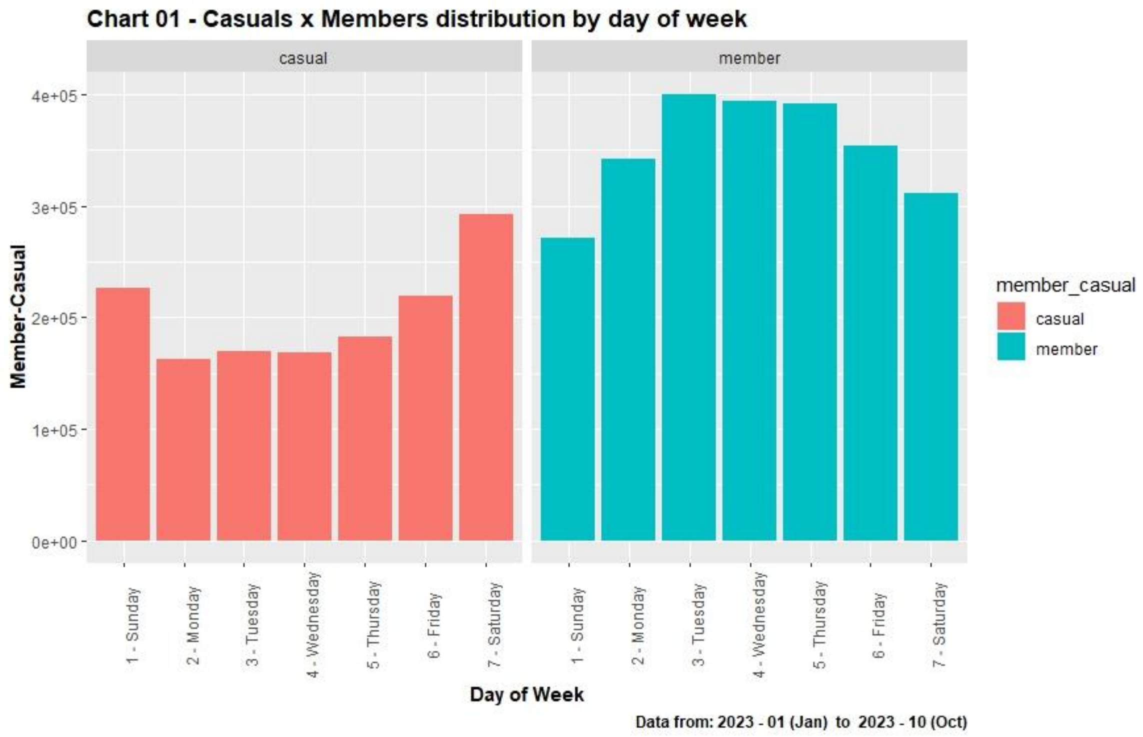
The share phase of the data analysis process typically involves communicating findings, summarizing results using data visualizations, and creating a slideshow to present to stakeholders.

3.5.1 Create data visualization with charts

Chart 01 - Casual-Member distribution by day of week.



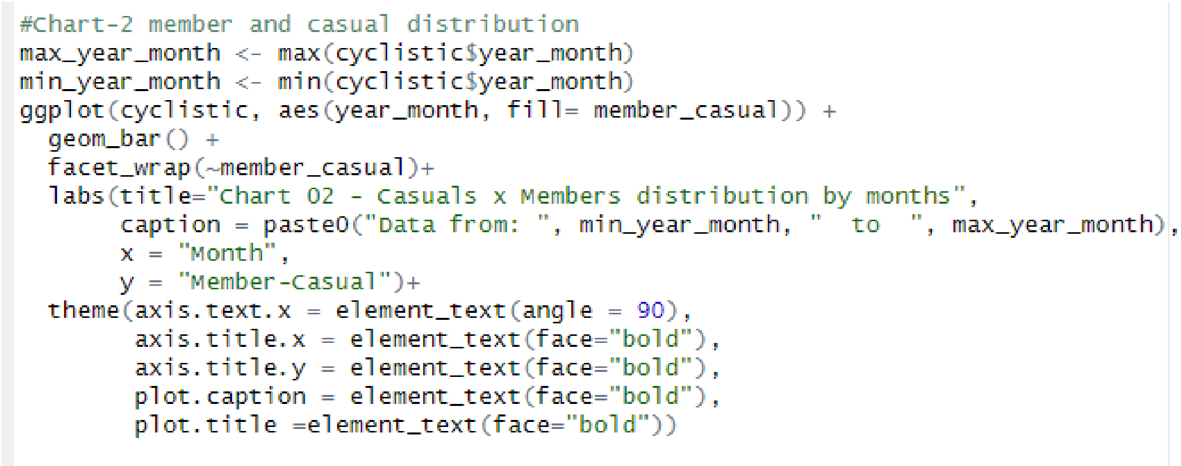
```
#extract year and month from start_at
cyclistic <- cyclistic %>%
  mutate(year_month = paste(strftime(cyclistic$started_at, "%Y"),
                            "-",
                            strftime(cyclistic$started_at, "%m"),
                            paste("(",strftime(cyclistic$started_at, "%b"), ")", sep="")))
view(cyclistic)

#Chart-1 member and casual distribution
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(day_of_week, fill= member_casual )) +
  geom_bar() +
  facet_wrap(~member_casual)+
  labs(title="Chart 01 - Casuals x Members distribution by day of week",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       x = "Day of week",
       y = "Member-Casual")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



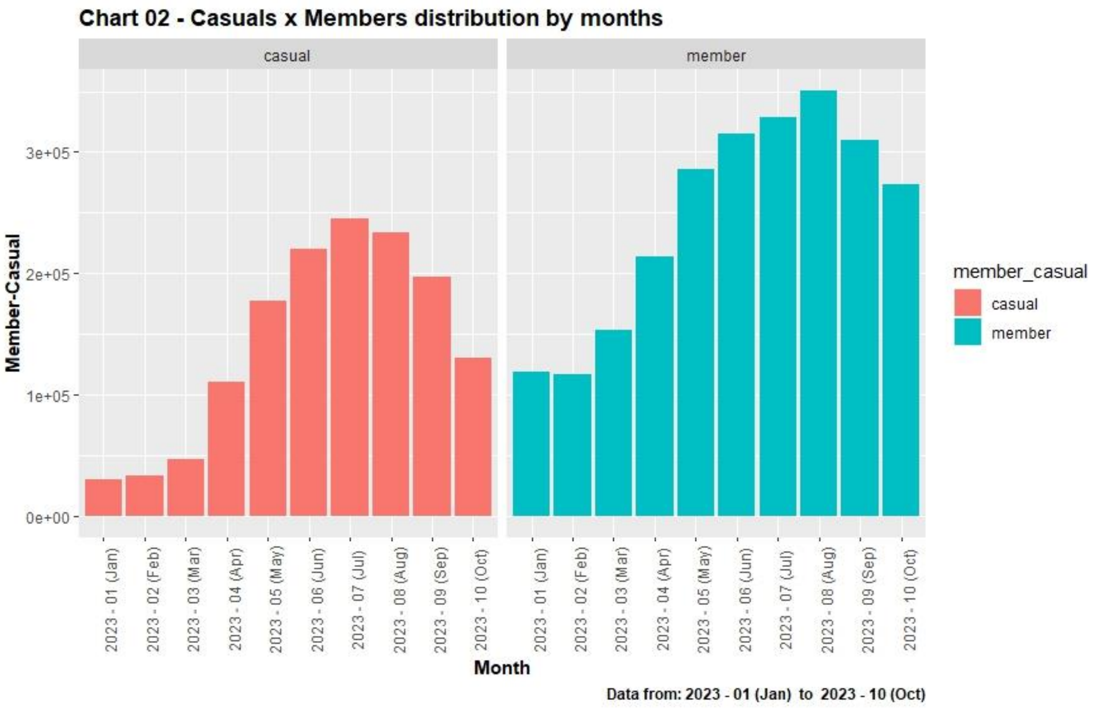
Some considerations can be taken by this chart:

- We have more members' rides than casual every day.
- The day with the smallest count of data points is Sunday with **~12.8%** of the dataset.
- The day with the biggest count of data points is Saturday with **~15.53%** of the dataset.
- The smallest different percentage of member-casual is **~3.204 %** on Saturday.
- The biggest different percentage of member-casual is **~40.33 %** on Tuesday.

Chart 02 - Casual-Member distribution by months.



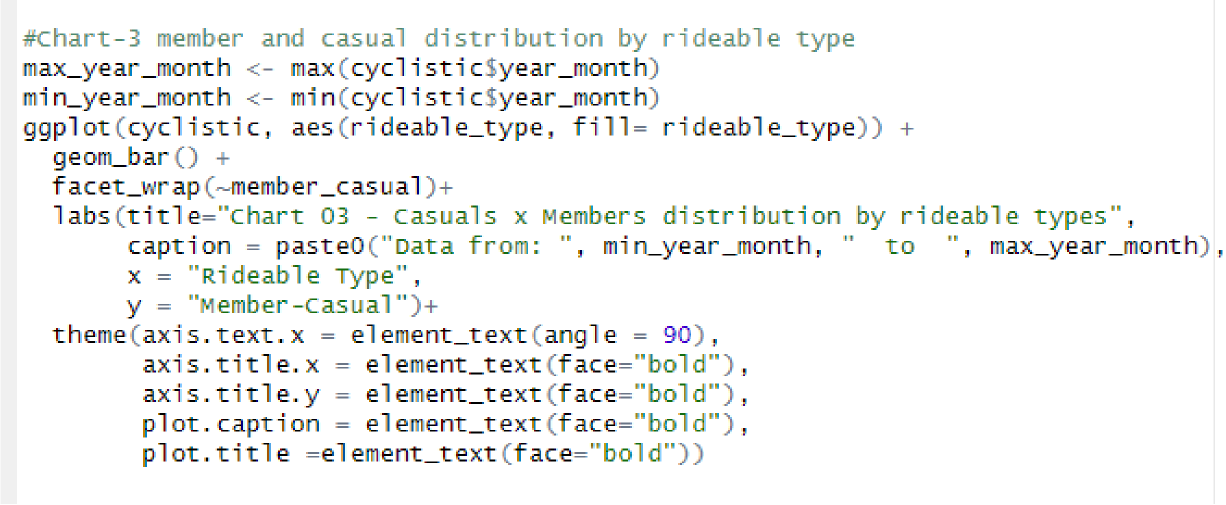
```
#Chart-2 member and casual distribution
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(year_month, fill= member_casual)) +
  geom_bar() +
  facet_wrap(~member_casual)+
  labs(title="Chart 02 - Casuals x Members distribution by months",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       x = "Month",
       y = "Member-Casual")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



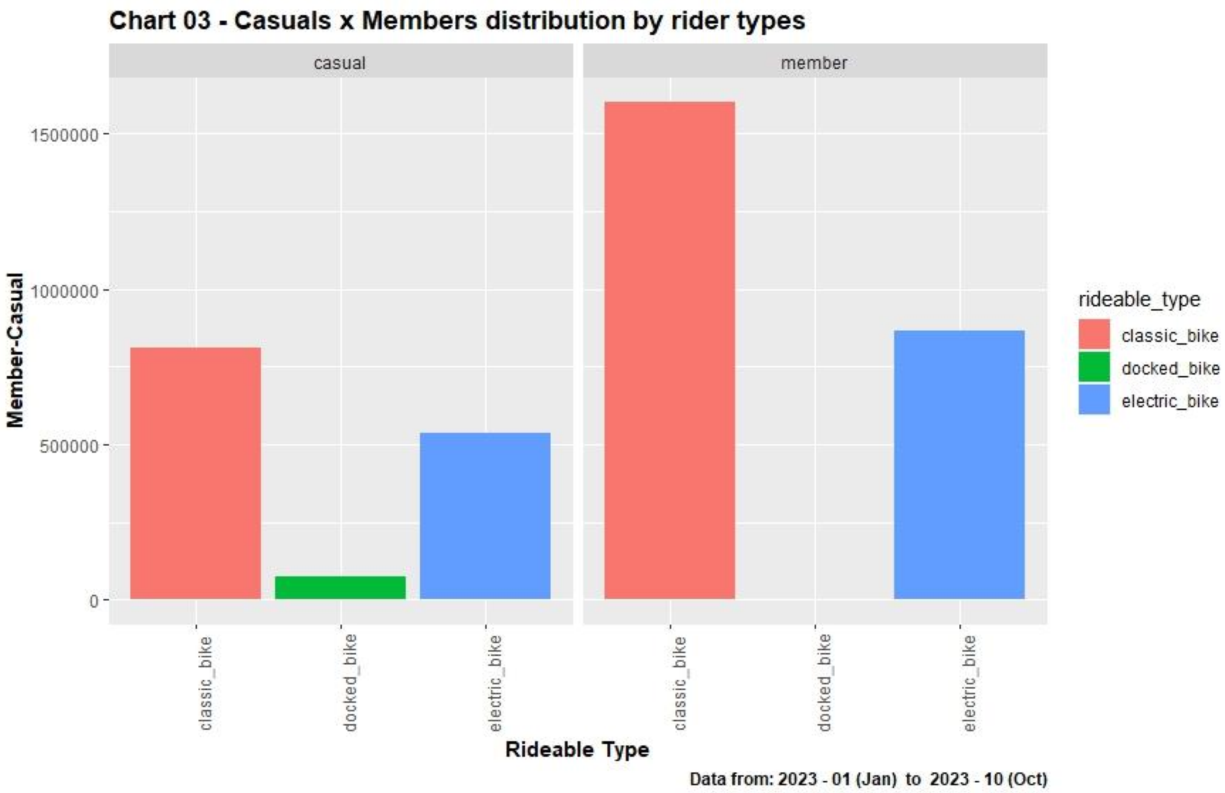
Some considerations can be taken by this chart:

- In all months we have more members' rides than casual.
- The month with the smallest count of data points is January with **~3.08%** of the dataset.
- The month with the biggest count of data points is August with **~15.04%** of the dataset.
- The smallest different percentage of member-casual is **~14.53 %** in July.
- The biggest different percentage of member-casual is **~60.05 %** in January.

Chart 03 - Casual-Member distribution by rideable types.



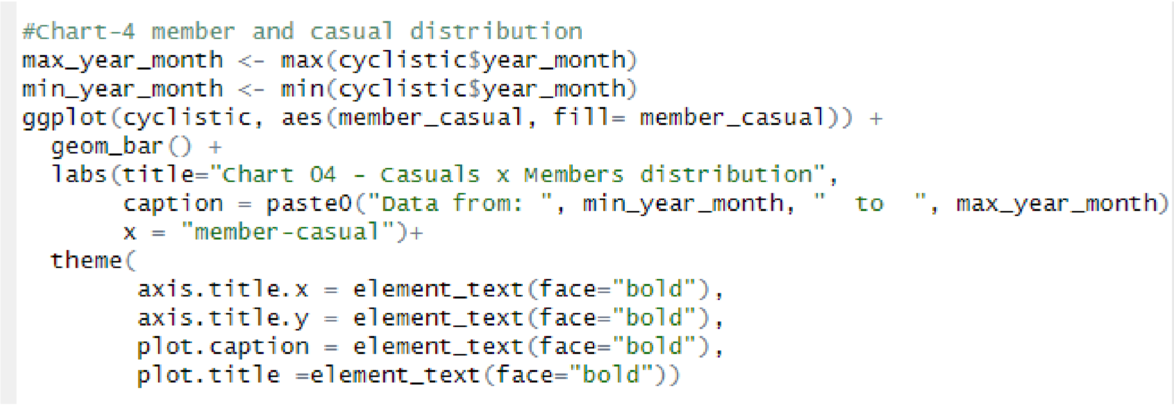
```
#Chart-3 member and casual distribution by rideable type
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(rideable_type, fill= rideable_type)) +
  geom_bar() +
  facet_wrap(~member_casual)+
  labs(title="Chart 03 - Casuals x Members distribution by rideable types",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       x = "Rideable Type",
       y = "Member-Casual")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



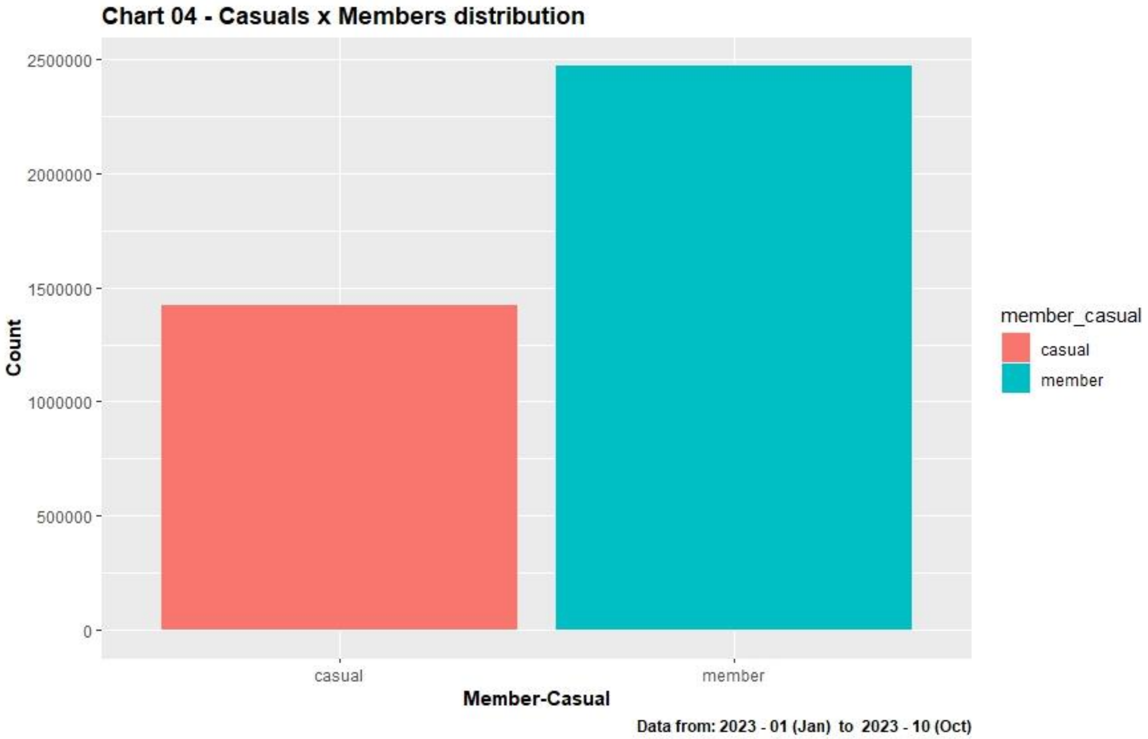
Some considerations can be taken by this chart:

- We have more member's rides than casual except docked bike because only casual riders take docked bikes.
- The bike with the smallest count of data points is docked bike with **~1.96%** of the dataset.
- The bike with the biggest count of data points is classic bike with **~62%** of the dataset.

Chart 04 - Over all Casual-Member distribution.

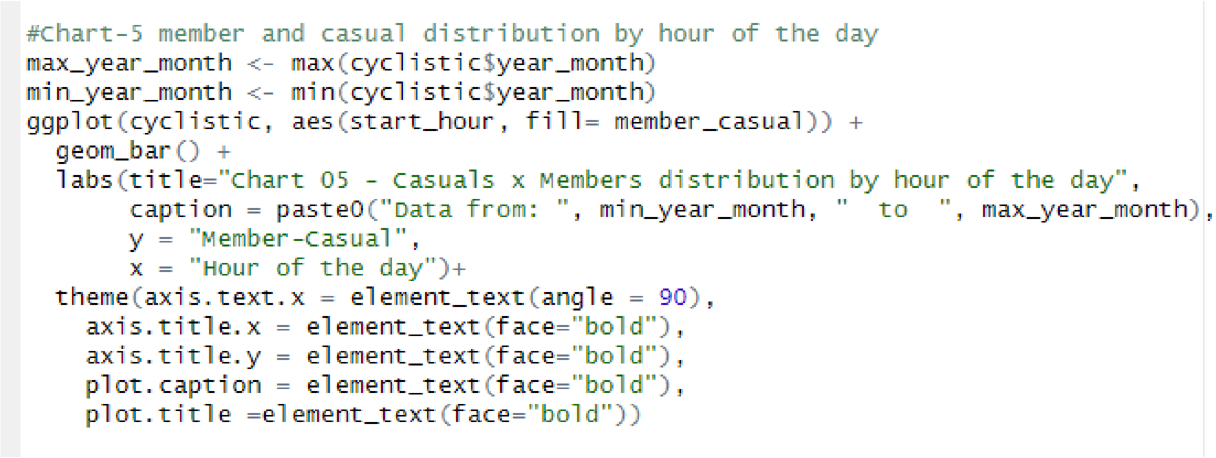


```
#Chart-4 member and casual distribution
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(member_casual, fill= member_casual)) +
  geom_bar() +
  labs(title="Chart 04 - Casuals x Members distribution",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       x = "member-casual")+
  theme(
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

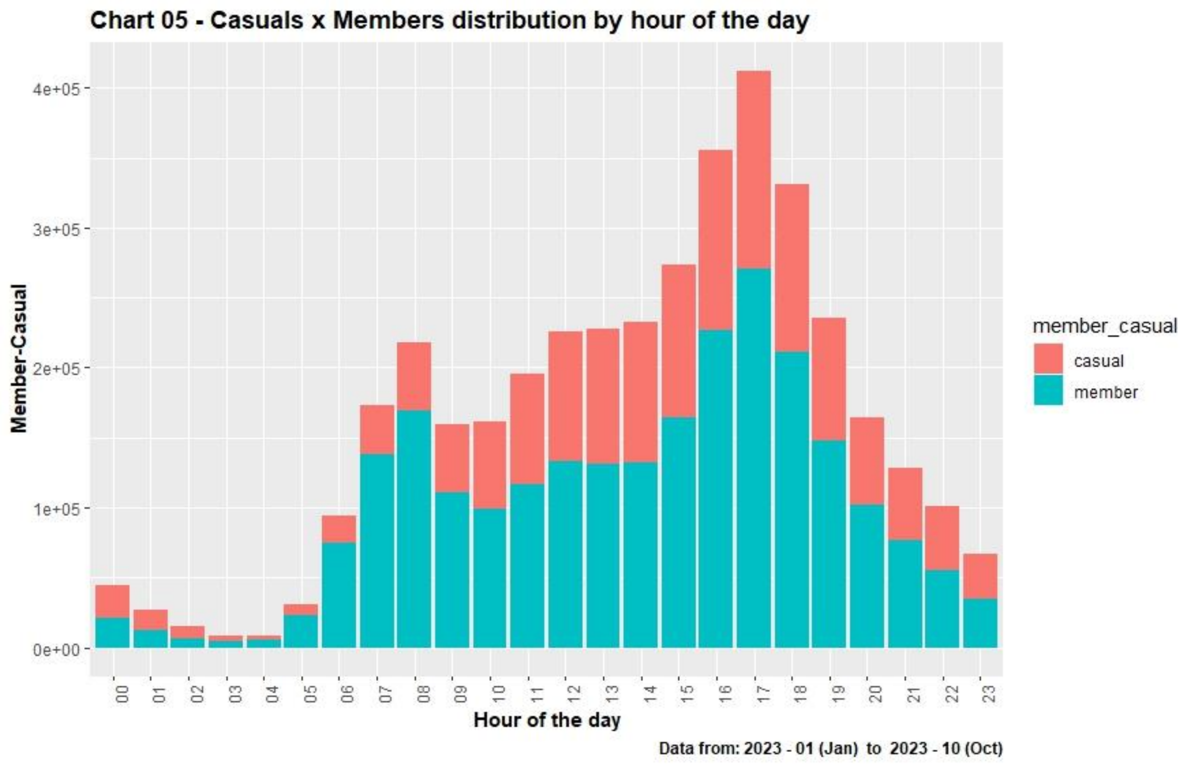


As we can see on the Member-Casual chart, members have a bigger proportion of the dataset, composing **~63.4%**, **~26.8%** bigger than the count of casual riders.

Chart 05 - Casual-Member distribution by hour of the day.



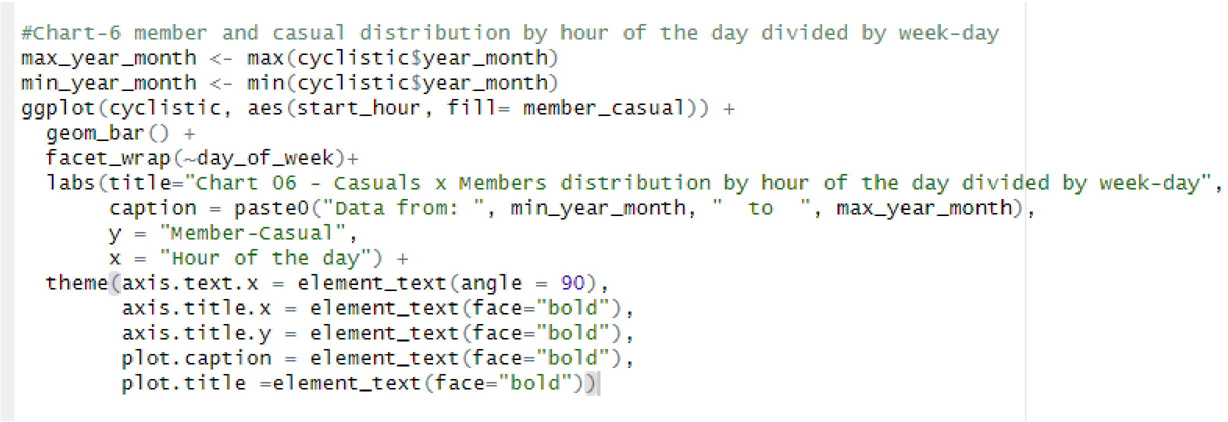
```
#Chart-5 member and casual distribution by hour of the day
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(start_hour, fill= member_casual)) +
  geom_bar() +
  labs(title="Chart 05 - Casuals x Members distribution by hour of the day",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       y = "Member-Casual",
       x = "Hour of the day")+
  theme(axis.text.x = element_text(angle = 90),
    axis.title.x = element_text(face="bold"),
    axis.title.y = element_text(face="bold"),
    plot.caption = element_text(face="bold"),
    plot.title =element_text(face="bold"))
```



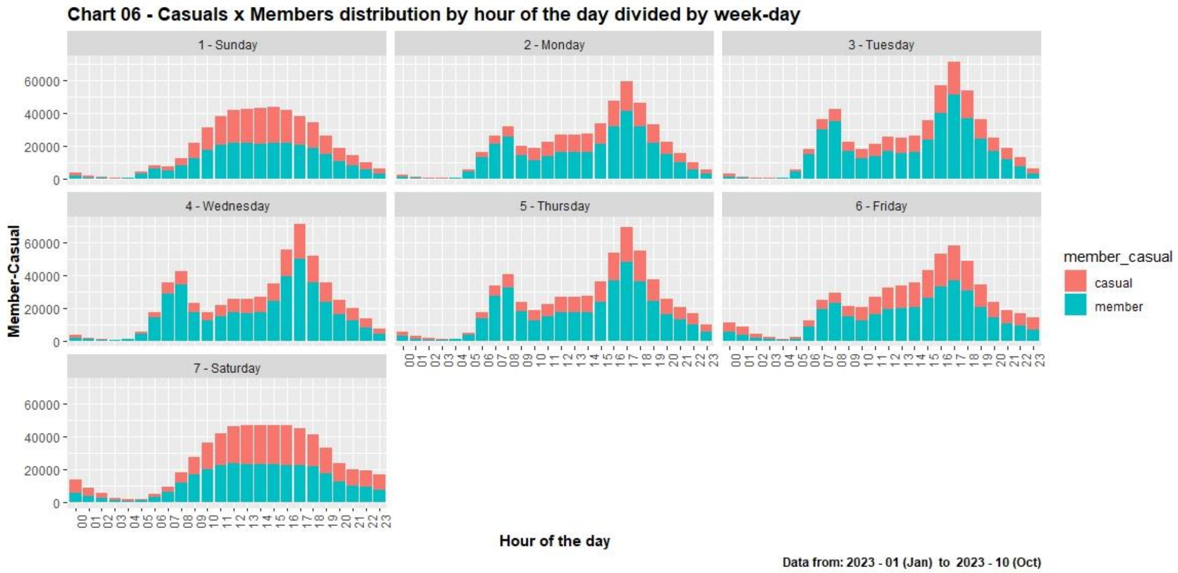
From this chart, we can see:

- There's a bigger volume of bikers between 4 pm and 6 pm.
- We have more casual riders after midnight to 3 am.
- And more member riders after 3 am to before midnight.

Chart 06 - Casual-Member distribution by hour of the day divided by weekday.

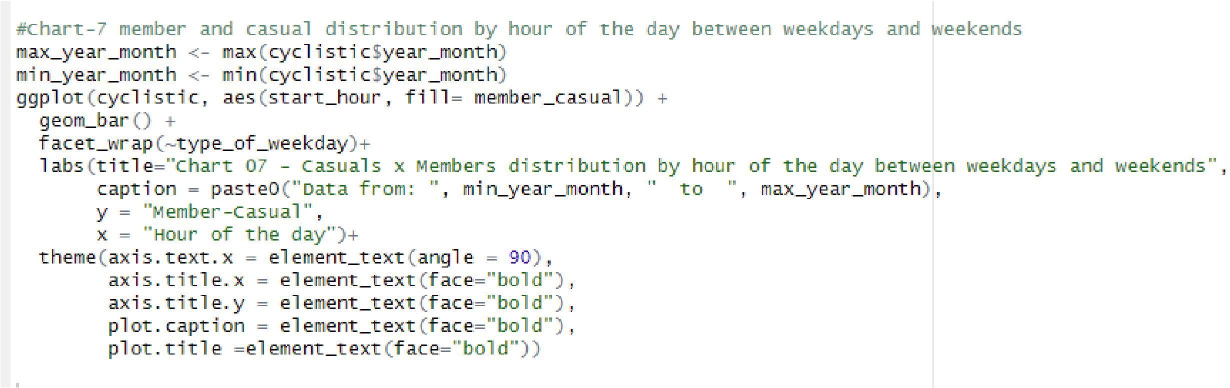


```
#Chart-6 member and casual distribution by hour of the day divided by week-day
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(start_hour, fill= member_casual)) +
  geom_bar() +
  facet_wrap(~day_of_week)+
  labs(title="Chart 06 - Casuals x Members distribution by hour of the day divided by week-day",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       y = "Member-Casual",
       x = "Hour of the day") +
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



There's a clear different between the weekdays and weekends. Let’s make it clear with another chart.

Chart 07 - Casual-Member distribution by hour of the day between weekdays and weekends.



```
#Chart-7 member and casual distribution by hour of the day between weekdays and weekends
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic, aes(start_hour, fill= member_casual)) +
  geom_bar() +
  facet_wrap(~type_of_weekday)+
  labs(title="Chart 07 - Casuals x Members distribution by hour of the day between weekdays and weekends",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       y = "Member-Casual",
       x = "Hour of the day")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

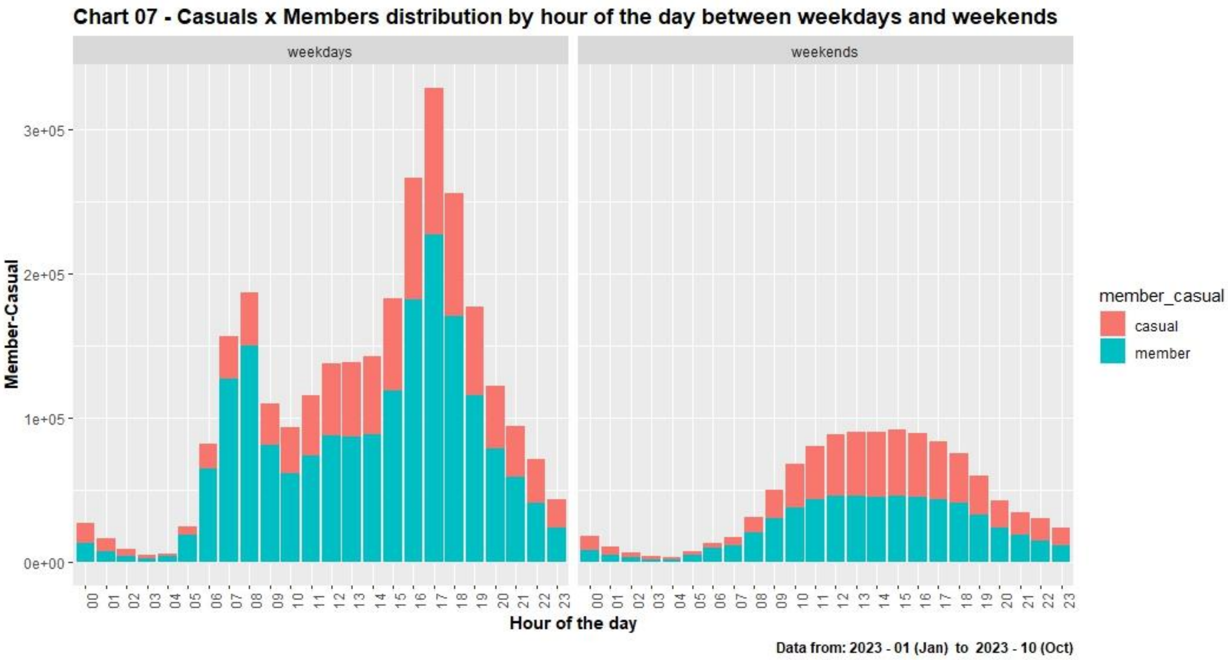


Chart 07 - Casuals x Members distribution by hour of the day between weekdays and weekends

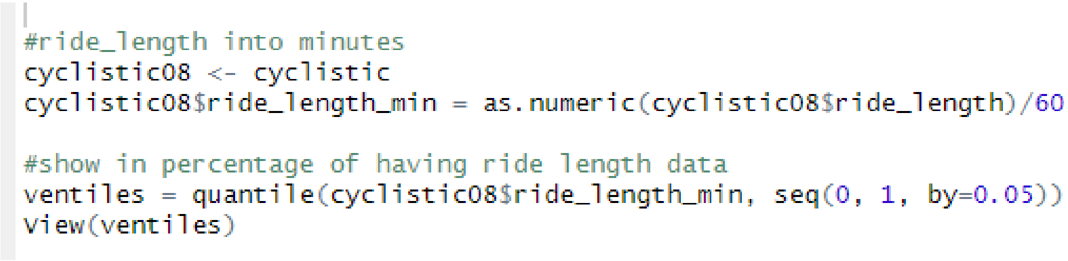
Data from: 2023 - 01 (Jan) to 2023 - 10 (Oct)

The two plots different on some key ways:

- While the weekends have a smooth flow of data points, the weekdays have a steeper flow of data.
- There are absolutely more riders on weekdays than on weekends
- During the weekdays we have a bigger flow of member riders.
- During the weekends we have a bigger flow of casual riders between midnight and 3 am, 2 pm to 4 pm and 10 pm to 11 pm.

It's fundamental to question who are the riders who use the bikes during this time of day. We can assume some factors, one is that member riders are who use the bikes during the daily routine activities, like go to work (data points between 6 am to 8 am in weekdays), go back from work (data points between 4 pm to 6 pm).

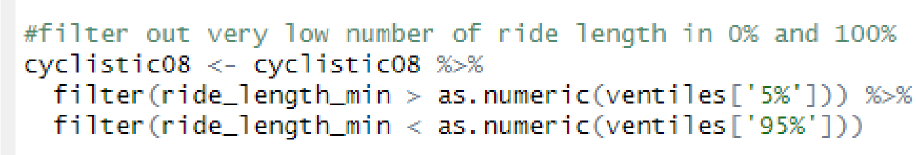
Chart 08 - Riding time for Casual-Member distribution by day of week.



```
#ride_length into minutes
cyclistic08 <- cyclistic
cyclistic08$ride_length_min = as.numeric(cyclistic08$ride_length)/60

#show in percentage of having ride length data
ventiles = quantile(cyclistic08$ride_length_min, seq(0, 1, by=0.05))
View(ventiles)
```

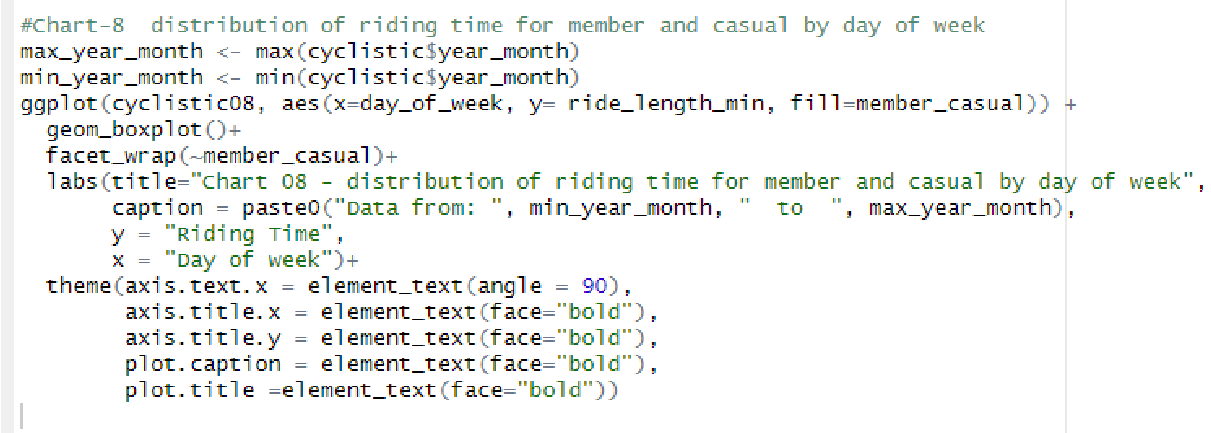
- In this case, we have a very far number of lowest and highest minutes of ride length data to include in a subset of the dataset.
- lowest minutes in 0% is **0.01666667** and highest minutes in 100% is **12136.3**.

So, we had to filter out some rows in 0% and 100% as follow.



```
#filter out very low number of ride length in 0% and 100%
cyclistic08 <- cyclistic08 %>%
  filter(ride_length_min > as.numeric(ventiles['5%'])) %>%
  filter(ride_length_min < as.numeric(ventiles['95%']))
```

Now, we can figure it out with box plot visualization.



```
#Chart-8  distribution of riding time for member and casual by day of week
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic08, aes(x=day_of_week, y= ride_length_min, fill=member_casual)) +
  geom_boxplot()+
  facet_wrap(~member_casual)+
  labs(title="Chart 08 - distribution of riding time for member and casual by day of week",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       y = "Riding Time",
       x = "Day of week")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

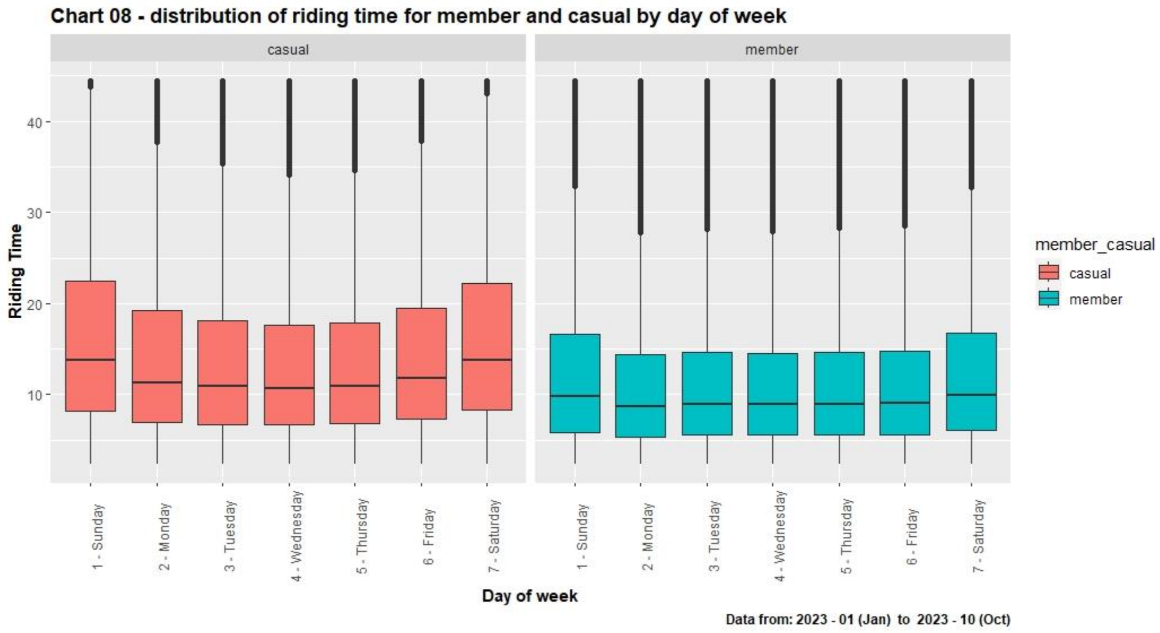


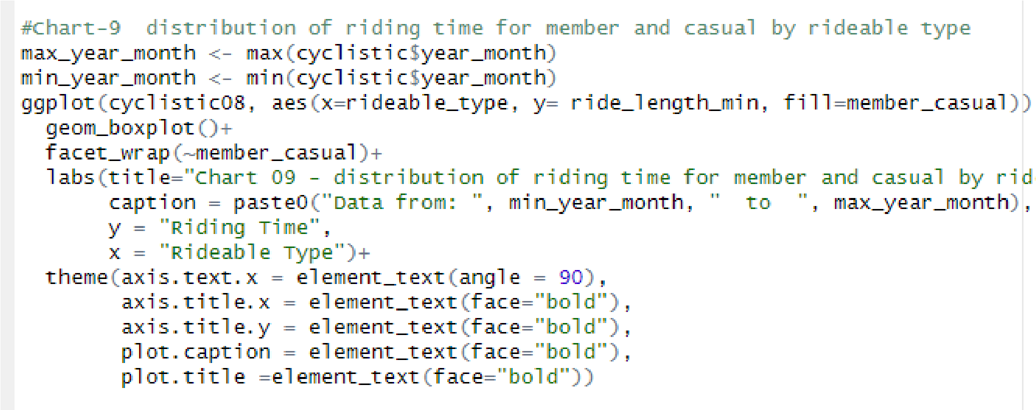
Chart 08 - distribution of riding time for member and casual by day of week

Data from: 2023 - 01 (Jan) to 2023 - 10 (Oct)

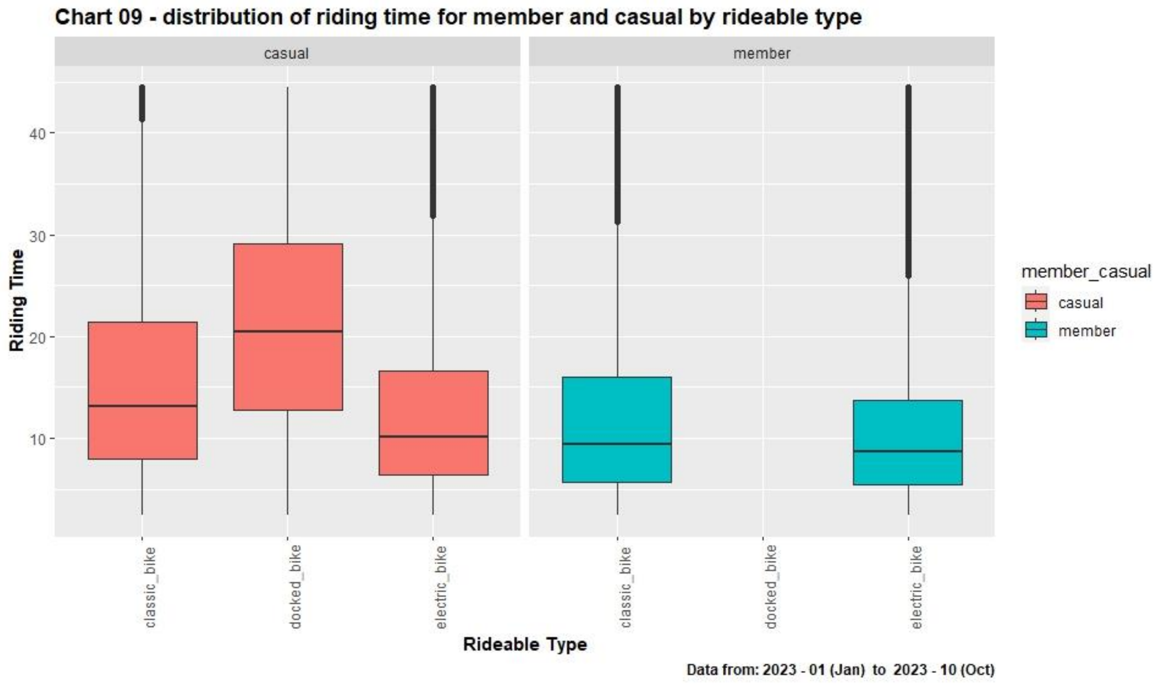
Some considerations can be taken by this chart:

- Casual have more riding time than members.
- Riding time for members keeps unchanged during the weekdays, increasing during weekends
- Casuals follow a more curved distribution, peaking on weekends and bottoming on Wednesdays.

Chart 09 - Riding time for Casual-Member distribution by ridable types.



```
#Chart-9  distribution of riding time for member and casual by rideable type
max_year_month <- max(cyclistic$year_month)
min_year_month <- min(cyclistic$year_month)
ggplot(cyclistic08, aes(x=rideable_type, y= ride_length_min, fill=member_casual)) +
  geom_boxplot()+
  facet_wrap(~member_casual)+
  labs(title="Chart 09 - distribution of riding time for member and casual by rideable type",
       caption = paste0("Data from: ", min_year_month, "  to  ", max_year_month),
       y = "Riding Time",
       x = "Rideable Type")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



Some considerations can be taken by this chart:

- There are no member riders in docked bike.
- Docked bike has more riding time for casual riders.
- Member riders seem like to take classic bike.

3.5.2 Summary of Analysis

Trends or relationship I found in the data are as follow.

- There are more members than casuals in the dataset, composing **~63.4%**, **~26.8%** bigger than the count of casual riders.
- Sunday is the lowest number of riders with **~12.8%** of the dataset.
- Saturday is the highest number of riders with **~15.53%** of the dataset.
- January is the lowest number of riders with **~3.08%** of the dataset.
- August is the highest number of riders with **~15.04%** of the dataset.
- There are more data points in the later month of 2023.
- Docked bike has no member riders but it has highest riding time for casual riders.
- Both riders tend to prefer classic bikes with **~62%** of the dataset.
- docked bike has lowest riders with **~1.96%** of the dataset.
- There's a bigger volume of bikers between 4 pm and 6 pm.
- We have more casual riders after midnight to 3 am and more member riders after 3 am to before midnight.
- There are absolutely more riders on weekdays than on weekends.
- Casual have more riding time than members.
- Riding time for members keeps unchanged during the weekdays, increasing during weekends.
- It looks like difference flow of members/casual taking bikes between weekdays and weekends.
- Members use bikes on schedules that differs from casual.

Conclusion.

There are more member riders in the dataset but less riding time. It looks like the riders don't want to take bikes in early months. According to charts, member riders like to use bikes in weekdays but casual riders like to use in weekends. During weekdays, the data has highest point in hours of going and returning from work. No member riders want to take docked bikes, the lowest data point of all bikes.

3.6 Act

In the Act phase, we have to provide three recommendations for marketing team to improve their working strategies.

Both riders have different habits when using the bikes. The conclusion is further stated on the share phase. The insights could be implemented when preparing a marketing campaign for turning casual into members.

Top three recommendations based on my analysis:

1. Build marketing campaign focus on how bikes help people to get to work at the start of the year.
2. Offer discount coupons for docked bikes which has the lowest data points and also provides extra supplements and promotion for member riders to attract casual riders.
3. Ads campaigns on social media could also be made showing people using the bikes for exercise during the weeks. The ads could focus on how practical and consistent the bikes can be.

4. Reference

1. **Snehith Tella** from Linked in.
2. **Jhelison Gabriel Lima Uchoa** from Kaggle.
3. **Akorez** from Github.