Google Data Analytics - Capstone - Cyclistic

Saman Musician

7/15/2022

Case Study:

How Does a Bike-Share Navigate Speedy Success?

Introduction

This project is the Cyclist bike-share case study, presented by **Saman Musician**, acting as a junior data analyst working in the marketing analyst team at Cyclist, a fictional bike-sharing company in Chicago.

Scenario

The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, our team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, our team will design a new marketing strategy to convert casual riders into annual members.

Characters and teams

- **Cyclistic**: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.
- **Lily Moreno**: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
- **Cyclistic marketing analytics team**: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. I joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how I, as a junior data analyst, can help Cyclistic achieve them.
- **Cyclistic executive team**: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

About the company

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs 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. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

This project is being completed following six phases of data analysis process:
1- Ask: Identifying the business task, considering the key stakeholders
2- Prepare: Collecting data, identifying its organization and credibility, sort and filter the data
3- Process: Checking for errors, choosing tools, transforming data and document data cleaning
4- Analyze: Aggregate, organize and format the data, performing calculations and identifying trends
5- Share: Determine the best way to share findings and presenting them by creating data visualizations
6- Act: Sharing conclusions and recommendations

Ask

The clear business task is maximizing the future growth of the company by converting casual riders into annual members.
The key stakeholders are Lily Moreno (The director of marketing), Cyclistic executive team (a detail-oriented executive team that decides whether to approve the recommended marketing program), Cyclist marketing analytics team (my team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy).

Prepare

For the purposes of this case study, the datasets are appropriate and will enable us to answer the business questions. This data is for sure credible.
Data is downloaded and stored appropriately in the project folder.

```
## Loading packages
###########################################################

if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(janitor)) install.packages("janitor", repos = "http://cran.us.r-project.org")
if(!require(lubridate)) install.packages("lubridate", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(scales)) install.packages("scales", repos = "http://cran.us.r-project.org")
if(!require(ggpubr)) install.packages("ggpubr", repos = "http://cran.us.r-project.org")

library(tidyverse)
library(janitor)
library(lubridate)
library(ggplot2)
library(scales)
library(ggpubr)
```

```r
## reading the CSV files and creating the data set
###########################################################

df1 <- read.csv("./Data/202106-divvy-tripdata.csv")
df2 <- read.csv("./Data/202107-divvy-tripdata.csv")
df3 <- read.csv("./Data/202108-divvy-tripdata.csv")
df4 <- read.csv("./Data/202109-divvy-tripdata.csv")
df5 <- read.csv("./Data/202110-divvy-tripdata.csv")
df6 <- read.csv("./Data/202111-divvy-tripdata.csv")
df7 <- read.csv("./Data/202112-divvy-tripdata.csv")
df8 <- read.csv("./Data/202201-divvy-tripdata.csv")
df9 <- read.csv("./Data/202202-divvy-tripdata.csv")
df10 <- read.csv("./Data/202203-divvy-tripdata.csv")
df11 <- read.csv("./Data/202204-divvy-tripdata.csv")
df12 <- read.csv("./Data/202205-divvy-tripdata.csv")
rides <- rbind(df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12)
rm(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
```

The organization of the data: there are 5.8 million observations and 13 variables.

```r
head(rides)
```

```
##              ride_id rideable_type          started_at            ended_at
## 1 99FEC93BA843FB20 electric_bike 2021-06-13 14:31:28 2021-06-13 14:34:11
## 2 06048DCFC8520CAF electric_bike 2021-06-04 11:18:02 2021-06-04 11:24:19
## 3 9598066F68045DF2 electric_bike 2021-06-04 09:49:35 2021-06-04 09:55:34
## 4 B03C0FE48C412214 electric_bike 2021-06-03 19:56:05 2021-06-03 20:21:55
## 5 B9EEA89F8FEE73B7 electric_bike 2021-06-04 14:05:51 2021-06-04 14:09:59
## 6 62B943CEAAA420BA electric_bike 2021-06-03 19:32:01 2021-06-03 19:38:46
##   start_station_name start_station_id end_station_name end_station_id start_lat
## 1                                                                         41.80
## 2                                                                         41.79
## 3                                                                         41.80
## 4                                                                         41.78
## 5                                                                         41.80
## 6                                                                         41.78
##   start_lng end_lat end_lng member_casual
## 1    -87.59   41.80  -87.60        member
## 2    -87.59   41.80  -87.60        member
## 3    -87.60   41.79  -87.59        member
## 4    -87.58   41.80  -87.60        member
## 5    -87.59   41.79  -87.59        member
## 6    -87.58   41.78  -87.58        member
```

```r
str(rides)
```

```
## 'data.frame':    5860776 obs. of  13 variables:
##  $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C0FE48C4122
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-06-13 14:31:28" "2021-06-04 11:18:02" "2021-06-04 09:49:35" "2021-0
##  $ ended_at          : chr  "2021-06-13 14:34:11" "2021-06-04 11:24:19" "2021-06-04 09:55:34" "2021-0
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.8 41.8 41.8 41.8 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.8 41.8 41.8 41.8 41.8 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
colnames(rides)
```

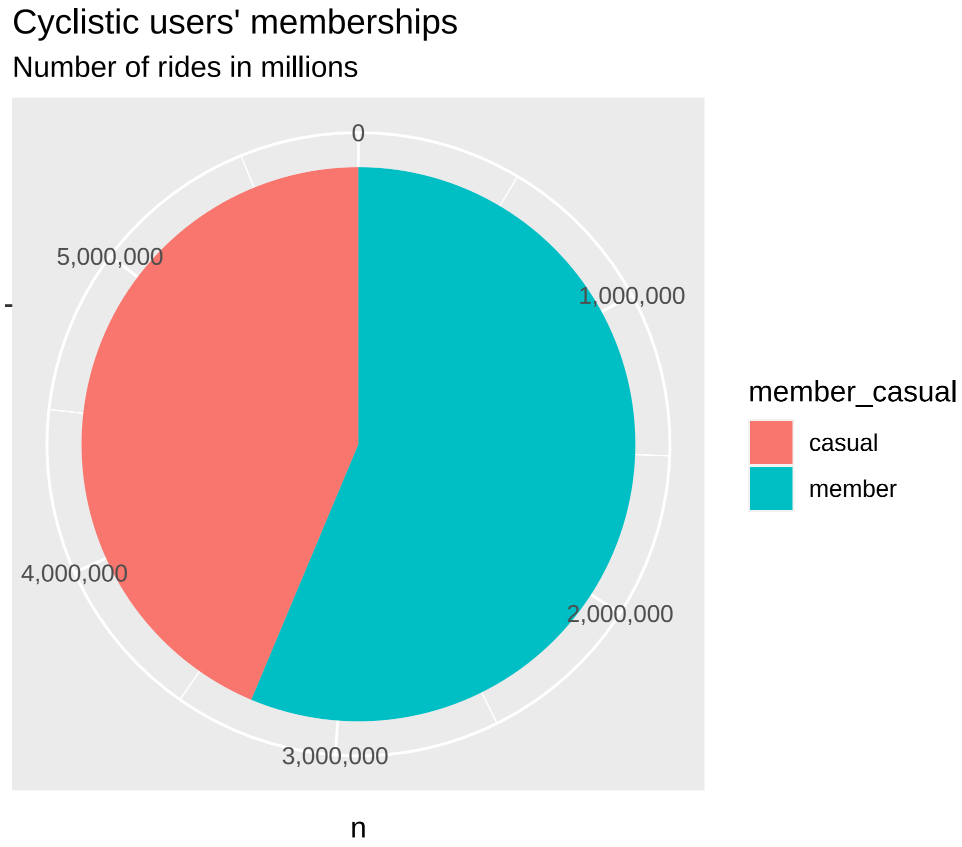
```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

As we see, there are many date and time features in the data set which is presented by a character data type. We will need to use proper functions to convert these columns into more usable data types. The last column in data represents the membership type of the ride, member and casual. Also, there are 3 different types of rideable bikes; Classic bikes, docked bikes, and electric bikes.

```
g1 <- rides %>% count(member_casual) %>%
  group_by(member_casual)
g1
```

```
## # A tibble: 2 x 2
## # Groups:   member_casual [2]
##   member_casual       n
##   <chr>           <int>
## 1 casual        2559857
## 2 member        3300919
```

```
g1 %>% ggplot(aes(y=n, x="", fill = member_casual)) + geom_col() + coord_polar(theta = "y") +
  labs(title = "Cyclistic users' memberships", subtitle = "Number of rides in millions", x = "") +
  scale_y_continuous(labels = comma)
```



```
# different rideable bike presented in the data
rides %>% count(rideable_type)
```

```
##   rideable_type       n
## 1  classic_bike 3217737
## 2   docked_bike  274447
## 3 electric_bike 2368592
```

```
# rideable type of the bike rides distribution
rides %>% count(rideable_type, member_casual)     #all docked_bike records are for casual riders
```

```
##   rideable_type member_casual       n
## 1  classic_bike        casual 1236535
## 2  classic_bike        member 1981202
## 3   docked_bike        casual  274447
## 4 electric_bike        casual 1048875
## 5 electric_bike        member 1319717
```

As the demonstration shows, 56.3 percent of the rides are done by the annual members. The main goal of this project is to increase this percentage. Also, docked bikes are only used by casual riders and all member-rides are either on a classic bike or an electric one.

Process

In this phase we process the data to be more useful for visualizing insights. For example, date and time features are buried in one character column. We need to convert this columns into proper formatted data. Then we extract the date and time for all trips by mutating new columns to the data for corresponding date and time. We also calculate the trip duration, the weekday and the month in which the trip has been recorded.

```
## converting data formats
##########################################################

rides$started_at <- ymd_hms(rides$started_at)
rides$ended_at <- ymd_hms(rides$ended_at)


## mutating new columns to data
##########################################################

# mutating dates and hour of the day
new_rides <- rides %>%
  mutate(start_date = as.Date(rides$started_at),
         end_date = as.Date(rides$ended_at),
         start_hour = hour(rides$started_at),
         end_hour = hour(rides$ended_at))

# calculating the trip duration
new_rides <- new_rides %>%
  mutate(ride_duration = difftime(ended_at, started_at, units = c("mins")))

  # Changing ride_duration data type to numeric
new_rides$ride_duration <- as.numeric(new_rides$ride_duration)

# calculating the trip weekday and month
new_rides <- new_rides %>%
  mutate(start_weekday = weekdays(started_at),
         month = month(started_at))
#new_rides %>% count(month)

  # Reordering the Weekdays
new_rides$start_weekday <- as.factor(new_rides$start_weekday)
summary(new_rides$start_weekday)
```

```
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##    819813    768193    987206    864907    810424    811829    798404
```

```
new_rides$start_weekday <- factor(new_rides$start_weekday,
                                  levels = c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Sa

  # Reordering the months
str(new_rides$month)
```

```
##  num [1:5860776] 6 6 6 6 6 6 6 6 6 6 ...
```

```
new_rides$month <- as.factor(new_rides$month)

levels(new_rides$month) <- list(January = "1", February = "2", March = "3", April = "4",
                                May = "5", June = "6", July = "7", August = "8", September = "9",
                                October = "10", November = "11", December = "12")
summary(new_rides$month)
```

```
##   January  February     March     April       May      June      July    August
##    103770    115609    284042    371249    634858    729595    822410    804352
## September   October  November  December
##    756147    631226    359978    247540
```

Data cleaning

We process the data and make it ready for analysis. We clean the data using tools in “janitor” package to remove empty rows and columns. Also, we investigate the data in each column to find the range of data and the probable errors that might be present within the data. In our case, no personal information or user ID is presented in the data. However, after calculating the ride duration and investigating this column, we need to remove some inaccurate information in which the start and end time were not presented with the correct sequence.

```
## Cleaning data
##########################################################

# Removing empty rows and columns
dim(rides)
```

```
## [1] 5860776      13
```

```
rides <- remove_empty(rides, which = c("rows", "cols"))

# Checking for data consistency
new_rides %>% count(member_casual)    # There are only two types of membership in the data as expected
```

```
##   member_casual       n
## 1        casual 2559857
## 2        member 3300919
```

```
new_rides %>%
  filter(end_date < start_date) %>%
  select(ride_id)                     # No start date is after the end date, as expected
```

```
## [1] ride_id
## <0 rows> (or 0-length row.names)
```

```
new_rides %>%
  filter(ride_duration < 0) %>%
  select(ride_id) %>% dim()          # There are 139 rides having a negative ride_duration
```

```
## [1] 139   1
```

```
#Removing incorrect observations from data
new_rides <- new_rides[!(new_rides$ride_duration < 0), ]

new_rides %>%
  filter(ride_duration < 0) %>%
  select(ride_id)                     # Data is now cleaned from rides with negative duration
```

```
## [1] ride_id
## <0 rows> (or 0-length row.names)
```

After cleaning the data I export the cleaned data to a CSV file for further use.

```
# Exporting the cleaned data as a CSV file for further use:

#write.csv(new_rides,"./New_rides.csv", row.names = TRUE)
```

Analyze and Share

In this section we analyze the data processed and present some graphs. The insight for each graph is explained after each data visualization.

Insights and visualizations

1- Ride Duration for members and casual riders

First, in order to better understand the behavior of the riders, we calculate the average ride duration for casual riders and for members.

```
## Graphs and data insights
###########################################################

# average ride duration by members and casual riders
new_rides %>%
  select(member_casual,ride_duration) %>%
  group_by(member_casual) %>%
  summarize(Average_ride_duration = mean(ride_duration))
```

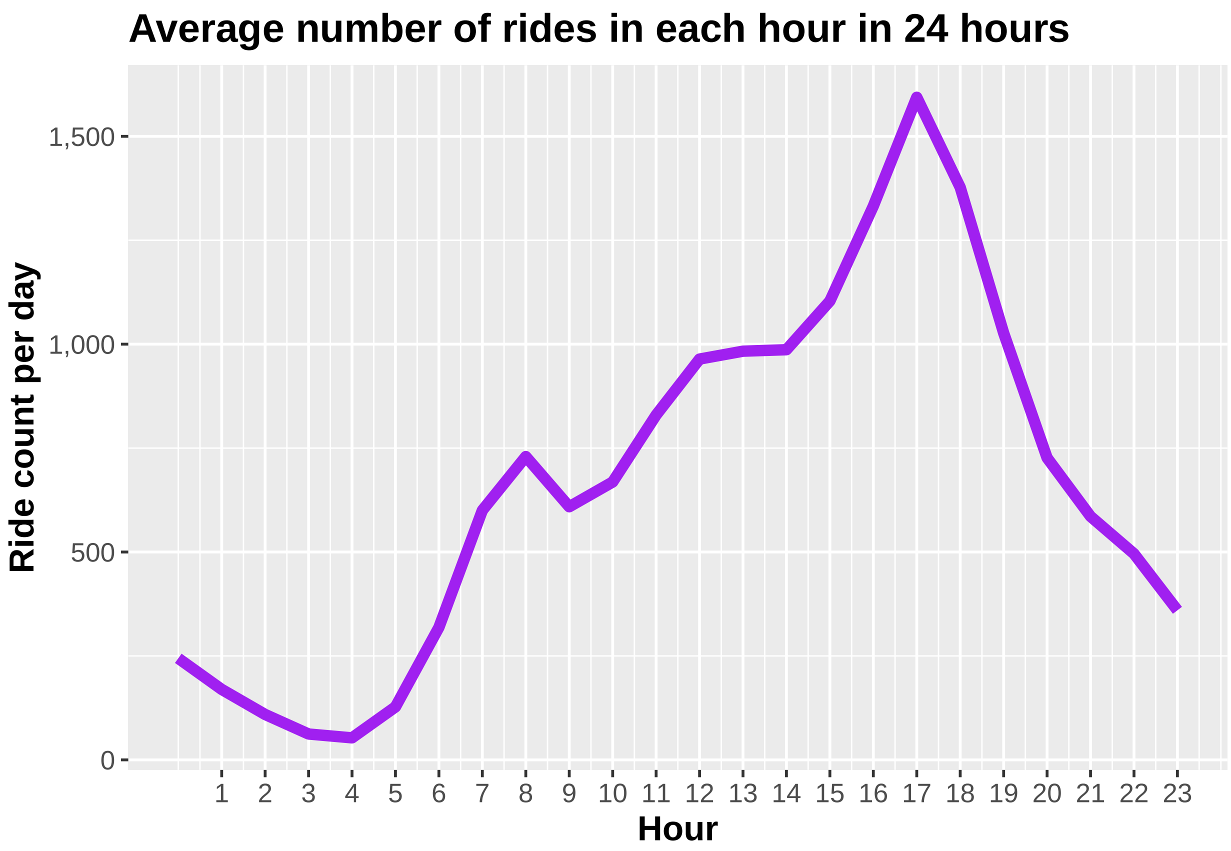
```
## # A tibble: 2 x 2
##   member_casual Average_ride_duration
##   <chr>                         <dbl>
## 1 casual                         30.5
## 2 member                         13.0
```

The average ride duration for a casual rider is more than twice the average ride duration for a member. Consequently, according to this data it is derived that casual riders ride longer than members.

2- Rides in each hour of a 24 hour period

Here we visualize the data on the hour that a ride has performed. Since the data we have is the total for a 12 month period, we calculate the average number of rides per day by dividing the total count by 365.

```
# rides based on the hour in each day
new_rides %>% count(start_hour) %>%
  ggplot(aes(start_hour,n/365 )) +
  geom_line(color = "purple", size = 2) +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(breaks = seq(1:23)) +
  labs(title = 'Average number of rides in each hour in 24 hours') +
  theme(plot.title = element_text(color="black", size=15, face="bold"),
        axis.title.x = element_text(color="black", size=13, face="bold"),
        axis.title.y = element_text(color="black", size=13, face="bold"),
        axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10)) +
  xlab("Hour") + ylab("Ride count per day")
```

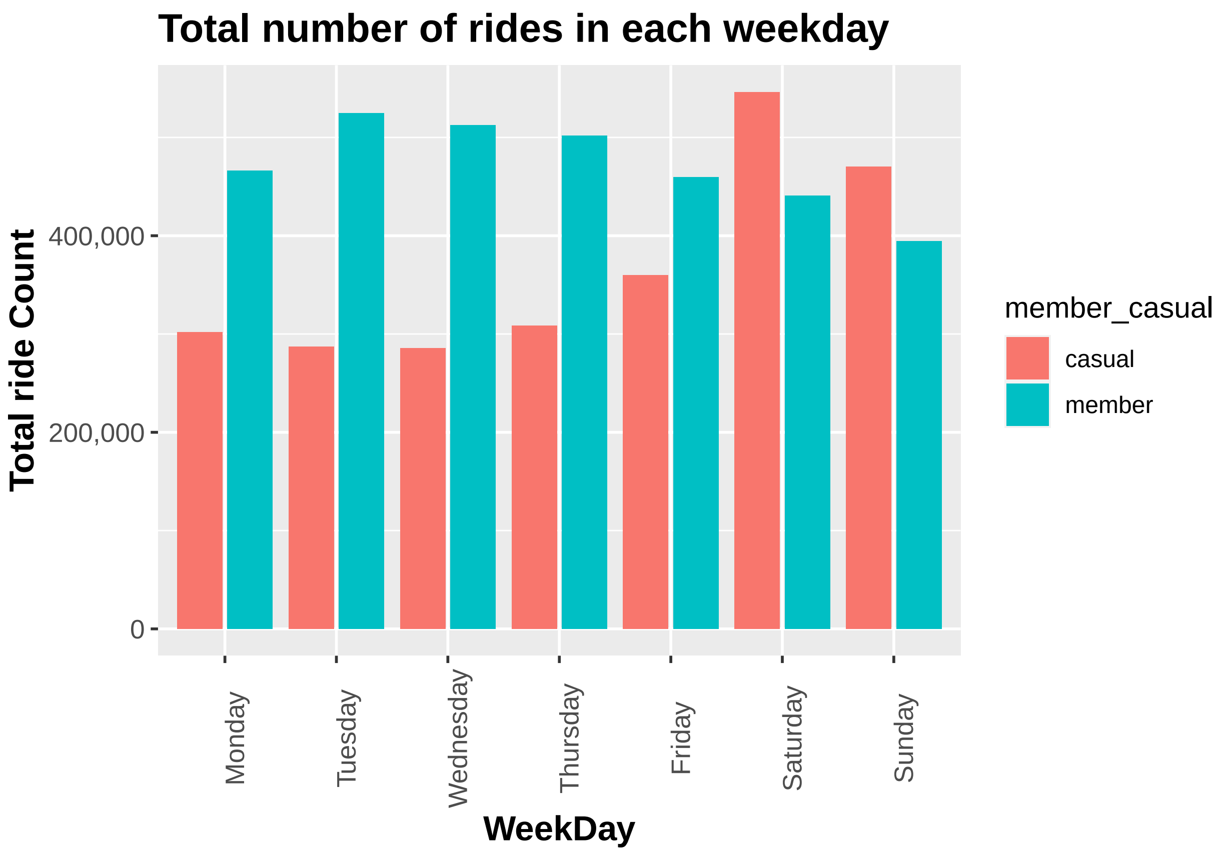


According to this plot, the number of rides steadily increases from 4 am (less than 100 rides) to 5 pm (more than 1600 rides) where it drops back with a sharper inclination. In fact it was logically expected that the number of rides will be in its maximum on rush-hour. In other words, the more the time passes rush-hour the less would be the number of rides on a single day.

3- The number of rides in each weekday

We calculated the weekday of each observation in our data in the process phase. Now we demonstrate the total number of rides in each weekday, grouped by different types of membership to evaluate the behavior of the Cyclistic users.

```
# rides based on the weekday
new_rides %>%
  ggplot(aes(start_weekday, fill = member_casual)) +
  geom_bar(position = "dodge2") +
  scale_y_continuous(labels = comma) +
  labs(title = 'Total number of rides in each weekday') +
  theme(plot.title = element_text(color="black", size=15, face="bold"),
        axis.title.x = element_text(color="black", size=13, face="bold"),
        axis.title.y = element_text(color="black", size=13, face="bold"),
        axis.text.x = element_text(angle = 90, size=10),
        axis.text.y = element_text(size=10)) +
  xlab("WeekDay") + ylab("Total ride Count")
```

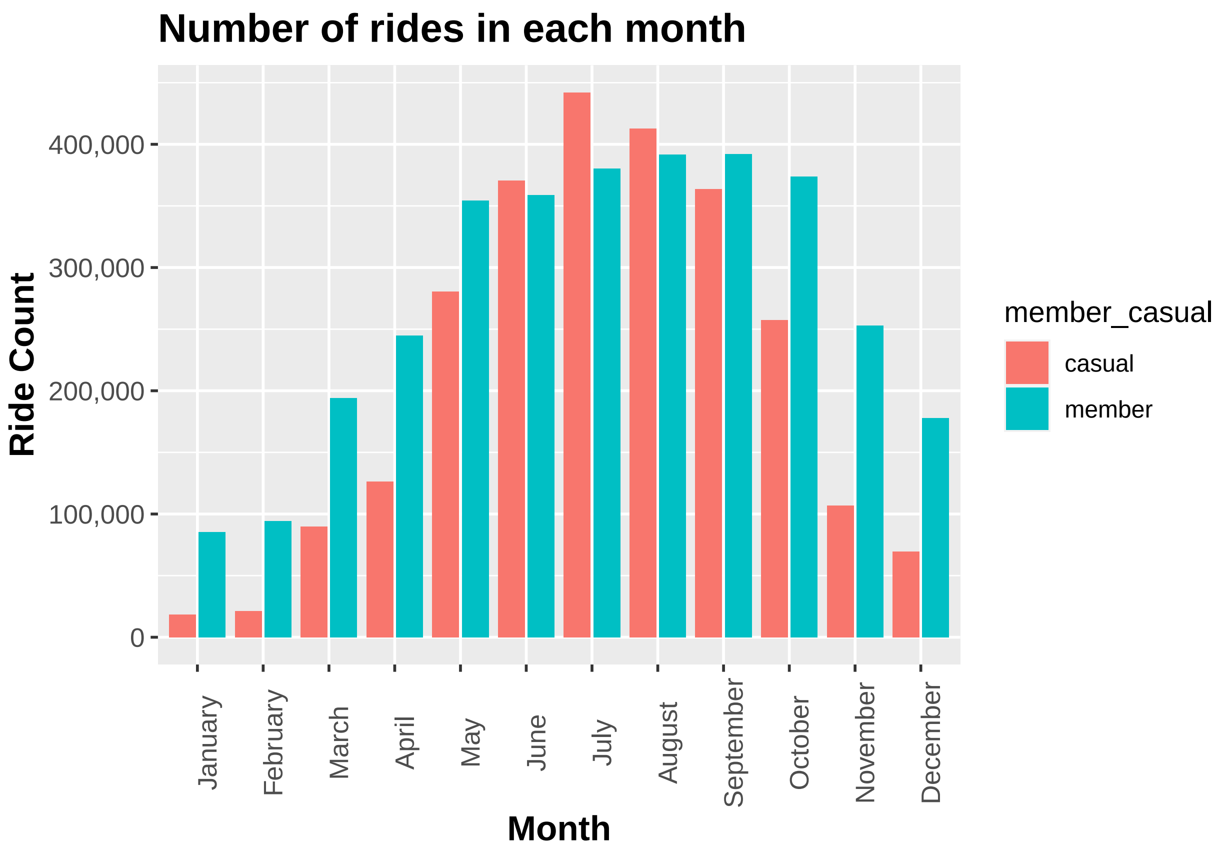


According to this data, a clear difference between the behavior of the annual members and casual users is obvious. Annual members use the bikes more on weekdays (probably for commuting to work), and less on weekends. However, casual riders, ride bikes dramatically more on weekends and less on weekdays. The casual users use Cyclistic on Saturday more than members on any weekdays. This behavior gives us, as data analysts, a perfect opportunity to suggest recommendations and pursuit the project goal.

4- The number of rides in each month during the year

We extracted a column for the month each observation has happened. We use this data to visualize the number of rides in each month by members and casual riders, in order to compare and derive some new insights on the data and on the behavior of users.

```r
# rides in each month
new_rides %>%
  ggplot(aes(month, fill = member_casual)) +
  geom_bar(position = "dodge2") +
  labs(title = "Number of rides in each month") +
  scale_y_continuous(labels = comma) +
  theme(plot.title = element_text(color="black", size=15, face="bold"),
        axis.title.x = element_text(color="black", size=13, face="bold"),
        axis.title.y = element_text(color="black", size=13, face="bold"),
        axis.text.x = element_text(angle = 90, size=10),
        axis.text.y = element_text(size=10)) +
  xlab("Month") + ylab("Ride Count")
```



According to this graph, we see a more subtle change in the number of rides by annual members than the number of rides by casual riders. Around summer, we see a rise in the number of casual rides. It can be concluded that more casual users of Cyclistic, use this bike-sharing company when the weather is suitable and probably for fun. On the other hand, casual rides are dropped significantly around winter season. One clear reason for this drop is the weather and the challenges bike riders would face on harsh conditions. This data visualization gives us a perfect insight towards the analysis goal of concentrating on casual riders and converting them into annual members.

5- Finding the most used stations

According to the rideable type data investigation, it is apparent that Cyclistic users are not necessarily forced to borrow bikes or leave them on the designated stations; however, there are many stations available for users. In this part of our analysis, we focus on the ride records which have data in the start/end station name columns. We count the number of rides started or ended at different stations and sort them descending.

```
# Busiest stations:

new_rides[c(new_rides$start_station_name != ""),] %>%
  count(start_station_name) %>% arrange(desc(n)) %>% head(10)
```

```
##          start_station_name     n
## 1   Streeter Dr & Grand Ave 84271
## 2     Wells St & Concord Ln 44079
## 3     Michigan Ave & Oak St 44009
## 4           Millennium Park 41276
## 5          Clark St & Elm St 40891
## 6          Wells St & Elm St 37556
## 7  Kingsbury St & Kinzie St 36107
## 8       Theater on the Lake 35949
## 9   Clark St & Armitage Ave 33047
## 10   Wabash Ave & Grand Ave 32986
```

```
new_rides[c(new_rides$end_station_name != ""),] %>%
  count(end_station_name) %>% arrange(desc(n)) %>% head(10)
```

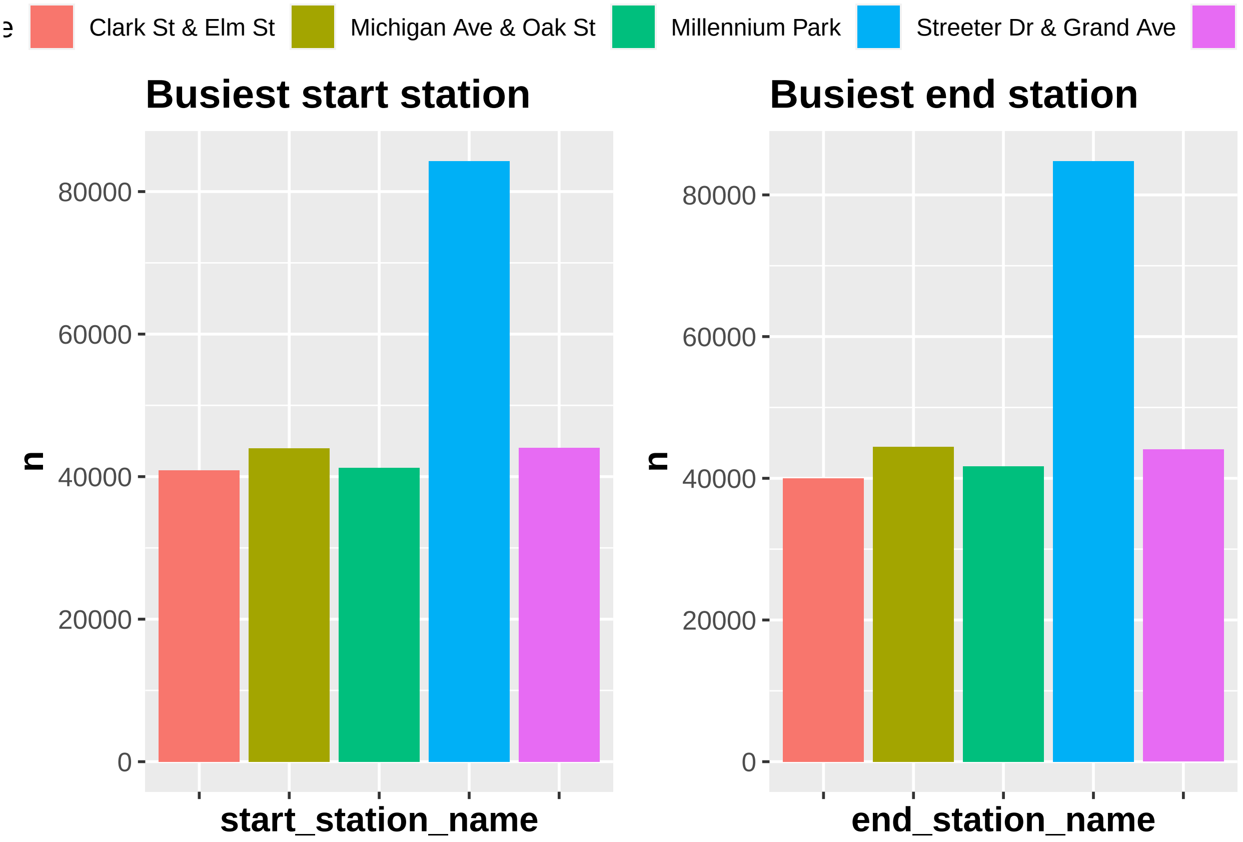
```
##                       end_station_name     n
## 1             Streeter Dr & Grand Ave 84763
## 2               Michigan Ave & Oak St 44431
## 3               Wells St & Concord Ln 44063
## 4                     Millennium Park 41692
## 5                   Clark St & Elm St 40011
## 6                   Wells St & Elm St 36937
## 7                 Theater on the Lake 36046
## 8  DuSable Lake Shore Dr & North Blvd 35808
## 9            Kingsbury St & Kinzie St 35148
## 10             Wabash Ave & Grand Ave 32887
```

Considering the top 10 most stations being used by Cyclistic users, we see a distinct popular station both for starting and for ending bike rides, the “Streeter Dr & Grand Ave” station. In order to find the most used stations, we choose the same top 5 stations for both starting and ending rides (more than 40000 ride records). As a result these 5 stations are statistically proven to be the busiest stations among all.

```
p1 <- new_rides[c(new_rides$start_station_name != ""),] %>%
  count(start_station_name) %>% arrange(desc(n)) %>% filter(n>40000) %>%
  ggplot(aes(start_station_name, n, fill=start_station_name)) + geom_col() +
  labs(title = 'Busiest start station') +
  theme(plot.title = element_text(color="black", size=15, face="bold"),
      axis.title.x = element_text(color="black", size=13, face="bold"),
      axis.title.y = element_text(color="black", size=13, face="bold"),
      axis.text.x = element_text(angle = 90, size=10, hjust = 1),
      axis.text.y = element_text(size=10)) +
  rremove("x.text")

p2 <- new_rides[c(new_rides$end_station_name != ""),] %>%
  count(end_station_name) %>% arrange(desc(n)) %>% filter(n>40000) %>%
  ggplot(aes(end_station_name, n, fill = end_station_name)) + geom_col() +
  labs(title = 'Busiest end station') +
  theme(plot.title = element_text(color="black", size=15, face="bold"),
        axis.title.x = element_text(color="black", size=13, face="bold"),
        axis.title.y = element_text(color="black", size=13, face="bold"),
        axis.text.x = element_text(angle = 90, size=10, hjust = 1),
        axis.text.y = element_text(size=10)) +
  rremove("x.text")

ggarrange(p1,p2, common.legend = TRUE, ncol = 2)
```



```
new_rides[c(new_rides$start_station_name != ""),] %>%
  count(start_station_name) %>% arrange(desc(n)) %>% filter(n>40000)
```

```
##        start_station_name     n
## 1 Streeter Dr & Grand Ave 84271
## 2   Wells St & Concord Ln 44079
## 3   Michigan Ave & Oak St 44009
## 4         Millennium Park 41276
## 5       Clark St & Elm St 40891
```

Act

Now that we have finished our visualizations, I will prepare recommendations based on the insights we gained on the data presented.

Summary of the insights:

- Casual riders, ride longer than annual members
- The peak of rides in each day is around the rush-hour
- Annual members use Cyclistic more during weekdays
- Casual riders use Cyclistic much more on Saturdays and Sundays
- Casual riders use Cyclistic dramatically more during summer and the adjoining months
- There are clearly 5 specific bike stations that are being used the most by Cyclistic users

Recommendations:

1- Run advertisements and marketing campaigns on the top 5 most used stations. It can include some presenters who can talk to the users and encourage them to opt for annual membership instead of single-ride or full-day passes.

2- Create an intuitive mobile application for collecting points on rides taken, based on the duration of the ride. Promoting longer rides and offering discounts on rides more than a fixed duration.

3- Focus on presenting promotions for membership on summers and on some special weekends.

4- Send limited coupons to casual riders based on their habit of using the service. These coupons should be sent either on weekends or during rush-hour time of the weekdays.