Cyclistic Bike Share: Analysis

Jeryl Gonsalves

Case Study 1: How Does a Bike-Share Navigate Speedy Success?

Introduction

My capstone project for the Google Data Analytics Certificate Course is the case study below. It entails analyzing historical data for a hypothetical company, Cyclistic, a Chicago-based bike sharing company, in order to make marketing campaign recommendations. Although the firm and the scenario are made up, the data for this project was acquired from a Chicago bike share program between June 2021 and May 2022 (a period of 12 months). The information is updated on a regular basis. In this project I am assuming the role of the junior analyst.

Scenario

Cyclistic is a fictional Chicago-based bike-sharing enterprise. It has a fleet of more than 5,800 bicycles available at over 600 docking stations across the city. Bikes can be leased from one docking station, ridden, and then returned to any of the system's docking stations. Marketing campaigns have been broad and targeted a cross-section of potential users over the years. Riders who pay for an annual membership are more profitable than casual riders, according to data research. The marketing department is considering developing a campaign to entice casual passengers to become members. The marketing analyst team is curious in the differences between yearly members and casual riders, as well as why casual riders would purchase a membership and how Cyclistic can leverage digital media to persuade casual riders to join. The team wants to look at the Cyclistic historical bike trip data to see whether there are any patterns in how casual and member users use bikes.

Further, for data analysis I have used 6 steps that is Ask, Prepare, Process, Analyze, Share, Act to figure out the scenario and come to a conclusion.

1. Ask

Business Objective

To enhance revenue by converting casual riders into annual members, through a targeted marketing approach,

Business Task for Junior Analyst

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

Stakeholders

The stakeholders for this project are Lily Moreno, Director of Marketing at Cyclistic, is in charge of the company's marketing campaign.

The marketing analytics team of Cyclistic. This group is in charge of gathering, analyzing, and reporting data for marketing efforts. This team's junior analyst is myself.

The Cyclistic management team. This group makes the ultimate decision on the marketing strategy that has been recommended. They are known for their meticulous attention to detail.

2. Prepare

Where is the Data located?

The data has been made available by Motivate International Inc., a company employed by the City of Chicago for bike-sharing.

How is the Data organized?

The data is organized in monthly csv files, which keep on updating. The most recent twelve months of data () were used for this project. The files consist of 13 columns containing information related to ride id, rideable type, ride time when it started, ride time when it ended, start station name, end station name, start station ID, end station ID, Start station Latitude, Start station Longitude, End station latitude, end station longitude, members and casuals.

Credibility of Data

Motivate, Inc., the corporation that manages the City of Chicago's Cyclistic Bike Share program, collects the data directly.There are some issues in variables though. The data is extensive in that it includes information from all of the rides taken on the system, rather than just a sample. The information is current. It is published on a monthly basis and was current as of May 2022. The data is made accessible to the public by the City of Chicago.

Licensing, privacy, security, and accessibility

All identifying information has been removed from this data, making it anonymous. This protects privacy, but also restricts the scope of the investigation. There is insufficient data to tell whether casual riders are repeat riders or if casual riders are Chicago residents. . The data has been made available by Motivate International Inc. under this license(https://ride.divvybikes.com/data-license-agreement (https://ride.divvybikes.com/data-license-agreement))

Data Integrity and Ability to be used to answer as Business Question

The type of rider is recorded in one of the fields in the data; casual riders pay for individual or daily rides, whereas member riders pay for an annual subscription. This data is necessary in order to identify differences in how the two groups use the bike share program.

Problems in the data

There are some problems with the data. Missing records and duplicate records in the data set as well as having the trip duration <=0

3. Process

Tools used

For this project I choose to use RStudio Desktop to clean, analyze the data and to create the visualizations. The data set was too large to be processed in spreadsheets and RStudio Cloud.

Data Cleaning

Data was examined to gain a general understanding of field content, data formats, and data integrity. Checking column names throughout the 12 original files and looking for missing values, trailing white spaces, duplicate records, and other data abnormalities were all part of the data assessment.

I first installed all the required libraries and then loaded them. After this, I set up the working directory and uploaded the csv files. The libraries that were loaded were tidyverse, janitor, lubridate, dplyr, readr, skimr, ggplot2 and gridExtra.

Installing and Loading the Libraries

```
rm(list=ls())

## Creating data sets

df1 <- read.csv("Data/202105-divvy-tripdata.csv")
df2 <- read.csv("Data/202106-divvy-tripdata.csv")
df3 <- read.csv("Data/202107-divvy-tripdata.csv")
df4 <- read.csv("Data/202108-divvy-tripdata.csv")
df5 <- read.csv("Data/202109-divvy-tripdata.csv")
df6 <- read.csv("Data/202110-divvy-tripdata.csv")
df7 <- read.csv("Data/202111-divvy-tripdata.csv")
df8 <- read.csv("Data/202112-divvy-tripdata.csv")
df9 <- read.csv("Data/202201-divvy-tripdata.csv")
df10 <- read.csv("Data/202202-divvy-tripdata.csv")
df11 <- read.csv("Data/202203-divvy-tripdata.csv")
df12 <- read.csv("Data/202204-divvy-tripdata.csv")
```

Once the initial review was completed, all twelve files were loaded into one data frame along with removing the empty data. The resulting amalgamated file consisted of 4,967,344 rows with 13 columns of character and numeric data. This matched the number of records in the twelve monthly files.

```
#Combining 12 data.frames into (1) data.frame
bike_rides <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
bike_rides <- janitor::remove_empty(bike_rides,which = c("cols"))
bike_rides <- janitor::remove_empty(bike_rides,which = c("rows"))
bike_rides <- bike_rides  %>% filter(start_station_name !="")
```

Several new columns were established and loaded with data along with converting the data from the started at date and time column to enable for more granular data analysis and insights. Day, month, year, time, and day of the week were the new columns.

```
## Convert Data/Time stamp to Date/Time ...
bike_rides$Ymd <- as.Date(bike_rides$started_at)
bike_rides$started_at <- lubridate::ymd_hms(bike_rides$started_at)
bike_rides$ended_at <- lubridate::ymd_hms(bike_rides$ended_at)


bike_rides$start_hour <- lubridate::hour(bike_rides$started_at)
bike_rides$end_hour <- lubridate::hour(bike_rides$ended_at)

##Separating the dates into month, day, year and day of week
bike_rides$Date<- as.Date(bike_rides$started_at)
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")
```

Another column was developed to keep track of how long the trip took (length of each trip). The difference in time between the ride's start and end times was used to generate the data for this column. The travel duration in minutes was then added to another version of this column.

```
## Finding the trip duration
bike_rides$Hours <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("hours"))
bike_rides$Minutes <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("mins"))
bike_rides$seconds<-difftime(bike_rides$ended_at,bike_rides$started_at)
df <- bike_rides %>% filter(Minutes >0) %>% na.omit()
summary(df$Minutes)
```

```
##   Length    Class     Mode
##  4961949 difftime  numeric
```

4. Analyze

In this step, the data is analysed to find descriptive statistics along with comparing the aggregates for members and casuals. Further, visualizations are done to give the better understanding of the data.

```
mean(df$Minutes)
```

```
## Time difference of 20.8008 mins
```

```
median(df$Minutes)
```

```
## Time difference of 11.73333 mins
```

```
str(df)
```

```
## 'data.frame':    4961949 obs. of  24 variables:
##  $ ride_id           : chr  "0F3AE375DEC608D9" "43E916C72D4C77B3" "DCFA1BA26E2431CD" "C2C4FF78B
860306E" ...
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-05-02 08:21:19" "2021-05-01 22:15:55" ...
##  $ ended_at          : POSIXct, format: "2021-05-02 09:20:46" "2021-05-01 22:51:27" ...
##  $ start_station_name: chr  "Michigan Ave & Oak St" "Michigan Ave & Oak St" "Dearborn St & Monr
oe St" "Michigan Ave & Oak St" ...
##  $ start_station_id  : chr  "13042" "13042" "TA1305000006" "13042" ...
##  $ end_station_name  : chr  "Michigan Ave & Oak St" "Michigan Ave & Oak St" "" "" ...
##  $ end_station_id    : chr  "13042" "13042" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 42 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
##  $ Ymd               : Date, format: "2021-05-02" "2021-05-01" ...
##  $ start_hour        : int  8 22 17 9 0 7 16 6 18 16 ...
##  $ end_hour          : int  9 22 17 10 0 8 16 7 18 16 ...
##  $ Date              : Date, format: "2021-05-02" "2021-05-01" ...
##  $ month             : chr  "05" "05" "05" "05" ...
##  $ day               : chr  "02" "01" "05" "30" ...
##  $ year              : chr  "21" "21" "21" "21" ...
##  $ day_of_week       : chr  "Sunday" "Saturday" "Wednesday" "Sunday" ...
##  $ Hours             : 'difftime' num  0.990833333333333 0.592222222222222 0.283055555555556 0.
766111111111111 ...
##   ..- attr(*, "units")= chr "hours"
##  $ Minutes           : 'difftime' num  59.45 35.5333333333333 16.9833333333333 45.9666666666667
...
##   ..- attr(*, "units")= chr "mins"
##  $ seconds           : 'difftime' num  3567 2132 1019 2758 ...
##   ..- attr(*, "units")= chr "secs"
##  - attr(*, "na.action")= 'omit' Named int [1:4766] 183 196 226 350 1323 1342 1370 1431 1449 146
7 ...
##   ..- attr(*, "names")= chr [1:4766] "183" "196" "226" "350" ...
```

```
## Comparing members and casuals
aggregate(df$Minutes~df$member_casual,FUN = mean)
```

df$member_casual <chr>	df$Minutes <drtn>
casual	30.72623 mins
member	12.97296 mins

2 rows

```
aggregate(df$Minutes~df$member_casual,FUN = median)
```

df$member_casual <chr>	df$Minutes <drtn>
casual	16.066667 mins

df$member_casual <chr>	df$Minutes <drtn>
member	9.333333 mins

2 rows

```
aggregate(df$Minutes~df$member_casual,FUN = max)
```

df$member_casual <chr>	df$Minutes <drtn>
casual	55944.150 mins
member	1499.933 mins

2 rows

```
aggregate(df$Minutes~df$member_casual,FUN = min)
```

df$member_casual <chr>	df$Minutes <drtn>
casual	0.01666667 mins
member	0.01666667 mins

2 rows

```
##Measuring the average ride time for each day for Members V/s Casuals
aggregate(df$Minutes~df$member_casual+df$day_of_week,FUN = mean)
```

df$member_casual <chr>	df$day_of_week <chr>	df$Minutes <drtn>
casual	Friday	28.79354 mins
member	Friday	12.68103 mins
casual	Monday	30.58387 mins
member	Monday	12.53732 mins
casual	Saturday	33.35031 mins
member	Saturday	14.59827 mins
casual	Sunday	35.94719 mins
member	Sunday	14.84954 mins
casual	Thursday	27.26467 mins
member	Thursday	12.27109 mins

1-10 of 14 rows

```
##As the days of the week are out of order, fixing it
df$day_of_week<-ordered(df$day_of_week,levels=c("Sunday","Monday","Tuesday","Wednesday","Thrusday"
,"Friday","Saturday"))
mean(df$Minutes)
```

```
## Time difference of 20.8008 mins
```

Data Visualization

The clean data set is now visualized and the following analysis is determined.

1. Number of trips by rider type

2. Mean Travel Time by rider type

3. Number of rides by user type during the week

4. Number of rides by user type during the year

5. Bike type usage by user type

6. Bike type usage by user type during a week

7. Top 10 start stations for members

8. Top 10 start station for casuals

Summary of Analysis

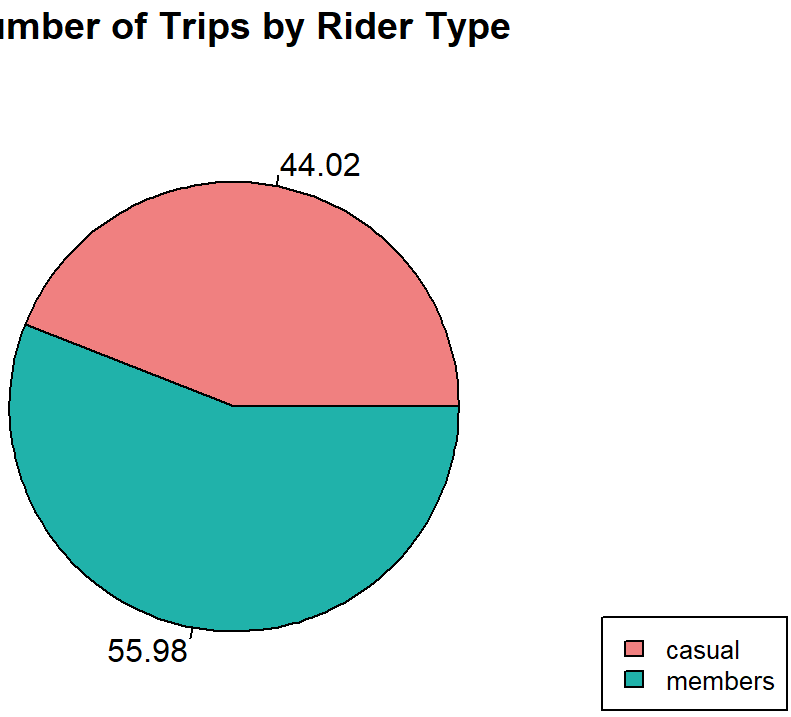
From the analysis we can see that there are several key differences between casual and member riders.

Number of Trips by Rider Type

```
##Finding out the total trips of members and casuals from a pie chart
table(df["member_casual"])
```

```
##
##  casual  member
## 2187842 2774107
```

```
  total_users <- c(2532243,3219890)
labels <- c("casual","members")
piepercent <- round(100*total_users/sum(total_users), 2)
colors <- c("lightcoral","lightseagreen")
pie(total_users,labels =  piepercent, main = "Number of Trips by Rider Type", col = colors)
legend("bottomright",c("casual","members"),cex = 0.8, fill = colors)
```



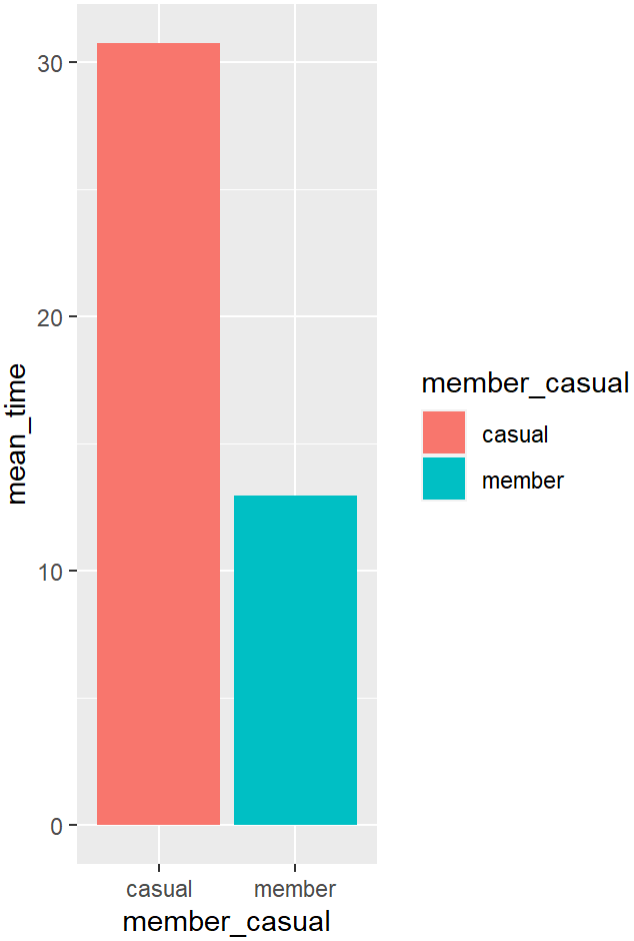
In the above pie chart, member riders take more trips that casual riders i.e the percentage of members is 55.98% and that of casual is 44.02%

Mean Travel Time by User Type

```
##Average Time for both members and casuals
userType_mean <- df %>% group_by(member_casual)%>% summarise(mean_time = mean(Minutes))
membervstime<-ggplot(userType_mean)+
  geom_col(mapping=aes(x=member_casual,y=mean_time,fill=member_casual))
  labs(title = "Mean Travel Time by User Type",x="User Type",y="Mean time in sec")
```

```
## $x
## [1] "User Type"
##
## $y
## [1] "Mean time in sec"
##
## $title
## [1] "Mean Travel Time by User Type"
##
## attr(,"class")
## [1] "labels"
```

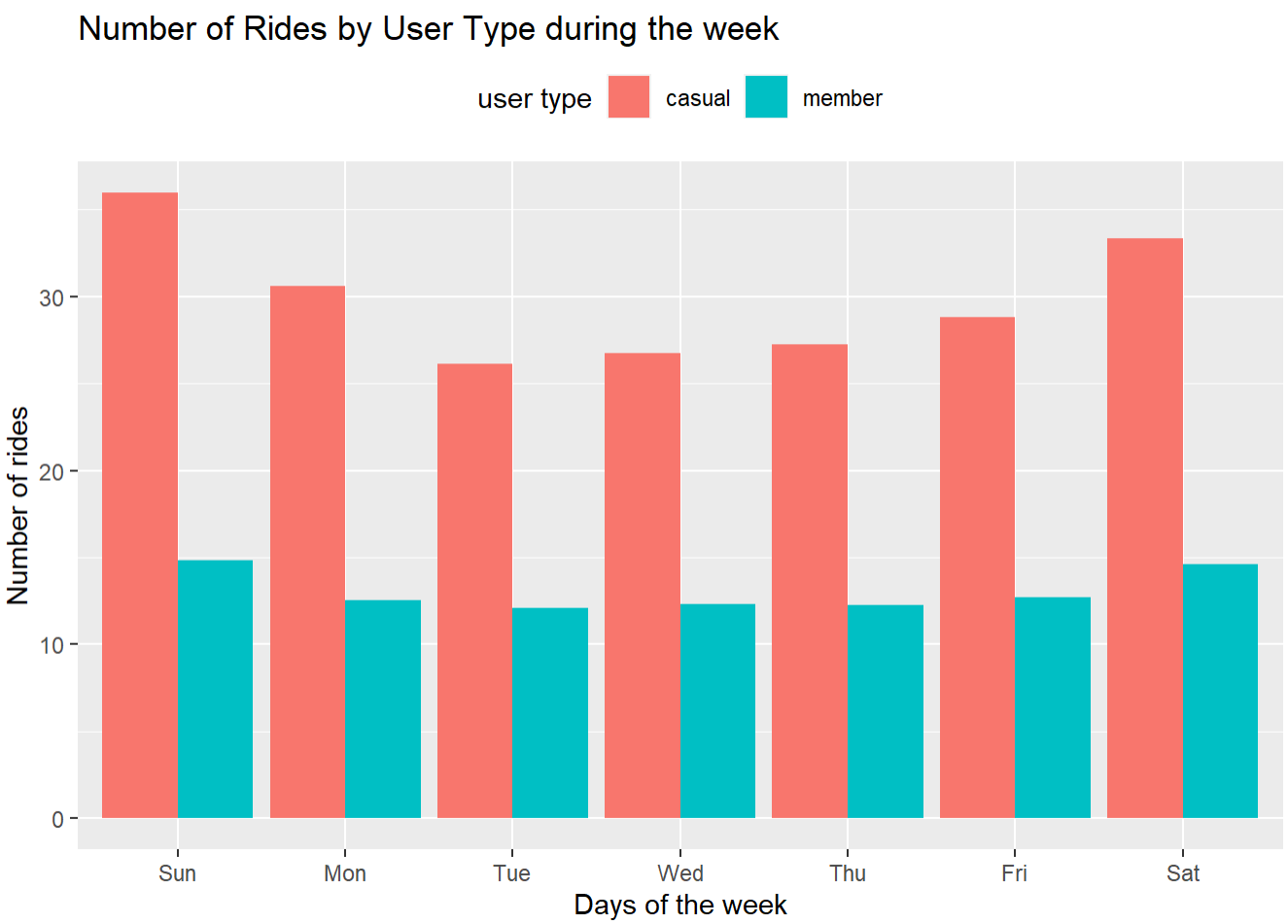
```
grid.arrange(membervstime,ncol=2)
```



Even if, there are more members than casual riders, the average time of casual riders is more than the members as seen in the figure.

Number of Rides by User Type during the week

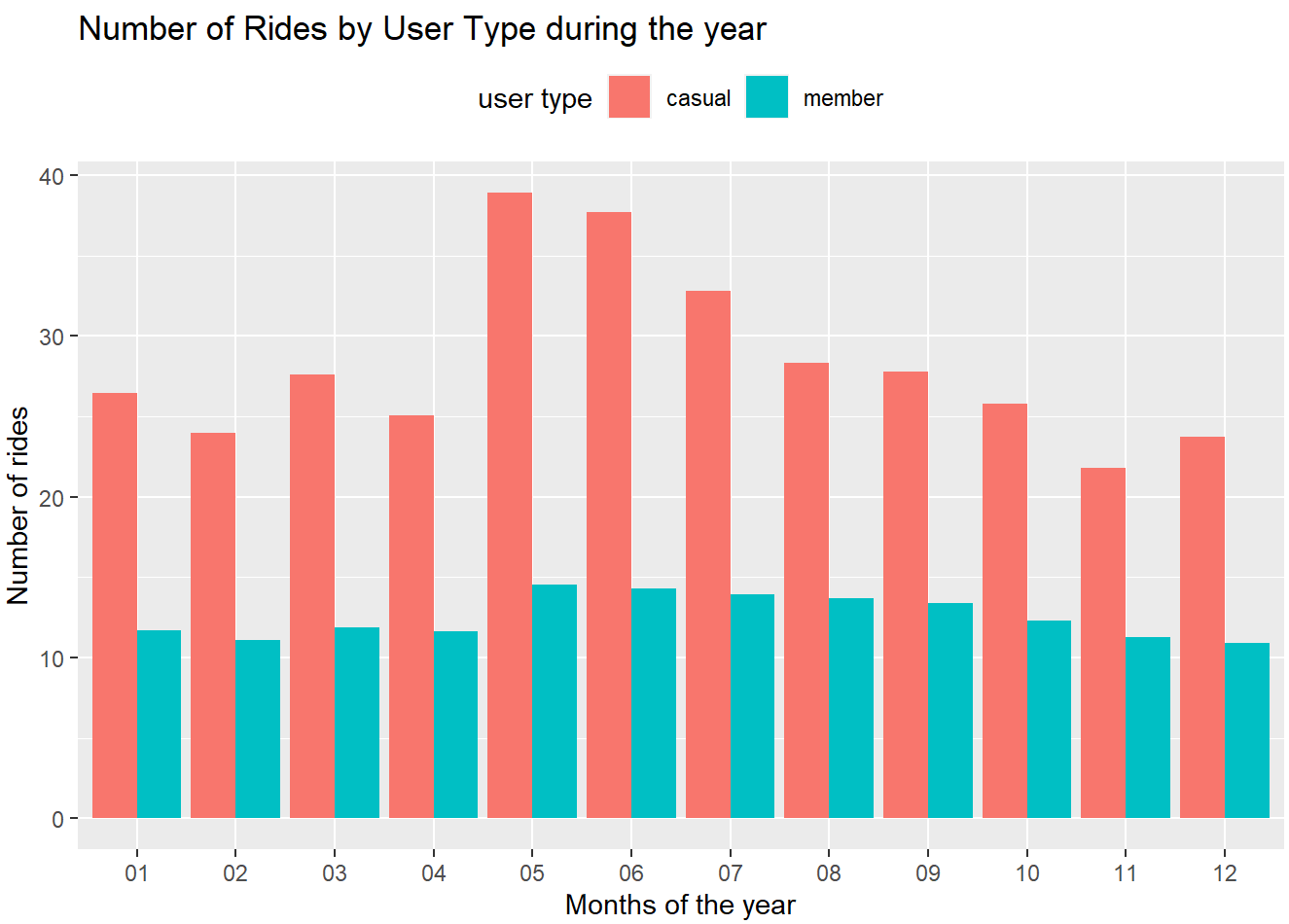
```
##Average duration of rides during the week
df %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>% group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(Minutes)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) + geom_col(position = "dodg
e")+ labs(title = "Number of Rides by User Type during the week",x="Days of the week",y="Number of
rides",fill="user type")+ theme(legend.position = "top")
```



In the figure, it shows that casual rides are more than member rides for all days. There is an increase in rides for both riders during the weekend.

Number of Rides by User Type during the year

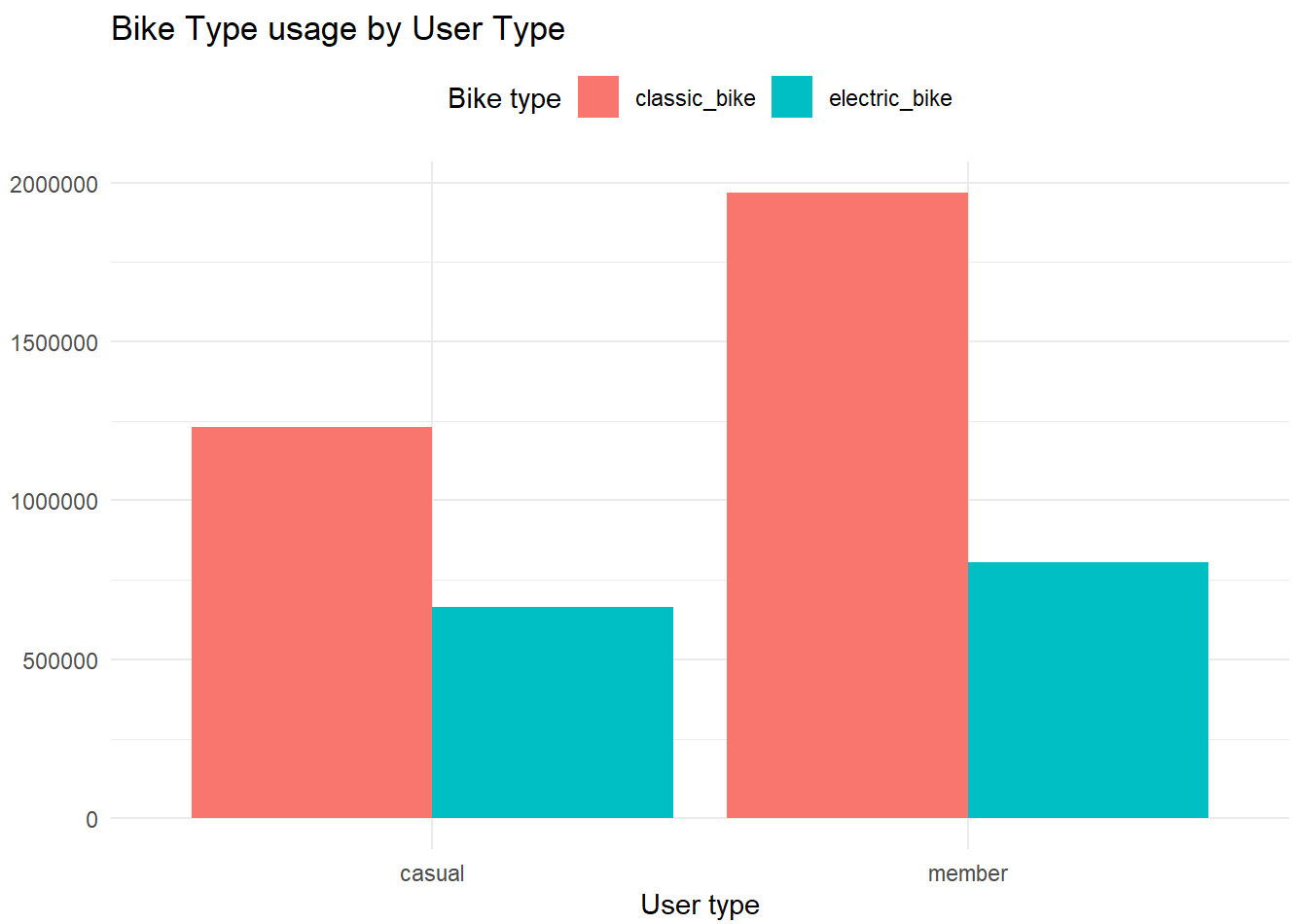
```
df%>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(Minutes)) %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(title = "Number of Rides by User Type during the year",x="Months of the year",y="Number of
 rides",fill="user type")+
  theme(legend.position = "top")
```



In this bar plot, As the months are from June 2021- May 2022 it is observed that the maximum rides are seen by casuals in the month of October 2021. Also, the rides are more in members in October, marginally more than November 2021. The least rides for casuals is April 2022 and for members it is May 2022.

Bike Type Usage by User Type

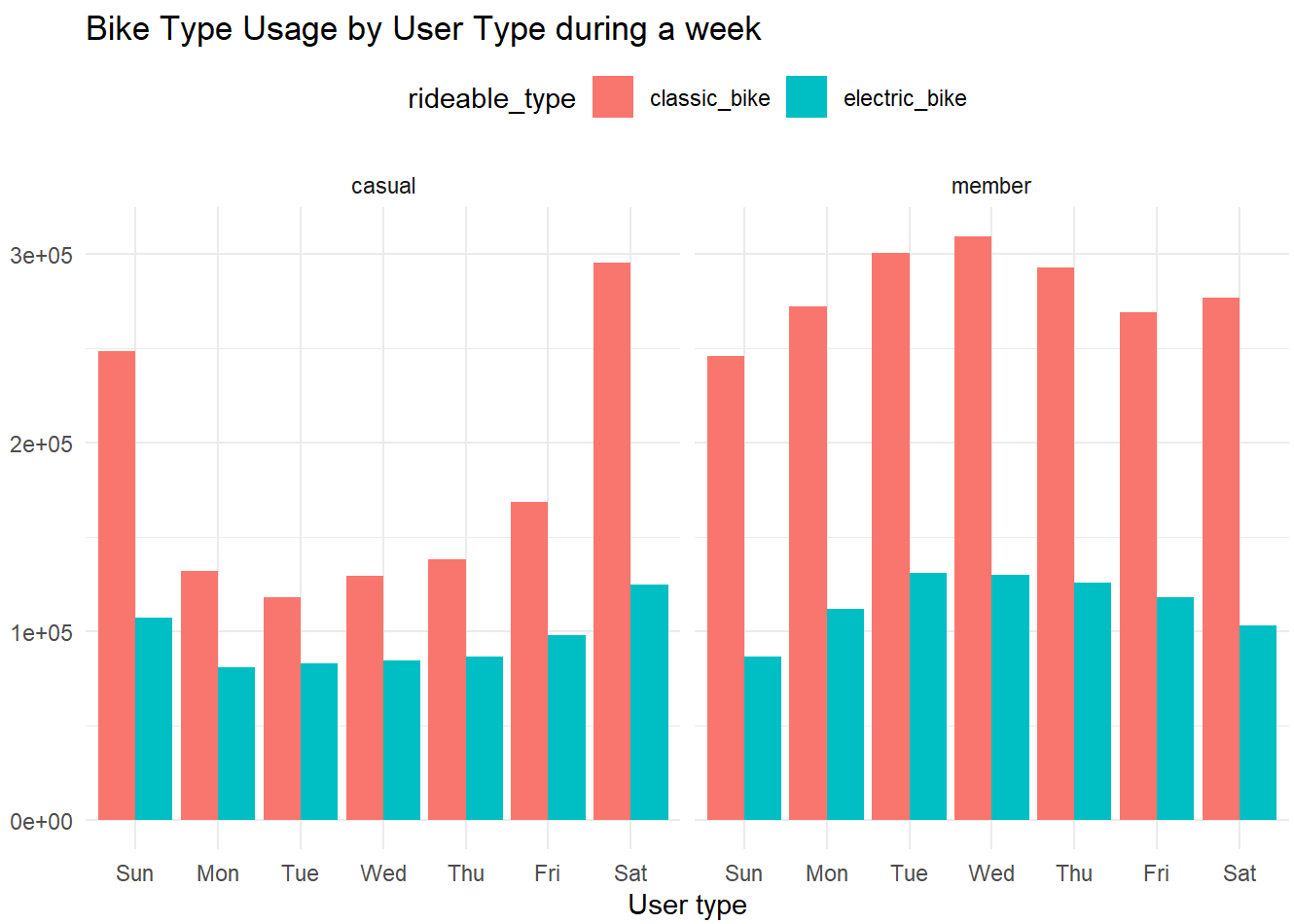
```
##Bike type members and casuals
with_bike_type<-df %>% filter(rideable_type=="classic_bike" | rideable_type=="electric_bike")

##Bike type usage by user type
with_bike_type %>%
  group_by(member_casual,rideable_type)%>%
  summarise(totals=n(), .groups = "drop")%>%
  ggplot()+
  geom_col(aes(x=member_casual,y=totals,fill=rideable_type),position = "dodge") +
  labs(title = "Bike Type usage by User Type",x="User type",y=NULL,fill="Bike type")+
  theme_minimal()+
  theme(legend.position = "top")
```



In the following figure, two bike types are considered that is classic and electric. Classic bike types are used more than electric bike types. Also, members prefer classic more than casual. On the other hand, the usage of electric bike types are less and also members use it more than casuals but the increase is just marginally for members from casuals.

Bike Type Usage by User Type during a week

```
with_bike_type %>%
  mutate(weekday = wday(started_at,label = TRUE))%>%
  group_by(member_casual,rideable_type,weekday)%>%
  summarize(totals=n(),groups="drop")%>%

  ggplot(aes(x=weekday,y=totals,fill=rideable_type))+
  geom_col(position = "dodge")+
  facet_wrap(~member_casual)+
  labs(title = "Bike Type Usage by User Type during a week",x="User type",y=NULL,caption = )+
  theme_minimal() +
  theme(legend.position = "top")
```



In the figure, for casual riders the classic bike usage is very high during weekends and along with electric bike usage. For members, classic bike usage is more on Wednesday and marginally less on Sunday and Electric bike usage is less on Sunday and more on Tuesday. Overall, for both charts classic bike usage is more in members as well as for electric bike usage.

Top 10 Start Stations for Members and Casuals

Top 10 Start Stations for Members

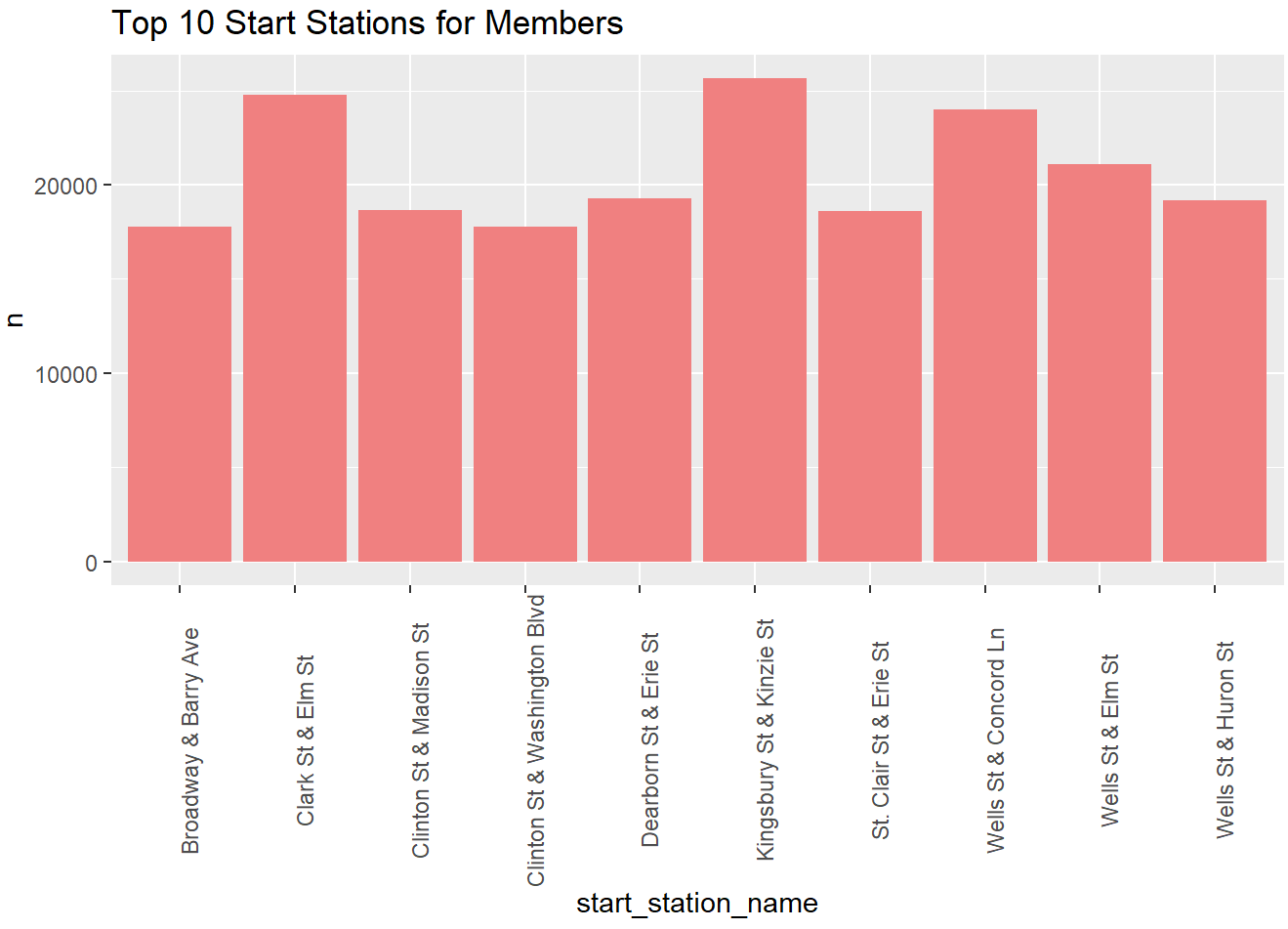
```
##First, analysis to find the top 10 stations
num_station <- df %>%
  group_by(start_station_name) %>%
  count(member_casual,sort = (decreasing = TRUE))

num_station <- data.frame(num_station)

top_member <- subset(num_station, member_casual == "member") %>%
  top_n (10)

top_member
```

start_station_name <chr>	member_casual <chr>	n <int>
Kingsbury St & Kinzie St	member	25634
Clark St & Elm St	member	24772
Wells St & Concord Ln	member	24006

start_station_name	member_casual	n
<chr>	<chr>	<int>
Wells St & Elm St	member	21087
Dearborn St & Erie St	member	19277
Wells St & Huron St	member	19162
Clinton St & Madison St	member	18651
St. Clair St & Erie St	member	18625
Clinton St & Washington Blvd	member	17774
Broadway & Barry Ave	member	17770

1-10 of 10 rows

```
##Plotting the top 10 start stations for members
ggplot(data=top_member)+
  geom_bar(stat="identity",mapping = aes(start_station_name, y=n), fill ="lightcoral")+
  theme(axis.text.x= element_text(angle = 90))+
  labs(title= "Top 10 Start Stations for Members")
```



By figuring, out the top 10 start stations for members by analysis, the further plot was formed.

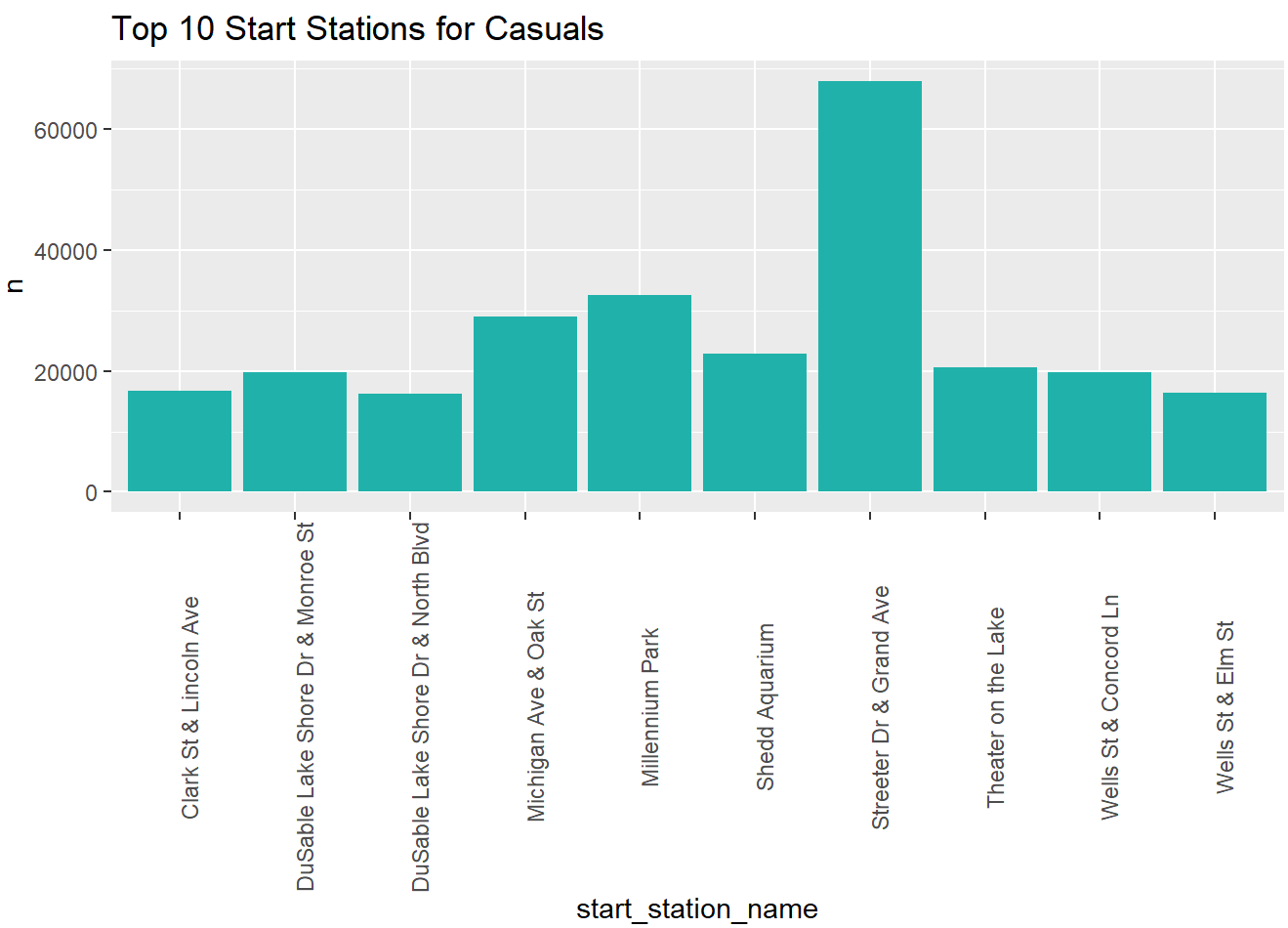
Top 10 Start Stations for Casuals

```
top_casual <- subset(num_station, member_casual == "casual") %>%
  top_n(10)
top_casual
```

start_station_name <chr>	member_casual <chr>	n <int>
Streeter Dr & Grand Ave	casual	67961
Millennium Park	casual	32565
Michigan Ave & Oak St	casual	29016
Shedd Aquarium	casual	22917
Theater on the Lake	casual	20607
DuSable Lake Shore Dr & Monroe St	casual	19700
Wells St & Concord Ln	casual	19696
Clark St & Lincoln Ave	casual	16654
Wells St & Elm St	casual	16383
DuSable Lake Shore Dr & North Blvd	casual	16286

1-10 of 10 rows

```
##Plotting the first 10 start stations for casuals
ggplot(data=top_casual)+
  geom_bar(stat="identity",mapping = aes(start_station_name, y=n), fill ="lightseagreen")+
  theme(axis.text.x= element_text(angle = 90))+
  labs(title= "Top 10 Start Stations for Casuals")
```



By figuring, out the top 10 start stations by analysis, the further plot was formed.

Members and casual riders also differ in the stations that are popular for starting their rides.

5. Share

This is the detailed documentation of R code

6. Act

Recommendations

Based on an analysis of the data, the following recommendations can be made to the Cyclistic stakeholders.

1. The marketing campaign should be targeted at the popular start and end stations for casual riders by providing discounts and certain benefits in order for them to enroll for membership.

2. To encourage casual users to enroll, weekends specials should be introduced which would cut down the cost of memberships to some extend as weekends is the busiest for casuals and more importantly the discounts should be applied more during the busiest time of the month for casuals that is October.

3. Introduce a “reservation” system allowing members to schedule a bike ahead of time during busy periods (e.g. weekends or October), effectively being able to skip the line during high-demand times.