CYCLISTIC CASE STUDY

Kishren Jay

Introduction

The director of Cyclistic, a bike-share company is convinced that the company's future growth depends on maximising the number of annual memberships. Hence, the analytics team wants to understand the different usage trends of Cyclistic bikes from subscribed annual members and casual riders. From the analysis, a new marketing strategy can be implemented to increase the conversion rate of casual riders into subscribed annual members.

Business task

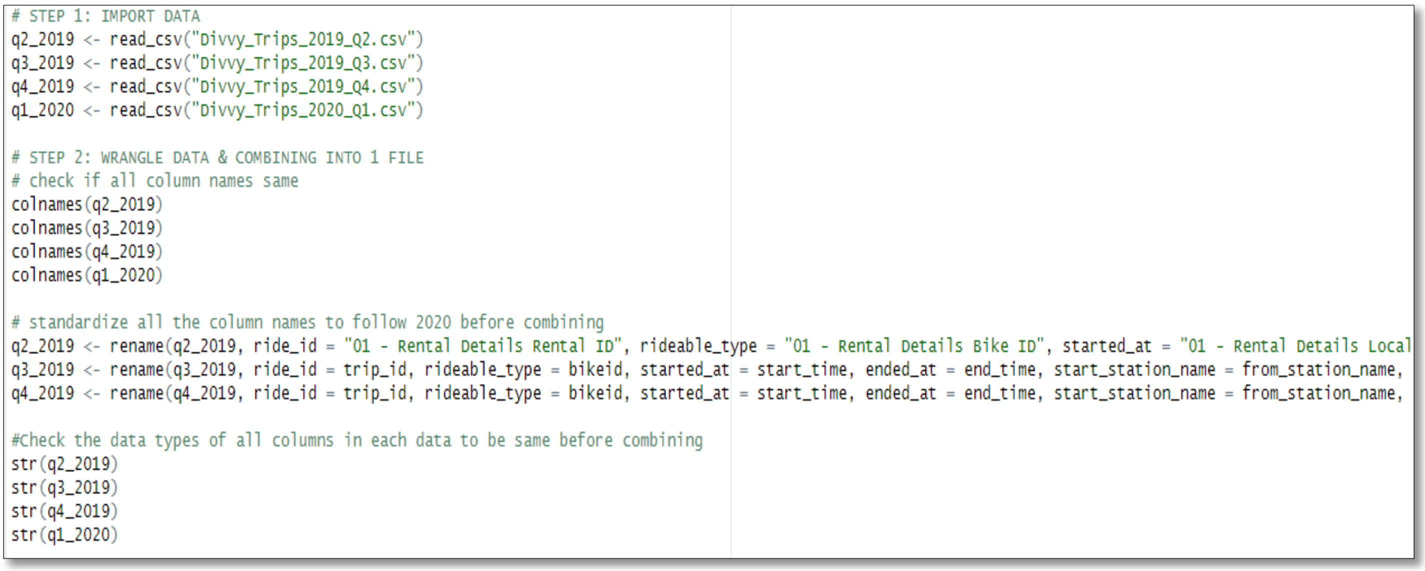
Analyse the Cyclistic data set from the 2nd Quarter of 2019 to the 1st Quarter of 2020 to understand how the various usage trends by subscribed annual members and casual riders.

Preparing the Data

- The **Ride Divvy Bike Data** that was used in this study was provided from Cyclistic management and licensed from https://ride.divvybikes.com/data-license-agreement.
- The dataset contained 10 CSV files, one file for the first quarter of 2020 data and nine files for each of the other months of the year 2019.
- Files that were selected for the analysis:
    - Divvy_Trips_2019_Q2.csv
    - Divvy_Trips_2019_Q3.csv
    - Divvy_Trips_2019_Q4.csv
    - Divvy_Trips_2020_Q1.csv

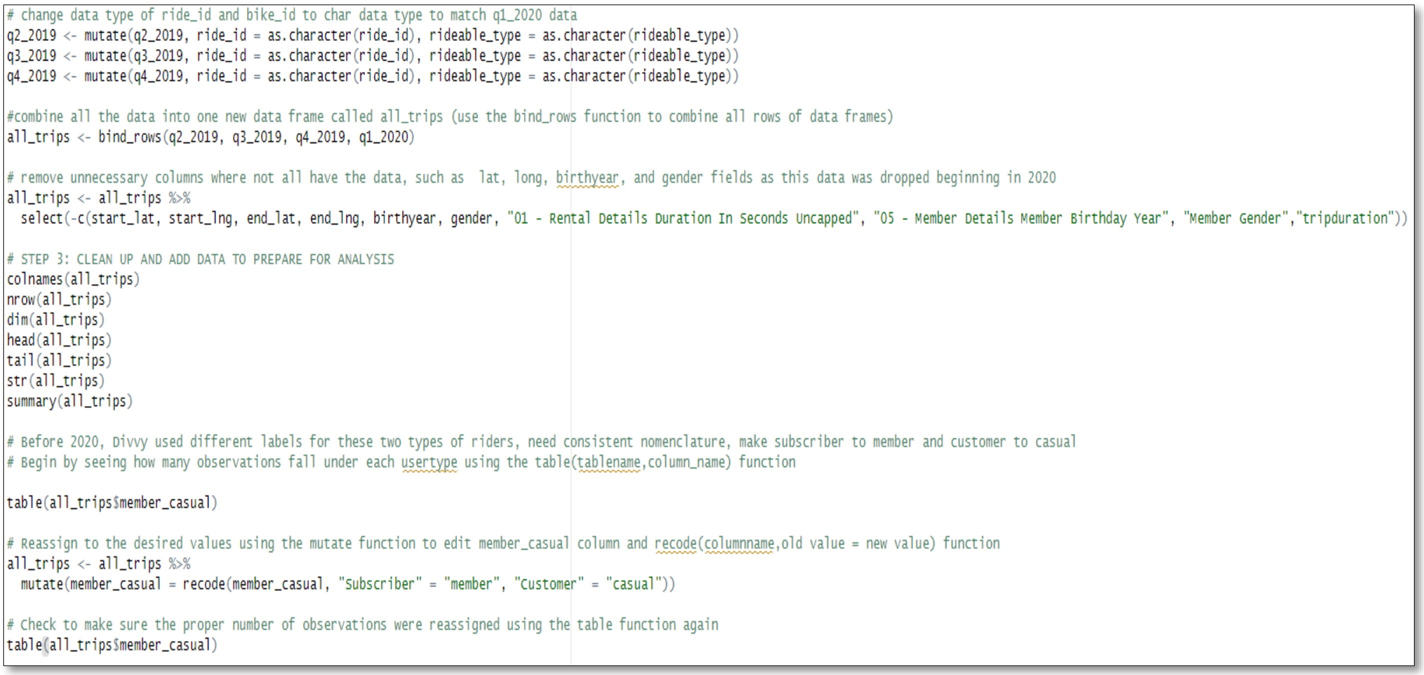
Processing the Data

- **Tool** : **R Studio**
  Reason : R programming application was able handle the large data amount which were millions of rows while providing easy data cleaning, merging and visualisation for the analysis.
- **Data Cleaning**
    - Checked if all column names and datatypes in all the 4 datasets are same before combining using the **colnames()** and **str()** functions.
    - Standardizing the column names in all files according to **Divvy_Trips_2020_Q1.csv** using **rename().**



```
# STEP 1: IMPORT DATA
q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
q1_2020 <- read_csv("Divvy_Trips_2020_Q1.csv")

# STEP 2: WRANGLE DATA & COMBINING INTO 1 FILE
# check if all column names same
colnames(q2_2019)
colnames(q3_2019)
colnames(q4_2019)
colnames(q1_2020)

# standardize all the column names to follow 2020 before combining
q2_2019 <- rename(q2_2019, ride_id = "01 - Rental Details Rental ID", rideable_type = "01 - Rental Details Bike ID", started_at = "01 - Rental Details Local
q3_2019 <- rename(q3_2019, ride_id = trip_id, rideable_type = bikeid, started_at = start_time, ended_at = end_time, start_station_name = from_station_name,
q4_2019 <- rename(q4_2019, ride_id = trip_id, rideable_type = bikeid, started_at = start_time, ended_at = end_time, start_station_name = from_station_name,

#Check the data types of all columns in each data to be same before combining
str(q2_2019)
str(q3_2019)
str(q4_2019)
str(q1_2020)
```

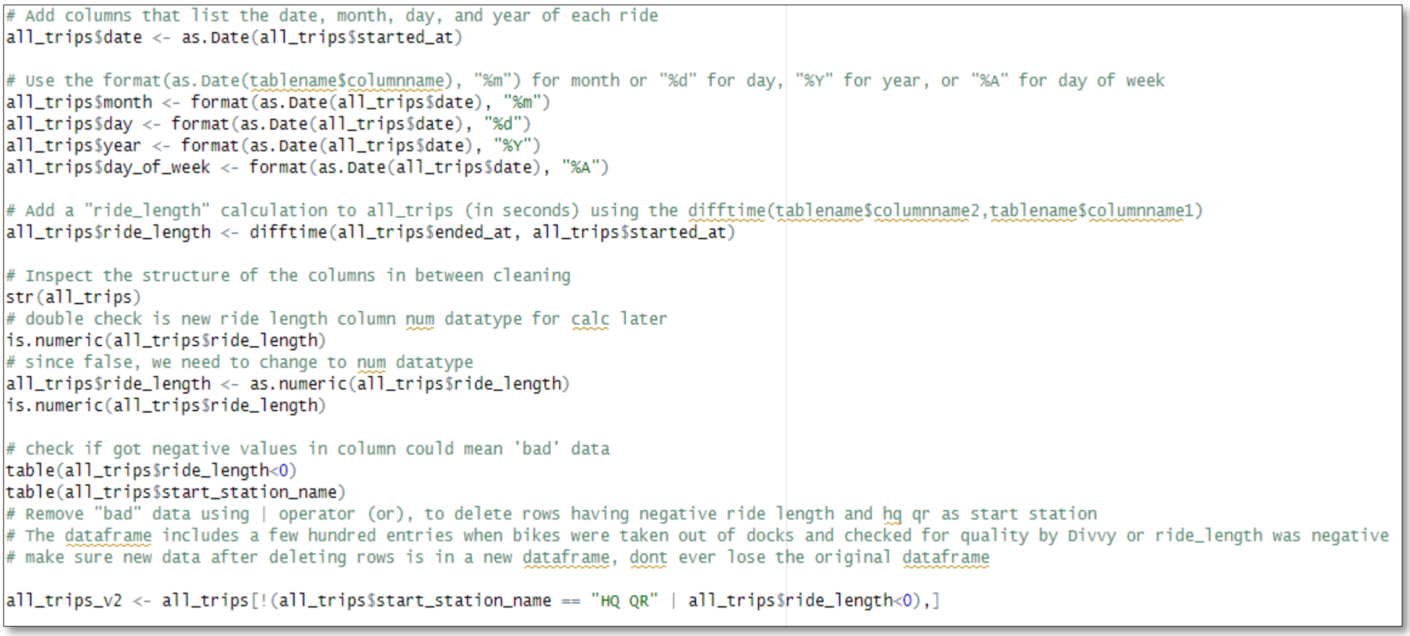
- Converted the data type of **ride_id** and **bike_id** to character data type using **mutate()** and **as.character()**.
- Combined data from all files into one dataset called **all_trips** using **bind_rows()**.
- Removed unnecessary columns where not all have the data, such as **lat**, **long**, **birthyear**, and **gender** fields as this data was dropped beginning in 2020.
- Standardized member_casual column data whereby **subscriber = member** and **customer = casual** using **mutate()** and **recode()**.



```
# change data type of ride_id and bike_id to char data type to match q1_2020 data
q2_2019 <- mutate(q2_2019, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
q3_2019 <- mutate(q3_2019, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
q4_2019 <- mutate(q4_2019, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))

#combine all the data into one new data frame called all_trips (use the bind_rows function to combine all rows of data frames)
all_trips <- bind_rows(q2_2019, q3_2019, q4_2019, q1_2020)

# remove unnecessary columns where not all have the data, such as  lat, long, birthyear, and gender fields as this data was dropped beginning in 2020
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, birthyear, gender, "01 - Rental Details Duration In Seconds Uncapped", "05 - Member Details Member Birthday Year", "Member Gender","tripduration"))

# STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS
colnames(all_trips)
nrow(all_trips)
dim(all_trips)
head(all_trips)
tail(all_trips)
str(all_trips)
summary(all_trips)

# Before 2020, Divvy used different labels for these two types of riders, need consistent nomenclature, make subscriber to member and customer to casual
# Begin by seeing how many observations fall under each usertype using the table(tablename,column_name) function

table(all_trips$member_casual)

# Reassign to the desired values using the mutate function to edit member_casual column and recode(columnname,old value = new value) function
all_trips <- all_trips %>%
  mutate(member_casual = recode(member_casual, "Subscriber" = "member", "Customer" = "casual"))

# Check to make sure the proper number of observations were reassigned using the table function again
table(all_trips$member_casual)
```

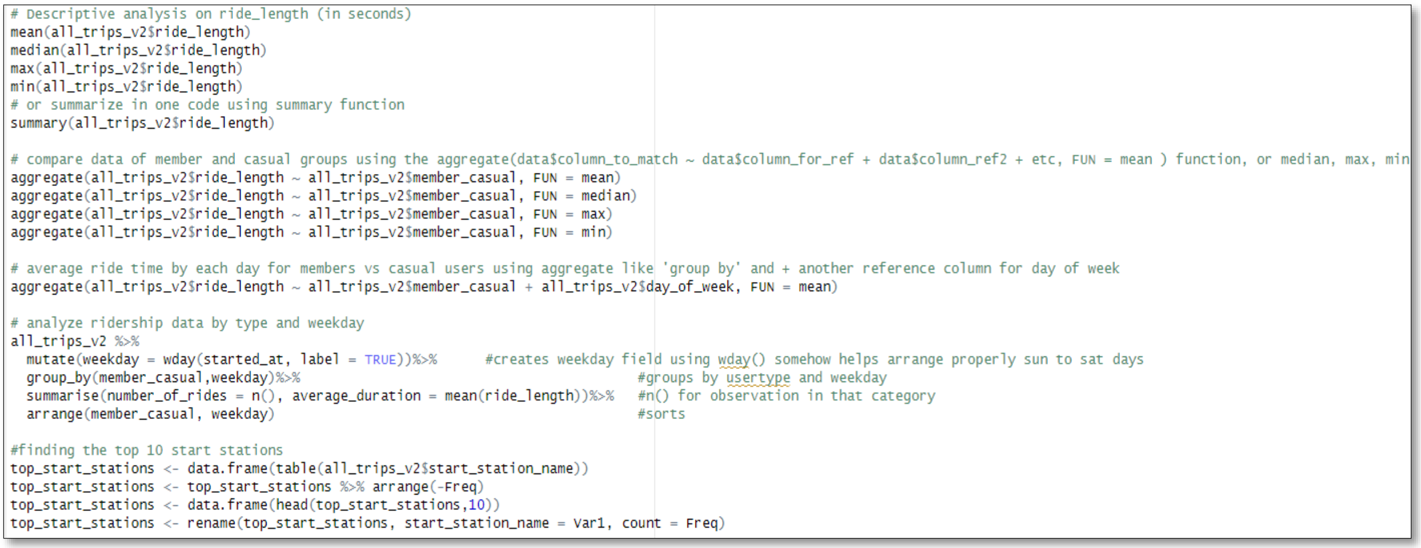
➢ Data Processing

- Added columns that list the date, month, day, and year for each ride using **as.Date()**.
- Added a **ride_length** calculation to **all_trips** (in seconds) using **difftime()** of start and end time of rides.
- Checked and converted data in ride_length column to numeric type using **as.numeric()**.
- Created new dataframe called **all_trips_v2** to store all the data from every column previous in **all_trips** excluding where **station_name** was **"HQ QR"** or **ride_length<0**.



```
# Add columns that list the date, month, day, and year of each ride
all_trips$date <- as.Date(all_trips$started_at)

# Use the format(as.Date(tablename$columnname), "%m") for month or "%d" for day, "%Y" for year, or "%A" for day of week
all_trips$month <- format(as.Date(all_trips$date), "%m")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year <- format(as.Date(all_trips$date), "%Y")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")

# Add a "ride_length" calculation to all_trips (in seconds) using the difftime(tablename$columnname2,tablename$columnname1)
all_trips$ride_length <- difftime(all_trips$ended_at, all_trips$started_at)

# Inspect the structure of the columns in between cleaning
str(all_trips)
# double check is new ride length column num datatype for calc later
is.numeric(all_trips$ride_length)
# since false, we need to change to num datatype
all_trips$ride_length <- as.numeric(all_trips$ride_length)
is.numeric(all_trips$ride_length)

# check if got negative values in column could mean 'bad' data
table(all_trips$ride_length<0)
table(all_trips$start_station_name)
# Remove "bad" data using | operator (or), to delete rows having negative ride length and hq qr as start station
# The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or ride_length was negative
# make sure new data after deleting rows is in a new dataframe, dont ever lose the original dataframe

all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<0),]
```

- Calculated **number_of_rides** and **average_duration** of **ride_length** and arranged according to **member_casual** and **day**.
- Created new data frame consisting of **top 10 start station names** where rides are started from.

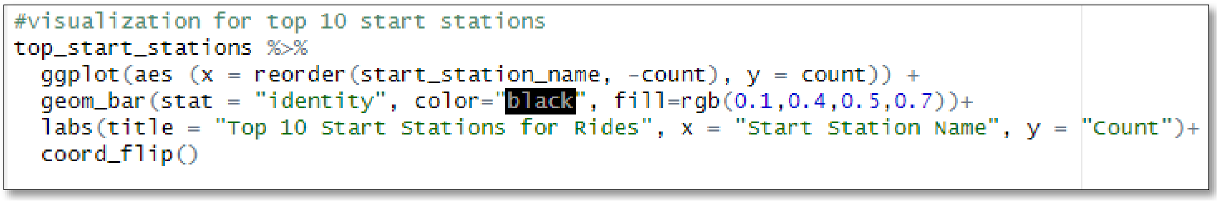


```
# Descriptive analysis on ride_length (in seconds)
mean(all_trips_v2$ride_length)
median(all_trips_v2$ride_length)
max(all_trips_v2$ride_length)
min(all_trips_v2$ride_length)
# or summarize in one code using summary function
summary(all_trips_v2$ride_length)

# compare data of member and casual groups using the aggregate(data$column_to_match ~ data$column_for_ref + data$column_ref2 + etc, FUN = mean ) function, or median, max, min
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)

# average ride time by each day for members vs casual users using aggregate like 'group by' and + another reference column for day of week
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)

# analyze ridership data by type and weekday
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE))%>%      #creates weekday field using wday() somehow helps arrange properly sun to sat days
  group_by(member_casual,weekday)%>%                       #groups by usertype and weekday
  summarise(number_of_rides = n(), average_duration = mean(ride_length))%>%   #n() for observation in that category
  arrange(member_casual, weekday)                          #sorts

#finding the top 10 start stations
top_start_stations <- data.frame(table(all_trips_v2$start_station_name))
top_start_stations <- top_start_stations %>% arrange(-Freq)
top_start_stations <- data.frame(head(top_start_stations,10))
top_start_stations <- rename(top_start_stations, start_station_name = Var1, count = Freq)
```

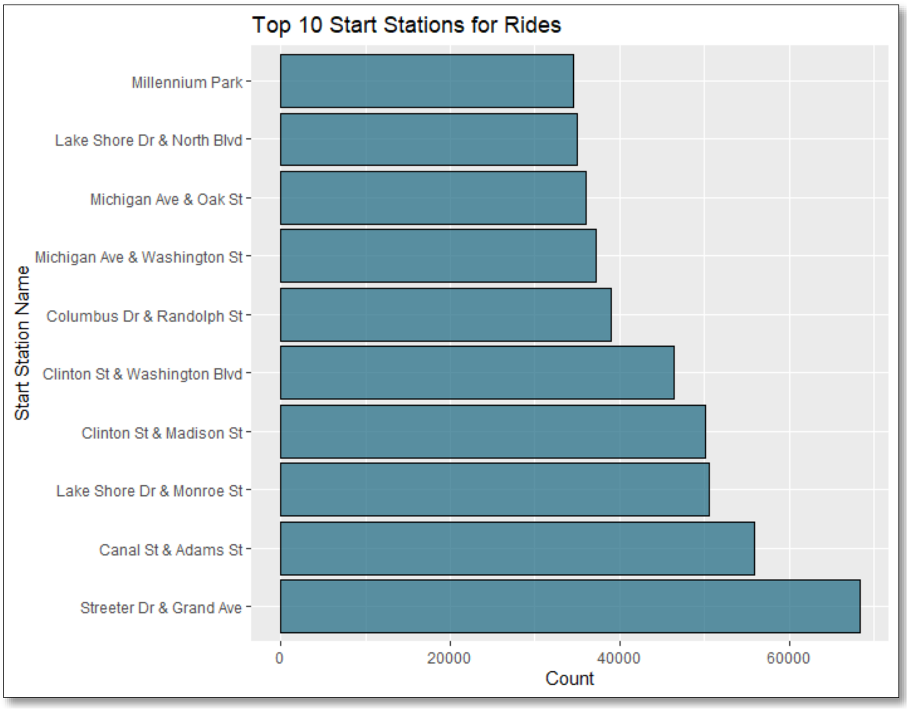
Data Insights & Visualisation

From the datasets, several hypotheses were initially made whereby there are relationships between:

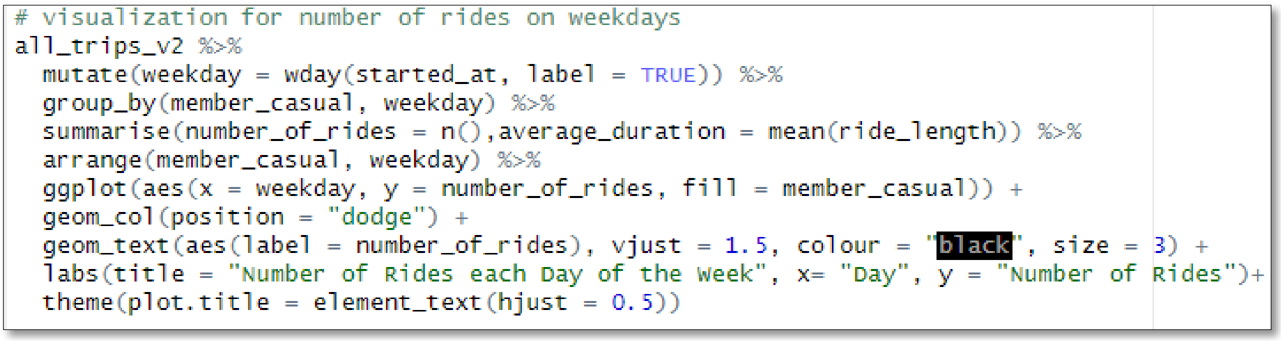
- From the 642 stations, certain stations are located nearer or more conveniently for certain members and casual riders.
- Members and casual users use the bike more on different times of the day.
- Duration of rides vary depending on the type of user.



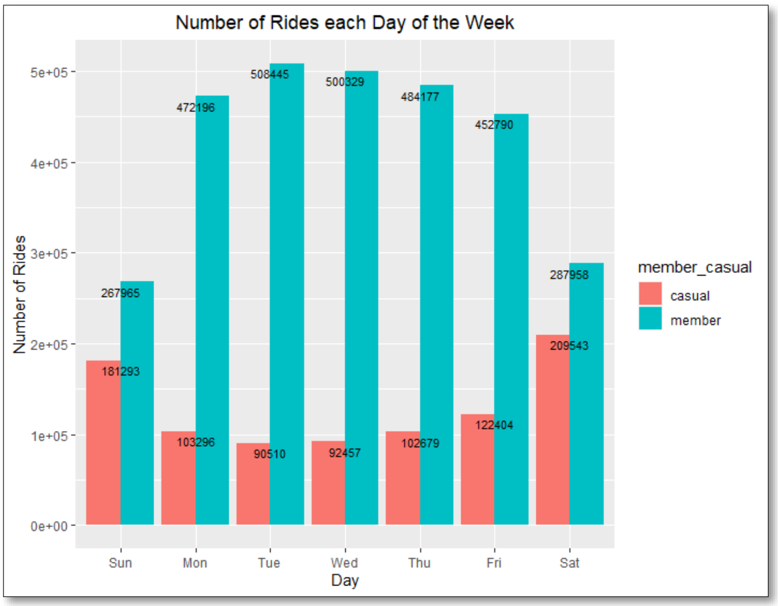
```
#visualization for top 10 start stations
top_start_stations %>%
  ggplot(aes (x = reorder(start_station_name, -count), y = count)) +
  geom_bar(stat = "identity", color="black", fill=rgb(0.1,0.4,0.5,0.7))+
  labs(title = "Top 10 Start Stations for Rides", x = "Start Station Name", y = "Count")+
  coord_flip()
```



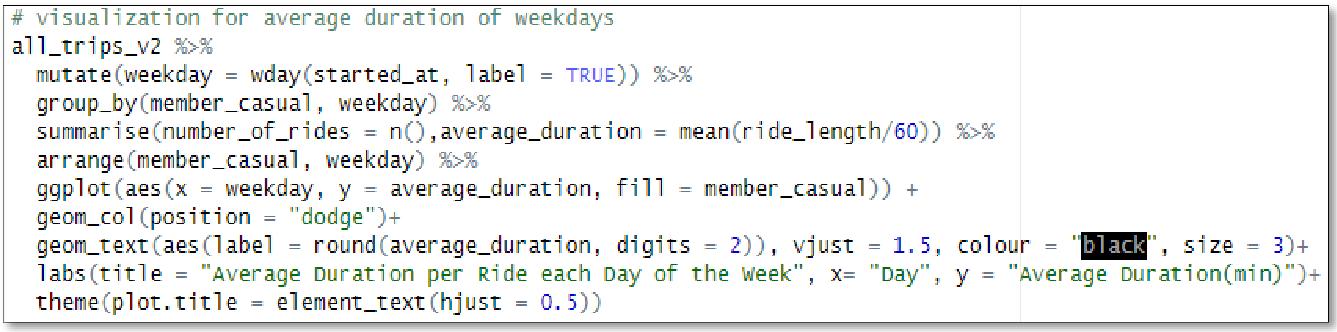
The station which most bikes depart from is Streeter Dr & Grand Ave, followed by Canal St & Adams St and others just as the bar plot above indicates. Among the 642 stations, most riders tend to use bikes from these 10 stations during the 2nd Quarter of 2019 till the 1st Quarter of 2020.



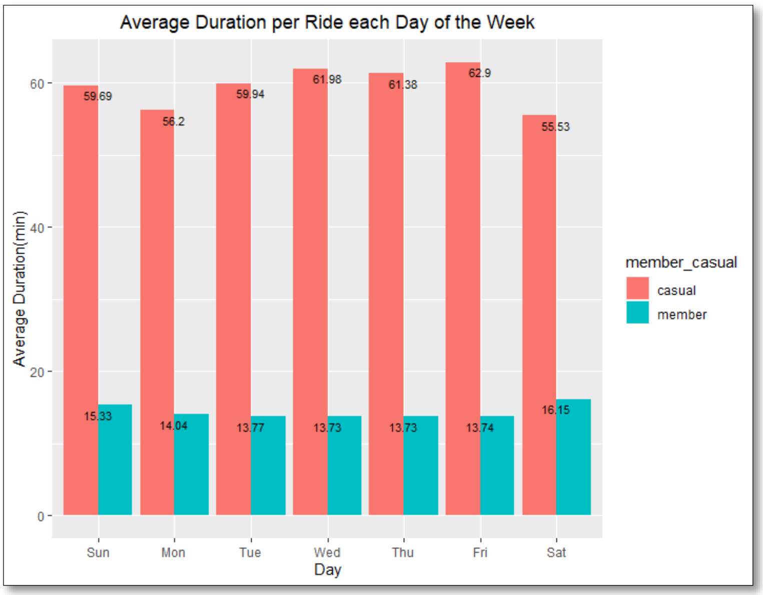
```
# visualization for number of rides on weekdays
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = number_of_rides), vjust = 1.5, colour = "black", size = 3) +
  labs(title = "Number of Rides each Day of the Week", x= "Day", y = "Number of Rides")+
  theme(plot.title = element_text(hjust = 0.5))
```



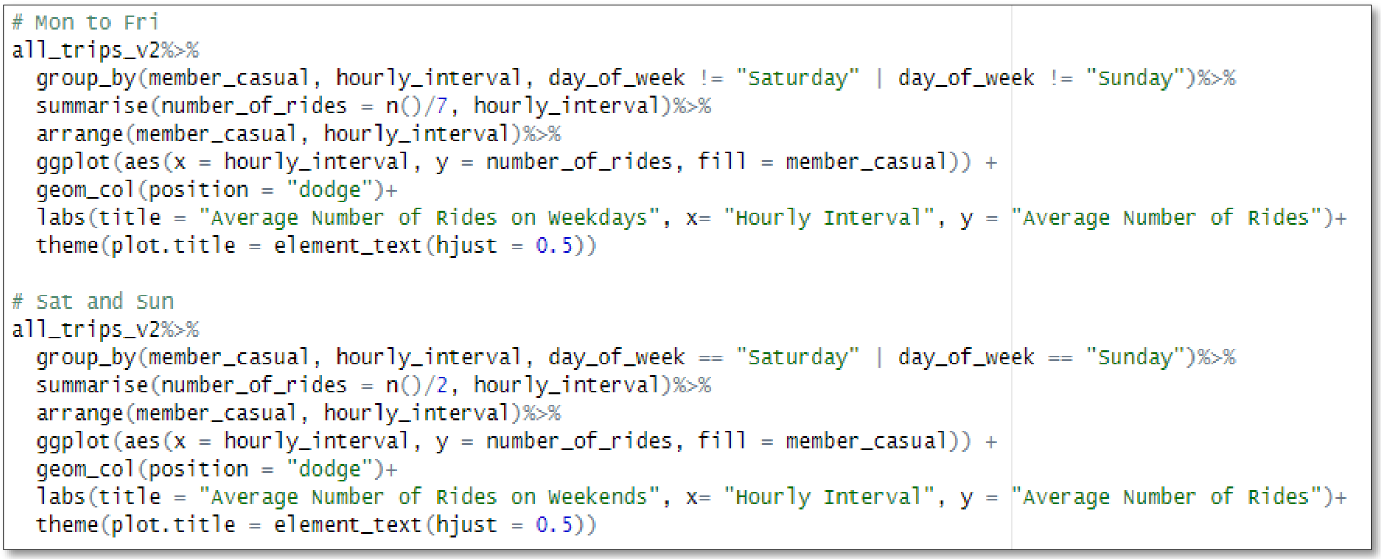
The visualisation above shows the number of rides on different days on the week by members and casual riders. For members, the peak number of rides occurs on Tuesday with 508445 rides followed by Wednesday, Thursday, Tuesday, Friday, Saturday and Sunday. As for casual users, Saturdays tend to be the day with the greatest number of rides which is about 209543 rides followed by Sunday with 181293 rides. Members tend to use the bike more on weekdays while casual riders utilize the bike more on weekends.



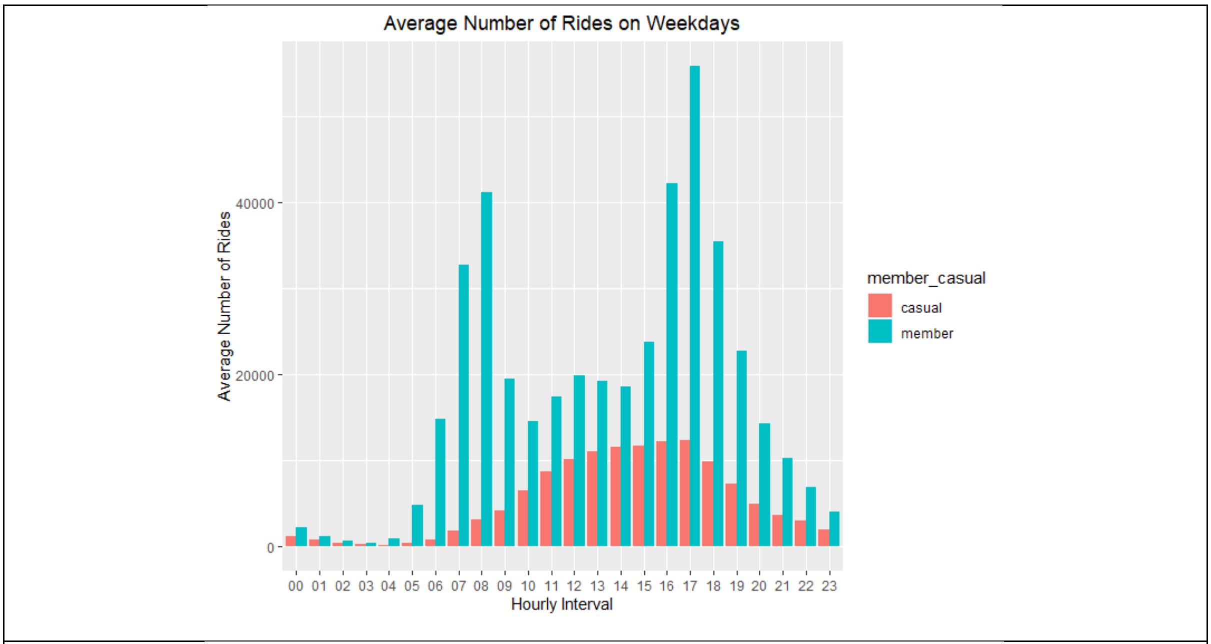
```
# visualization for average duration of weekdays
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),average_duration = mean(ride_length/60)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")+
  geom_text(aes(label = round(average_duration, digits = 2)), vjust = 1.5, colour = "black", size = 3)+
  labs(title = "Average Duration per Ride each Day of the Week", x= "Day", y = "Average Duration(min)")+
  theme(plot.title = element_text(hjust = 0.5))
```

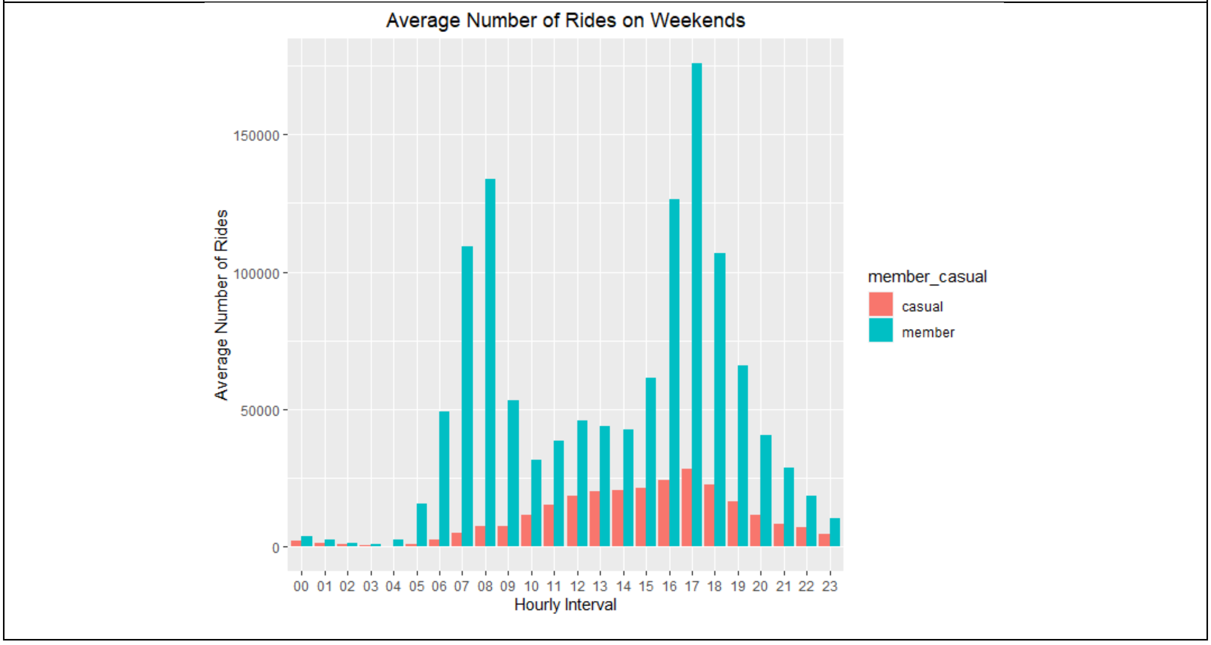


The plot above shows the average duration of each ride on different days of the week. From the visualisation, it is clear that casual users tend to spend much more time on the bike as compared to members as the data indicates that casual users spend on average 3.5 to 4 times the duration on the bike compared to members.



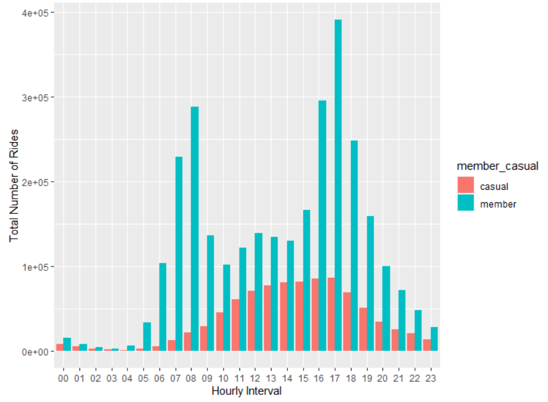
```
# Mon to Fri
all_trips_v2%>%
  group_by(member_casual, hourly_interval, day_of_week != "Saturday" | day_of_week != "Sunday")%>%
  summarise(number_of_rides = n()/7, hourly_interval)%>%
  arrange(member_casual, hourly_interval)%>%
  ggplot(aes(x = hourly_interval, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(title = "Average Number of Rides on Weekdays", x= "Hourly Interval", y = "Average Number of Rides")+
  theme(plot.title = element_text(hjust = 0.5))

# Sat and Sun
all_trips_v2%>%
  group_by(member_casual, hourly_interval, day_of_week == "Saturday" | day_of_week == "Sunday")%>%
  summarise(number_of_rides = n()/2, hourly_interval)%>%
  arrange(member_casual, hourly_interval)%>%
  ggplot(aes(x = hourly_interval, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(title = "Average Number of Rides on Weekends", x= "Hourly Interval", y = "Average Number of Rides")+
  theme(plot.title = element_text(hjust = 0.5))
```



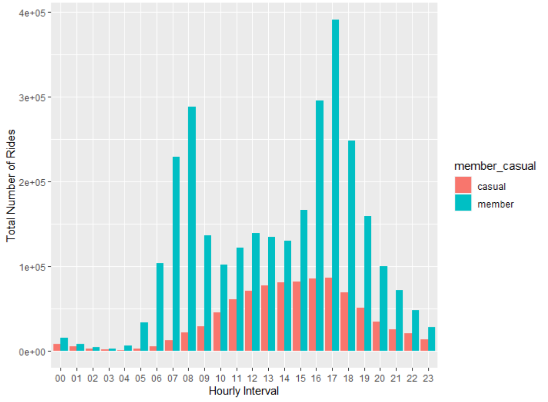


The overall comparison for the average number of rides on weekdays and weekends can be seen from the above 2 bar plots. On weekends, the average rides throughout the day are more than on the weekdays for both members and casuals. Meanwhile, the trend of usage of the bikes throughout the have almost a similar pattern on weekends and weekdays for both members and casuals. For members, peak hours or pre and post work hours (7am to 9am & 4pm to 7pm) are when the number of rides begin to drastically increase. As for casual riders, non-peak hours or working hours tend to be favoured as the number or rides are increase from 10am to 5pm and then the usage declines.

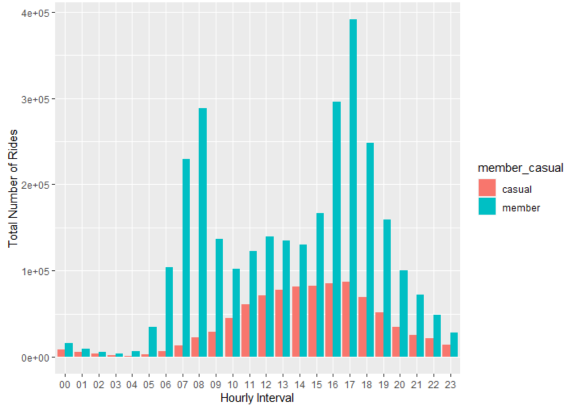
Total Number of Rides on Mondays



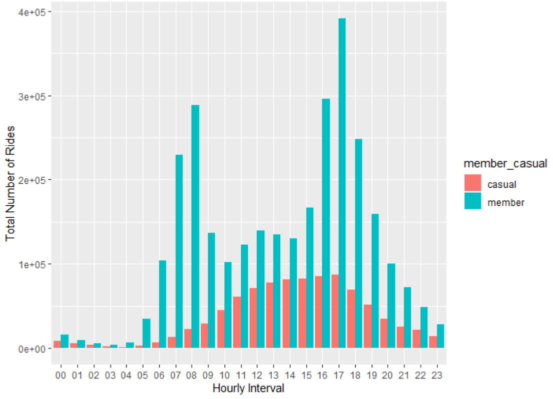
Total Number of Rides on Tuesdays



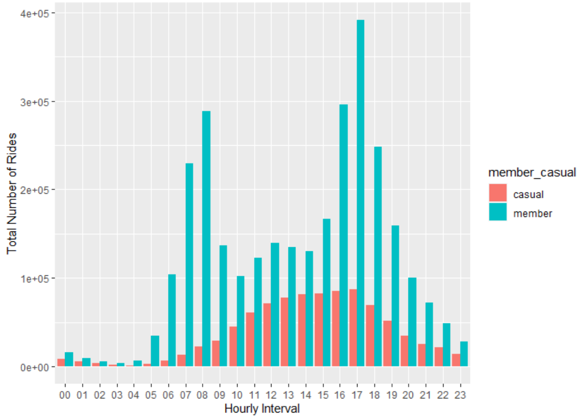
Total Number of Rides on Wednesdays



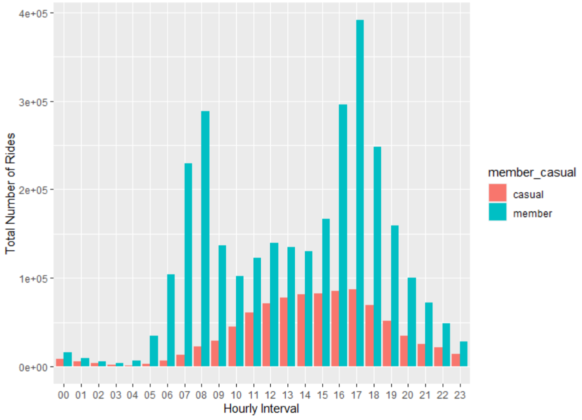
Total Number of Rides on Thursdays



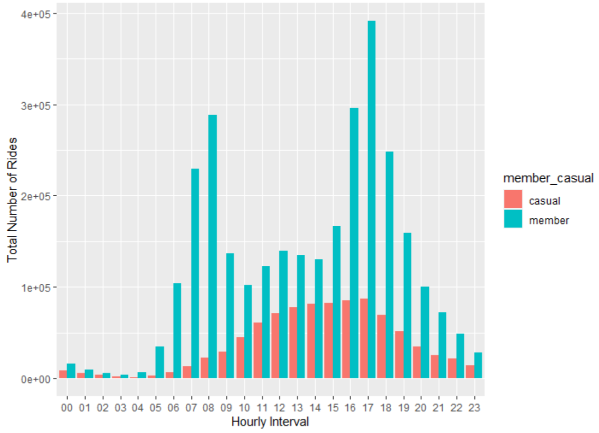
Total Number of Rides on Fridays



Total Number of Rides on Saturdays



Total Number of Rides on Sundays



The individual day plots for number of rides each day warrants the similar pattern of usage and reinforces the previous weekday and weekend plots.

Conclusion

From the analysis on the available ride data, the main points discovered were:

- The top 10 stations out of the 642 stations starting with Streeter Dr & Grand Ave are the stations closest and most convenient for the riders.
- Members tend to use the bike more on weekdays while casual riders utilize the bike more on weekends.
- Casual users tend to spend on average 3.5 to 4 times the duration per ride on the bike compared to members.
- For casual riders, non-peak hours or working hours tend to be favoured as the number or rides are increase from 10am to 5pm and then the usage declines.

Recommendations

In order for Cyclistic to grow and thrive, the goal to increase the conversion rate of casual users to annual members must be achieved by utilizing the analysis on the ride data and implementing suggested recommendations based off the analysis such as:

- Focus on promotional and advertising activity about the perks of becoming a member in the top 10 stations that users tend to start their rides. This method allows efficient use of resources and a higher exposure to users.
- Increase the rate of renting a bike on weekends if the user is not a member and consider giving discounts for members on weekends to attract the casual users that use the bikes more weekends.
- Introduce a point system for using the bikes for longer periods of time which can later be exchanged for prizes or more ride time in the future after a minimum duration of being a member is achieved by the user.
- Conduct seasonal promotions during non-peak/working hours with a condition that the user has to be a member to encourage the users that belong in the younger or older age groups that do not work and rather use the bikes during these times of the day.