# Output

Lim Wei Hern

11/21/2021

Data used in this project can be downloaded from here

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(readr)
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library("here")

## here() starts at C:/Users/limwe/OneDrive/Desktop/Case Study 1

library("skimr")
library("janitor")

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

nov20 <- read_csv("202011-divvy-tripdata.csv")

## Rows: 259716 Columns: 13
## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
dec20 <- read_csv("202012-divvy-tripdata.csv")
```

```
## Rows: 131573 Columns: 13

## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
jan21 <- read_csv("202101-divvy-tripdata.csv")
```

```
## Rows: 96834 Columns: 13

## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
feb21 <- read_csv("202102-divvy-tripdata.csv")
```

```
## Rows: 49622 Columns: 13
## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

mar21 <- read_csv("202103-divvy-tripdata.csv")

## Rows: 228496 Columns: 13

## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

apr21 <- read_csv("202104-divvy-tripdata.csv")

## Rows: 337230 Columns: 13

## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

may21 <- read_csv("202105-divvy-tripdata.csv")

## Rows: 531633 Columns: 13
## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
jun21 <- read_csv("202106-divvy-tripdata.csv")
```

```
## Rows: 729595 Columns: 13

## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
jul21 <- read_csv("202107-divvy-tripdata.csv")
```

```
## Rows: 822410 Columns: 13

## -- Column specification ---------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
aug21 <- read_csv("202108-divvy-tripdata.csv")
```

```
## Rows: 804352 Columns: 13
## -- Column specification -----------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
sep21 <- read_csv("202109-divvy-tripdata.csv")
```

```
## Rows: 756147 Columns: 13
```

```
## -- Column specification -----------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
oct21 <- read_csv("202110-divvy-tripdata.csv")
```

```
## Rows: 631226 Columns: 13
```

```
## -- Column specification -----------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

## Table of Contents

Deliverables:

- A clear statement of the business task
- A description of all data sources used
- Documentation of any cleaning or manipulation of data
- A summary of your analysis
- Supporting visualizations and key findings
- Your top three recommendations based on your analysis

## Business Task

In this report, I will examine how Cyclistic, a bike-share company in Chicago, can maximize the number of annual memberships. To do so, I will answer the following research questions: *How do annual members and casual riders use Cyclistic bikes differently?* Why casual riders would buy a membership? *How digital media could affect their marketing tactics?

To that end, the report will conclude with suggestions for a new marketing strategy to convert casual riders into annual members.

This report defines customers who purchase single-ride or full-day passes as casual riders while customers who purchase annual memberships are defined as Cyclistic members.

## Prepare and Process the Data

The data used in this report is the previous 12 months of Cyclistic trip data which has been made available by Motivate International Inc. under this license. For privacy purposes, no personally identifiable information was used.

To analyze the data, the following ROCCC framework was used:

- **Reliable:** This data is reliable as there is unlikely to be sampling bias given that I am using the data wholesale from Cyclistic.
- **Original:** This data is considered original as it is downloaded directly from Cyclistic (first-party data).
- **Comprehensive:** The data is comprehensive given that it contains many details for each individual trip.
- **Current:** The data is current as it is the previous twelve months of data.
- **Cited:** The data is first-hand data.

### Structure

Each month's trip data is downloaded individually as a .csv file. Each file contains 13 columns:

*A ride_id to identify each ride individually* rideable_type which identifies the type of bike used. There are 3 options: electric_bike, docked_bike, and classic_bike *start and end date and time: started_at, ended_ at* The starting and ending station names and ID - start_nation_name, start_station_id, end_station_name, end_station_id *Detailed geographical coordinates of the starting and ending stations: start_lat,start_lng, end_lat, end_lng* A boolean value detailing if a ride was by a casual rider or a member

## The Process

After combining the 12 csv files into the tibble "ttm", I created 2 columns: ride_length and day_of_week. I then removed all rows where the trip duration was negative and dropped all rows with NULL values.

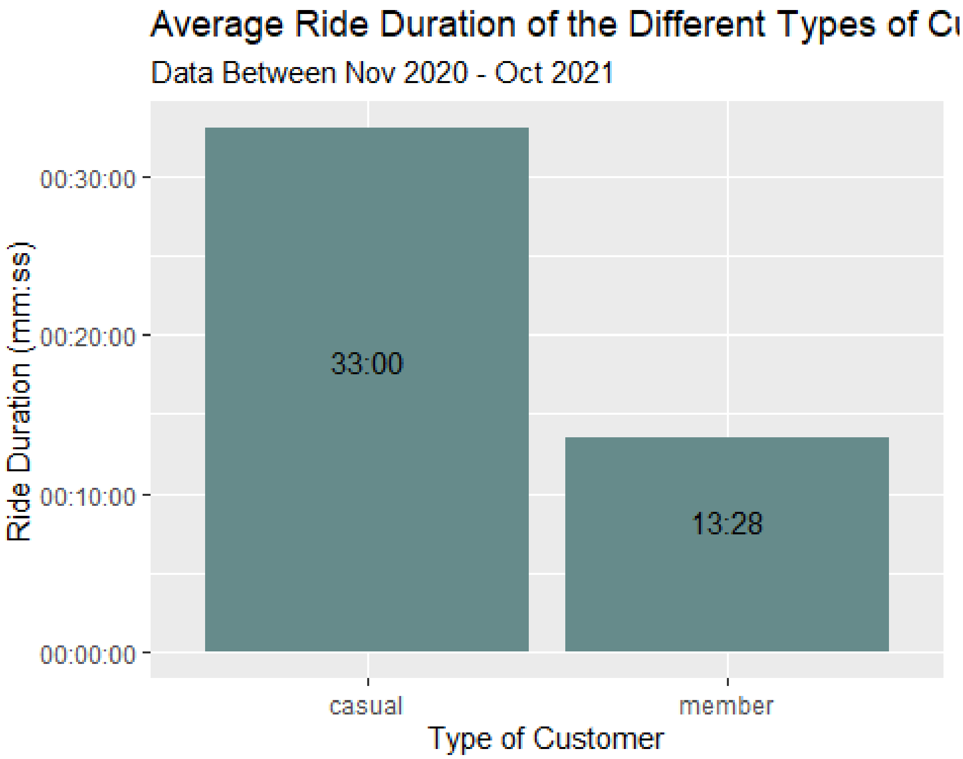
I also created 2 separate tibbles filtered for members and casual for easier reference in the analysis process.

```
ttm <- rbind(nov20, dec20, jan21, feb21, mar21, apr21, may21, jun21, jul21,
aug21,
             sep21, oct21)
rm(nov20, dec20, jan21, feb21, mar21, apr21, may21, jun21, jul21, aug21,
sep21, oct21)
ttm<-ttm %>% mutate(year=year(started_at),month=month(started_at),
day=day(started_at))
ttm <- mutate(ttm, ride_length=ended_at - started_at)
ttm[['ride_length']] <- hms::hms(seconds_to_period(ttm[['ride_length']]))
ttm <- mutate(ttm,day_of_week=weekdays(started_at))
ttm <- ttm %>% filter(ride_length>"0")
ttm <- mutate(ttm, month= month.abb[month(started_at)])
ttm <- ttm %>% drop_na()
member <- ttm %>% filter(member_casual == "member")
casual <- ttm %>% filter(member_casual == "casual")
```

## Analysis

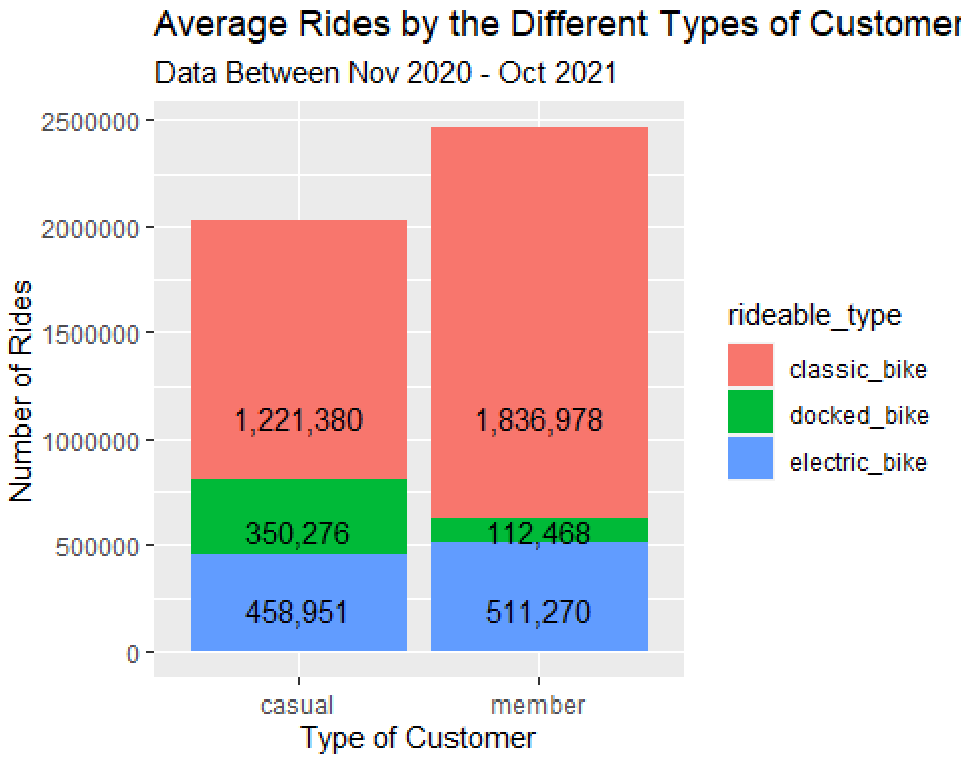
First, lets examine the statistical distribution of bike rides over the past 12 months, comparing between members and casuals.

```
member_stats <- member %>% summarize(mean = mean(ride_length), sd =
sd(ride_length),
                                     number = nrow(member), max =
max(ride_length),
                                     min=min(ride_length), member_casual
="member")
casual_stats <- casual %>% summarize(mean = mean(ride_length), sd =
sd(ride_length),
                                     number = nrow(member),max = 
max(ride_length),
                                     min=min(ride_length), member_casual = 
"casual")
stats <- rbind(member_stats, casual_stats)
rm(member_stats, casual_stats)
stats[["mean"]] <- hms::hms(seconds_to_period(stats[['mean']]))
stats[["max"]] <- hms::hms(seconds_to_period(stats[['max']]))
stats[["min"]] <- hms::hms(seconds_to_period(stats[['min']]))

ggplot(data=ttm, aes(x=member_casual,y=ride_length))+
  geom_bar(stat="summary",fun="mean",fill="paleturquoise4")+
  labs(x="Type of Customer", y="Ride Duration (mm:ss)",
       title = "Average Ride Duration of the Different Types of Customers",
       subtitle ="Data Between Nov 2020 - Oct 2021")+
  annotate("text",x=1,y=1100,label="33:00")+
  annotate("text",x=2,y=500,label="13:28")
```



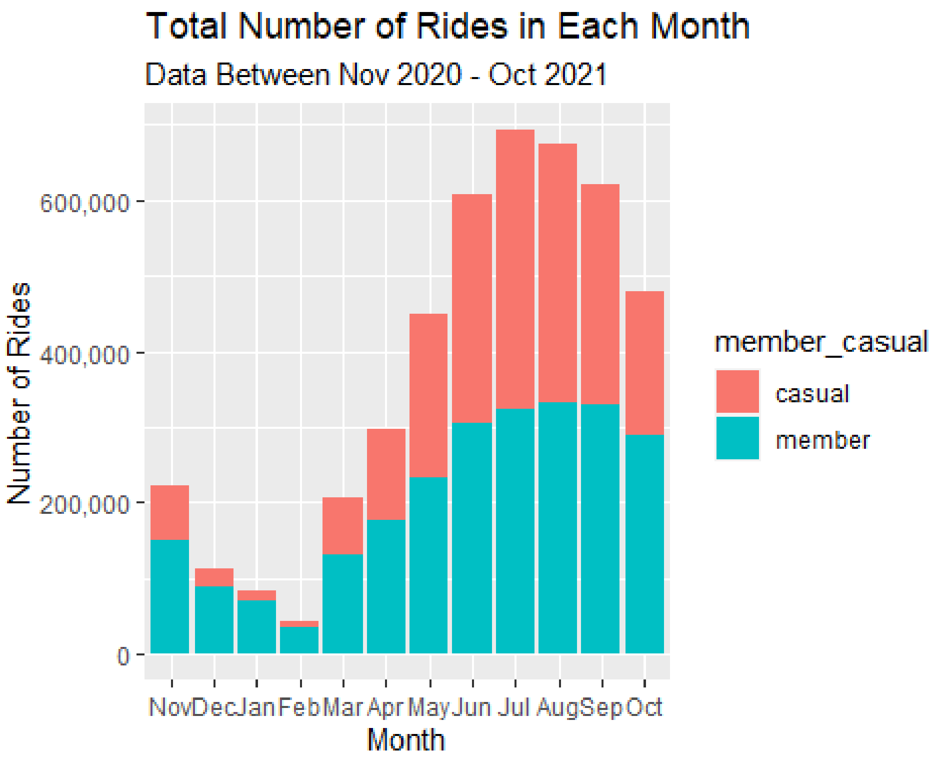
From the above chart, we can see that the average casual customer rides Cyclistic bikes for nearly twice the duration of members. Yet, on an overall percentage, we can see that there are roughly 21% more rides by members than casual riders.

```
ggplot(data=ttm)+
  geom_bar(mapping=aes(x=member_casual, fill=rideable_type))+
  labs(x="Type of Customer", y="Number of Rides",
       title = "Average Rides by the Different Types of Customers",
       subtitle ="Data Between Nov 2020 - Oct 2021")+
  annotate("text",x=1,y=1100000,label="1,221,380")+
  annotate("text",x=1,y=570000,label="350,276")+
  annotate("text",x=1,y=200000,label="458,951")+
  annotate("text",x=2,y=1100000,label="1,836,978")+
  annotate("text",x=2,y=570000,label="112,468")+
  annotate("text",x=2,y=200000,label="511,270")
```



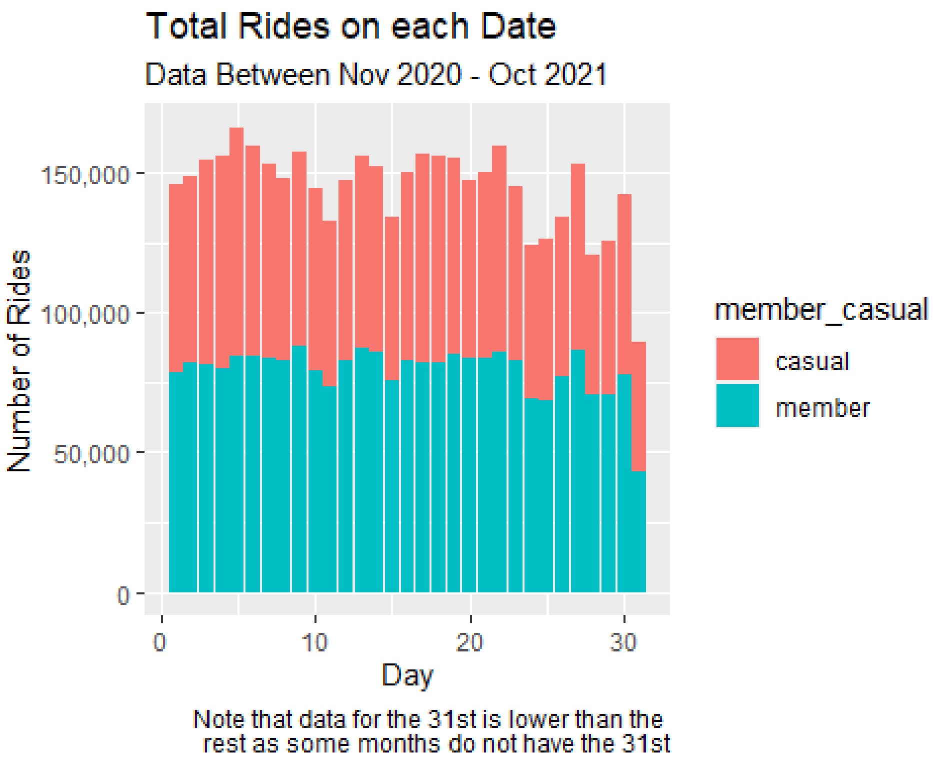
The second step is to understand the temporal distribution of rides using Cyclistic Bikes. From the chart below, there is clear seasonality where the second half of the year saw a dramatic increase in number of rides compared to the months of January and February. Further, the proportion of total rides by casual riders becomes extremely small in the months of January, February and March.

```
ttm$month <-factor(ttm$month,levels=c("Nov","Dec","Jan","Feb","Mar",

"Apr","May","Jun","Jul","Aug","Sep","Oct"))
ggplot(ttm,aes(x=month,fill=member_casual))+
  geom_bar()+
  labs(x="Month", y="Number of Rides", title = "Total Number of Rides in Each
Month",
       subtitle ="Data Between Nov 2020 - Oct 2021")+
  scale_y_continuous(labels= scales::comma)
```



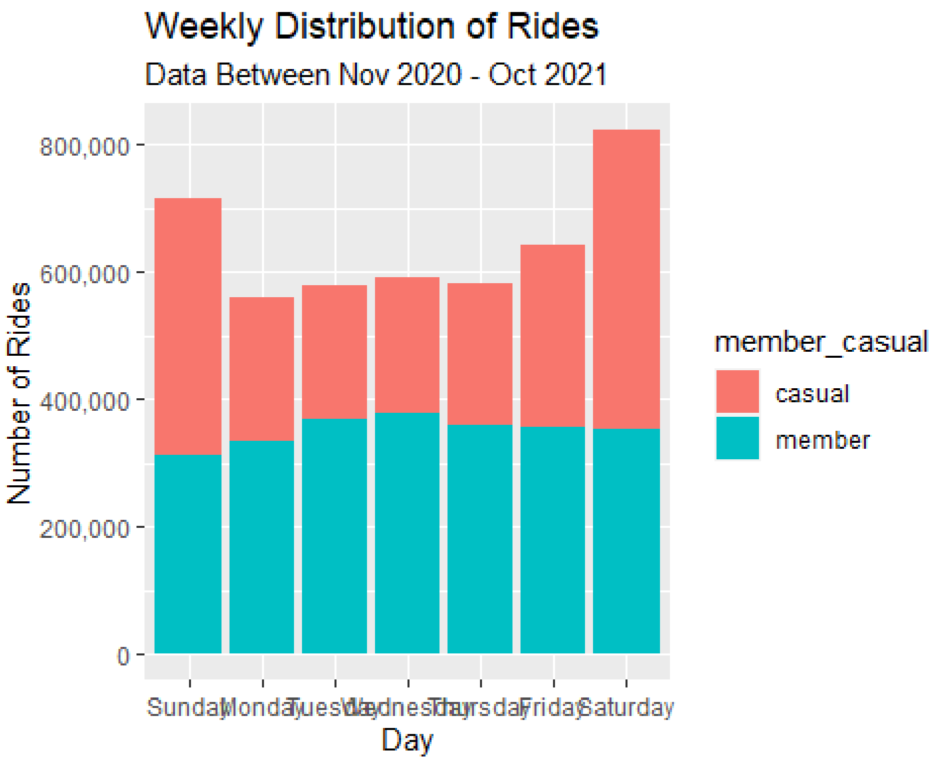
On a monthly timeframe, the chart below shows there is not much difference across individual dates, and the number of rides remain relatively uniform throughout the month.

```
ggplot(ttm,mapping=aes(x=day,fill=member_casual))+
  geom_bar()+
  labs(x="Day", y="Number of Rides", title = "Total Rides on each Date",
       subtitle ="Data Between Nov 2020 - Oct 2021",
       caption="Note that data for the 31st is lower than the
       rest as some months do not have the 31st")+
  scale_y_continuous(labels= scales::comma)
```



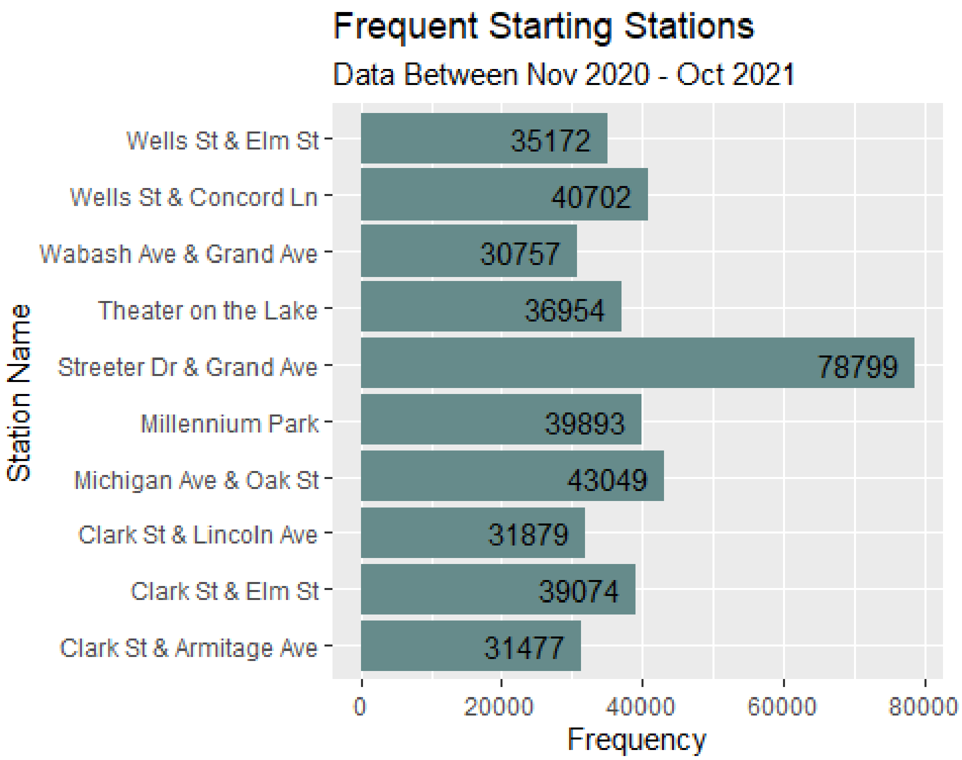
On a weekly time frame, the number of rides on a given day is similarly uniform.

```
ttm$day_of_week <-
factor(ttm$day_of_week,levels=c("Sunday","Monday","Tuesday",

"Wednesday","Thursday","Friday", "Saturday"))
ggplot(ttm,aes(x=day_of_week,fill=member_casual))+
  geom_bar()+
  labs(x="Day", y="Number of Rides", title = "Weekly Distribution of Rides",
       subtitle ="Data Between Nov 2020 - Oct 2021")+
  scale_y_continuous(labels= scales::comma)
```

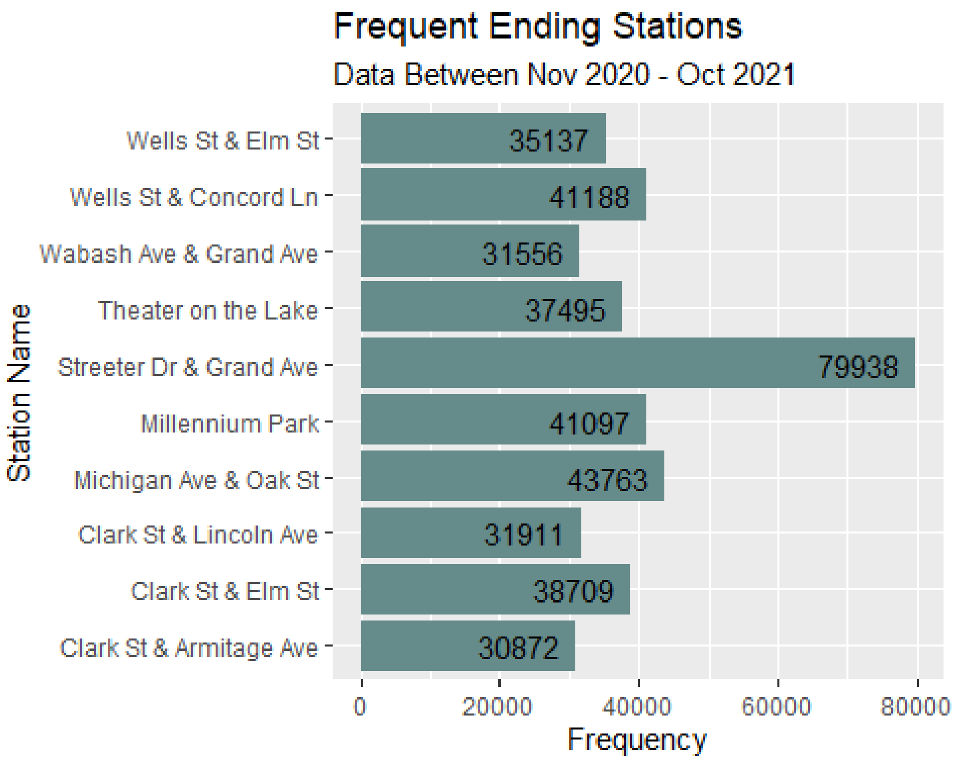


To identify where to target our marketing strategies at consumers, I examined the popularity of the starting and ending stations. The charts below show the top 10 starting and ending stations.

```
start_station_count <- ttm%>% select(start_station_name) %>%
  count(start_station_name,sort =TRUE)
start_station_count <- head(start_station_count,10)
end_station_count <- ttm%>% select(end_station_name) %>%
  count(end_station_name,sort =TRUE)
end_station_count <- head(end_station_count,10)

ggplot(data=start_station_count, aes(x=start_station_name, y =n))+
  geom_col(fill="paleturquoise4")+
  coord_flip()+
  labs(x="Station Name", y= "Frequency", title = "Frequent Starting
Stations",
       subtitle = "Data Between Nov 2020 - Oct 2021")+
  geom_text(aes(label=n),hjust=1.2)
```



```r
ggplot(data=end_station_count, aes(x=end_station_name, y =n))+
  geom_col(fill="paleturquoise4")+
  coord_flip()+
  labs(x="Station Name", y= "Frequency", title = "Frequent Ending Stations",
       subtitle = "Data Between Nov 2020 - Oct 2021")+
  geom_text(aes(label=n),hjust=1.2)
```



As we can see, the top 10 stations are practically the same, regardless if they are starting or ending stations. Of note, Street Dr & Grand Ave Station is by far the most popular, with almost twice the number of trips starting and ending from it compared to any other station.

## Summary of Analysis

### Key Finding 1

The most prominent feature of the rider demographic is that whilst there are more member rides than casual ones, the duration of the casual rides is far longer than that of Cyclistic Members. This implies that members often use Cyclistic for short but often commutes while casual riders use Cyclistic for longer, one-off journeys.

### Key Finding 2

While there is no real trends in the weekly and monthly distribution of rides, there is clear seasonality through the months. It is important to note however that given the COVID-19 Pandemic, the increase in trips in the second half of 2021 is highly likely to be due to the re-opening measures introduced into many cities. This might also explain the extremely small proportion of casual riders during these months as only the members feel compelled to ride to make full use of their membership. Further data would be required to determine if there was seasonality.

### Key Finding 3

Street Dr & Grand Ave Station is by far the most popular station. However, with that being said, it is important to note that the popularity of the other stations are very much equal with Street Dr & Grand Ave Station being the outlier in this dataset.

## Recommendations

### Offering Membership Discounts for Longer Trips

Given that casual riders are more likely to engage in longer trips, it would be advisable to provide discounts for riders who have completed a trip longer than 15:00 minutes. This is longer than the average for members and is thus unlikely to affect existing revenue from members. At the same time, this would incentivise casual riders to buy a membership as they have the most to benefit given their longer trip duration.

### Timed Discounts

To fully incentivise casual riders to buy a membership, a discount for new registering members could be provided in the months towards the end of the year. This is due to the fact that there are more casual riders during this period and would be more likely to see a response. However, it remains important to remember that more data would be required to determine if the seasonality is a trend or a by-product of the pandemic

### Targeted Marketing Efforts

To best utilise resources, in-person and digital marketing teams should focus their attention at Street Dr & Grand Ave Station as it is the most popular station. They are thus likely to have a larger impact given the large number of riders that would utilise that station.

## Change Log

- 21/11/2021
  - Combined all to form a single Data Frame (ttm)
  - Split started_at into year, month and date columns
  - Created a new column called "ride_length" which is the difference between "start_at" and "ended_at"
  - Convert ride_length to  data type
  - Created day_of_week using started_at
  - Removed all negative ride_length. Rows decreased from 5,378,834 to 5,376,953
  - Dropped all rows with NULL values. Rows decreased from 5,376,953 to 4,491,323
  - Created casual tibble of dimensions 2,030,607 x 18
  - Created member tibble of dimensions 2,460,716 x 18
  - Created stats tibble of dimensions 2 x 6

- Created start_station_count and end_station_count tibbles of dimensions 10 x 2 each