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

Capstone Project for the Google Data Analytics Professional Certificate
By Quentin Ferreira

1. Identify the business task:

In an effort to increase revenue, the marketing department, led by Lily Moreno, would like to find a way to maximise the number of annual memberships to our service. In order to better understand the market, I have been asked to analyse the ways in which annual members and casual riders use Cyclistic bikes differently. Data-driven insight into these trends should help the marketing team determine what might make casual riders more likely to buy an annual membership, and this will ultimately shape their digital media strategy. Given that the executive team must approve the marketing strategy, I will include some recommendations that align with the goal of increasing annual memberships.

In particular, I am interested in the following:

1. What is the total number of trips for members and casuals, and what proportion of total trips do they represent?
2. What are the average ride lengths for members and casuals respectively?
3. What are the most common starting and ending stations for each?
4. Are there particular days of the week on which most rides take place by category?
5. Is there a preference for the rideable type for members and casuals?

2. Prepare Phase:

The data used for this analysis was collected by Cyclistic, and pertains to rider patterns over the past twelve months from April 2020 to March 2021. The data was located on Cyclistics server, and was organised as separate files by month and year. The data was saved as .csv files within .zip folders. The server is secured, and only accessible by the relevant Cyclistic personnel (for the purposes of this exercise, we will pretend this is the case.) During the analysis, I stored original copies of the data on a secured hard drive, and worked with copies of the data on my pc.

The data included the following fields:

ride_id - a unique ID per ride
rideable_type: the type of bicycle used
started_at: the date and time that the bicycle was checked out
ended_at: the date and time that the bicycle was checked in

start_station_name: the name of the station at the start of the trip
start_station_id: a unique identifier for the start station
end_station_name: the name of the station at the end of the trip
end_station_id: a unique identifier for the end station
Start_lat: the latitude of the start station
start_lng: the longitude of the start station
end_lat: the latitude of the end station
end_lng: the longitude of the end station
member_casual: a field indicating whether the bicycle was taken about by a member or a casual

During the analysis, the following fields were added:
Ride_length: the length of the ride calculated as ended_at - started_at.
day_of_week: the day of the week for started_at

Given that this was internal data, it is safe to assume that is is unbiased and credible, although I did notice that some of the ended_at times were before the started_at times, and that some started_at and ended_at fields seemed very close together in time (a manner of seconds in some cases) which leads me to believe there is an error in the way this data is being collected. There was also an issue of missing start and end station names in some datasets, despite there being some geolocation coordinates available for these fields.

Privacy is protected by using the ride_id as opposed to the riders personal information, although it might have been useful to have a rider_id and total spent per ride so we could track the differences in spending between members and casuals.

This data should be sufficient to answer the business problem, considering the caveat mentioned above

In addition to the data above, I made use of an open source shapefile for the city of Chicago, which I obtained here:
https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

3. Process Phase:

I used several tools for the analysis:

EXCEL: I used this for the initial cleaning and processing of the individual csv files, as it provided a quick and simple way to transform the data and create the new columns I needed. I also used it to analyse the average ride length per month, mode and number of rides per month, and to create a single visualisation showing the percentage of users in each category.

Tableau: Used to create two visualisations from EXCEL spreadsheets, as the platform lends itself to “drag and drop” functionality and allows one to create simple yet clear visuals and to join data from various sources.

RStudio: I used RStudio for the bulk of the manipulation, analysis and visualisation due to its ability to handle a large amount of data and it’s very logical approach to functions and syntax. I found the sqldf library to be a great resource too, as it allowed me to use SQL queries on the large dataframe.

EXCEL Cleaning & Manipulation:

Since the data was delivered in a .csv format, I used the “text to columns” function to format the data into rows and columns. I then saved the csv file as an EXCEL workbook.

I then removed any duplicates of the data.

I converted the columns “started_at” and “ended_at” to date time format in the dd/mm/yyyy - hh:mm format.

I created a ride_length column with a formula to subtract the started_at value from the ended_at value. I formatted the data in column as hh:mm:ss. I then copied and pasted the values only into a new ride_length column. Then, I sorted the sheet by ride length in ascending order, and deleted several rows where ride length was negative (indicating an issue where the “ended at” value was before the “started at” value.) I considered deleting the rides where ride length = 0 but decided against it.

I created a “day_of_week” column, and used the WEEKDAY function to create a numerical representation of the day of the week that each bicycle was checked out, where Sunday was represented by 1.

Sorted sheet by “started at date” in ascending order.

I then created pivot tables, to calculate the number of rides and the average ride length per month for members and casuals.

In order to calculate and represent the average monthly ride length I created a formula to convert the time to minutes (=HOUR(A2)*60+MINUTE(A2)+SECOND(A2)/60.)

RStudio Cleaning and manipulation.

Firstly, I read all the csv files into separate dataframes in RStudio, and glimpsed dataframe to see whether the data types I defined in EXCEL were preserved.

Given that the data types were not preserved, I decided to merge all the dataframes together into one, after which I would convert some data types and then clean the rest as needed once I progressed through my analysis.

```{r}
df_1 <- do.call("rbind", list(apr_20, may_20, jun_20, jul_20, aug_20, sep_20, oct_20, nov_20, dec_20, jan_21, feb_21, mar_21))
```

Before merging the dataframes, I calculated the number of rows in each dataframe and wrote a function to compare the sum of all rows to the final dataframe. Once I merged the dataframes, I ran this function to check if all rows were preserved., which they were.

```{r}
tot_rows <- nrow(apr_20) + nrow(may_20) + nrow(jun_20) + nrow(jul_20) + nrow(aug_20) + nrow(sep_20) + nrow(oct_20) + nrow(nov_20) + nrow(dec_20) + nrow(jan_21) + nrow(feb_21) + nrow(mar_21)
```

```{r}
if (tot_rows == nrow(df_1)){
  print("Binding complete, data verified.")
} else{
  print("Error, please verify your data.")
}
```

I then mutated the started_at and ended_at columns into datetime format using the as_datetime function, and I mutated the ride_length column into a time difference format using as.difftime which returned the time in seconds.

```{r}
df_1 <- df_1 %>%
  mutate(started_at = as_datetime(df_1$started_at, format = "%d/%m/%Y %H:%M")) %>%
  mutate(ended_at = as_datetime(df_1$ended_at, format = "%d/%m/%Y %H:%M")) %>%
  mutate(ride_length = as_hms(df_1$ride_length))

glimpse(df_1)
```

I then moved on to create some more dataframes to dig a bit deeper into the data.

Using the sqldf function, I created two queries that returned the name, trip count per station and latitude and longitude of the top 5 starting stations for members and casuals respectively. Noticing that this initially returned a null value in station names, I reformulated the query to exclude values with no names, and then grouped the data by station name, ordered it by trip count, and then limited it to 5. I then bound these two dataframes into one.

```{r}
mem_start_geo <- sqldf("SELECT member_casual, start_station_name AS Start,
                start_lat AS Starting_Latitude,
                start_lng As Starting_Longitude, count(start_station_name) AS Num_Trips
                FROM df_1
                WHERE start_station_name IS NOT ''
                AND member_casual = 'member'
                GROUP BY start_station_name
                ORDER BY count(start_station_name) DESC
                LIMIT 5", method = "auto")
```

####Top 5 starting geolocations for casuals

```{r}
cas_start_geo <- sqldf("SELECT member_casual, start_station_name AS Start,
                start_lat AS Starting_Latitude, start_lng As Starting_Longitude,
                count(start_station_name) AS Num_Trips
                FROM df_1
                WHERE start_station_name IS NOT ''
                AND member_casual = 'casual'
                GROUP BY start_station_name
                ORDER BY count(start_station_name) DESC
                LIMIT 5", method = "auto")
```

Next, I did exactly the same thing for the end stations.

```{r}

mem_end_geo <- sqldf("SELECT member_casual, end_station_name AS End,
                end_lat AS Ending_Latitude,
                end_lng As Ending_Longitude, count(end_station_name) AS Num_Trips
                FROM df_1
                WHERE end_station_name IS NOT ''
                AND member_casual = 'member'
                GROUP BY end_station_name
                ORDER BY count(end_station_name) DESC
                LIMIT 5", method = "auto")
```

####Top 5 ending geolocations for casuals

```{r}
cas_end_geo <- sqldf("SELECT member_casual, end_station_name AS End,
                end_lat AS Ending_Latitude, end_lng As Ending_Longitude,
                count(end_station_name) AS Num_Trips
                FROM df_1
                WHERE end_station_name IS NOT ''
                AND member_casual = 'casual'
                GROUP BY end_station_name
                ORDER BY count(end_station_name) DESC
                LIMIT 5", method = "auto")
```

The separate dataframes were then combined for plotting, but after attempting some plots unsuccessfully, I realised that the latitude and longitudes needed to be converted into numbers, but my initial attempts returned null values until I realised that the commas in the strings were causing the issue, so I used the gsub function to substitute all the periods for commas and combined that with the as.numeric function and this did the trick.

```
###Binding the two tables into a dataframe, and viewing it
```{r}
start_geo <- rbind(mem_start_geo, cas_start_geo)

View(start_geo)
```

Changing the datatype of the coordinates to real numbers to use for plots
```{r}
start_geo$Starting_Latitude = as.numeric(gsub(",",".",start_geo$Starting_Latitude,fixed=TRUE))
start_geo$Starting_Longitude = as.numeric(gsub(",",".",start_geo$Starting_Longitude,fixed=TRUE))
```
```

I then created a dataframe with the data for my shapefile, so that I could use it to plot my visualisations.

I then plotted the top five starting and ending locations, using a facetwrap to separate the plots by member_casual.

For the yearly mode, I ran another sqldf query, to return the day of the week, a count of the days of the week and grouped them by member_casual, and ordered them in descending order by day of the week.
I then replaced each numerical day of the week entry with the names of the day of the week, because I wanted to plot them.

```
## SQL Queries for the yearly Mode of day_of_week (total, members, casuals)
```{r}
mode_t <- sqldf("SELECT day_of_week, member_casual, COUNT(day_of_week) AS Total
                FROM df_1
                GROUP BY member_casual, day_of_week
                ORDER BY day_of_week DESC", method = "auto")
```

Replacing the numerical values with names of weekdays
```{r}
mode_t$day_of_week[mode_t$day_of_week == "1"] <- "Sunday"
mode_t$day_of_week[mode_t$day_of_week == "2"] <- "Monday"
mode_t$day_of_week[mode_t$day_of_week == "3"] <- "Tuesday"
mode_t$day_of_week[mode_t$day_of_week == "4"] <- "Wednesday"
mode_t$day_of_week[mode_t$day_of_week == "5"] <- "Thursday"
mode_t$day_of_week[mode_t$day_of_week == "6"] <- "Friday"
mode_t$day_of_week[mode_t$day_of_week == "7"] <- "Saturday"

```
```

I used a function to assign factor levels to the day of the week variables, so that RStudio did not reorder them when I plotted them on the x-axis.

```
```{r}
mode_t$day_of_week <- factor(mode_t$day_of_week, levels = rev(unique(mode_t$day_of_week)), ordered=TRUE)
```
```

I then plotted the mode of the day of the week with a stacked bar plot, and included the totals for members/casuals.

Lastly, I looked at the rideable types used by members and casuals. Again, I used a sqldf query, which returned the rideable types with a count of the rideable types as the member_casual field and then grouped them firstly by member_casual and then by rideable type, and ordered them by the count. I changed the names of the rideable types just to remove the underscores, then created a side-by-side plot.

```
## A query to return results related to rideble types used by members
```{r}
bike_df <- sqldf("SELECT rideable_type, member_casual, count(rideable_type) as number_of_uses
                 FROM df_1
                 GROUP BY member_casual, rideable_type
                 ORDER BY count(rideable_type) DESC", method = "auto" )
```

Changing the names of the rideable type to remove the underscore
```{r}
bike_df$rideable_type[bike_df$rideable_type == "classic_bike"] <- "Classic Bike"
bike_df$rideable_type[bike_df$rideable_type == "docked_bike"] <- "Docked Bike"
bike_df$rideable_type[bike_df$rideable_type == "electric_bike"] <- "Electric Bike"
```
```

4. Analyse Phase:

A summary of my analysis is as follows:

Total trips:

Overall, these are the total valid rides for the year, calculated using Pivot Tables in EXCEL:

Members: 2052077
Casuals: 1427121
Total: 3479198

59% of all trips were made by members, while the remaining 41% of trips were undertaken by casual riders.

The month with the highest trip was August 2020, while the month with the fewest trips was February 2021. Perhap not unexpectedly, the summer months overall had far more trips than the winter months.

Ride Lengths:

A quick R analysis (shown below) returned the following

```{r}
mean_r_length <- as.numeric(mean(df_1$ride_length))/60
cat("The average ride length over the year is:",mean_r_length,"minutes")

max_r_length <- as.numeric(max(df_1$ride_length))/3600
cat("The longest ride for the year was:",max_r_length,"hours")
```

Mean ride length for the year was: 24 minutes
Max ride length for the year was: 23.99 hours (perhaps someone left their bike overnight without checking it back in?)
The average ride lengths per month were also calculated using Pivot Tables, but it would be best to show them visualised in the next section.

Some important insights to consider from this data are as follows:

Overall, the trips for casuals were longer on average than those for members, often being twice as long on average. This indicates that casual riders were using the service more for longer, leisurely rides, whilst members seem to use it more for commuting. This hypothesis should be kept in mind in the following section.

July was the month with the longest average trip length, and January had the shortest average trip length: this was in keeping with the trend of longer trips during the warmer months and shorter trips during the colder months.

Busiest Stations:

<u>The top 5 starting stations for members:</u>

Clark St. &. Elm St.
Broadway & Barry Ave.
St. Clair St. & Erie St.
Dearborn St. & Erie St.
Wells St. & Concord Ln.

<u>The top 5 starting stations for casuals:</u>

Streeter Dr. & Grand Ave.
Millennium Park
Lake Shore Dr. & Monroe St.
Theater on the Lake.
Michigan Ave. & Oak St.

The top 5 ending stations for members:

Clark St. &. Elm St.
Broadway & Barry Ave.
St. Clair St. & Erie St.
Dearborn St. & Erie St.
Wells St. & Concord Ln.

The top 5 ending stations for casuals:

Streeter Dr. & Grand Ave.
Millennium Park
Lake Shore Dr. & Monroe St.
Theater on the Lake.
Lake Shore Dr.. & North Blvd.

It seemed that most trips took place right in the centre of the city, although the trips by casual riders were concentrated in an area slightly South of where most trips by members were

Busiest Weekdays (Mode):

Over the year, Saturday was the busiest day of the week for both members and casual riders, while Mondays had the fewest rides by members and Tuesday had the fewest rides by casual riders.

Weekends were busier overall than week days, and it seemed that far fewer casuals use the service during weekdays, perhaps indicating that they prefer to use it for leisure as opposed to commuting, while members were much more likely to use the service to commute. Nevertheless, there are a number of casual members who use the service during the week, and this might represent an area for growth.

Rideable Type Preference:

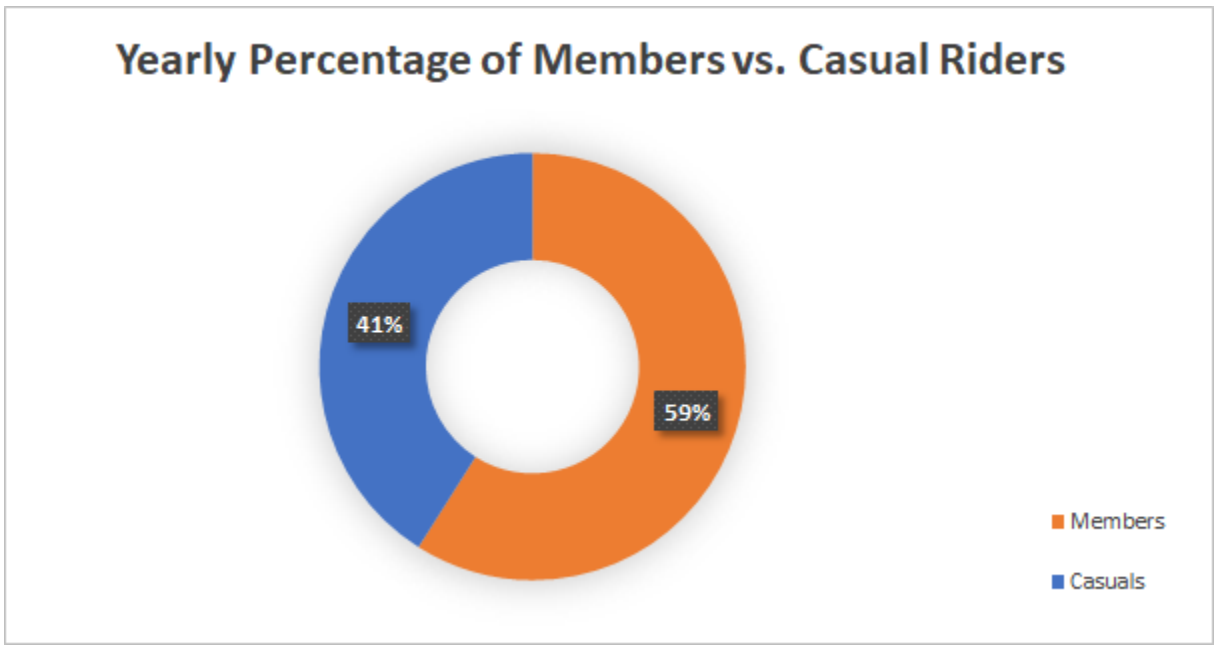
Lastly, the analysis of the rideable types indicated that docked bikes were the most popular all around, followed by electric bikes and classical bikes in last place. This trend was consistent across members and casual riders.

5. Share Phase:

Total trips:

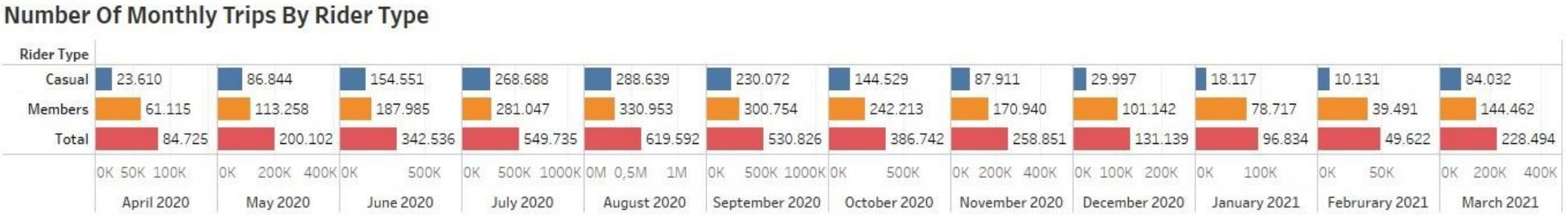
Given that the primary goal of the analysis is to find ways to convert casual riders to members, it is important to understand the current makeup of our customer base.

As you can see below, members accounted for 59% of the total rides taken during the year, while casual riders comprised the remaining 41% of trips. This indicates that there is significant room for converting casual riders into members, and all further analysis should take this proportion into account.



In terms of monthly trips, there is a clear relationship between the number of trips and seasonality; there are far fewer rides during the colder months than in the warmer months, and this trend is particularly pronounced when comparing members to causal riders as per the graph below.

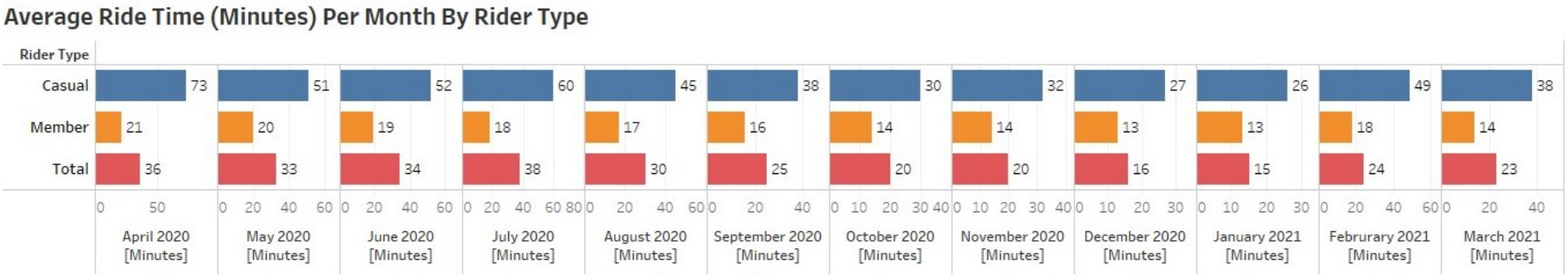
However, while members are much more likely to brave the cold than casual riders, there are still a few thousand casual riders who are undeterred by the cold weather



* Created using Tableau

Ride Lengths:

The length of the average ride for the year was 24 minutes, but this figure is less than the average length for casual riders and more than the average ride length for members. Casuals seem to take rides which are twice as long on average as those of members, indicating that their trips either cover more distance, are undertaken at a leisurely pace, or that perhaps casual riders tend to ride to a destination, keep their bikes “checked out” until they are done and then ride them back. The average ride length of members seems to fit the hypothesis that they are using the service primary to commute to work.

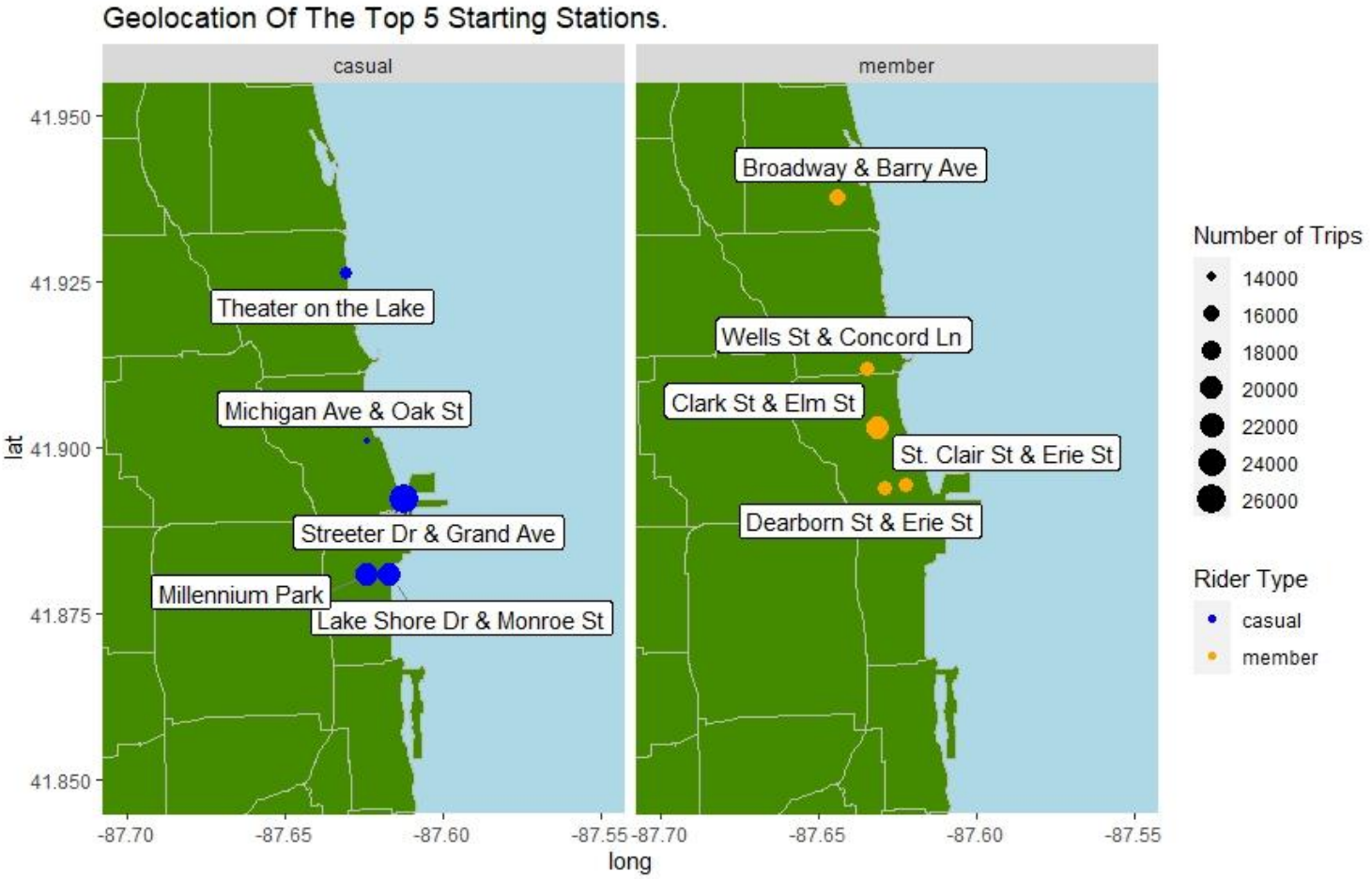


* Created using Tableau

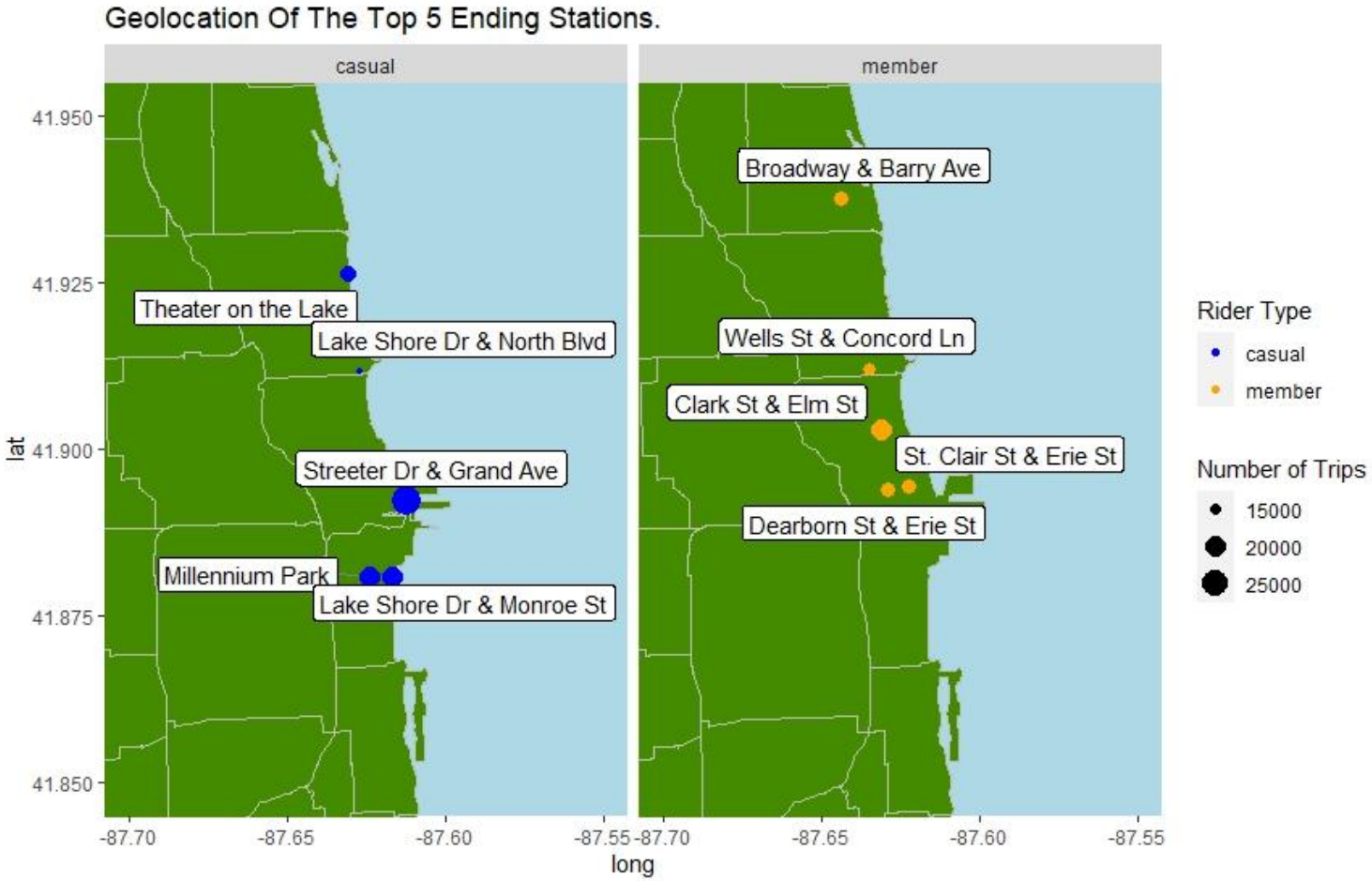
Busiest Stations:

It is clear from these visualisations that the busiest station for casual riders is Streeter Drive and Grand Avenue, while for members it is Clark Street and Elm Street.

As alluded to above, the busiest stations for casuals are concentrated slightly to the South of those of members, and given the fact that Millenium Park features significantly in the busiest stations for casual members this adds even more weight to the assumption that they use the service mostly for leisure activities as opposed to work commutes.



* Created using RStudio

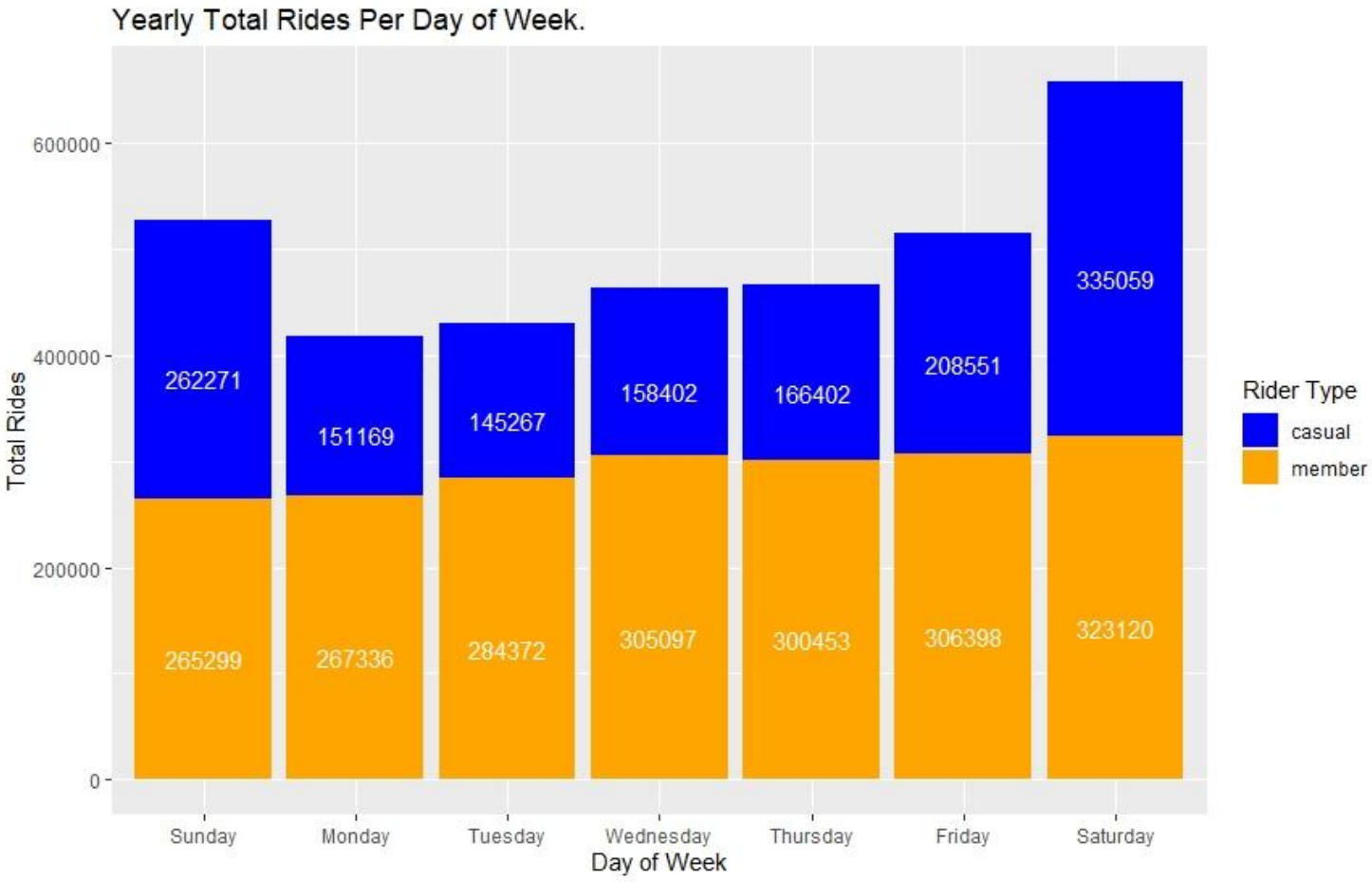


* Created using RStudio

Busiest Weekdays (Mode):

The data indicates that Saturday is the most popular day of the week for both members and casual riders, and weekends were busier overall than week days. Monday is the quietest day overall and for members, while Tuesday is the quietest day for casual riders.

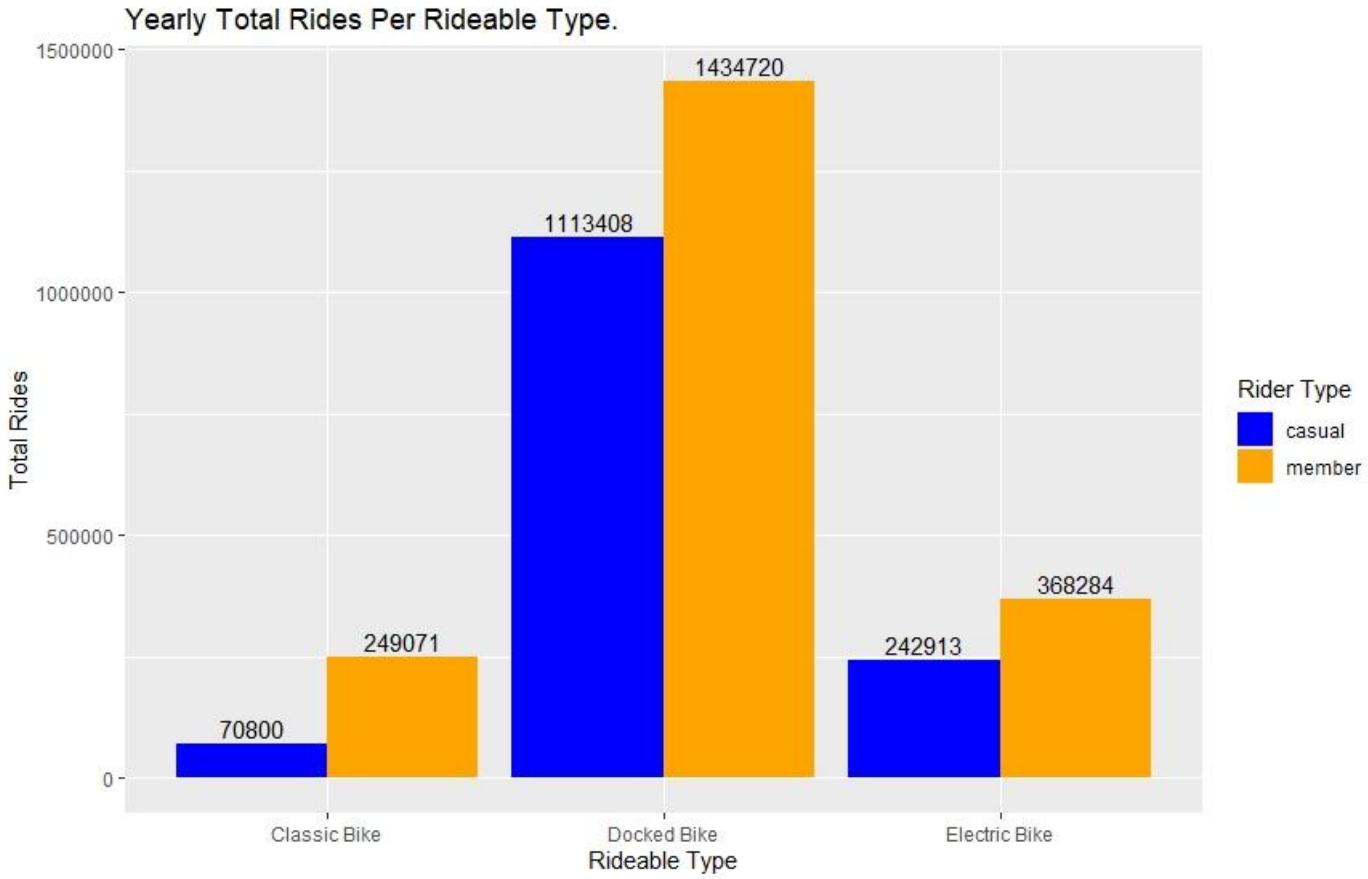
From the graph below, it is clear that there is a significant drop in use by casual riders during the week, while members seem to ride more consistently throughout. Again, this fits the hypothesis that casual riders are mostly using the service for leisure, though there are still a decent number of trips being undertaken during the week.



* Created using RStudio

Rideable Type Preference:

Besides the fact that docked bikes are the most popular option of all, it is also clear that classic bikes are rather unpopular with casual members.



* Created using RStudio

6. Act Phase:

Based on the analysis above, these are my top recommendations:

1. In order to appeal more to the casual rider demographic, any advertising campaigns could focus on the benefits of using the service for work commutes. While it is true that casual riders tend to use the service more over the weekends, a promotion targeting those who regularly use the service during the week could be more successful than one that tries to convert weekend only casual riders into members.
2. Given the seasonal nature of the service, it would be prudent to advertise it during the warmer months, when people are out enjoying the weather. Considering the higher volumes of riders on the weekends, this also presents an opportunity for the marketing team to reach higher numbers of people with any “on-the-ground” type of promotions.
3. “On the ground” promotions or those making use of print media should focus around the stations which see the highest number of starts and ends. A targeted strategy in these areas will reach the maximum number of riders as opposed to a blanket approach that casts a wider net.
4. Any appeals to the casual rider demographic should avoid the use of the classic bike in their promotional material, and should focus instead on the docked or electric bikes.

Going forward, it might be a good idea to assign each rider with a unique rider ID, as this would allow the company to better track their riding patterns in order to offer them further services. This will help identify any casual riders whose riding patterns approximate those of members, and would allow more targeted advertising options. Additional data on amounts spent per month on trips could also open the door to promotions focussing on how much a rider could save by transitioning from a casual rider to a member.