Google Capstone Case Study: Cyclistic

28th April 2023

How Does a Bike-Share Navigate Speedy Success?

Done By:

Lim Ryan (rlim31449@gmail.com)

1. Introduction

Over the course of a few months, I have learnt a lot from the Google Data Analytics Course. There will be 6 sections for each phase of data analysis where I will be applying what I have learnt from the course. This would include the usage of Tableau, R, SQL, Excel Sheets.

2. Scenario

Before diving into the business task, I would like to provide the scenario of the case study. Cyclistics is a bike-share company in Chicago whereby I am part of the data analytics team. The director of marketing believes that annual memberships are the way to grow the company. Hence, we are tasked to understand how casual and annual members differ in their usage of Cyclistic bikes. These insights would help design a new marketing strategy that focuses on converting casual riders to annual members.

3. Ask Phase: The Business Task

The business task can be summarised into one sentence: How can Cyclistic convert casual riders into annual members through identifying the difference between annual and casual riders?

4. Prepare Phase: Understanding Data Sources

The data has been made available publicly by Motivate International Inc. under this licence. Since the data is public, personal details of the rider are not recorded in the dataset. This serves as a limitation of the dataset as it is impossible to identify specific riders to the rides. The team will be using historical trip data of Cyclistic for the previous 12 months. This would be from the range of 2023 March to 2022 April located here.

After looking through the dataset, the columns of all 12 months are uniform. I also noticed that there are empty fields for start_station_name, start_station_id, end_station_name and end_station_id columns. There is also inconsistency in the

station ids. One version has a 2 letter prefix: “TA1308000050”, “WL-008” the other is just numbers of different lengths: “20231”, “623”.

The key columns in the data set would be the rider_id, rideable_type, started_at, ended_at, start_lat, start_lng, end_lat, end_lng and member_casual.

5. Process Phase: Cleaning & Manipulation of Data

In the Process Phase, I had 3 ways to clean and manipulate the data: Excel, SQL or R. Since the data contains thousands of rows, I felt that SQL or R would be better suited for the case study. After attempts to clean data in both languages, I felt that R was more apt for the case study due to the huge storage provided by R Studio and ease of documentation through R markdown file.

These are the key steps I performed for the Process Phase: removing duplicated ride_id, removing inconsistent started_at and ended_at timings, removing incomplete data rows, validating different columns and adding columns based on calculations. Any cleaning or manipulation done in R can be found in the R markdown file found here.

6. Analyse Phase & Share Phase

The overview of the cleaned data table can be found in figure 1 below.

```
  ride_id          rideable_type        started_at
Length:4397954     Length:4397954     Min.   :2022-04-01 00:02:30.00
Class :character   Class :character   1st Qu.:2022-06-18 13:21:58.00
Mode  :character   Mode  :character   Median :2022-08-11 17:18:52.50
                                      Mean   :2022-08-24 10:50:22.02
                                      3rd Qu.:2022-10-13 09:58:30.00
                                      Max.   :2023-03-31 23:59:28.00
   ended_at                      start_station_name start_station_id   end_station_name
Min.   :2022-04-01 00:13:29.00   Length:4397954     Length:4397954     Length:4397954
1st Qu.:2022-06-18 13:48:58.50   Class :character   Class :character   Class :character
Median :2022-08-11 17:35:56.50   Mode  :character   Mode  :character   Mode  :character
Mean   :2022-08-24 11:07:19.26
3rd Qu.:2022-10-13 10:10:11.50
Max.   :2023-04-01 15:00:11.00
end_station_id       start_lat       start_lng         end_lat         end_lng
Length:4397954     Min.   :41.65   Min.   :-87.83   Min.   :41.65   Min.   :-87.83
Class :character   1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
Mode  :character   Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
                   Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.65
                   3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
                   Max.   :42.06   Max.   :-87.53   Max.   :42.06   Max.   :-87.53
member_casual       trip_length        day_of_week       distance
Length:4397954     Min.   :    1.00   Min.   :1.000   Min.   :    0.0
Class :character   1st Qu.:    6.12   1st Qu.:2.000   1st Qu.:  908.6
Mode  :character   Median :   10.55   Median :4.000   Median : 1579.2
                   Mean   :   16.95   Mean   :4.091   Mean   : 2110.9
                   3rd Qu.:   18.82   3rd Qu.:6.000   3rd Qu.: 2745.0
                   Max.   :32035.45   Max.   :7.000   Max.   :30349.4
```

Figure 1: Summary of Cleaned Data

At first glance, I noticed that there is some data with a distance of 0 metres. Since distance is calculated by the start and end points, I am assuming that the rider used it as a round trip.

After reviewing the data summary and considering the business task, I have decided to analyse trends in trip length and distance for different member types. This analysis will allow you to better understand the behaviour of different customer segments and identify any differences or similarities in how they use the bike-sharing service.

In addition to analysing trip length and distance for different member types, I have also decided to analyse the differences in rideable types across member types, trip length, and distance. This will allow you to gain further insights into the preferences and behaviours of different customer segments, and identify any potential opportunities for improving the bike-sharing service.

The steps done for analysis and its relevant visualisations can be found in the following R markdown file located here.

The following section will be an overview of my analysis of the dataset.

6.1 Summary of Analysis

The following subsections summarise the following trends:"Trip Length/Distance", "Rideable Type" , "Day of the Week" and "Top 5 Locations".

6.1.1 Trends for Trip Length/Distance

Based on the graph shown in figure 2. The trend line of annual members & casual members depicts that annual members usually travel shorter distances and have shorter trip lengths. This could be due to the fact that annual members have a subsidised fee for any type of ride. This allows annual members to utilise Cyclistic more often. The graph also shows that outliers which have very long trip lengths and distances belong to casual members as highlighted in red.

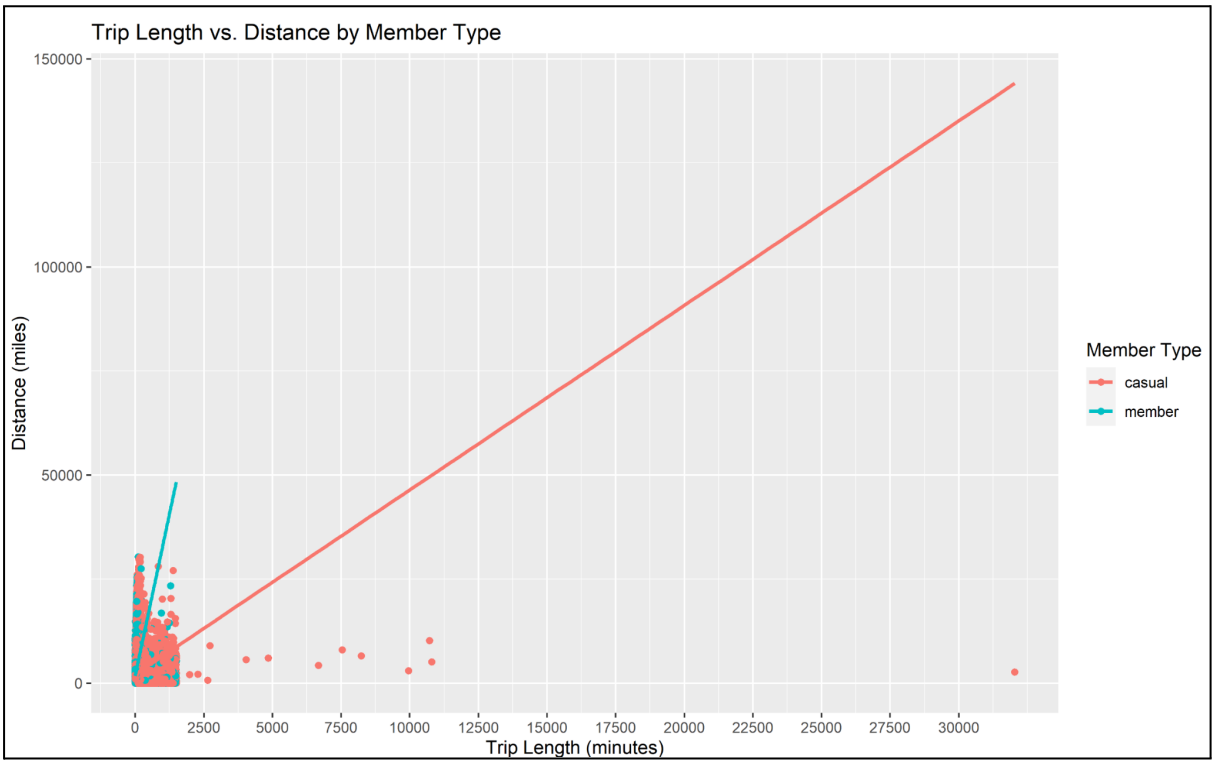

Figure 2: Trip Length Vs Distance For Each Member Type

The graph created does not show a clear representation of the dataset due to the outliers. This led me to remove outliers using IQR (interquartile range) as shown in figure 3.

```
### Removing outliers
Finding Q1 and Q3 of the dataset and finding its IQR.
From that i can find out the upper and lower quartile to further clean the data
```{r}
q1 <- quantile(clean_12_months$trip_length, 0.25)
q3 <- quantile(clean_12_months$trip_length, 0.75)
iqr <- q3 - q1

upper <- q3 + 1.5*iqr
lower <- q1 - 1.5*iqr

clean_12_months <- clean_12_months[clean_12_months$trip_length >= lower & clean_12_months$trip_length
<= upper, ]
```
```

Figure 3: Removing Outliers From Dataset

After removing the outliers, I performed the same analysis again. It produces the following graph as shown in figure 4. The distance that casual members travel on average per trip may not vary as much based on the length of the trip, compared to annual members. This could be due to a variety of factors, such as differences in trip purpose or preferences in route selection.

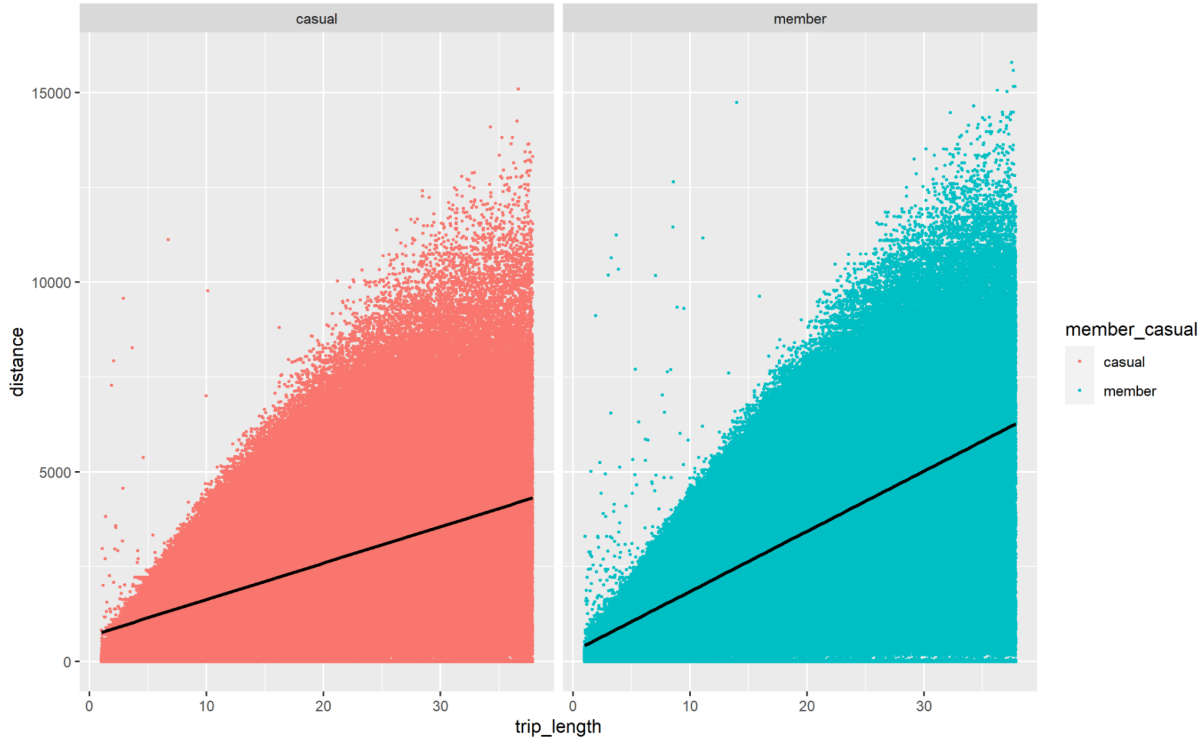

Figure 4: Trend Line for Casual and Annual Members

6.1.2 Trends for Rideable Type

After analysing the trip lengths and distances between each member type, I am now looking into how rideable types affect them as well. The graph shown in figure 5 depicts that there is a stronger relationship for trip lengths and distance travelled for both casual and annual members for electric bikes. This could be due to the fact that it is less tiring to use electric bikes.

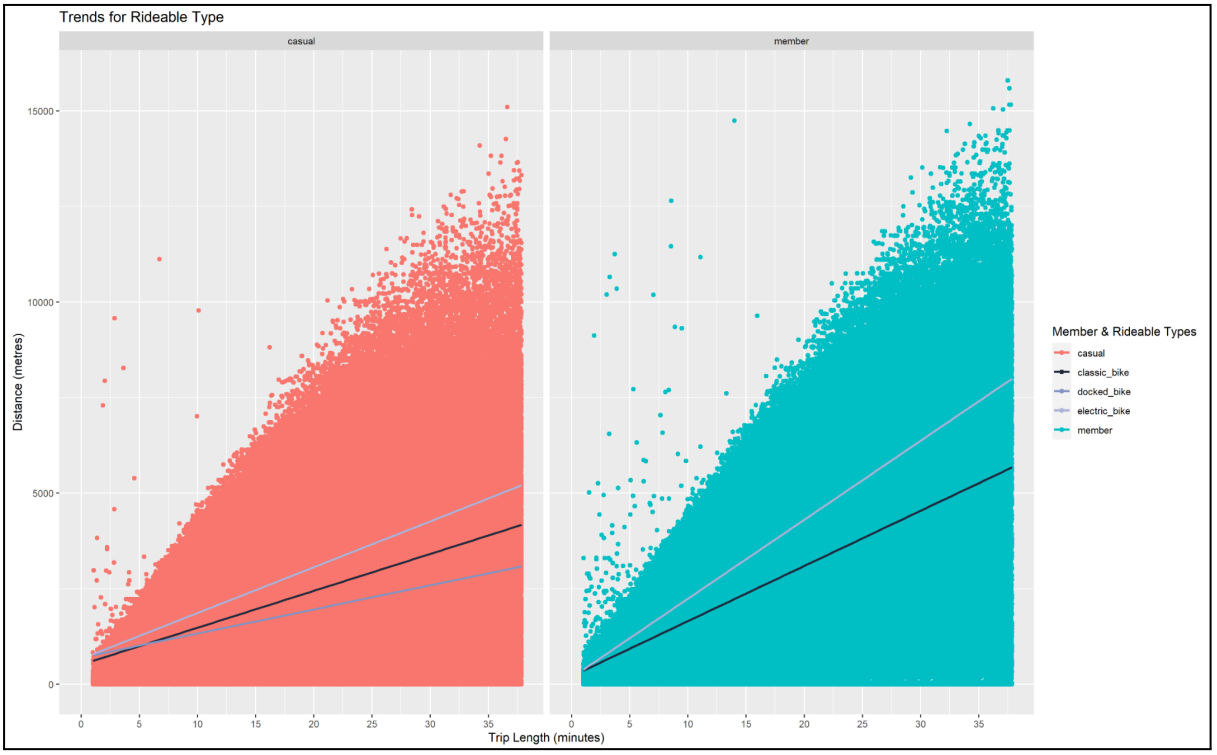

Figure 5: Trend Line for Rideable Types

6.1.3 Days of the Week

Before analysis could be conducted, I created two new data frames: “avg_ridership” and “avg_ridership_member”. The “avg_ridership” data frame consists of the average trip lengths, distance and the amount of riders for each day of the week The “avg_ridership” consists of the same data but is broken further into the different rideable types. There are two approaches to my analysis. First would be using the “avg_ridership_member” data frame. The second would be using the “avg_ridership “which consists of riderable type.

6.1.3.1 Avg_Ridership_Member Analysis

In this section I would be analysing the trends for each member type according to the days of the week. This would be in the order of Sunday to Saturday. The bar graph shown in figure 6 depicts the number of riders per day for each member type. It shows that casual members usually ride on the weekends while annual members ride more frequently on weekdays. This could be due to a myriad of reasons. One reason could be that an annual member utilises Cyclistic as a form of transportation while casual members use it as a form of exercise. Other than that, annual members could be environmentally friendly and choses bike sharing instead of other forms of transportation.

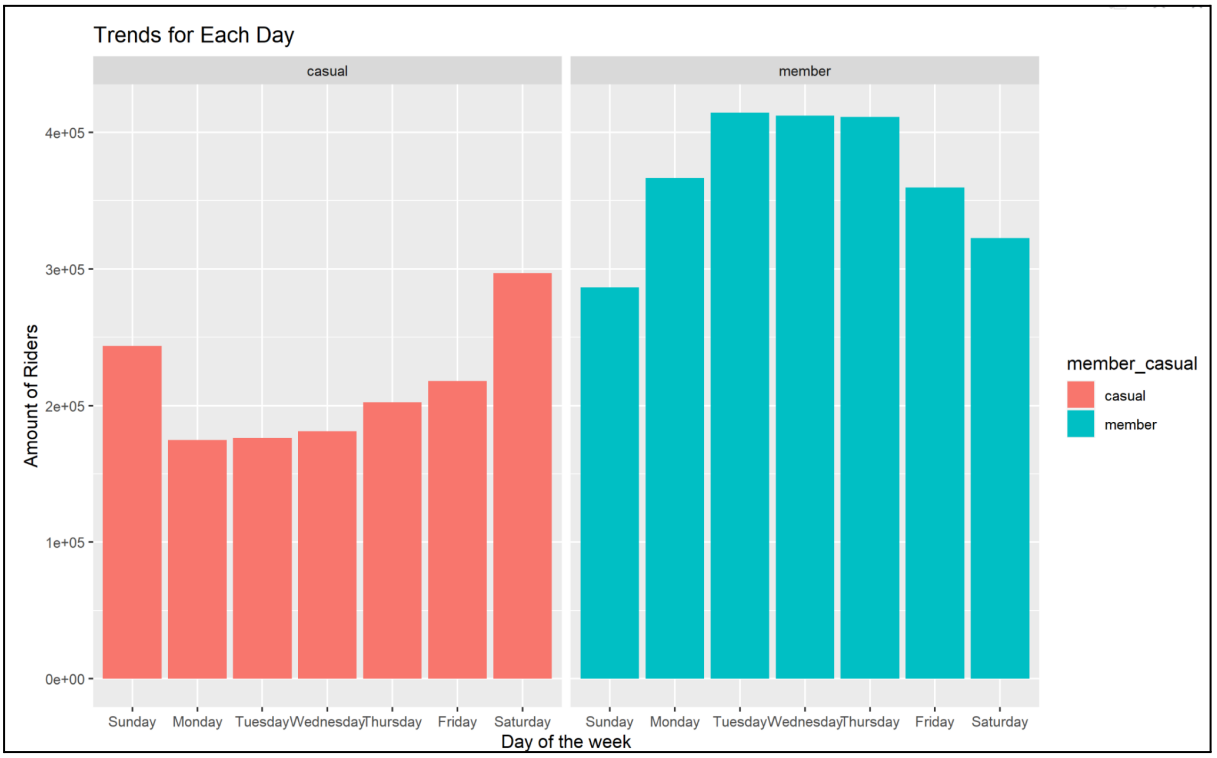

Figure 6: Trends for Each Day

I decided to dive deeper into the analysis by comparing the total average trip lengths and distance for each member type. From figure 7, it shows that casual members have a higher average trip length.

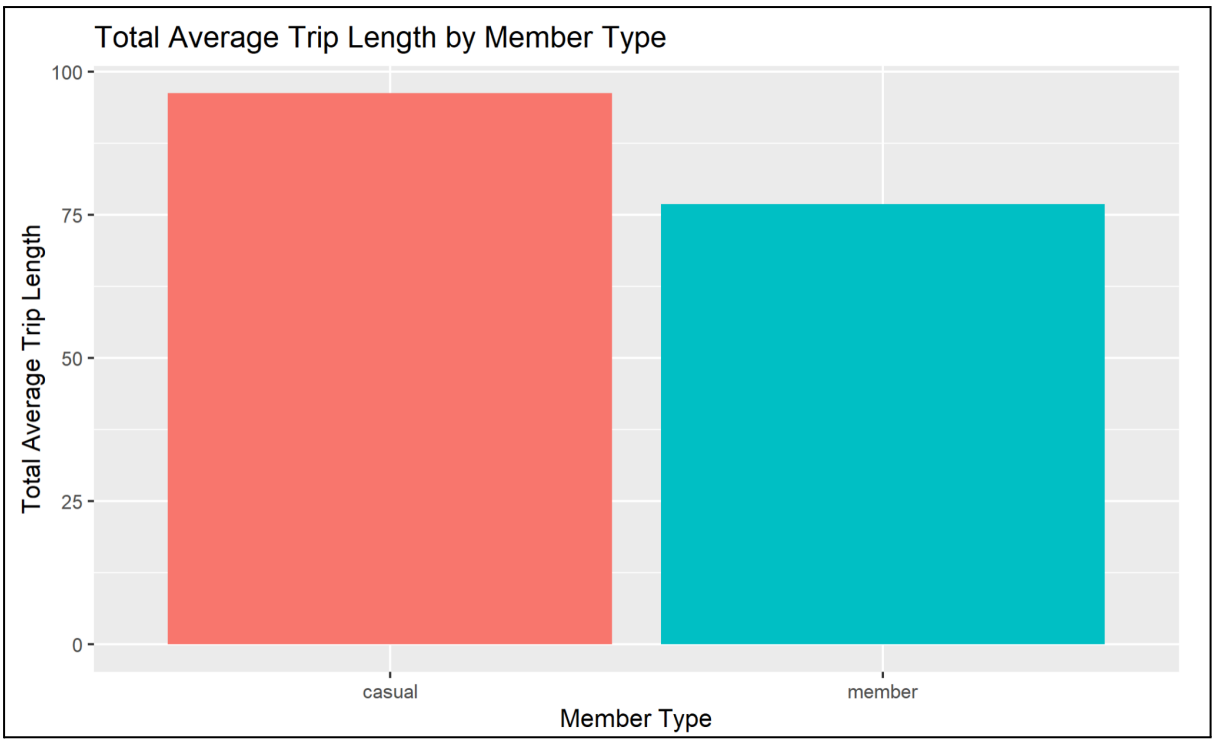

Figure 7: Total Average Trip Length by Member Type

The following figure 8, depicts the total distance travelled by each member type. Although casual members have a higher total trip length, surprisingly both types of members have around the same total average distance travelled.

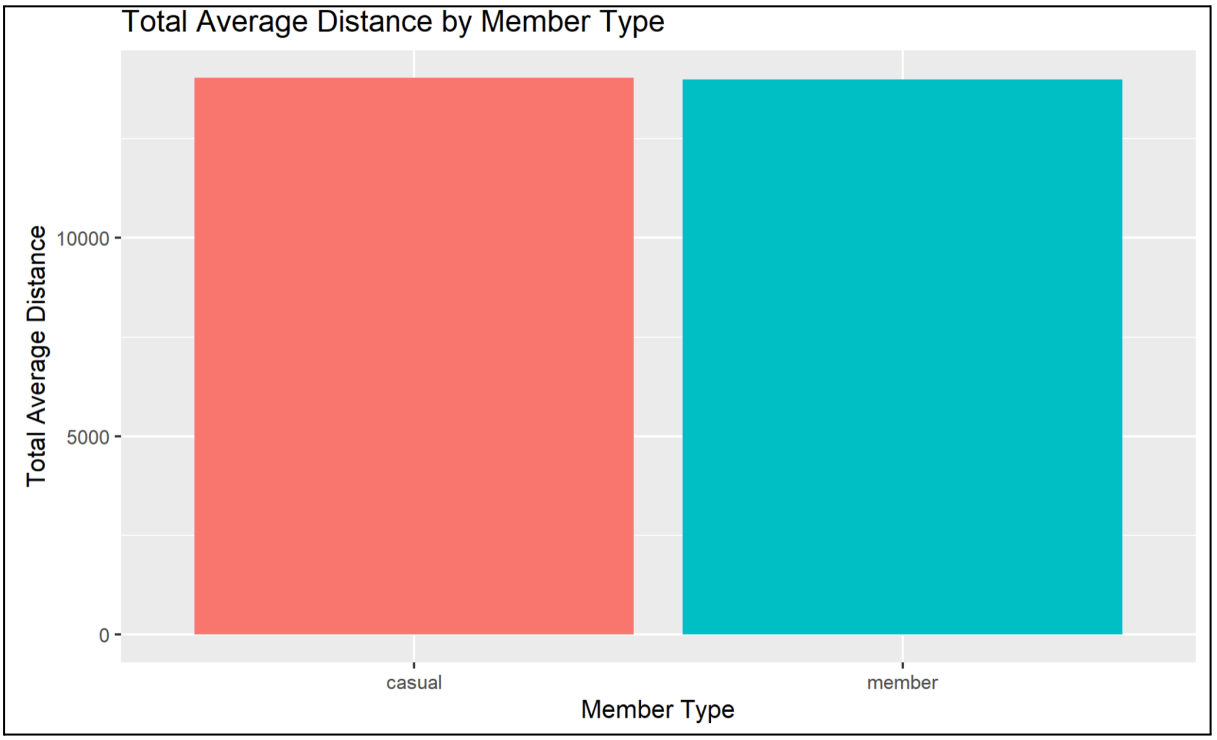

Figure 8: Total Average Distance by Member Type

This could be due to the fact that casual members are using it as a form of leisure. In addition, casual members could be tourists visiting the area and utilising Cylistic for sightseeing.

6.1.3.2 Avg_Ridership Analysis

In this section I would be analysing the trends for each rideable type according to the days of the week. This would be in the order of Sunday to Saturday. The bar graph shown in figure 9 depicts the number of riders per day for each member type. It shows that both casual and annual members prefer casual bikes over the other two types of bikes. This could be due to the accessibility of classic bikes as compared to the other bikes. In addition, classic bikes do not need to be charged or docked.

There is also an interesting insight where only casual members utilise the docked bikes. A possible reason could be that docked bikes are found more frequently at areas of interest where tourists can utilise Cyclistic.

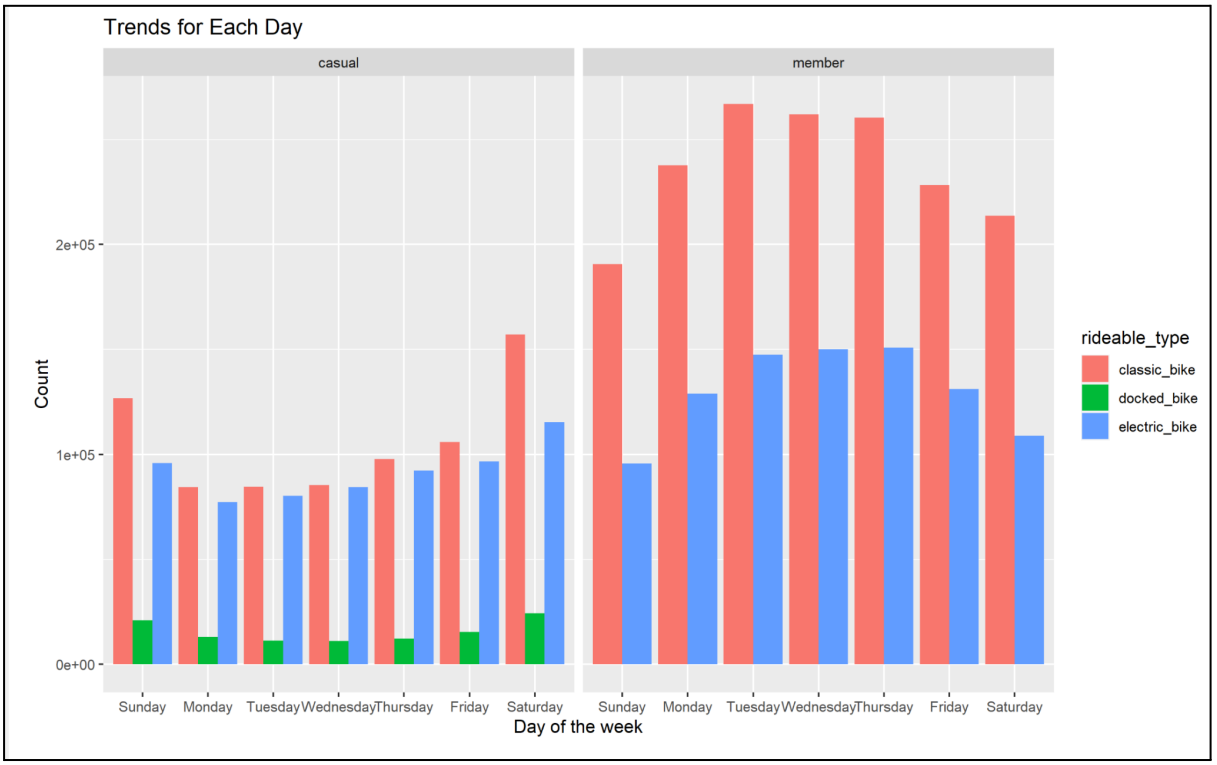

Figure 9: Trends for Each Day for Each Rideable Type

I decided to dive deeper into the analysis by comparing the total average trip lengths and distance for each rideable type. From figure 10, it shows that docked bikes have the average trip length for casual members. Other than that, there is not much difference for both classic and electric bikes for annual members. The high total average trip length for docked bikes could be due to the fact that tourists are using it to sight see.

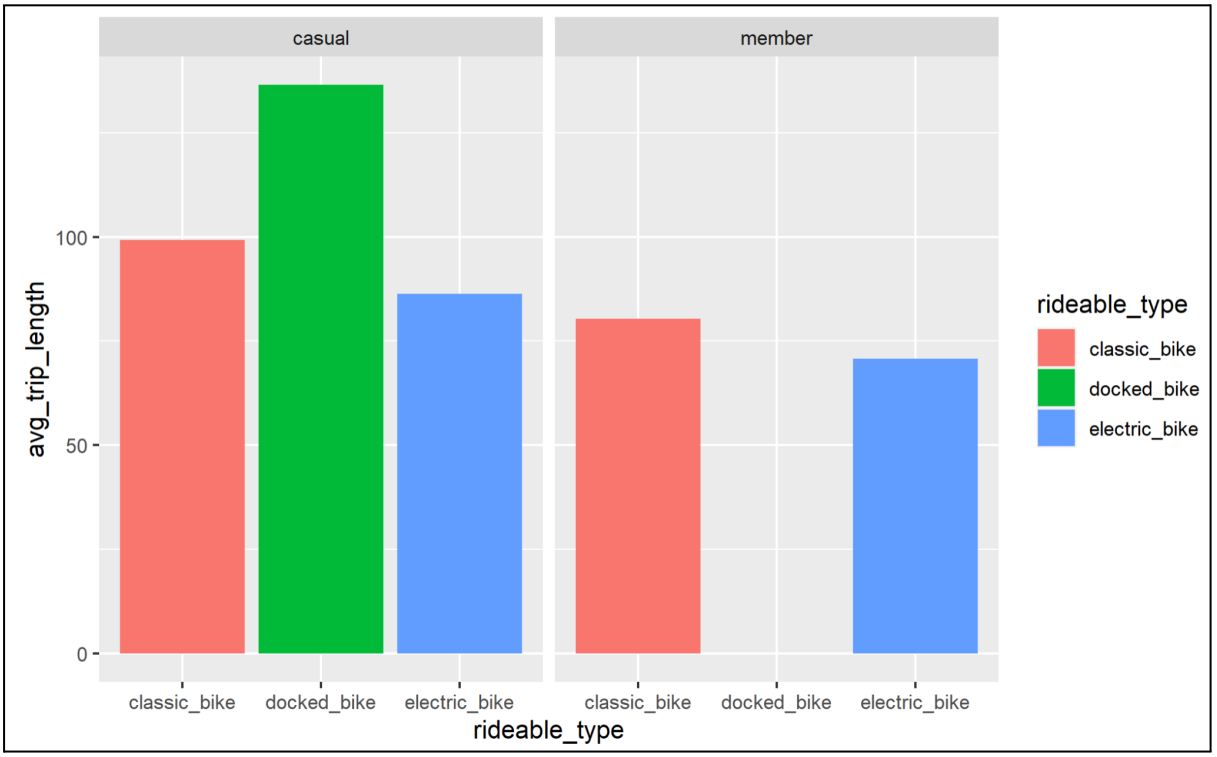

Figure 10:Total Average Trip Length by Rideable Type

The following figure 11, depicts the total distance travelled by each rideable type. It shows that electric bikes for both types of members travelled the most distance.

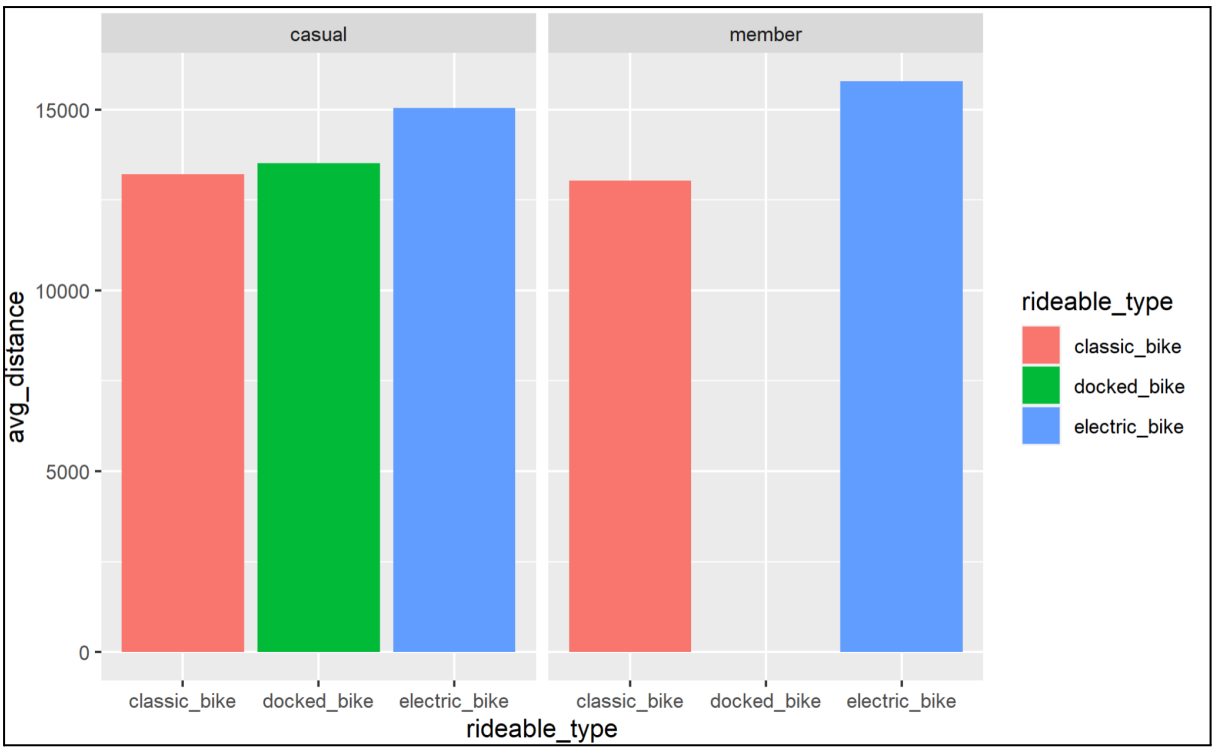

Figure 11:Total Average Distance by Rideable Type

6.1.4 Top 5 Starting Locations

I counted the frequency of each start location from the dataset and saved it into a location_counts data frame. After that I manipulated the data frame to show only the top 5 locations. The top 5 locations are:

1. Streeter Dr & Grand Ave
2. Wells St & Concord Ln
3. DuSable Lake Shore Dr & North Blvd
4. Clark St & Elm St
5. Kingsbury St & Kinzie St

This is shown in the figure 12 which is done using the leaflet package.

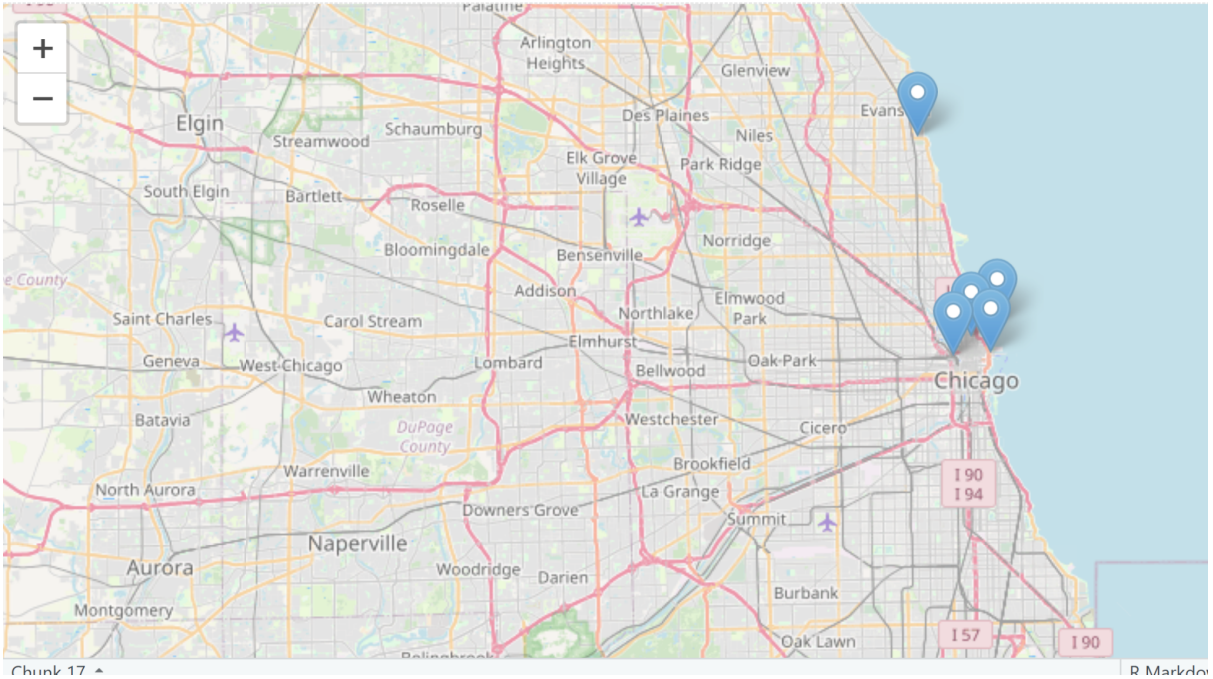

Figure 12: Map of Top 5 Starting Locations

7. Act Phase: Recommendations

Upon reviewing the analysis and business task, Cyclistic aims to increase annual memberships by identifying the differences between casual and annual riders. However, it is important to note some limitations in the dataset that may impact the accuracy of the recommendations. Specifically, there is a lack of personal details linked to ride IDs, which makes it impossible to determine the frequency of an individual's rides. Additionally, the station IDs for both start and end are inconsistent, raising questions about the accuracy of the station names. Despite these limitations, I will provide recommendations on how Cyclistic can potentially convert more casual riders into annual members.

While keeping in mind the limitations of the analysis, it is recommended that Cyclistic increase the number of areas where they have docked bikes. Currently, casual members are the only ones using docked bikes. Therefore, increasing the capacity of docked bikes could be a way to attract more casual members to apply for membership. To implement this recommendation, Cyclistic could invest in additional docking stations in areas where casual members are likely to ride. By providing more

options for docked bikes, Cyclistic can increase the convenience and accessibility of their service, which may encourage more casual members to sign up

Other than that, it is recommended that Cyclistic increase marketing campaigns in the top five locations identified: Streeter Dr & Grand Ave, Wells St & Concord Ln, DuSable Lake Shore Dr & North Blvd, Clark St & Elm St, and Kingsbury St & Kinzie St. To attract casual members to join, Cyclistic could offer different membership options, such as day passes or exclusive passes for classic bikes. Given that the majority of casual members use classic bikes, offering a classic bike pass would likely be especially appealing to this group.

An insight gained from the analysis is that both casual and annual members utilise the bikes for the same amount of time, but casual members tend to travel longer distances. This suggests that a subscription model based on total distance travelled, rather than time, could be a more attractive option for casual members. By implementing a subscription that charges based on total distance, Cyclistic could better cater to the needs of their casual members and provide a more affordable and flexible pricing option. For example, a casual member who wants to take a longer ride would not be penalised for going over a certain time limit. By offering this option, Cyclistic could attract more casual members and increase overall ridership.

In conclusion, by tailoring their membership options and marketing campaigns to the preferences and behaviours of casual members, Cyclistic can attract new members and increase not only overall ridership but increase in subscription of memberships.