Cyclistic Customer Analysis

Benjamin Allen

7/31/22

Introduction

Cyclistic is a bike-share company based in Chicago and operates with a fleet of nearly 6,000 bicycles. Cyclistic has two main subsets of customers: casual and members. Casual riders are those who use Cyclistic bicycles by purchasing single-ride or full-day passes. Members are those who have purchased an annual membership to use Cyclistic's services.

Cyclistic's financial analysts have identified members as being much more profitable than casual customers. Given this, it has been decided to pivot our marketing efforts toward the converting existing casual customers over to annual memberships.

To aid with this pivot, this analysis is aimed at identifying actionable differences between casual riders and members. Actionable differences refer to those that could be leveraged in upcoming Cyclistic marketing.

Data Sources & Setting up SQL

Our primary data source is the past 12 months (Data from July 2021 to June 2022) of Cyclistic trip data which was retrieved from: https://divvy-tripdata.s3.amazonaws.com/index.html

This data consisted of information such as type of bike used, start and end time, customer type, latitude and longitude, and start and end location names. This data was initially in the form of 12 individual CSV files corresponding to each month of trip data. For convenience, these files were combined into a single SQL database table as well as exported into a single CSV file.

Official geographic data on Chicago neighborhoods from the municipal government of Chicago was also used for the geographical visualizations. This was retrieved from: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9

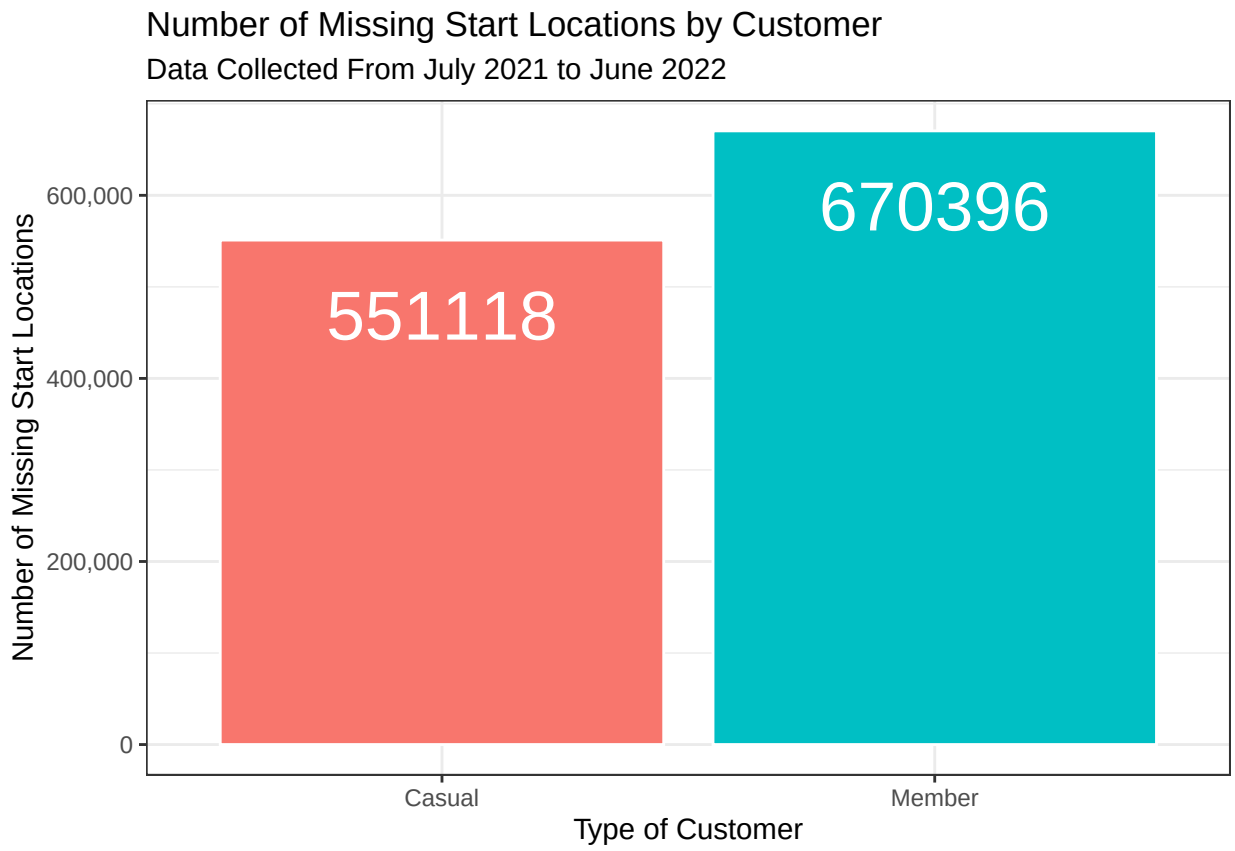
Data Cleaning & Manipulation

Missing Data by Customer

First, the data set only contained missing values for the start and end station names, and these missing values were also only seen in entries where an electric bike was used.

```
ggplot(data = Missing_By_Customer) +
  geom_bar(mapping = aes(x = member_casual, y = num_missing, fill = member_casual),
           color = "#FFFFFF", position = "dodge", stat = 'identity') +
  geom_text(aes(x = member_casual, y = num_missing, label = num_missing),
```

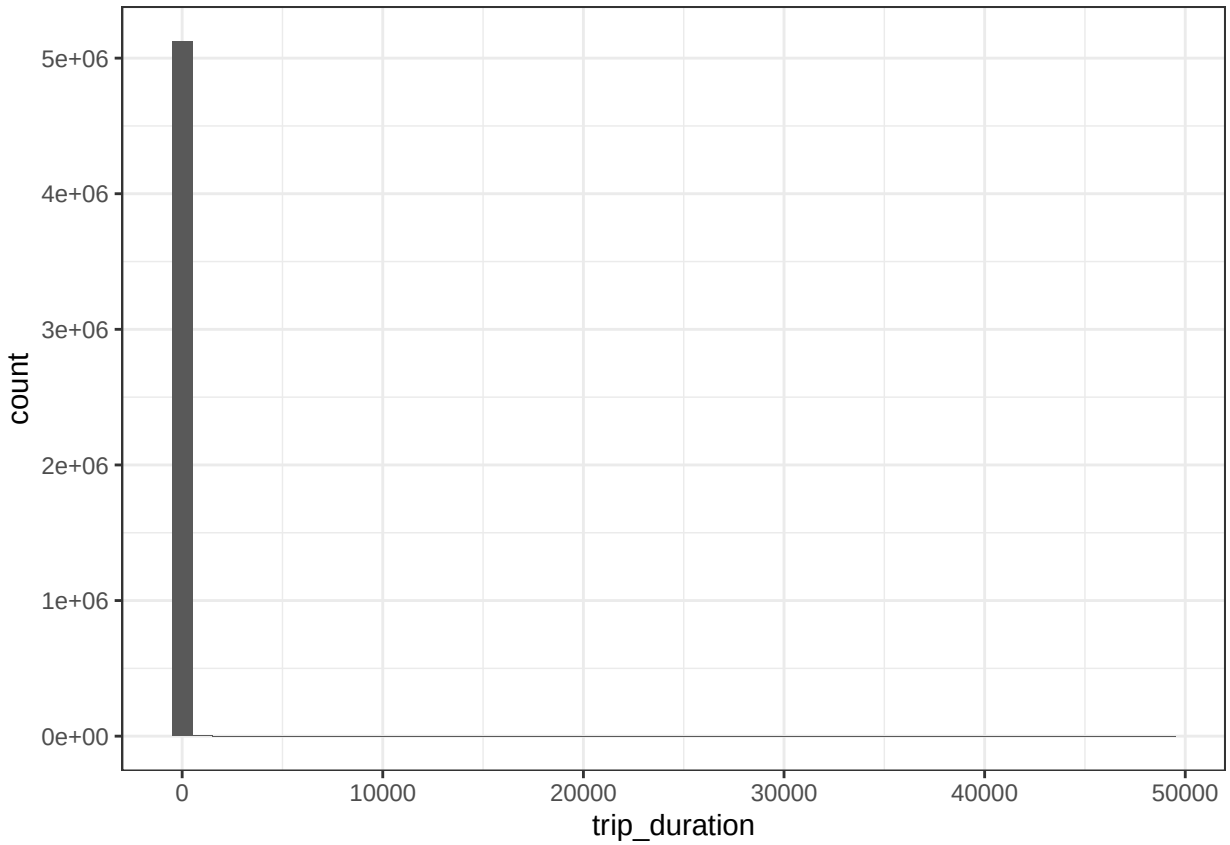
```
          vjust = 2, colour = "white", cex = 9) +
theme_bw() + labs(title = "Number of Missing Start Locations by Customer",
                  subtitle = "Data Collected From July 2021 to June 2022  ----  Note:
                  ↪  All Missing Data From Electric Bike Users",
                  x = "Type of Customer", y = "Number of Missing Start Locations",
                  fill = "Type of Customer") +
theme(legend.position = "none") +
scale_x_discrete(name = "Type of Customer", labels = c("Casual", "Member")) +
scale_y_continuous(name="Number of Missing Start Locations", labels = scales::comma)
```



Applying absolute value because the data features negative values for trip duration that most likely resulted from flipped start and end times.

```
trip_duration_data$trip_duration <- sapply(trip_duration_data$trip_duration, abs)

ggplot(data = trip_duration_data, mapping = aes(x = trip_duration)) +
  geom_histogram(binwidth = 1000)
```



The plot makes it clear that there are major outliers in the data, likely due to bikes being stolen or not returned for whatever reason.

To fix this, let's look at the number of entries with extreme trip durations.

```
print(length(trip_duration_data$trip_duration[trip_duration_data$trip_duration > 10000]))
```

```
## [1] 234
```

```
print(length(trip_duration_data$trip_duration[trip_duration_data$trip_duration > 100]))
```

```
## [1] 72148
```

```
print(length(trip_duration_data$trip_duration[trip_duration_data$trip_duration == 0]))
```

```
## [1] 446
```

Once we do this we can see that there are a few hundred entries with a duration of longer than 10,000 minutes, which seems rather unlikely to be correct. We can also see that there only about 70,000 entries that are greater than a 100 minutes, and given that our data set has over 5,000,000 entries, this is negligible. With this information, it is permissible to trim out these entries for visualizations.

Moving on, there are just a few entries with a trip duration of 0 minutes, which we could likely remove from our dataset, although it is unlikely they will skew results significantly at all.

Trimming outliers from data

As stated before, we will trim out entries with a trip duration greater than 100 minutes because these are infrequent and large enough to be considered outliers.

```
trip_duration_trim <- trip_duration_data[trip_duration_data$trip_duration < 100, ]
```

Analysis & Visualizations

```
casual_trip_duration <- filter(trip_duration_data, member_casual == "casual")

summary(casual_trip_duration$trip_duration)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     8.50    14.93    29.44    27.28 49107.15
```

```
sd(casual_trip_duration$trip_duration)
```

```
## [1] 217.9933
```

```
member_trip_duration <- filter(trip_duration_data, member_casual == "member")

summary(member_trip_duration$trip_duration)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    5.183    8.933   12.846   15.517 1499.967
```

```
sd(member_trip_duration$trip_duration)
```

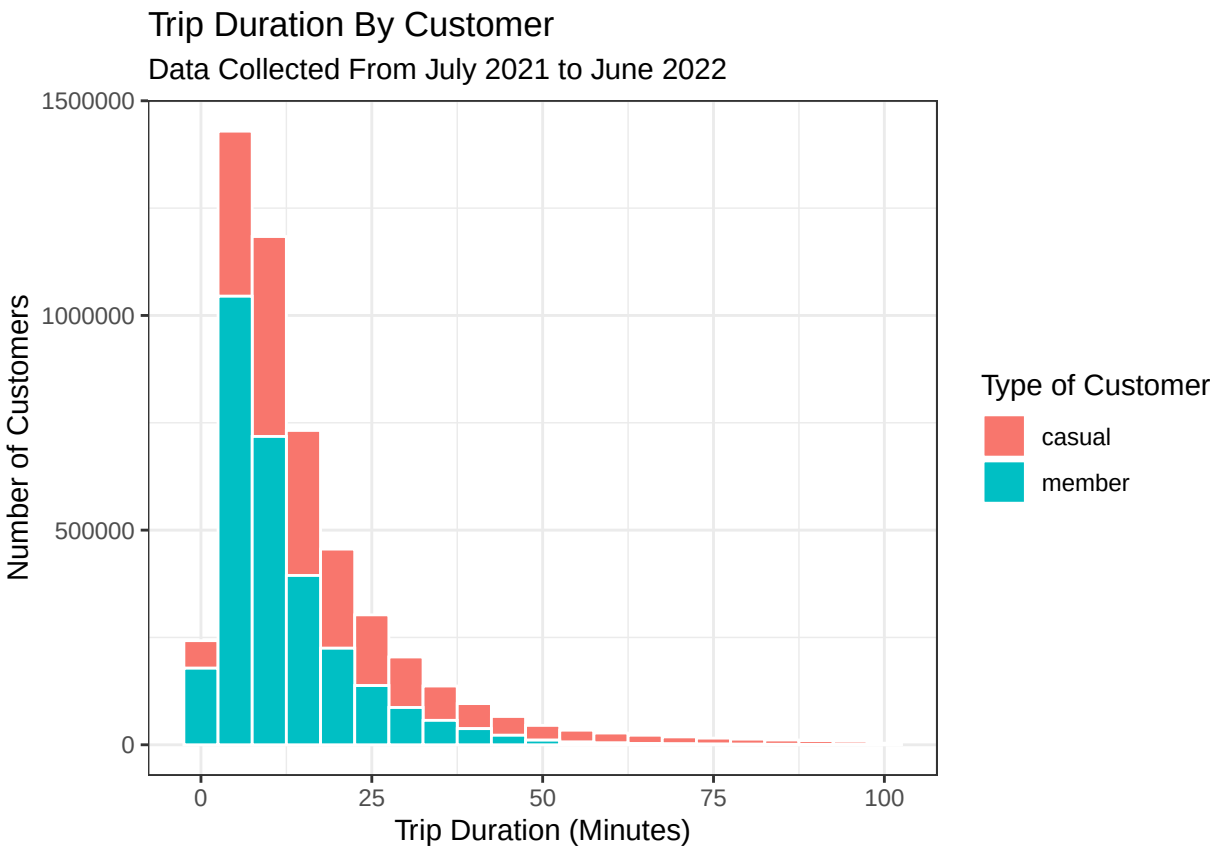
```
## [1] 27.0703
```

Starting our analysis by looking at summary statistics for each customer type, we can see that the extreme outliers in trip duration came from casual customers, evidenced by the maximum value for members being roughly 1500 minutes, compared to the almost 50,000 minute max of casual riders. It can also be seen from both the median and the mean that members have shorter trip durations on average. Finally, and unsurprisingly, the standard deviation of trip durations for casual customers is significantly larger than for members. This is due to the more extreme outliers present for casual customers.

Density of Trip Duration And Differences Between Customers

Next, we will take a look at the distribution of the trip durations by customer.

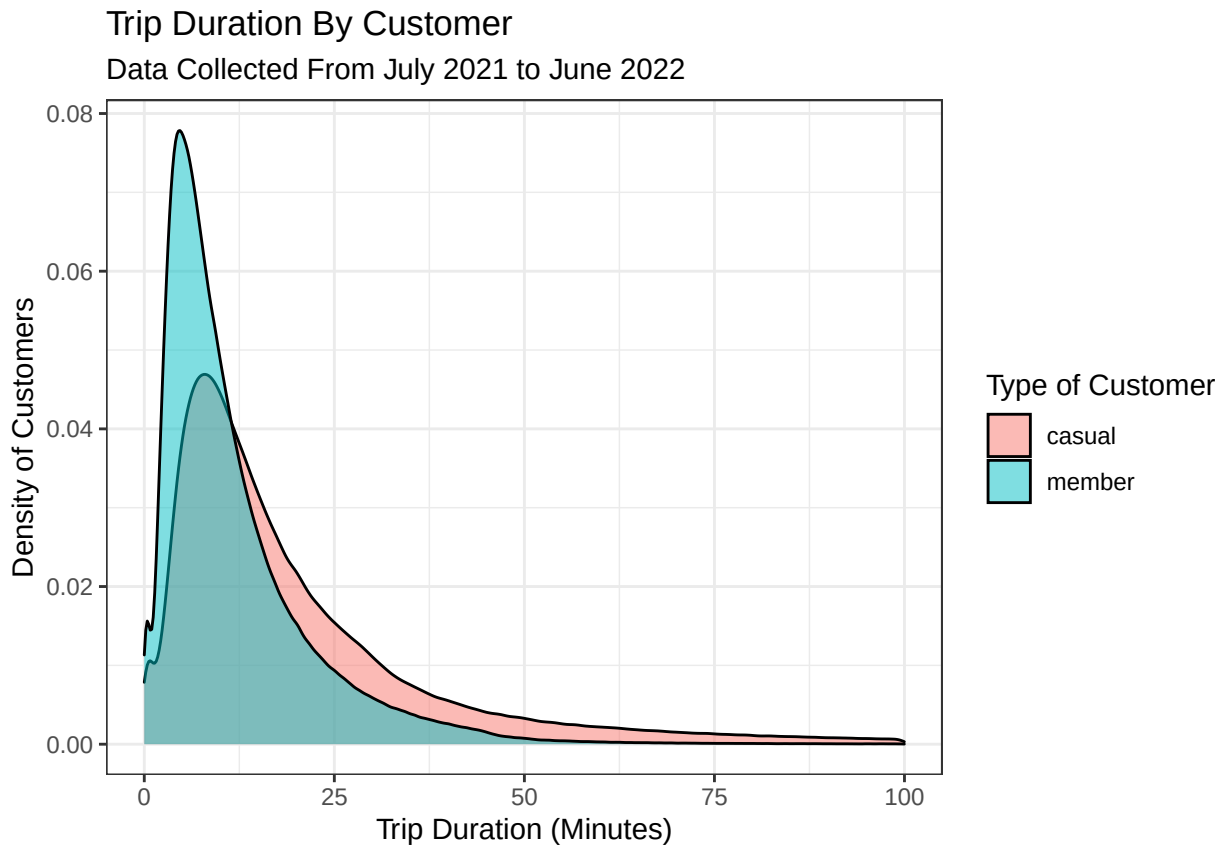
```
ggplot(data = trip_duration_trim) +
  geom_histogram(mapping = aes(x = trip_duration, fill=member_casual),
                 color = "#FFFFFF", binwidth = 5) +
  theme_bw() + labs(title = "Trip Duration By Customer",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Trip Duration (Minutes)", y = "Number of Customers",
                    fill = "Type of Customer")
```



From this first graph we can see that most trips take less than 12.5 minutes for all customers, and we see how the number of trips rapidly decreases as trip duration increases. We can also notice how the trip durations for casual customers are more spread out, and tend to be higher than for members.

Density plot separated by customer

```
ggplot(data = trip_duration_trim, mapping = aes(x = trip_duration, fill=member_casual)) +
  geom_density(alpha = 0.5) + theme_bw() + labs(title = "Trip Duration By Customer",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Trip Duration (Minutes)", y = "Density of Customers",
                    fill = "Type of Customer")
```

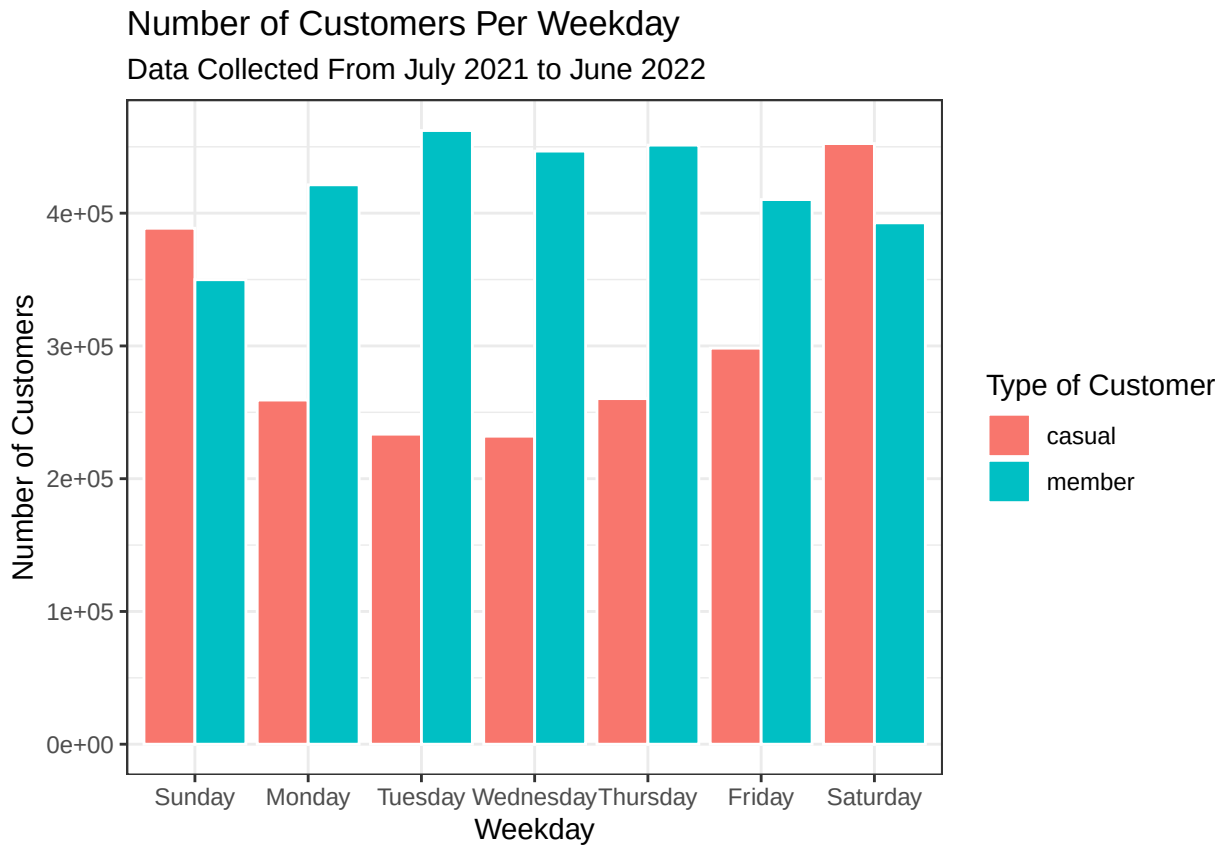


This next graph makes the differences in distribution more obvious, and makes it clear to see the tighter spread of the member distribution, and how members tend to take shorter trips. It is also worth mentioning that both distributions appear to only decrease after reaching their peaks as trip duration increases. This adds further support to the general relationships of the number of trips decreasing with trip duration.

Comparison of Weekdays

We can also compare the number of customers by day of the week, and further break it down by customer type.

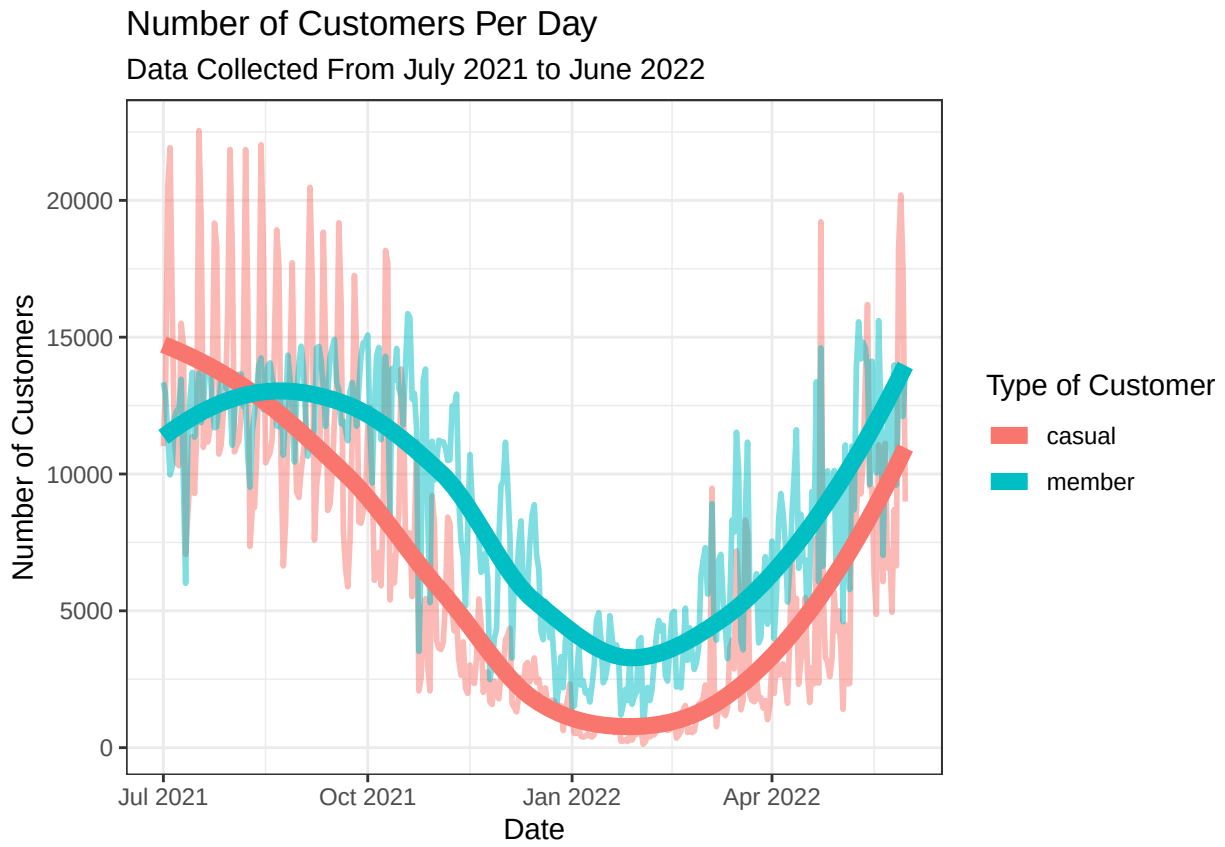
```
ggplot(data = trip_duration_trim, mapping = aes(x = weekday, fill = member_casual)) +
  geom_bar(color = "#FFFFFF", position = 'dodge') +
  theme_bw() + labs(title = "Number of Customers Per Weekday",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Weekday", y = "Number of Customers",
                    fill = "Type of Customer")
```



Doing this, we see that casual customers greatly prefer to take trips on Saturday and Sunday, while members are bit more even though clearly have more preference towards weekdays. This hints at casual customers tending to use Cyclistic for recreational riding, whereas members seem to use it more for commuting to work. Additional evidence of this hypothesis is that members ride on Sunday by far the least, and this can be explained by the fact that many places of work are closed on Sunday.

Seasonal Look at Customers

```
ggplot(data = grouped_trip_data,
       mapping = aes(x = started_at, y = count, color = member_casual)) +
  geom_line(size = 1, alpha = 0.5) +
  geom_smooth(se = FALSE, size = 3, method = 'loess', formula = 'y ~x') +
  theme_bw() + labs(title = "Number of Customers Per Day",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Date", y = "Number of Customers",
                    color = "Type of Customer")
```



Moving on to a view of ridership over the entire year, we can see the basic trend of ridership being highest during the summer and early fall, while dropping substantially during the winter and early spring months. This makes sense as Chicago is well-known for its brutal winters, and that makes for a poor cycling environment.

Looking at the differences between customers, we can see that casual ridership is a lot more volatile both on a day-to-day scale, but also over the course of a year. This is likely because casual customers tend to use bikes recreationally and therefore are less likely to want to ride a bike in snow and terrible weather. Members on the other hand are more stable and appear to peak later in September. This may be explained by the commuter hypothesis because members need to be at work, and therefore are more willing to brave worse weather.

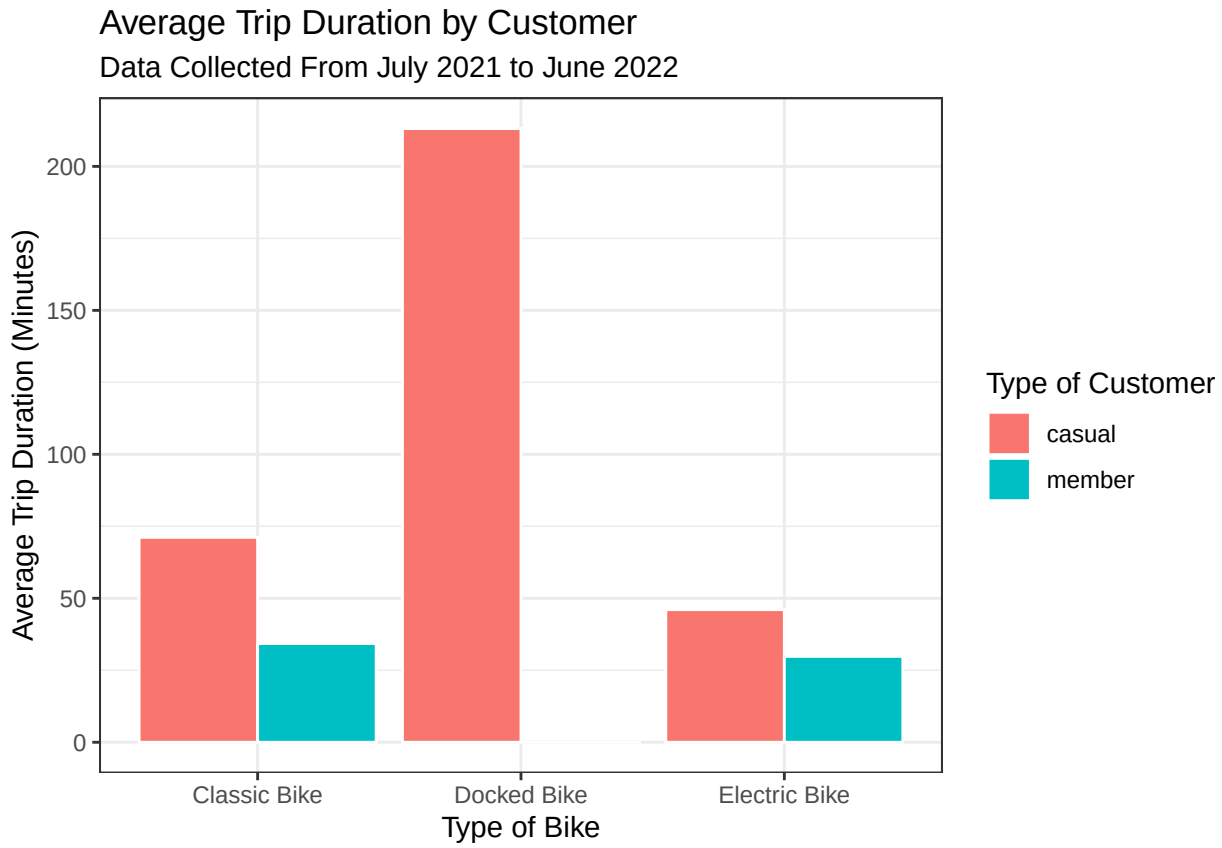
Graphs from initial SQL queries

The following graphs have all been generated from basic SQL queries of the whole data set, however these queries were performed in a database software and not R.

Average Trip Duration by Customer and Bike

```
Average_Trip_Duration_By_Customer_And_Bike <-
↪  rbind(Average_Trip_Duration_By_Customer_And_Bike, c(0, "docked_bike", "member"))
Average_Trip_Duration_By_Customer_And_Bike$average_trip_duration <-
↪  as.numeric(Average_Trip_Duration_By_Customer_And_Bike$average_trip_duration)
```

```
ggplot(data = Average_Trip_Duration_By_Customer_And_Bike) +
  geom_bar(mapping = aes(x = rideable_type, y = average_trip_duration,
                         fill = member_casual), color = "#FFFFFF",
           position = "dodge", stat = 'identity') +
  theme_bw() + labs(title = "Average Trip Duration by Customer",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Type of Bike", y = "Average Trip Duration (Minutes)",
                    fill = "Type of Customer") +
  scale_x_discrete(name = "Type of Bike",
                   labels = c("Classic Bike", "Docked Bike", "Electric Bike"))
```



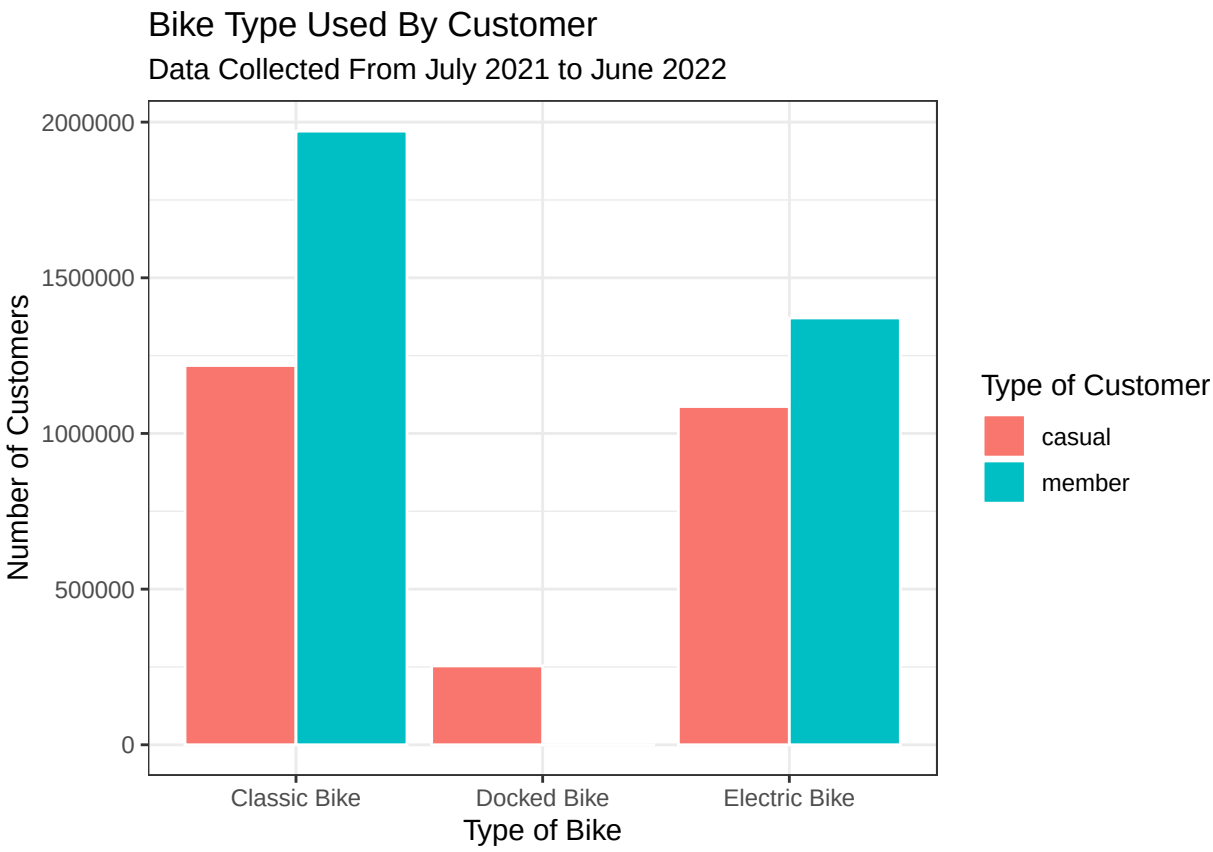
From this graph, we can see that docked bikes have by far the highest average trip duration. This may be a major contributing factor to the greater spread of the casual distribution. We can also see that for both casual riders and members, classic bikes have longer trip durations on average compared to electric bikes.

Bike Type by Customer

```
Bike_Type_By_Customer <- Bike_Type_By_Customer %>%
  rbind(c(0, "member", "docked_bike"))
Bike_Type_By_Customer$COUNT... <- as.numeric(Bike_Type_By_Customer$COUNT...)

ggplot(data = Bike_Type_By_Customer) +
  geom_bar(mapping = aes(x = rideable_type, y = COUNT...,
                         fill = member_casual), color = "#FFFFFF",
```

```
          position = "dodge", stat = 'identity') +
  theme_bw() + labs(title = "Bike Type Used By Customer",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Type of Bike", y = "Number of Customers",
                    fill = "Type of Customer") +
  scale_x_discrete(name = "Type of Bike",
                   labels = c("Classic Bike", "Docked Bike", "Electric Bike"))
```



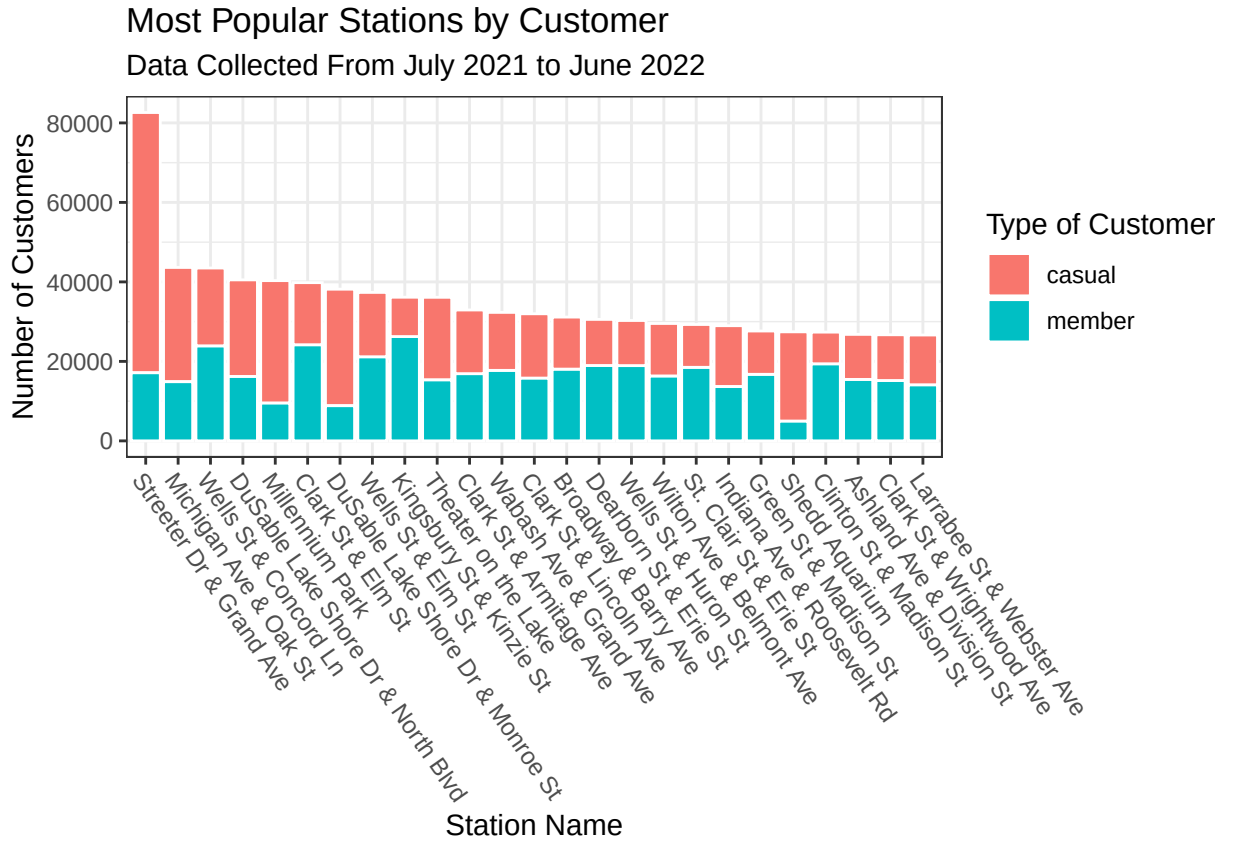
When we break down bike type by the customer type, we can first see that classic bikes are the preferred type by both types of customers. However, members clearly prefer classic bikes over electric bikes, while the difference for casual customers is less stark. Casual customers are also the only type to use docked bikes, although it is by far the least popular option.

Most Popular Stations by Customer

```
Start_Location_By_Customer <-
↪  Start_Location_By_Customer[is.na(Start_Location_By_Customer$start_station_name) ==
↪  F,]
Most_Popular_Start <- Most_Popular_Start[-1,]
Start_Location_By_Customer <- left_join(Start_Location_By_Customer, Most_Popular_Start)
```

```
## Joining, by = "start_station_name"
```

```
Start_Location_By_Customer %>%
  filter(start_num >= 26642) %>%
  ggplot() +
  geom_bar(mapping = aes(x = reorder(start_station_name, -start_num), y = num,
                         fill = member_casual), color = "#FFFFFF", stat = 'identity') +
  theme_bw() + labs(title = "Most Popular Stations by Customer",
                    subtitle = "Data Collected From July 2021 to June 2022",
                    x = "Station Name", y = "Number of Customers",
                    fill = "Type of Customer") +
  theme(axis.text.x=element_text(angle = -55, hjust = 0))
```



Lastly, we can see that Streeter Drive & Grave Avenue is by far the most popular station for customers in total, however the breakdown by customer shows that this is due to the huge presence of casual customers, while for members it is a less important station. Besides this station, none of the others stand out in particular, however we do consistently see stations that favor one type of customer over the other.

Geographic Analysis

To finish off this analysis, we will take a look at the geographic distribution of customers, both in total and by customer type.

Geographic Distribution of Total Rides

```
summary(chicago_data$n_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    2024   10110   51783   53310  463672
```

```
summary(chicago_data$n_member)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     626    4920   29721   27401  291019
```

```
summary(chicago_data$n_casual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1122    5067   22062   24286  195688
```

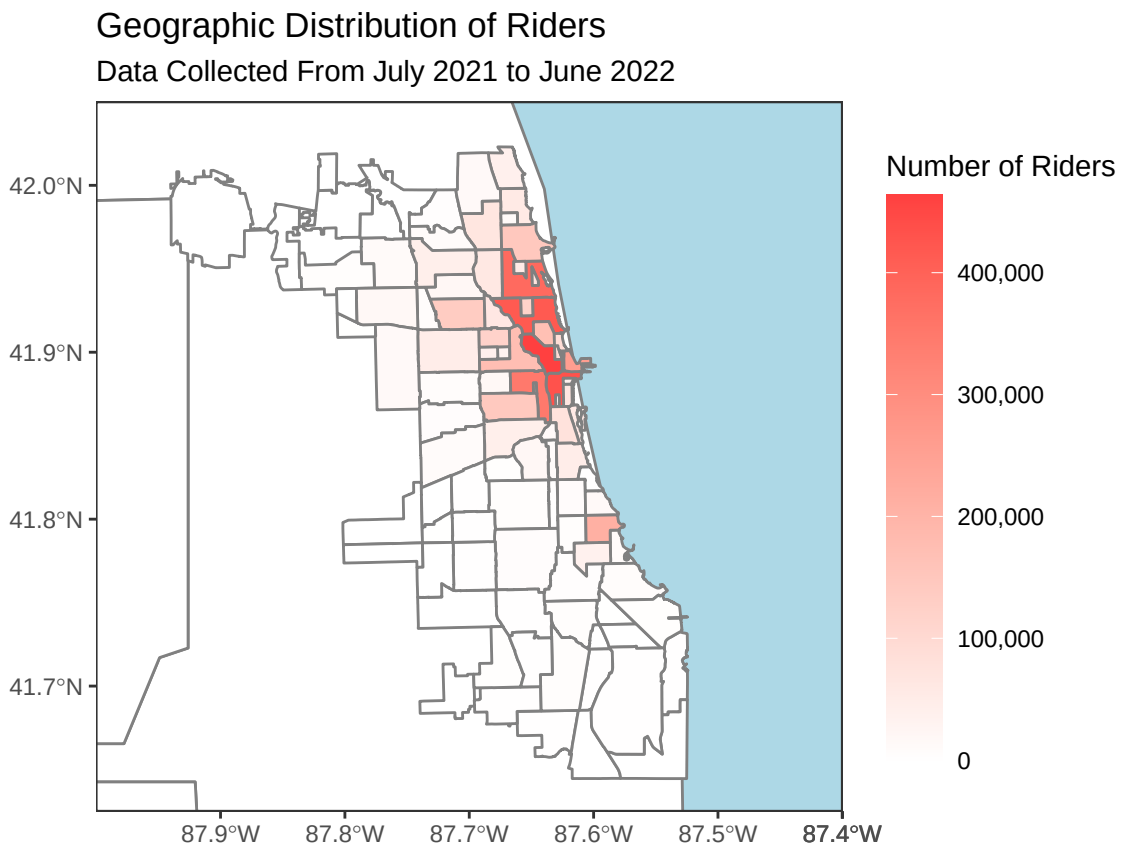
```
summary(chicago_data$mc_difference)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -23681    -502      17    7660    5637  118366
```

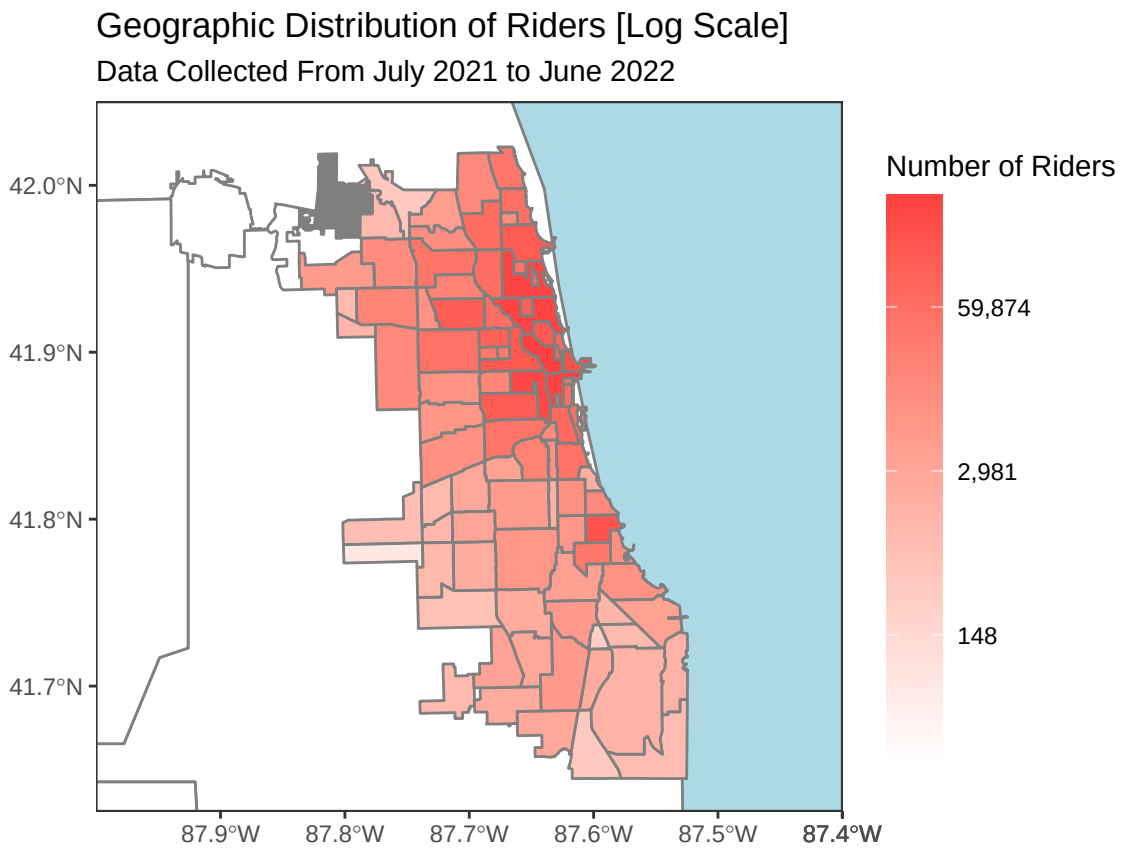
Looking at the summary statistics for the total count of trips per neighborhood, we can see that there is a wide range of values going from zero all the way up to almost 475,000. If we look at summary for both customer types we can see that members tend to be more centralized. This is indicated by the fact that members have a higher mean and maximum than casual customers, however the median is actually lower. This means that if we were to take out a few of the largest neighborhoods, then we would expect to see members have much lower mean. Essentially, members are centralized within larger neighborhoods compared to casual customers. This supports the commuter hypothesis because it would make since that member trips would be centralized in the larger, more commercial neighborhoods if they were commuting to work.

To check these numbers, let's create a few maps to see the geographic distribution of ridership.

```
ggplot(data = world) +
  geom_sf(fill = 'lightblue') +
  geom_sf(data = counties, fill = 'white', color = gray(0.5)) +
  geom_sf(data = chicago_data, aes(fill = n_total), color = gray(0.5)) +
  coord_sf(xlim = c(-88, -87.40), ylim = c(41.625, 42.05), expand = FALSE) +
  labs(title = "Geographic Distribution of Riders",
       subtitle = "Data Collected From July 2021 to June 2022",
       fill = "Number of Riders") +
  scale_fill_gradient(low = 'white', high = '#ff4040', labels = scales::comma) +
  theme(legend.key.height = unit(1.5, 'cm'),
        legend.key.width = unit(0.75, 'cm'))
```




```
ggplot(data = world) +
  geom_sf(fill = 'lightblue') +
  geom_sf(data = counties, fill = 'white', color = gray(0.5)) +
  geom_sf(data = chicago_data, aes(fill = n_total), color = gray(0.5)) +
  coord_sf(xlim = c(-88, -87.40), ylim = c(41.625, 42.05), expand = FALSE) +
  labs(title = "Geographic Distribution of Riders [Log Scale]",
       subtitle = "Data Collected From July 2021 to June 2022",
       fill = "Number of Riders") +
  scale_fill_gradient(low = 'white', high = '#ff4040', labels = scales::comma, trans =
  ↪  'log') +
  theme(legend.key.height = unit(1.5, 'cm'),
        legend.key.width = unit(0.75, 'cm'))
```

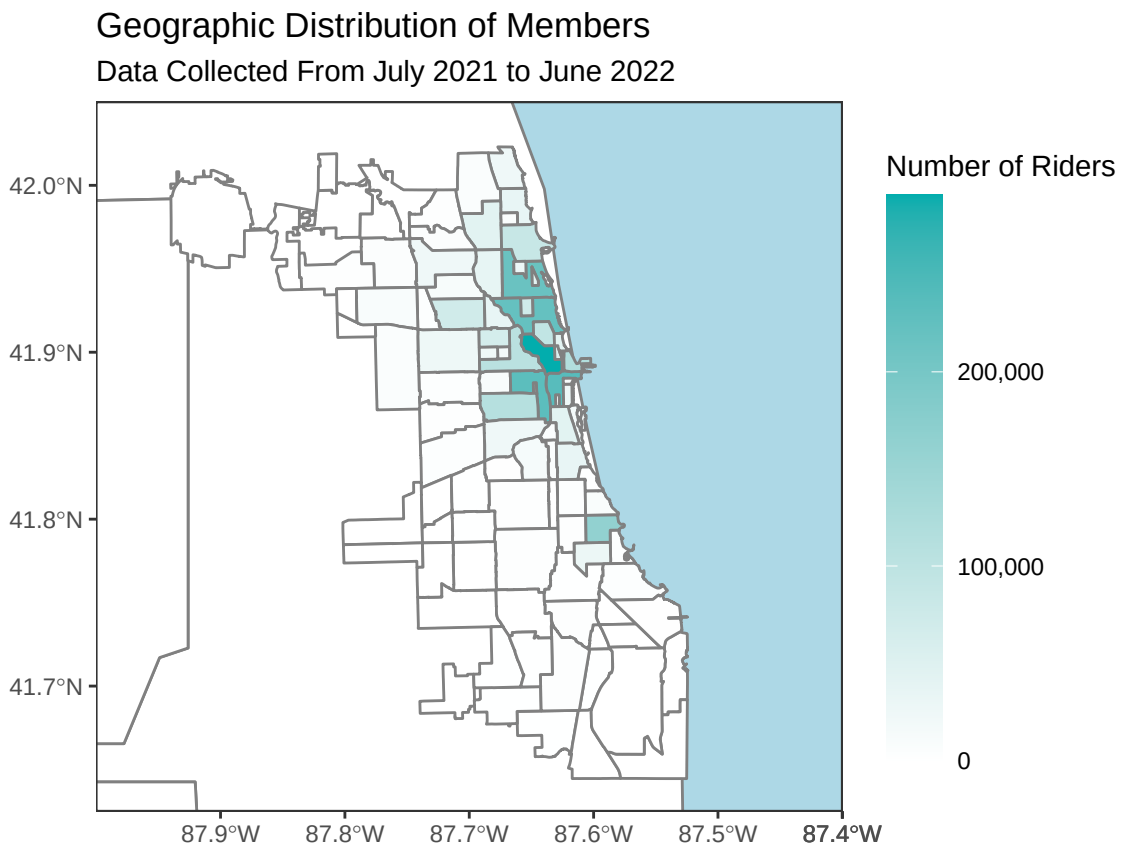


From the first graph we can see that total ridership is centralized in downtown Chicago, and that it decreases gradually moving away from this area. If we use a logarithmic scale, then we can more clearly see this gradual decrease in ridership.

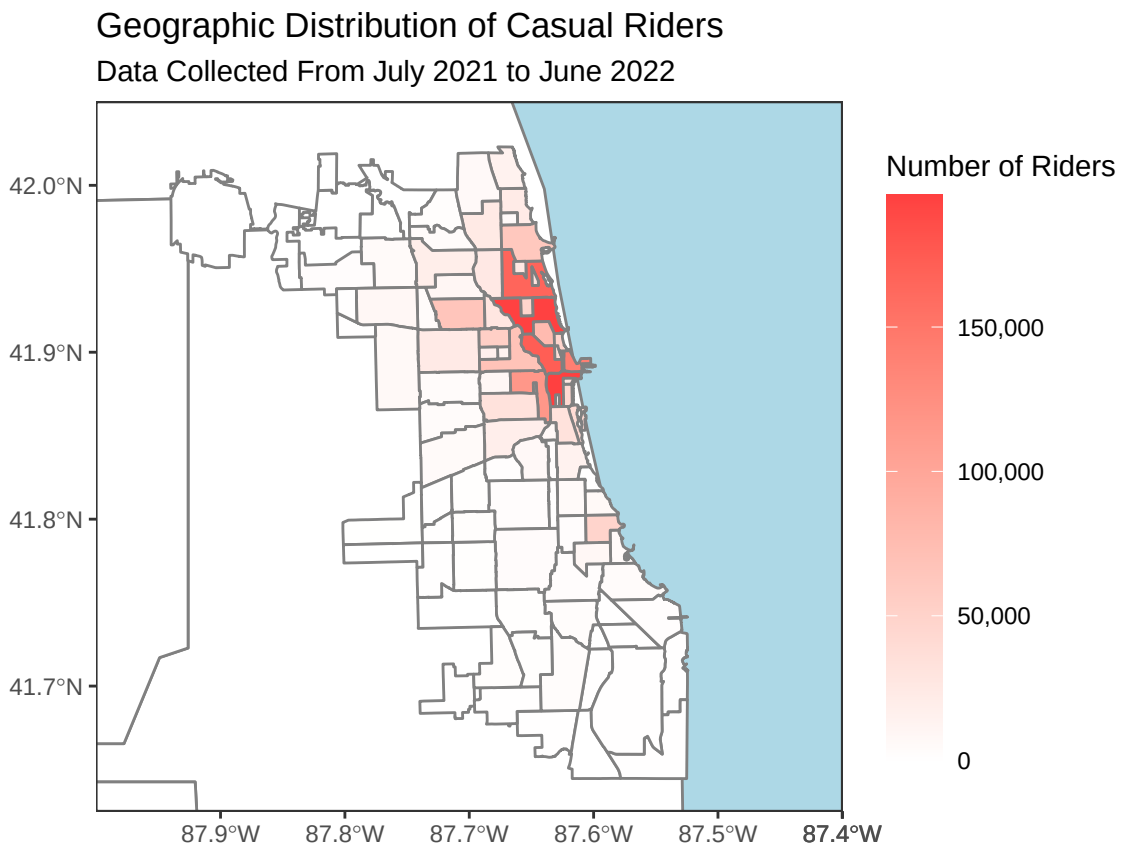
Geographic Distribution by Customer

Finally, it would be valuable to see the geographic breakdown of ridership by customer type.

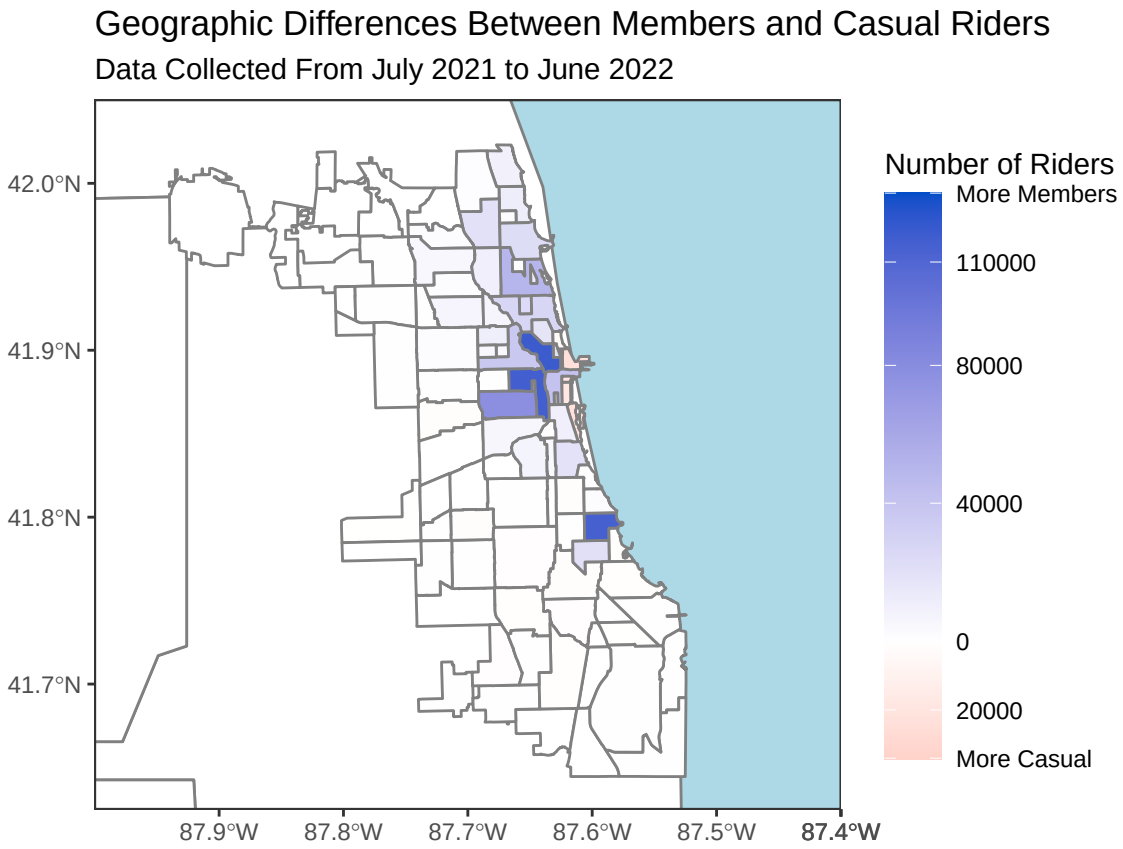
```
ggplot(data = world) +
  geom_sf(fill = 'lightblue') +
  geom_sf(data = counties, fill = 'white', color = gray(0.5)) +
  geom_sf(data = chicago_data, aes(fill = n_member), color = gray(0.5)) +
  coord_sf(xlim = c(-88, -87.40), ylim = c(41.625, 42.05), expand = FALSE) +
  labs(title = "Geographic Distribution of Members",
       subtitle = "Data Collected From July 2021 to June 2022",
       fill = "Number of Riders") +
  scale_fill_gradient(low = 'white', high = '#02adad', labels = scales::comma) +
  theme(legend.key.height = unit(1.5, 'cm'),
        legend.key.width = unit(0.75, 'cm'))
```




```
ggplot(data = world) +
  geom_sf(fill = 'lightblue') +
  geom_sf(data = counties, fill = 'white', color = gray(0.5)) +
  geom_sf(data = chicago_data, aes(fill = n_casual), color = gray(0.5)) +
  coord_sf(xlim = c(-88, -87.40), ylim = c(41.625, 42.05), expand = FALSE) +
  labs(title = "Geographic Distribution of Casual Riders",
       subtitle = "Data Collected From July 2021 to June 2022",
       fill = "Number of Riders") +
  scale_fill_gradient(low = 'white', high = '#ff4040', labels = scales::comma) +
  theme(legend.key.height = unit(1.5, 'cm'),
        legend.key.width = unit(0.75, 'cm'))
```




```
ggplot(data = world) +
  geom_sf(fill = 'lightblue') +
  geom_sf(data = counties, fill = 'white', color = gray(0.5)) +
  geom_sf(data = chicago_data, aes(fill = mc_difference), color = gray(0.5)) +
  coord_sf(xlim = c(-88, -87.40), ylim = c(41.625, 42.05), expand = FALSE) +
  labs(title = "Geographic Differences Between Members and Casual Riders",
       subtitle = "Data Collected From July 2021 to June 2022",
       fill = "Number of Riders") +
  scale_fill_gradient2(low = '#ff4040', mid = "white", high = '#024ec9',
                       limits = c(-34000, 130000),
                       breaks = c(-34000, -20000, 0, 40000, 80000, 110000, 130000),
                       labels = c("More Casual", 20000, 0, 40000, 80000, 110000, "More
                       ↪  Members")) +
  theme(legend.key.height = unit(1.5, 'cm'),
        legend.key.width = unit(0.75, 'cm'))
```



From the first two graphs, there appears to be slight differences in where the two groups of customers are concentrated, however the same general centralization around downtown Chicago is present for both groups. Lastly, if we look at the map of the differences between m embers and casual customers, we can see that members have significantly higher numbers in a few locations, with a general t rend o f m embers being the larger group in the downtown area. It is worth noting that there are a couple of neighborhoods with more casual customers along the coast in the downtown area. This may make sense as coasts tend to be more recreational areas, which fits with the hypothesis that casual customers use Cyclistic for recreational bike rides.

Recommendations

From this analysis, the top three recommendations are as follows:

1. **Identify how memberships can be beneficial for recreational users**

2. **Enhance membership value for less frequent riders**

3. **Increase advertising in areas with a heavier casual presence**