Bike_Trip_Report

By: Kai Jensen

7/28/2022

Cyclistic Data Analysis Case Study

Ask

Business Task

Using a years worth of Bike Trip data we will explore how Cyclistic members (Annual Members), and Non-members (Casual Riders) use the bike rental system differently. Using these findings we will make recommendations on designing marketing strategies aimed at converting casual riders into annual members.

Key Stakeholders

- Lily Moreno the director of marketing for Cyclistic
- Detail oriented Cyclistic executive team
- Cyclistic marketing analytics team

Prepare

The Data

We will be using the Divy Trips (https://divvy-tripdata.s3.amazonaws.com/index.html) data from 2019 for our analysis. The relevant data is stored in 4 ".csv" files which represent the 4 quarters of 2019.

The following code will load necessary libraries as well as importing the data onto the working environment.

```
library(tidyverse)
library(lubridate)

setwd("~/R") #  After downloading these csv files you can set the working directory using this l
ine. You will need to change the quoted text to reflect where you have stored these files.

Q1 <- read.csv("Divvy_Trips_2019_Q1.csv")
Q2 <- read.csv("Divvy_Trips_2019_Q2.csv")
Q3 <- read.csv("Divvy_Trips_2019_Q3.csv")
Q4 <- read.csv("Divvy_Trips_2019_Q4.csv")
```

Data Organization and Filtering

These 4 data frames all contain the same types of information however the column names in Q2 are not consistent. The following code will:

1. Change the column names of Q2.

2. Drop columns from all four dataframes that are not necessary for our analysis.
3. Combine the remaining data into a single dataframe ‘df’

```
# Change column names of Q2
colnames(Q2) <- colnames(Q1)

# Delete trip_id, bikeid, and gender from dataframes. This info is no useful for this task.

Q1 <- subset(Q1, select = -c(trip_id, bikeid, gender, birthyear))
Q2 <- subset(Q2, select = -c(trip_id, bikeid, gender, birthyear))
Q3 <- subset(Q3, select = -c(trip_id, bikeid, gender, birthyear))
Q4 <- subset(Q4, select = -c(trip_id, bikeid, gender, birthyear))


# Combine data into a single dataframe
df <- rbind(Q1, Q2, Q3, Q4)
```

Process

Check for Errors/Missing Data

The following prints a summary of df and counts the total number of NAs in each column. Note that start_time, end_time, and duration are all classified as character columns. When we use them in the future we will need to cast them as some other appropriate data type.

```
summary(df)
```

```
##   start_time          end_time          tripduration        from_station_id
##  Length:3818004     Length:3818004     Length:3818004     Min.   :  1.0
##  Class :character   Class :character   Class :character   1st Qu.: 77.0
##  Mode  :character   Mode  :character   Mode  :character   Median :174.0
##                                                           Mean   :201.7
##                                                           3rd Qu.:289.0
##                                                           Max.   :673.0
##  from_station_name  to_station_id   to_station_name      usertype
##  Length:3818004     Min.   :  1.0   Length:3818004     Length:3818004
##  Class :character   1st Qu.: 77.0   Class :character   Class :character
##  Mode  :character   Median :174.0   Mode  :character   Mode  :character
##                     Mean   :202.6
##                     3rd Qu.:291.0
##                     Max.   :673.0
```

```
# Count NA by column
df %>%
  select(everything()) %>%
  summarise_all(funs(sum(is.na(.))))
```

```
##   start_time end_time tripduration from_station_id from_station_name
## 1          0        0            0               0                 0
##   to_station_id to_station_name usertype
## 1             0               0        0
```

Their are no NAs in the data so it is almost ready to analyze.

Data Transformation

Since the trip duration column is measured in seconds we will create a new column trip_min that is numeric, and is measured in minutes.

```
# Create column representing trip duration in minutes
trip_min <- gsub(",","",df$tripduration)
trip_min <- as.numeric(trip_min) / 60
df$trip_min <- trip_min

summary(df$trip_min)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##      1.02      6.85     11.82     24.17     21.38 177140.00
```

Analyze

Data Organization and formatting

Notice that the max trip duration is very large, over 123 days. Since their is relatively a very small number of trips that last longer than two and a half hours we will filter out trips that last longer than 150 minutes.

We will also create columns for the day of the week and the month that the trip started on. These will help us compare different trends between casual riders and annual members.

```
# filter out trips that are longer than 150 min.
# adding weekday column
# adding month column

df_filtered <- df %>%
  filter(trip_min < 150) %>%
  mutate(wkday = wday(as_datetime(start_time), label = TRUE)) %>%
  mutate(month = month(as_datetime(start_time), label = TRUE))

summary(df_filtered)
```

```
##   start_time          end_time         tripduration       from_station_id
##  Length:3793927     Length:3793927     Length:3793927     Min.   :  1.0
##  Class :character   Class :character   Class :character   1st Qu.: 77.0
##  Mode  :character   Mode  :character   Mode  :character   Median :174.0
##                                                           Mean   :201.6
##                                                           3rd Qu.:289.0
##                                                           Max.   :673.0
##
##  from_station_name  to_station_id   to_station_name      usertype
##  Length:3793927     Min.   :  1.0   Length:3793927     Length:3793927
##  Class :character   1st Qu.: 77.0   Class :character   Class :character
##  Mode  :character   Median :174.0   Mode  :character   Mode  :character
##                     Mean   :202.6
##                     3rd Qu.:291.0
##                     Max.   :673.0
##
##     trip_min        wkday            month
##  Min.   :  1.017   Sun:421748   Aug    :584898
##  1st Qu.:  6.833   Mon:557476   Jul    :552553
##  Median : 11.733   Tue:583151   Sep    :490126
##  Mean   : 17.365   Wed:581567   Jun    :471906
##  3rd Qu.: 21.117   Thu:585463   Oct    :369961
##  Max.   :149.983   Fri:574855   May    :365202
##                    Sat:489667   (Other):959281
```

This eliminates approximately .63% of total trips so it should not affect our analysis.

Identifying Trends

To begin to compare Annual Members to Casual riders we calculate several summary statistics.

```
# Average trip length in minutes separated by user type.
aggregate(df_filtered$trip_min ~ df_filtered$usertype, FUN= mean)
```

```
##   df_filtered$usertype df_filtered$trip_min
## 1             Customer             33.88194
## 2           Subscriber             12.52655
```

```
# Median trip length in minutes separated by user type.
aggregate(df_filtered$trip_min ~ df_filtered$usertype, FUN= median)
```

```
##   df_filtered$usertype df_filtered$trip_min
## 1             Customer            25.300000
## 2           Subscriber             9.783333
```

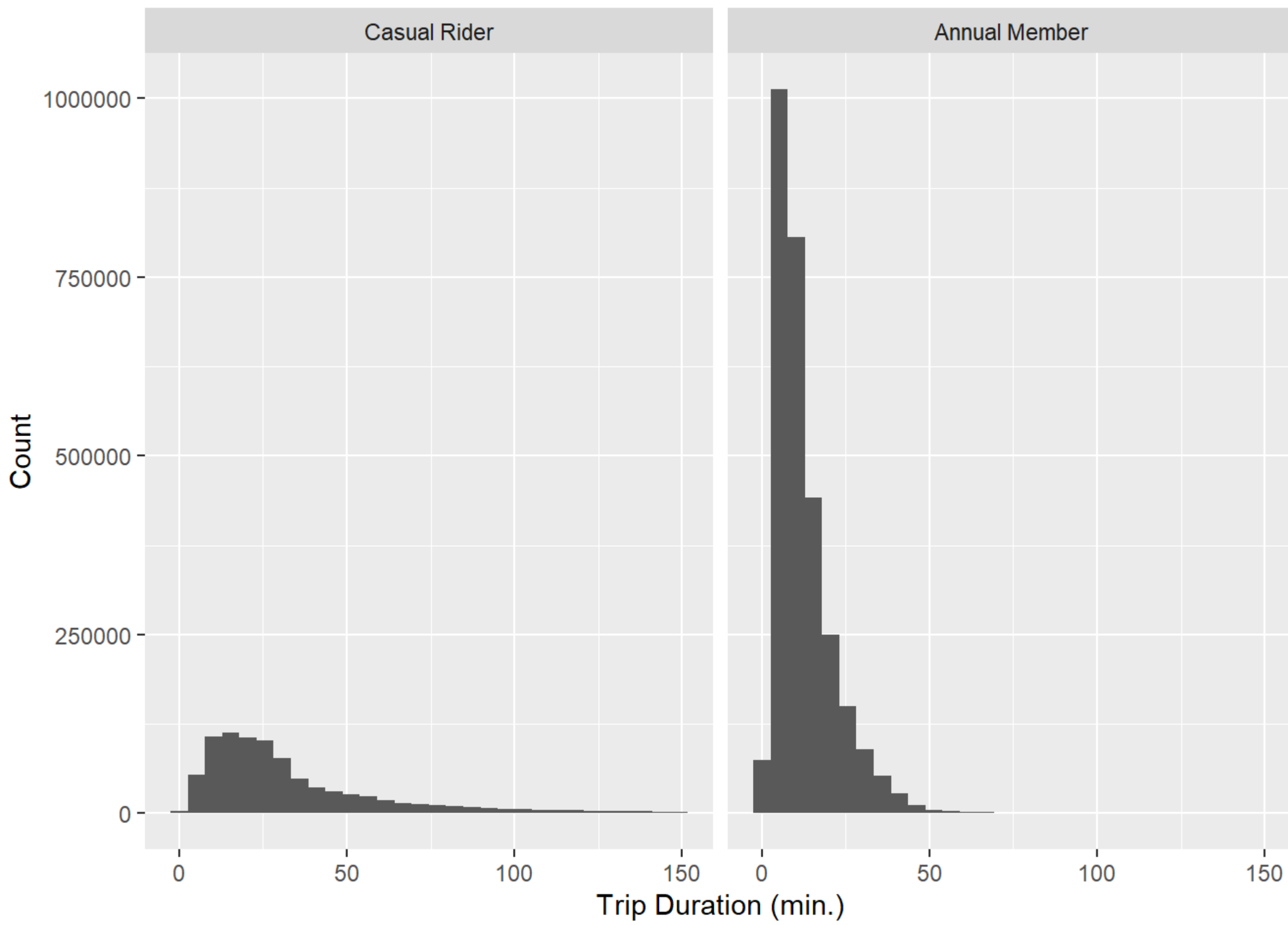
Based on these summary statistics we can see a clear difference between the 2 categories of bike users. We will explore these further using some visualizations.

Share

Visualizations

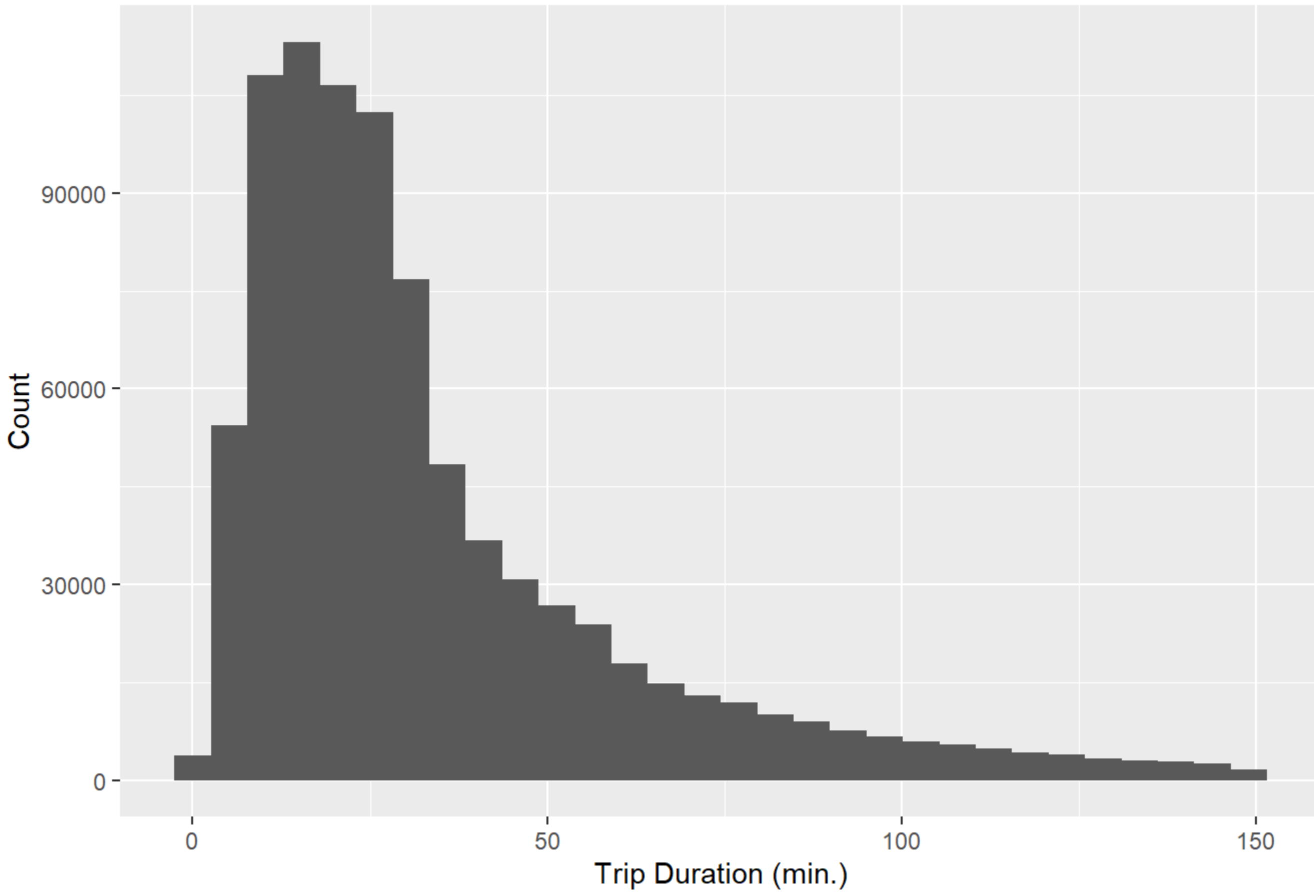
We will first look at the differences between trip length for the 2 types of riders.

```
# histogram of trip duration split by usertype
labels <- c("Customer" = "Casual Rider", "Subscriber" = "Annual Member")
ggplot(data=df_filtered, aes(x=trip_min))+
  geom_histogram(bins = 30)+
  labs(x = "Trip Duration (min.)",
       y = "Count") +
  facet_wrap(~usertype,
             labeller = labeller(usertype = labels)
             )
```

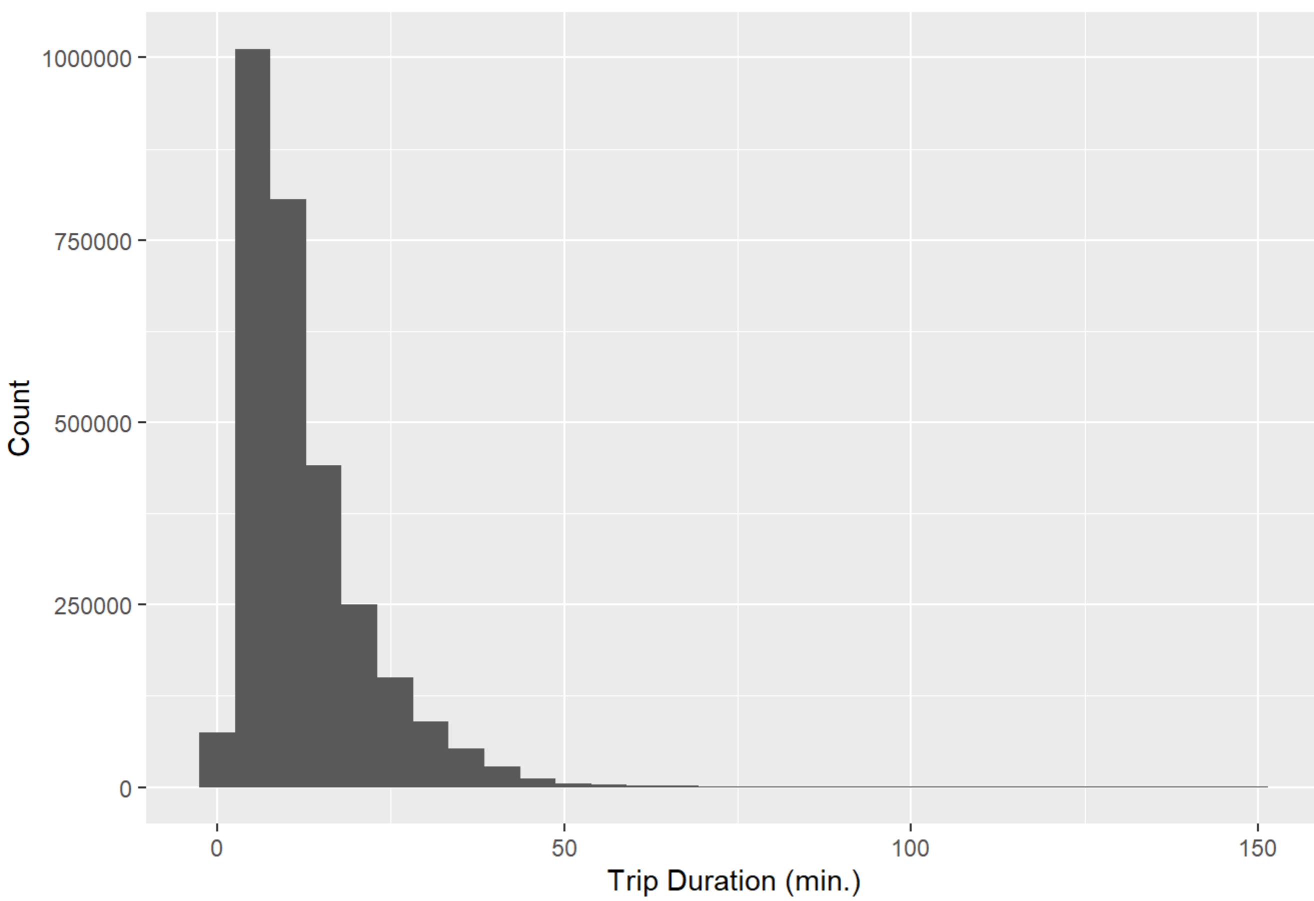


It is difficult to gain much insight into this graph because their are quite a few more annual members than casual riders so we can create seperate plots for each type of rider. We will also skip printing the R code for the remainder of the visualizations, so we can focus on analyzing the graphs.

Histogram of Trip Duration for Casual Riders



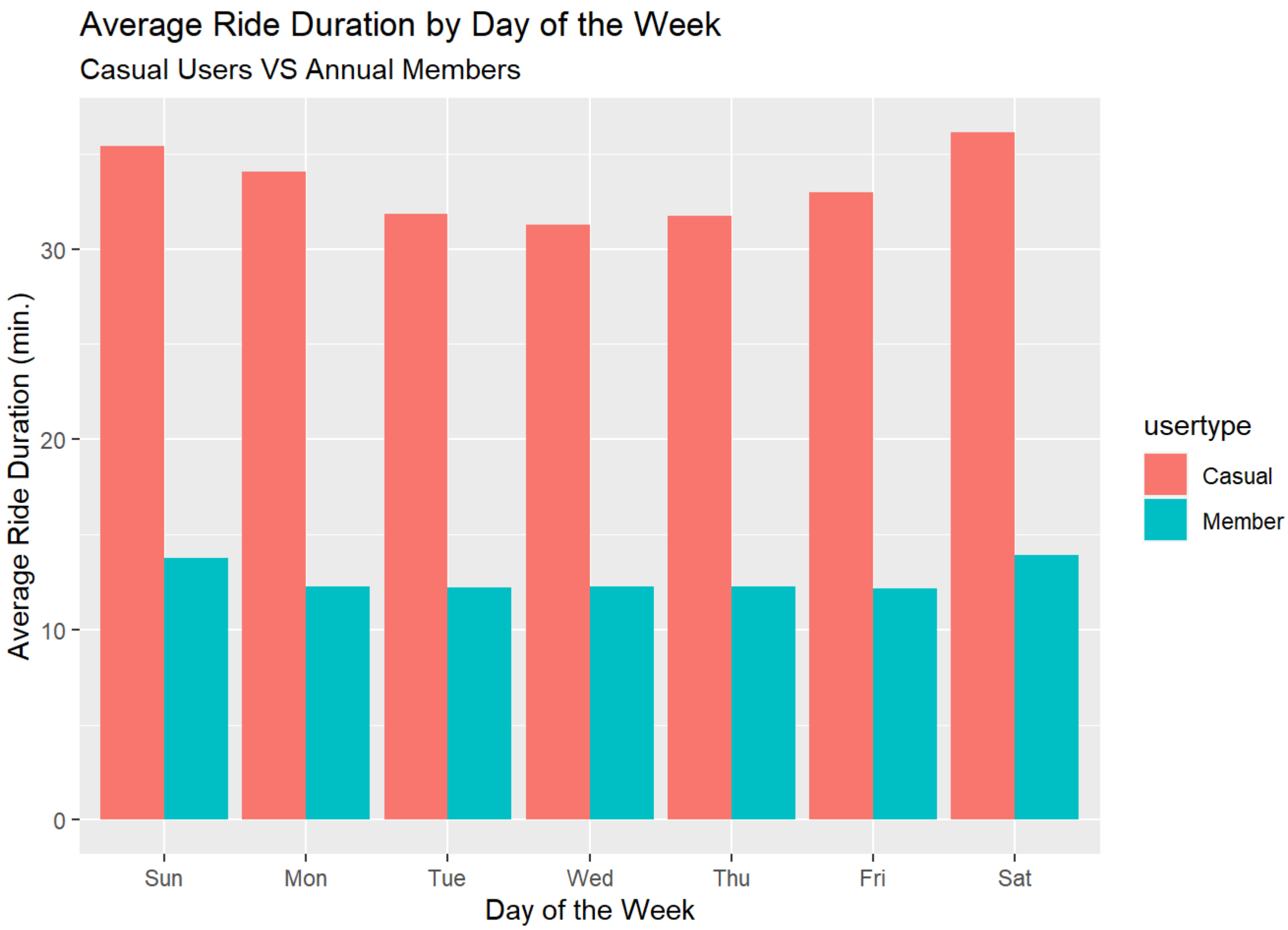
Histogram of Trip Duration for Annual Members



We can see that Annual members are much more likely to take shorter rides and the majority of these rides are less than 20 minutes.

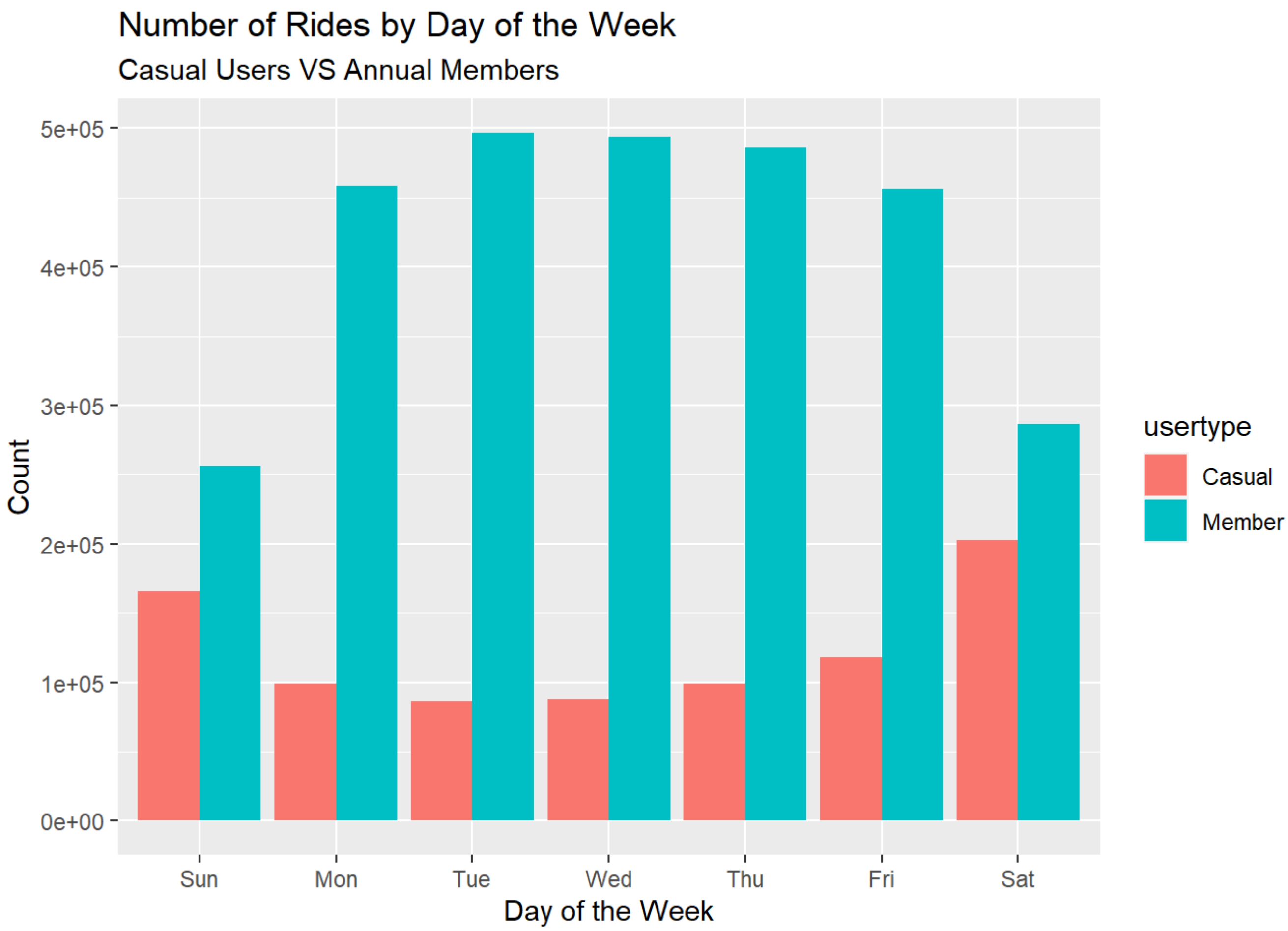
Casual rides are mostly grouped between 15 and 30 minutes however their is a significant number of rides that are longer than an hour. A possible explanation for this being that annual members are using the bikes as commuter vehicles and casual riders are using them for recreation.

To help visualize how these 2 types of riders use this service differently throughout the week we can look at the average ride duration for each group separated by the day of the week.



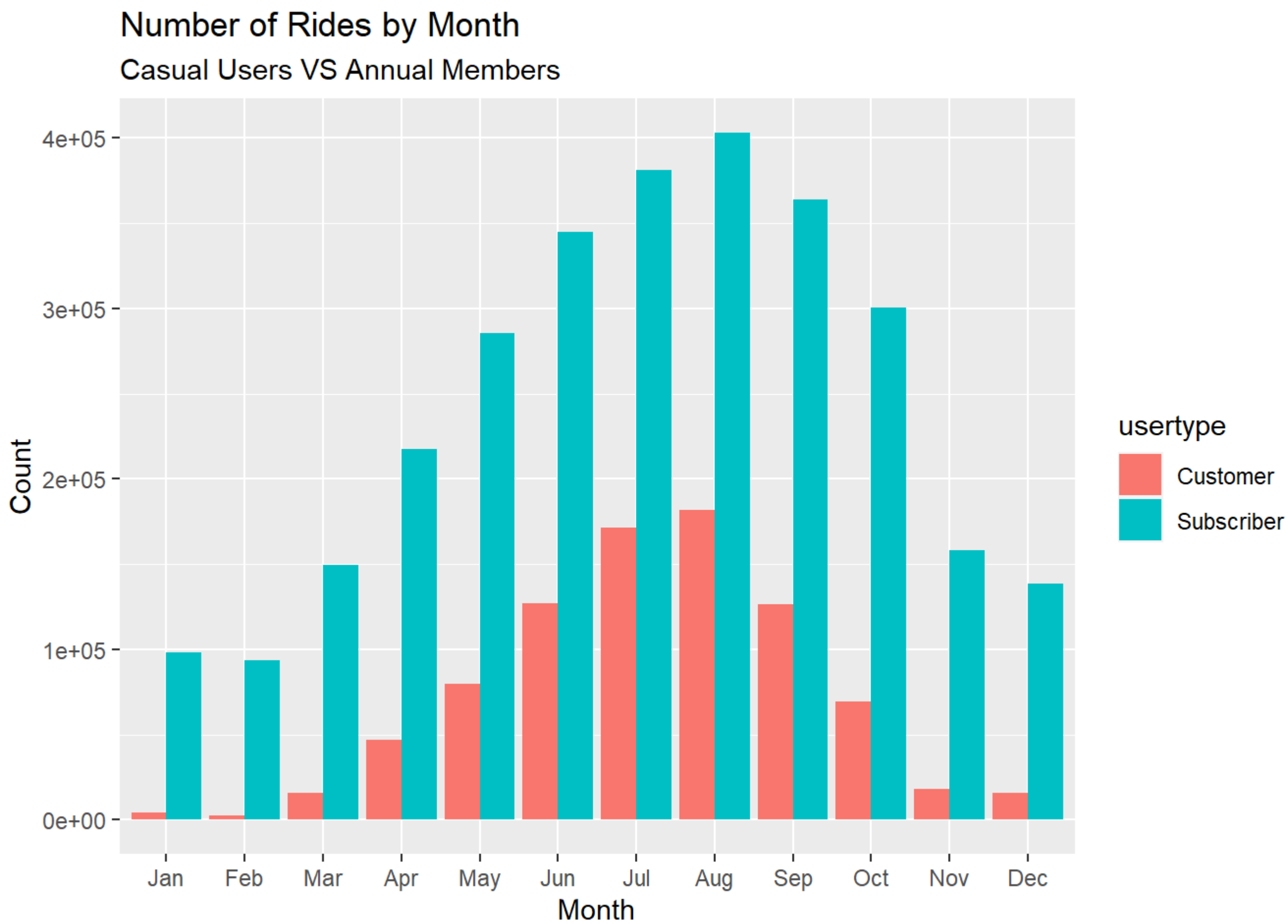
We can see that casual riders average ride length across the week is more than twice that of members. Also, average ride length dips slightly in the middle of the week for both user types. However this dip is slightly more significant for the casual riders.

Looking at the total number of rides taken on each day of the week for each category will help us see their difference more explicitly.



Notice that the annual members have a consistently high count Monday through Friday and their usage drops off on the weekends, while the casual riders are more likely to ride on the weekend. This helps to confirm our earlier suspicion that annual members use the bikes as commuter vehicles and casual riders use the bikes mostly for recreation.

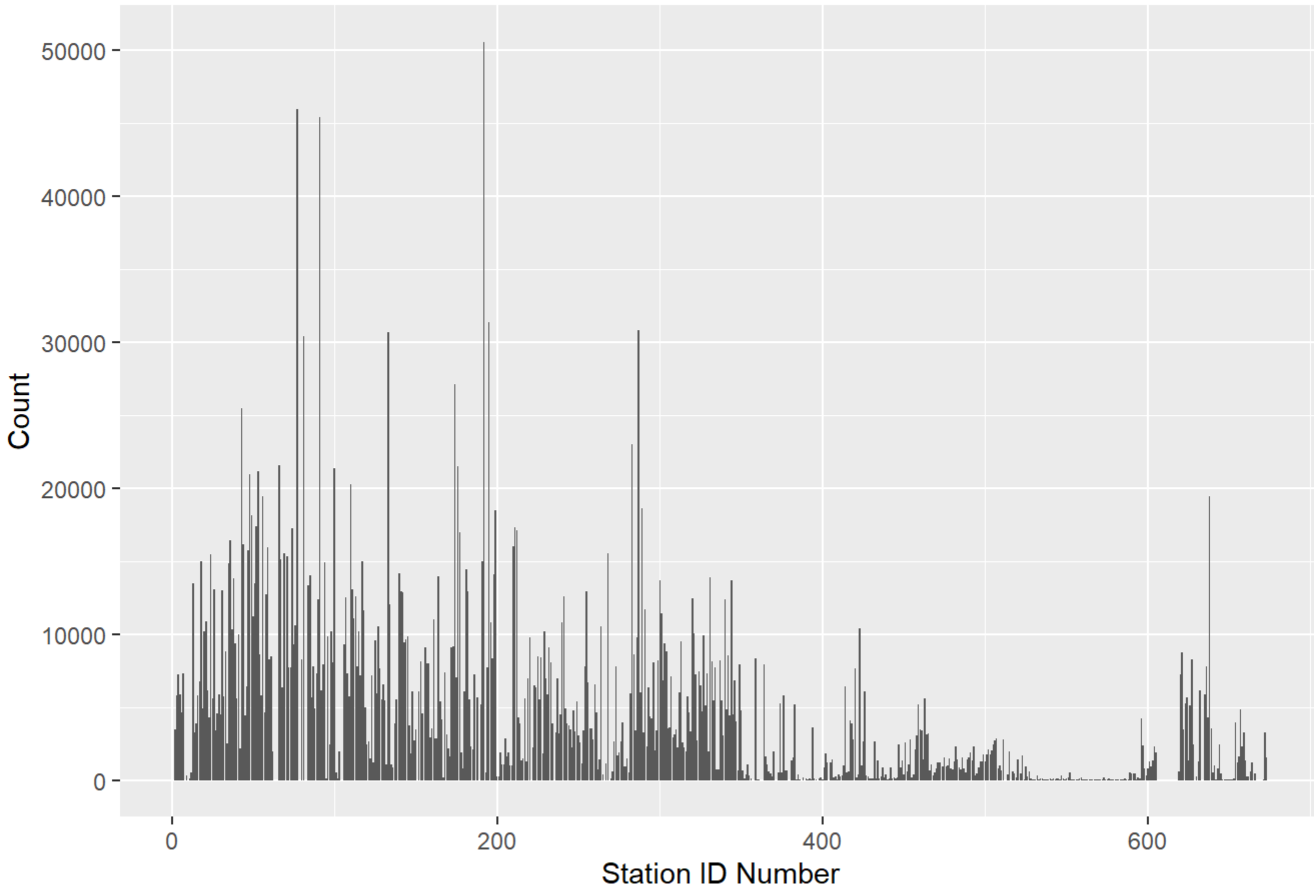
We can also break down the number of rides per month to find a possible trend in bike use.



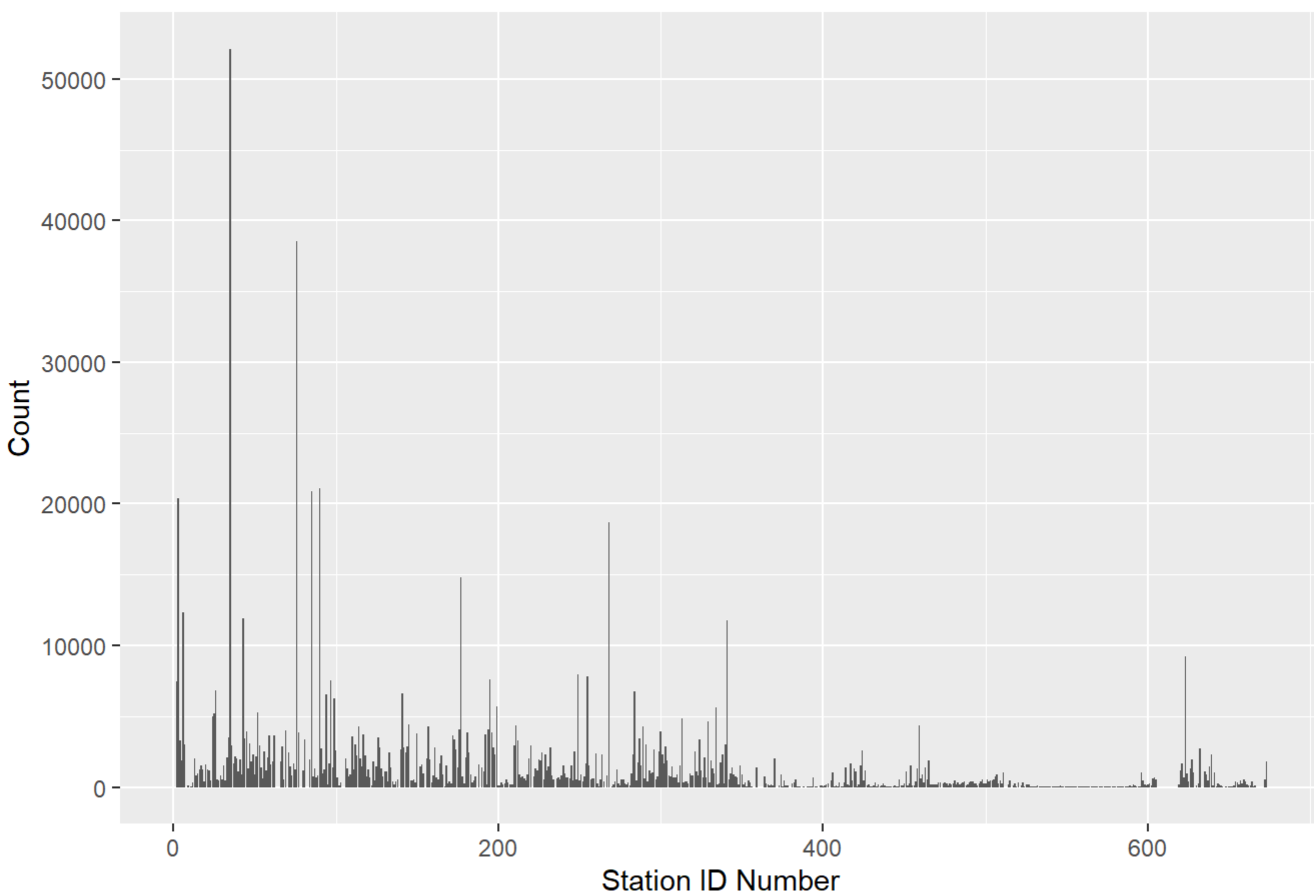
We see that the shape of the distribution is similar for both types of rider. However, during the winter months the number of annual member trips makes up more than 90% of all rides but during the summer peak months the percentage of casual rider trips grows close to 30% of all trips. This further points towards casual riders not relying on these bikes to commute to work.

It may also be helpful to consider where each type of rider is starting/ending their trip. Each bike station has a unique ID number. For the next graphs this is how we will divide the x axis. Since their are over 600 of these stations the following graphs will only be helpful to illustrate any possible trends. We will not pick out specific stations IDs yet.

Start Station Count for Annual members



Start Station Count for Casual Riders



From these graphs we can see that the casual rider traffic is much more centralized around several stations than the annual member traffic. So, focusing our marketing around these stations should help maximize its effectiveness. The following tables list the top 20 stations used for each user type.

```
##                         Annual Members frequency
## 1                  Canal St & Adams St     50540
## 2              Clinton St & Madison St     45952
## 3         Clinton St & Washington Blvd     45349
## 4            Columbus Dr & Randolph St     31347
## 5              Franklin St & Monroe St     30813
## 6             Kingsbury St & Kinzie St     30626
## 7                  Daley Center Plaza     30388
## 8                Canal St & Madison St     27114
## 9         Michigan Ave & Washington St     25447
## 10           LaSalle St & Jackson Blvd     23010
## 11                Clinton St & Lake St     21551
## 12                   Clark St & Elm St     21461
## 13 Orleans St & Merchandise Mart Plaza     21335
## 14                  Wells St & Huron St     21121
## 15          Larrabee St & Kingsbury St     20897
## 16               Dearborn St & Erie St     20210
## 17            Desplaines St & Kinzie St     19413
## 18               Wells St & Concord Ln     18605
## 19              Wabash Ave & Grand Ave     18451
## 20             Dearborn St & Monroe St     18131
```

```
##                  Casual Riders frequency
## 1       Streeter Dr & Grand Ave     52138
## 2     Lake Shore Dr & Monroe St     38545
## 3               Millennium Park     21125
## 4         Michigan Ave & Oak St     20859
## 5                Shedd Aquarium     20353
## 6    Lake Shore Dr & North Blvd     18665
## 7            Theater on the Lake     14805
## 8                Dusable Harbor     12308
## 9  Michigan Ave & Washington St     11915
## 10            Adler Planetarium     11740
## 11        Michigan Ave & 8th St      9209
## 12              Montrose Harbor      7938
## 13   Indiana Ave & Roosevelt Rd      7828
## 14    Columbus Dr & Randolph St      7585
## 15                 Field Museum      7531
## 16     McClurg Ct & Illinois St      6798
## 17  Michigan Ave & Jackson Blvd      6724
## 18       Clark St & Lincoln Ave      6578
## 19      Clark St & Armitage Ave      6503
## 20      Lake Shore Dr & Ohio St      6232
```

Findings

- Casual rider trip duration is more than twice as long as annual members on average.

- Annual members use the bike service primarily to commute to and from work, while casual riders use the service for recreation.
- This means that casual riders mostly use the bikes on weekends.
- The majority of casual rides happen between May and October.
- Many of the casual rides originate from the same 10 - 11 stations.

Act

Recomendations

- Plan marketing campaign to start in the spring and go throughout the summer.
- Partner with a map/directions app to create unique sight seeing routes around town. One for every week during the peak months.
    - These would need to be randomly shown to customers to prevent supply shortage of bikes at specific locations.
- Offer a type of subscription that costs slightly less but benefits casual riders.
    a. A subscription that is only good on Friday, Saturday, and Sunday.
    b. A subscription that you can opt into/out of at the beginning of each season.