Cyclistic Case Study

Brady Fisher

9/15/2023

Introduction

The Capstone Project of the Google Data Analytics course is to complete a case study. For this case study, I am acting as a junior data analyst. I am working with a fictional Bike-Sharing company in Chicago called Cyclistic. Cyclistic features more than 5,800 bicycles and 600 docking stations. Cyclistic's current marketing strategy relies on appealing to broad consumer segments. In alignment with this marketing strategy Cyclistic offers flexible pricing plans including single-ride passes, full-day passes, and annual memberships. Single-ride and full-day pass customers are referred to as **casual riders**. Customers who purchase annual memberships are referred to as **Cyclistic members**.

Although the pricing flexibility helps Cyclistic attract more customers, Lily Moreno, the director of marketing for Cyclistic, believes maximizing the number of annual memberships will lead to the best success for the company. Therefore my team is tasked with supplying powerful data observations and data visualizations that provide insight on the difference between casual riders and annual members. From these insights a new marketing strategy will be designed and shared to convert casual riders into annual members.

In summary, the problem for the business is to design a new marketing strategy with the intent of converting casual riders into annual members. Three questions were raised to help solve this problem. For the extent of this project I will only be focusing on the first question.

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Business Task and Stakeholders

For this case study, the business task is to analyze Cyclistic bike data from the previous 12 months to identify patterns in how annual members and casual riders use Cyclistic bikes differently.

Lily Moreno and the Cyclistic executive team will be considered the primary stakeholders for this project. The Cyclistic marketing analytics team will be considered secondary stakeholders.

Preparing the Data

For this project, I will be analyzing historical Cyclistic bike trip data. This Data is open source and has been made publicly available by Motivate Intentional Inc. This data also does not include personally identifiable information of the riders to ensure their privacy.

It is important to ensure a data source is credible and lacks bias before it is used in analysis. This data was collected by Cyclistic themselves meaning it is first-party data. With this being the Cyclistic team's own data there is a low chance of bias and high credibility in the data. The ROCCC data system is another test that can be used to determine the confidence in using a data source. This data source passes the ROCCC test because it is reliable, original, comprehensive, current, and cited.

Process the Data

At the time of this analysis, this public data set has data ranging from January 2013 to June 2023, and is available here. In order to keep our analysis current while also maintaining a substantial amount of data, we will be limiting the scope to the past 12 months.

I will now document the steps I am taking to process the data. All the available data is organized into ZIP files, so first I download the following 12 ZIP files containing the data from July 2022 to June 2023.

202207-divvy-tripdata.zip	Aug 5th 2022, 05:27:33 pm	29.51 MB	ZIP file
202208-divvy-tripdata.zip	Sep 8th 2022, 05:20:19 pm	27.13 MB	ZIP file
202209-divvy-tripdata.zip	Oct 11th 2022, 10:59:39 am	25.31 MB	ZIP file
202210-divvy-tripdata.zip	Nov 8th 2022, 04:47:10 pm	20.08 MB	ZIP file
202211-divvy-tripdata.zip	Dec 5th 2022, 12:17:32 pm	12.36 MB	ZIP file
202212-divvy-tripdata.zip	Jan 3rd 2023, 02:19:01 pm	6.75 MB	ZIP file
202301-divvy-tripdata.zip	Feb 7th 2023, 01:58:38 pm	6.78 MB	ZIP file
202302-divvy-tripdata.zip	Mar 7th 2023, 04:28:12 pm	7.08 MB	ZIP file
202303-divvy-tripdata.zip	Apr 6th 2023, 03:38:59 pm	10.27 MB	ZIP file
202304-divvy-tripdata.zip	May 4th 2023, 02:43:25 pm	15.40 MB	ZIP file
202305-divvy-tripdata.zip	Jun 8th 2023, 05:17:13 pm	23.44 MB	ZIP file
202306-divvy-tripdata.zip	Jul 13th 2023, 04:22:44 pm	25.66 MB	ZIP file

Next I unzip each file I downloaded. Each ZIP file contains a folder and a CSV (comma-separated values) file with a name containing the year and month of the data. For example the data for June 2023 is named "202306-divvy-tripdata". In totality our data set for the scope of this project contains the following 12 CSV files:

Name	Status	Date modified	Type	Size
202207-divvy-tripdata		8/13/2023 7:57 PM	CSV File	149,306 KB
202208-divvy-tripdata		8/13/2023 7:57 PM	CSV File	142,148 KB
202209-divvy-publictripdata		8/13/2023 7:57 PM	CSV File	138,135 KB
202210-divvy-tripdata		8/13/2023 7:57 PM	CSV File	109,293 KB
202211-divvy-tripdata		8/13/2023 7:57 PM	CSV File	66,348 KB
202212-divvy-tripdata		8/13/2023 7:57 PM	CSV File	35,612 KB
202301-divvy-tripdata		8/13/2023 7:57 PM	CSV File	37,551 KB
202302-divvy-tripdata		8/13/2023 7:57 PM	CSV File	37,691 KB
202303-divvy-tripdata		8/13/2023 7:57 PM	CSV File	51,112 KB
202304-divvy-tripdata		8/13/2023 7:57 PM	CSV File	83,762 KB
202305-divvy-tripdata		8/13/2023 7:58 PM	CSV File	118,966 KB
202306-divvy-tripdata		8/13/2023 7:58 PM	CSV File	140,974 KB

Next, I create a folder on my Desktop called Cyclistic Case Study Data. Now, within that folder I make two sub-folders called Original Cyclistic Data and Analyzing Cyclistic Data. I save a copy of all 12 pre-processed CSV files into the Original Cyclistic Data folder. It is important to keep a copy of the original data like this, so that if data is lost or corrupted later in the cleaning or analyzing steps there is data to start over with.

Next, I open each CSV file in Excel and save them as Excel Workbook files in the Analyzing Cyclistic Data folder.

The data in each of these files is organized into rows and columns, this is known as structured data. Each row represents one bike trip record. Each column of the table represents different fields or pieces of information for each record.

Here are the column names and brief description of what they represent:

Column Name	Description
ride_id	Identification number of the bike trip.
rideable_type	Type of Bike used on the bike trip. (classic, electric or docked bike)
started_at	Starting time of the bike trip.
ended_at	Ending time of the bike trip.
start_station_name	Station the bike trip started at.
start_station_id	Identification number of the starting station.
end_station_name	Station the bike trip ended at.
end_station_id	Identification number of the ending station.
start_lat	Latitude of the start of the bike trip.
start_lng	Longitude of the start of the bike trip.
end_lat	Latitude of the end of the bike trip.
end_lng	Longitude of the end of the bike trip.
member_casual	The type of rider. (member or casual rider)

Right away as I open the Excel files, I see there are some changes or manipulations I want to make to the data to make it more organized and easier to read. I will now make the following manipulations:

1. Changing the format for column C by using Format > More Number Formats > Date > 3/7/01 12:00 AM.
2. Changing the format for column D by using Format > More Number Formats > Date > 3/7/01 12:00 AM.

I will now look over my data and ensure it is clean. For each of the 12 Google sheets files I am checking the following:

1. That there is no missing data. In order to do this I will use the filter tool on first row of data which contains the column names. One by one, I will filter each column by "blank" cells to find any missing data. As a result I see that there is data only missing from the start_station_name, start_station_id, end_station_name, end_station_id, end_lat, and end_lng columns. I will keep this in mind in my analysis, but do not expect it to cause any issue as those columns might not give the best insight in the analysis.

2. That all the data is in the correct data type and has a value that makes sense. It is important that none of the data is incorrect, so that calculations and analysis on the data will be accurate. In order to do this I will check each column to ensure they only contain values that are valid for that field. In this case I will check that columns that should contain dates only contain dates, columns that contain numbers only contain numbers and columns that have a limited number of values only contain those values. As a result, I see that the data types and values all make sense.

For my analysis, I want to add a couple of fields or columns to the data. I will add a "ride_length" column which expresses the amount of time each ride took. A "day_of_week" column will also be added to the data to represent what day of the week each ride took place on. Both of these fields are easy to extract from the data we currently have and could provide great insight on the difference between annual members and casual riders. I will also check that all the values in the cells are logical.

Column D contains the end date and time of the ride and column C contains the beginning date and time of the ride each bike ride. Therefore, to add a "ride_length" column to all 12 data sets in column N, I am writing the formula "=HOUR(D2-C2)x3600+MINUTE(D2-C2)x60+SECOND(D2-C2)" in cell N2. This will give the length of each bike ride in seconds. I will now fill in the formula for the rest of rows by clicking on the bottom right corner of cell N2. Using Format > General, I manipulate column N to be in an appropriate format. Now, I will check that all the values in the cells are logical. In this case, I will check that each value is positive, since you cannot have a bike ride for a negative amount of time. There are a couple of these values in each data set, and I deleted the whole row for these instances.

Next, I will add a "day_of_week" column to all 12 data sets in column O by utilizing the WEEKDAY command. The WEEKDAY command produces an integer that represents a day of the week when given a date as a parameter, so this formula will result in a 1 for Sunday, 2 for Monday, 3 for Tuesday, and all the way to 7 for Saturday. In cell O2, I write the formula "=WEEKDAY(C2,1))" and fill in the formula to the rest of rows by clicking on the bottom right corner of cell N2. I will now use Format as General on the column, since the column was originally showing in a Time Format. Now, I will check that each cell makes logical sense. I confirmed that each cell in the column is an integer from 1 to 7, meaning each cell is properly represented by a day of the week.

Analyze the Data

To better analyze the Cyclistic Data that I have processed and cleaned, I work in R Studio which uses the programming tool R. R is a free open source programming language that can process a lot of data quickly, create easily reproducible and shareable analysis, and create high quality visualizations. For these reasons R is very popular among Data Analysts.

Install R Packages

To begin, I will install some packages in R that will help me with my analysis. I will download the readxl, tidyverse, lubridate and ggplot2 as they will help import, wrangle and visualize the Cyclistic Data.

```
library(readxl)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Import Cyclistic Data

Now I will import the Cyclistic Data I have already processed.

```
Jul2022 <- read_excel("202207-divvy-tripdata.xlsx")
Aug2022 <- read_excel("202208-divvy-tripdata.xlsx")
Sep2022 <- read_excel("202209-divvy-tripdata.xlsx")
Oct2022 <- read_excel("202210-divvy-tripdata.xlsx")
Nov2022 <- read_excel("202211-divvy-tripdata.xlsx")
Dec2022 <- read_excel("202212-divvy-tripdata.xlsx")
Jan2023 <- read_excel("202301-divvy-tripdata.xlsx")
Feb2023 <- read_excel("202302-divvy-tripdata.xlsx")
Mar2023 <- read_excel("202303-divvy-tripdata.xlsx")
Apr2023 <- read_excel("202304-divvy-tripdata.xlsx")
May2023 <- read_excel("202305-divvy-tripdata.xlsx")
Jun2023 <- read_excel("202306-divvy-tripdata.xlsx")
```

Combine into a Single Data Set

To combine all these Data sets into 1 data set we must make sure all the column names are the same and have the same type. Using the str function in R we can see the column names and types for each of the month data sets. Here is an example for the month of July 2022.

```
options(width=60)
str(Jul2022)
```

```
## tibble [823,482 x 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:823482] "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6
##  $ rideable_type     : chr [1:823482] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:823482], format: "2022-07-05 08:12:47" ...
##  $ ended_at          : POSIXct[1:823482], format: "2022-07-05 08:24:32" ...
##  $ start_station_name: chr [1:823482] "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buc
##  $ start_station_id  : chr [1:823482] "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr [1:823482] "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan A
##  $ end_station_id    : chr [1:823482] "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num [1:823482] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:823482] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:823482] 41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num [1:823482] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:823482] "member" "casual" "casual" "casual" ...
##  $ ride_length       : num [1:823482] 705 113 463 3509 1578 ...
##  $ day_of_week       : num [1:823482] 3 3 1 1 4 6 2 5 1 1 ...
```

```
str(Jul2022)
str(Aug2022)
str(Sep2022)
str(Oct2022)
str(Nov2022)
str(Dec2022)
str(Jan2023)
str(Feb2023)
str(Mar2023)
str(Apr2023)
str(May2023)
str(Jun2023)
```

Looking at each month, I see that the month of September 2022 needs its end_station_id column converted from a num type to a chr type.

```
Sep2022$end_station_id <- as.character(Sep2022$end_station_id)
```

Now that all the column names and types are matching, I will combine them into a single data set "CompleteDataSet".

```
CompleteDataSet <- bind_rows(Jul2022,Aug2022,Sep2022,Oct2022,Nov2022,Dec2022,
                             Jan2023,Feb2023,Mar2023,Apr2023,May2023,Jun2023)
```

In this data set, the day_of_week column is numeric with 1 representing Sunday, 2 representing Monday all the way to 7 representing Saturday. I will now convert these numbers into the day of the week they represent. I will also add a "Months" column to signify the month of the bike ride.

```
CompleteDataSet$day_of_week <- recode(CompleteDataSet$day_of_week,
       "1"="Sunday",
       "2"="Monday",
       "3"="Tuesday",
       "4"="Wednesday",
       "5"="Thursday",
       "6"="Friday",
       "7"="Saturday")
CompleteDataSet <- within(CompleteDataSet,
                        Month <- month.abb[month(CompleteDataSet$started_at)])
```

We can now see the day_of_week column has been converted and the Month column has been added.

```
head(CompleteDataSet[15:16])
```

```
## # A tibble: 6 x 2
##   day_of_week Month
##   <chr>       <chr>
## 1 Tuesday     Jul
## 2 Tuesday     Jul
## 3 Sunday      Jul
## 4 Sunday      Jul
## 5 Wednesday   Jul
## 6 Friday      Jul
```

Now to begin my analysis, I will look at how many of our bike trip observation are from **casual riders** and how many are from **Cyclistic members**.

```
table(CompleteDataSet$member_casual)
```

```
##
##  casual  member
## 2244212 3535153
```

From this I see that there are 2,244,212 casual rider trips and 3,535,153 member trips.

Next I will compare the ride length of Cyclistic members and casual riders. In particular, I will compare their mean, median, and max ride length in seconds.

```
aggregate(ride_length ~ member_casual,data = CompleteDataSet, mean)
```

```
##   member_casual ride_length
## 1        casual   1263.6121
## 2        member    723.7589
```

```
aggregate(ride_length ~ member_casual,data = CompleteDataSet, median)
```

```
##   member_casual ride_length
## 1        casual         720
## 2        member         513
```

```r
aggregate(ride_length ~ member_casual,data = CompleteDataSet, max)
```

```
##   member_casual ride_length
## 1        casual       86396
## 2        member       86390
```

This analysis shows that the mean and median ride length for casual riders is greater than members. On average, casual riders ride for 1,264 seconds or 21 minutes compared to 724 seconds or 12 minutes for members. The maximum ride length for both casual riders and members are similar at around 86,390 seconds or 24 days.

Next I will create a table that shows how the average ride length differs by the day of the week and member type. I will also create a table that looks at how the number of rides change throughout the week for each group.

```r
Table1 <- aggregate(ride_length ~ member_casual + day_of_week,
                    data = CompleteDataSet, NROW)
Table2 <- aggregate(ride_length ~ member_casual + day_of_week,
                    data = CompleteDataSet, mean)
colnames(Table1)[3] = "number_of_rides"
Table1
```

```
##    member_casual day_of_week number_of_rides
## 1         casual      Friday          347107
## 2         member      Friday          518876
## 3         casual      Monday          253085
## 4         member      Monday          477618
## 5         casual    Saturday          459983
## 6         member    Saturday          464391
## 7         casual      Sunday          351385
## 8         member      Sunday          387924
## 9         casual    Thursday          298278
## 10        member    Thursday          565672
## 11        casual     Tuesday          257098
## 12        member     Tuesday          549032
## 13        casual   Wednesday          277263
## 14        member   Wednesday          571625
```

```r
Table2
```

```
##    member_casual day_of_week ride_length
## 1         casual      Friday   1220.6544
## 2         member      Friday    717.6161
## 3         casual      Monday   1248.5749
## 4         member      Monday    689.2322
## 5         casual    Saturday   1446.3811
## 6         member    Saturday    812.7388
## 7         casual      Sunday   1460.2654
## 8         member      Sunday    798.9351
## 9         casual    Thursday   1108.3113
## 10        member    Thursday    696.9066
## 11        casual     Tuesday   1113.4473
```

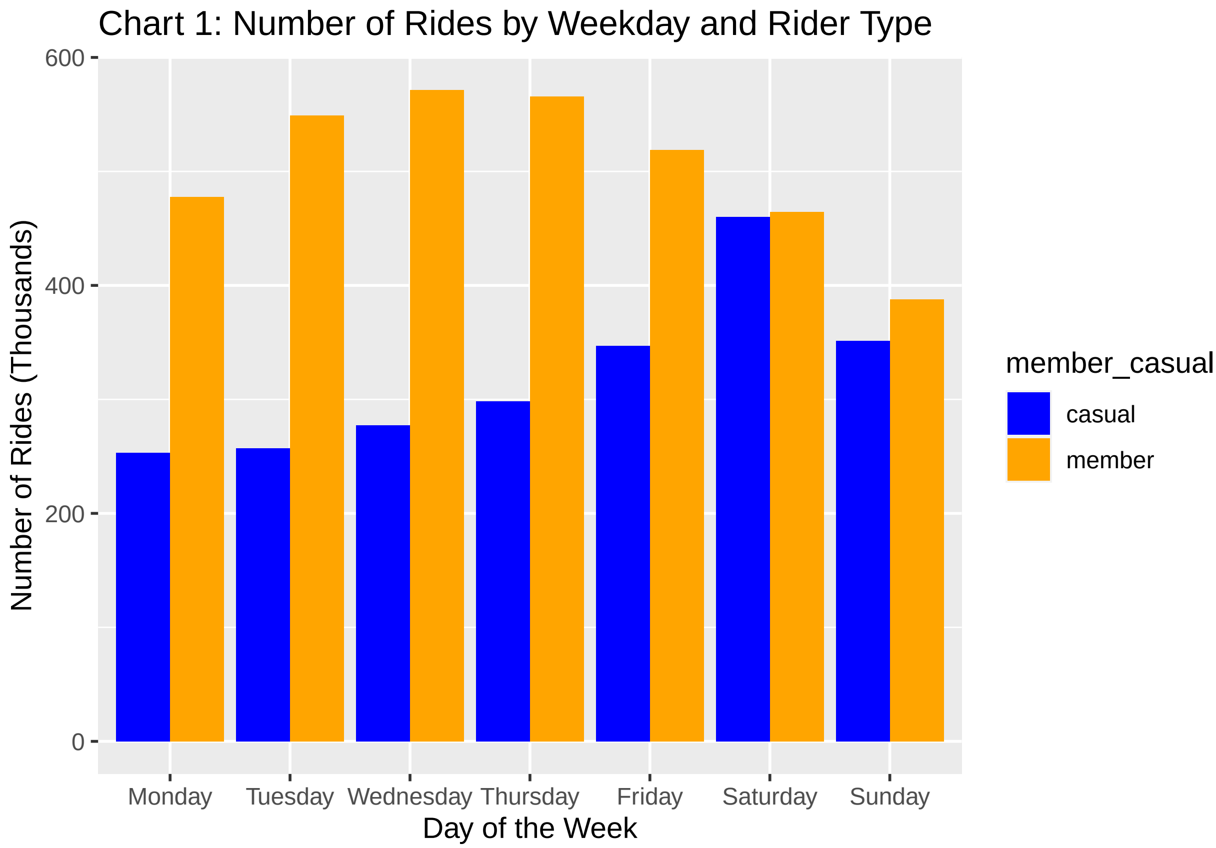
```
## 12         member     Tuesday    690.0666
## 13         casual   Wednesday   1084.9906
## 14         member   Wednesday    693.8121
```

The first table shows that casual riders have the least amount of rides on Monday and that number gradually increases throughout the week, maxing on Saturday and slightly decreasing on Sunday. However for members, the number of rides is lowest on Sunday and Saturday. For members, the number of rides is much higher during the week.

The second table shows that both members and casual riders spend the most time on average riding on Sundays and Saturdays. However, the groups differ in their lowest average ride length day. Casual riders spend only an average of 1084 seconds or 18.07 minutes during rides on Wednesday, while members spend only an average of 689 seconds or 11.48 minutes during rides on Monday.

We will now take a look at this first table comparing the number of rides throughout the week between the 2 groups.

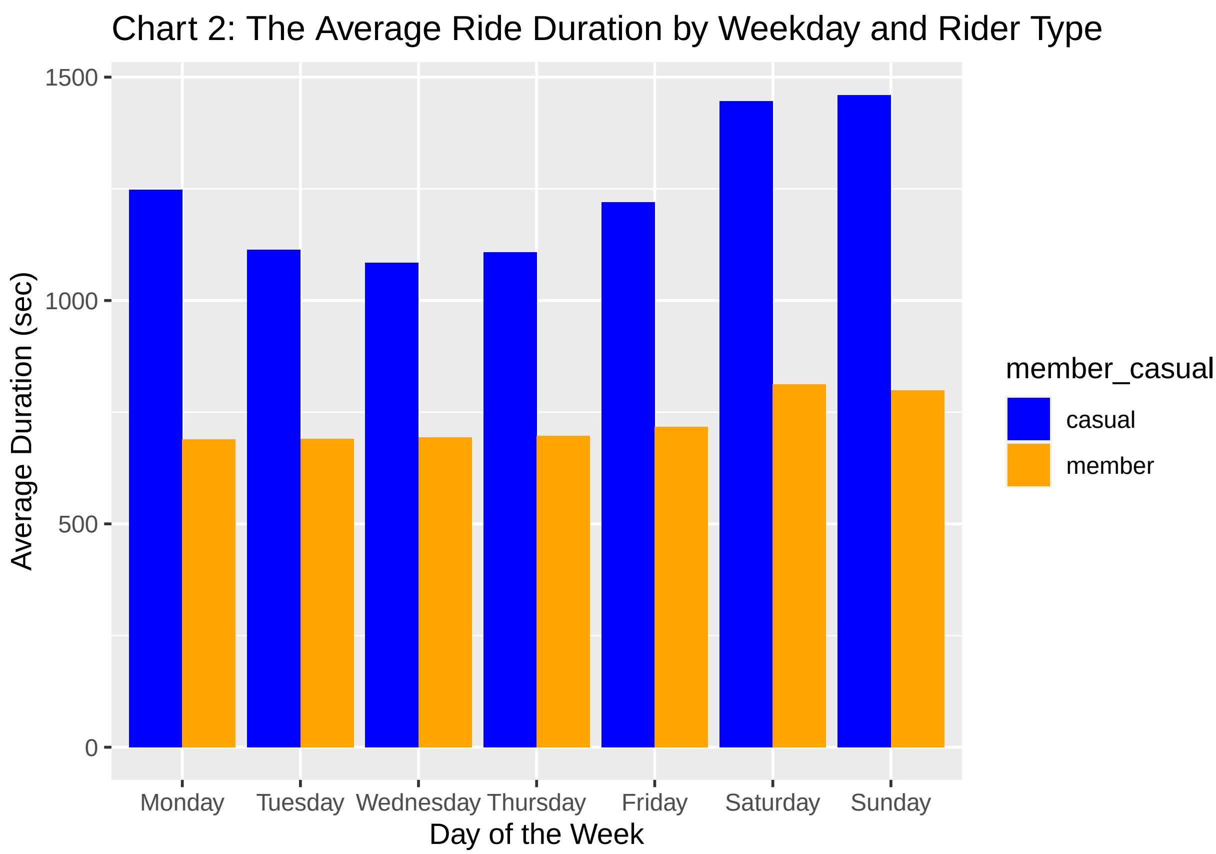
```
Table1$day_of_week <- factor(Table1$day_of_week,
                             levels = c("Monday", "Tuesday", "Wednesday",
                                        "Thursday", "Friday",
                                        "Saturday", "Sunday"))
Chart1 <- ggplot(data = Table1, aes(x = day_of_week,
                                    y = (number_of_rides)/1000,
                                    fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("blue", "orange")) +
  labs(title = "Chart 1: Number of Rides by Weekday and Rider Type") +
  ylab("Number of Rides (Thousands)") + xlab("Day of the Week")
Chart1
```



From this chart, we can clearly see that members tend to ride more on the weekdays and casual riders ride more on the weekend. It is interesting that member rides nearly double casual rides on Tuesday, Wednesday and Thursday, but are about the same on Saturday and Sunday.

Next, we will look at the average length of bike rides for each group throughout the week.

```
Table2$day_of_week <- factor(Table2$day_of_week,
                             levels = c("Monday", "Tuesday", "Wednesday",
                                        "Thursday", "Friday",
                                        "Saturday", "Sunday"))
Chart2 <- ggplot(data = Table2, aes(x = day_of_week, y = ride_length,
                                    fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("blue", "orange")) +
  labs(title = "Chart 2: The Average Ride Duration by Weekday and Rider Type")+
  ylab("Average Duration (sec)") + xlab("Day of the Week")
Chart2
```



From this chart, it is clear that casual rides spend on average more time on bike rides compared to members. Throughout the week, the average length of rides for members stays fairly consistent, with only a small uptick on the weekend. However, casual rides spend on average much more time riding on the weekends.

Next let us examine how the number of bike rides and average ride time change by the month.

```
Table3 <- aggregate(ride_length ~ member_casual + Month,
                    data = CompleteDataSet, NROW)
Table4 <- aggregate(ride_length ~ member_casual + Month,
                    data = CompleteDataSet, mean)
colnames(Table3)[3] = "number_of_rides"
Table3
```

```
##    member_casual Month number_of_rides
## 1         casual   Apr          147284
## 2         member   Apr          279302
## 3         casual   Aug          358917
## 4         member   Aug          427000
## 5         casual   Dec           44894
## 6         member   Dec          136912
## 7         casual   Feb           43016
## 8         member   Feb          147428
## 9         casual   Jan           40008
## 10        member   Jan          150293
## 11        casual   Jul          406046
## 12        member   Jul          417426
## 13        casual   Jun          301226
```

```
## 14        member   Jun          418385
## 15        casual   Mar           62201
## 16        member   Mar          196477
## 17        casual   May          234178
## 18        member   May          370639
## 19        casual   Nov          100747
## 20        member   Nov          236947
## 21        casual   Oct          208988
## 22        member   Oct          349693
## 23        casual   Sep          296694
## 24        member   Sep          404636
```

```
Table4
```

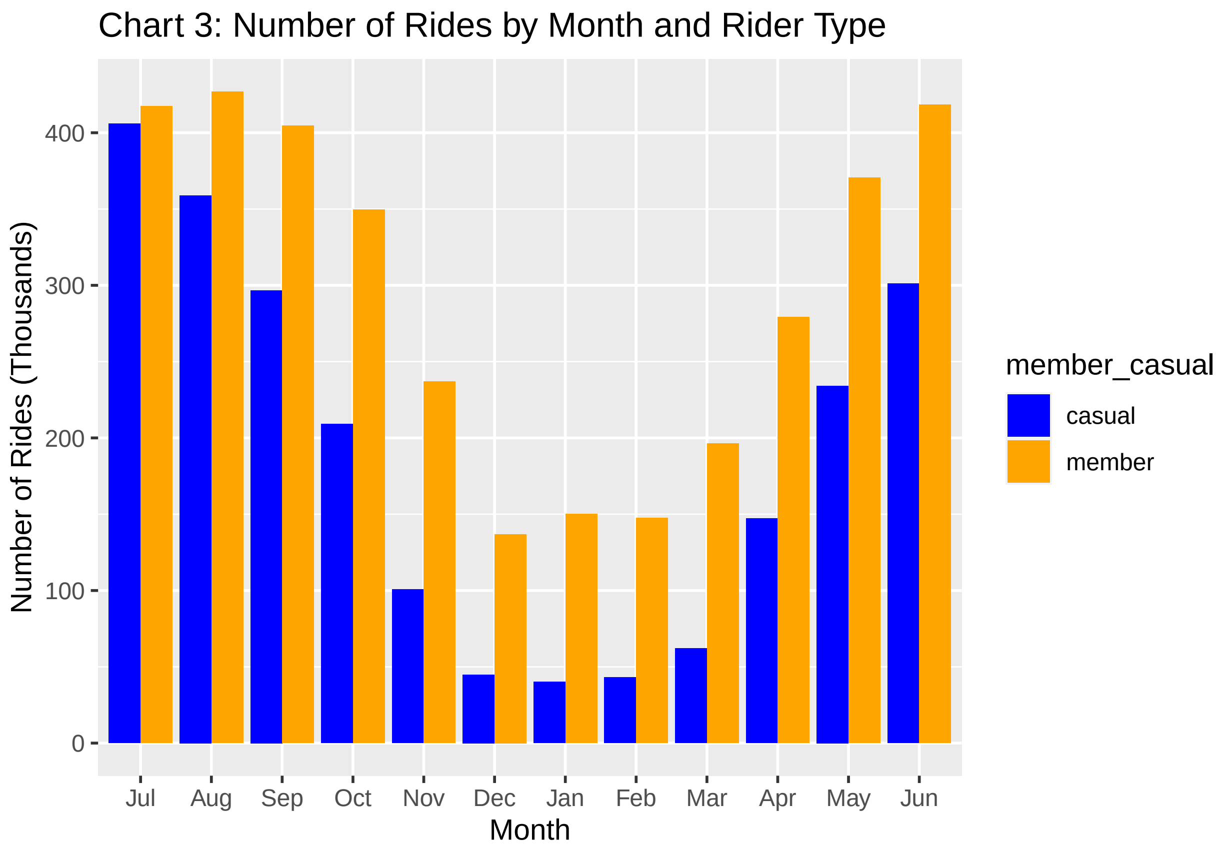
```
##    member_casual Month ride_length
## 1         casual   Apr   1279.6592
## 2         member   Apr    688.0230
## 3         casual   Aug   1324.1319
## 4         member   Aug    785.0584
## 5         casual   Dec    840.8443
## 6         member   Dec    620.7616
## 7         casual   Feb    997.8742
## 8         member   Feb    628.2107
## 9         casual   Jan    856.5941
## 10        member   Jan    604.4595
## 11        casual   Jul   1413.9361
## 12        member   Jul    805.9390
## 13        casual   Jun   1339.3374
## 14        member   Jun    771.7784
## 15        casual   Mar    961.0890
## 16        member   Mar    612.0217
## 17        casual   May   1359.2335
## 18        member   May    754.6729
## 19        casual   Nov    964.1477
## 20        member   Nov    651.8729
## 21        casual   Oct   1138.4123
## 22        member   Oct    692.8180
## 23        casual   Sep   1235.0337
## 24        member   Sep    758.0572
```

From a quick glance at these tables it looks like the number and average length of bike rides is lowest in the winter months and highest in the summers.

To get a better look at the data, let's create some charts. This first chart will show how the number of rides changes by month for both groups.

```
Table3$Month <- factor(Table3$Month,
                       levels = c("Jul", "Aug", "Sep", "Oct", "Nov", "Dec",
                                  "Jan", "Feb", "Mar", "Apr", "May", "Jun"))
Chart3 <- ggplot(data = Table3, aes(x = Month, y = (number_of_rides)/1000,
                                    fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("blue", "orange")) +
  labs(title = "Chart 3: Number of Rides by Month and Rider Type") +
```

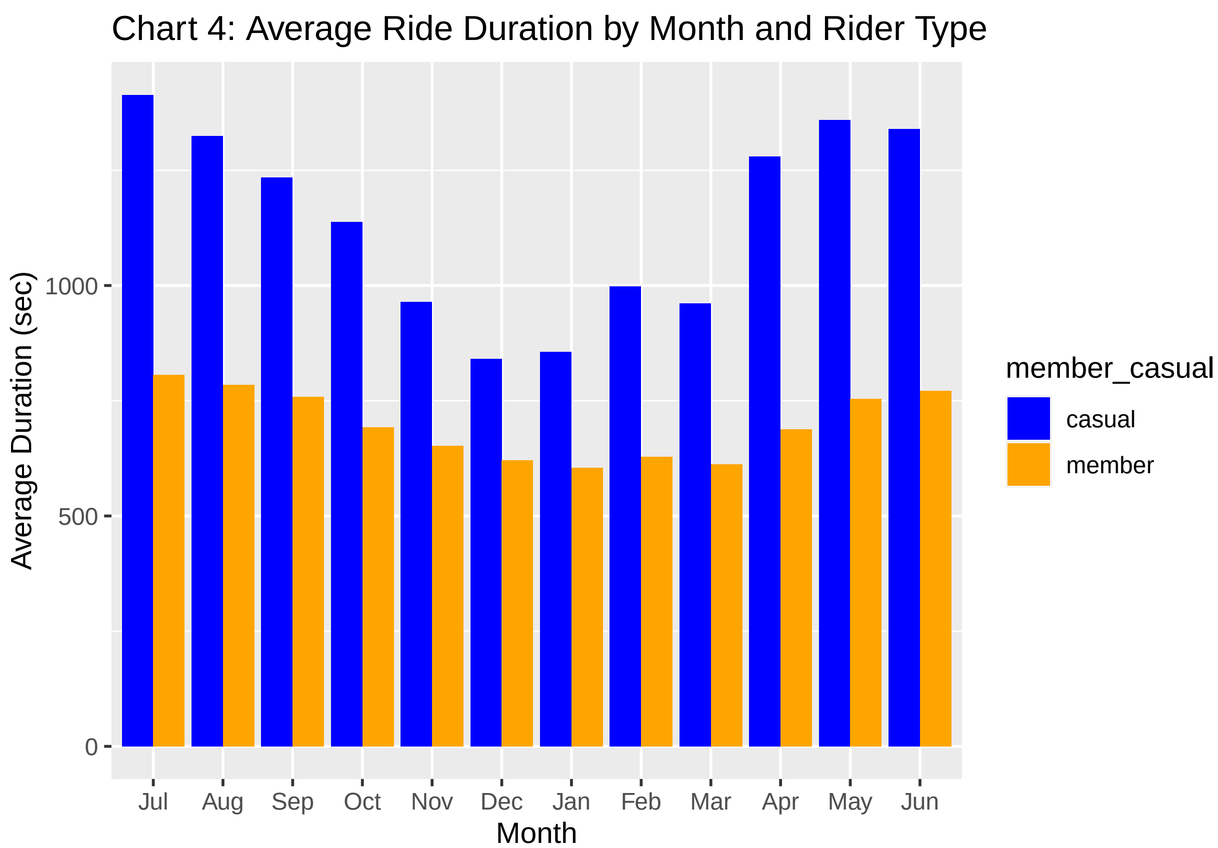
```
  ylab("Number of Rides (Thousands)") + xlab("Month")
Chart3
```



From this chart, we can see that both groups have about the same tendencies. Both members and casual riders have more rides in the summer decreasing during the fall into winter and then rising from the winter during spring and into summer. Members and causal riders differ in that the number of casual riders drops off much more dramatically during the winter.

Next we will take a look at how the average ride duration changes by month for both rider groups.

```
Table4$Month <- factor(Table3$Month,
                       levels = c("Jul", "Aug", "Sep", "Oct", "Nov", "Dec",
                                  "Jan", "Feb", "Mar", "Apr", "May", "Jun"))
Chart4 <- ggplot(data = Table4, aes(x = Month, y = ride_length,
                                    fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("blue", "orange")) +
  labs(title = "Chart 4: Average Ride Duration by Month and Rider Type") +
  ylab("Average Duration (sec)") + xlab("Month")
Chart4
```



Again from this graph, we can see that both rider groups share a similar tendency of biking less in the winter months and more in the summer months. However, again the members' average duration does not dip down quite as dramatically as casual members.

Now I will look at how members and casual riders differ in the stations they start at. I will create 2 tables that show the top 10 starting stations for each group.

```
TableStations <- aggregate(ride_length ~ member_casual + start_station_name,
                           data = CompleteDataSet, NROW)
Table5 <- filter(TableStations, member_casual == "member")
Table6 <- filter(TableStations, member_casual == "casual")
colnames(Table5)[3] = "number_of_rides"
colnames(Table6)[3] = "number_of_rides"
Table5 <- Table5[order(Table5$number_of_rides , decreasing = TRUE),]
Table5 <- Table5[1:10,]
Table6 <- Table6[order(Table6$number_of_rides , decreasing = TRUE),]
Table6 <- Table6[1:10,]
Table5
```

```
##      member_casual           start_station_name
## 516         member      Kingsbury St & Kinzie St
## 193         member             Clark St & Elm St
## 220         member Clinton St & Washington Blvd
## 1494        member         Wells St & Concord Ln
## 1456        member      University Ave & 57th St
## 605         member      Loomis St & Lexington St
## 326         member           Ellis Ave & 60th St
```

```
## 1495        member            Wells St & Elm St
## 216         member      Clinton St & Madison St
## 85          member         Broadway & Barry Ave
##      number_of_rides
## 516            25293
## 193            23442
## 220            22439
## 1494           21797
## 1456           21147
## 605            20695
## 326            20123
## 1495           19503
## 216            19412
## 85             18579
```

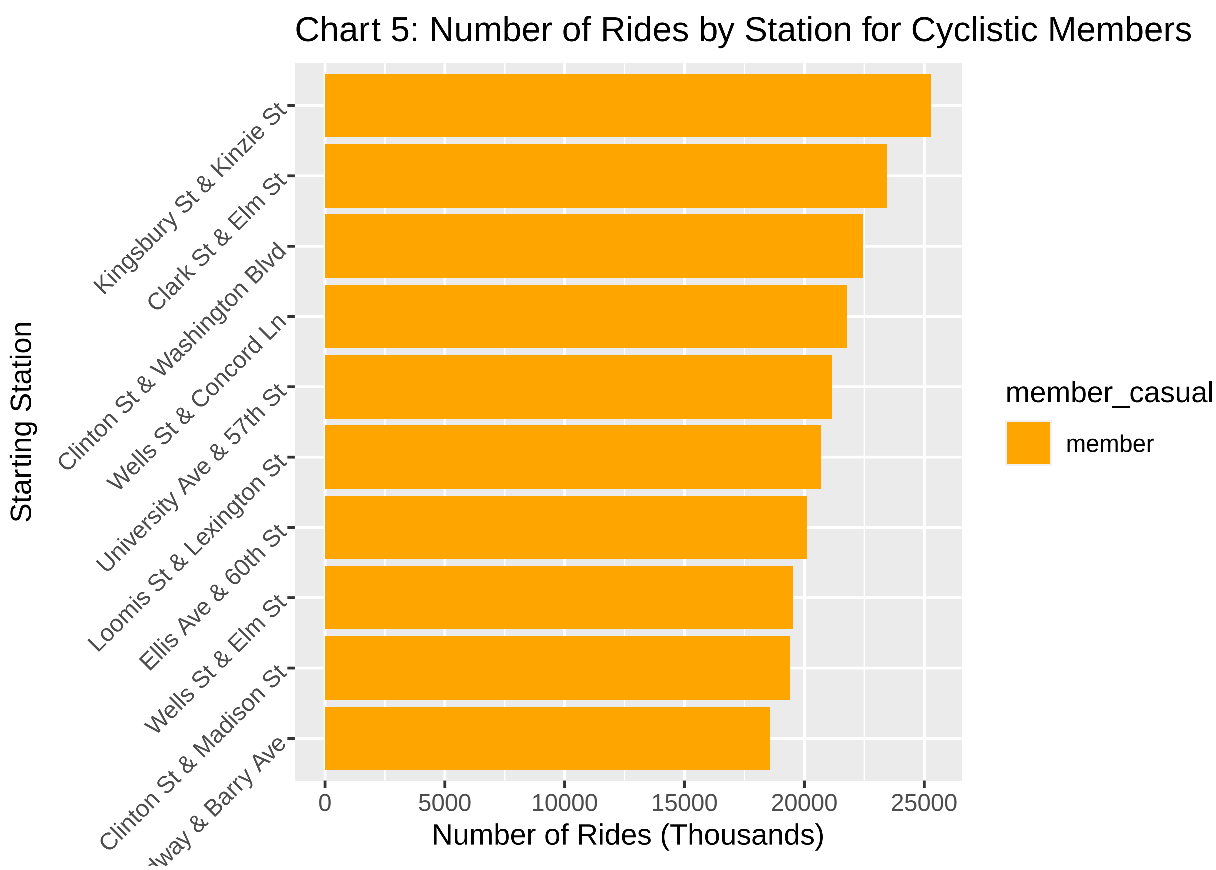
```
Table6
```

```
##      member_casual                 start_station_name
## 1599        casual            Streeter Dr & Grand Ave
## 315         casual  DuSable Lake Shore Dr & Monroe St
## 689         casual              Michigan Ave & Oak St
## 696         casual                    Millennium Park
## 316         casual DuSable Lake Shore Dr & North Blvd
## 1519        casual                     Shedd Aquarium
## 1604        casual                 Theater on the Lake
## 1665        casual              Wells St & Concord Ln
## 312         casual                     Dusable Harbor
## 472         casual         Indiana Ave & Roosevelt Rd
##      number_of_rides
## 1599           52920
## 315            30825
## 689            23978
## 696            23519
## 316            21925
## 1519           19418
## 1604           17377
## 1665           14979
## 312            14141
## 472            12932
```

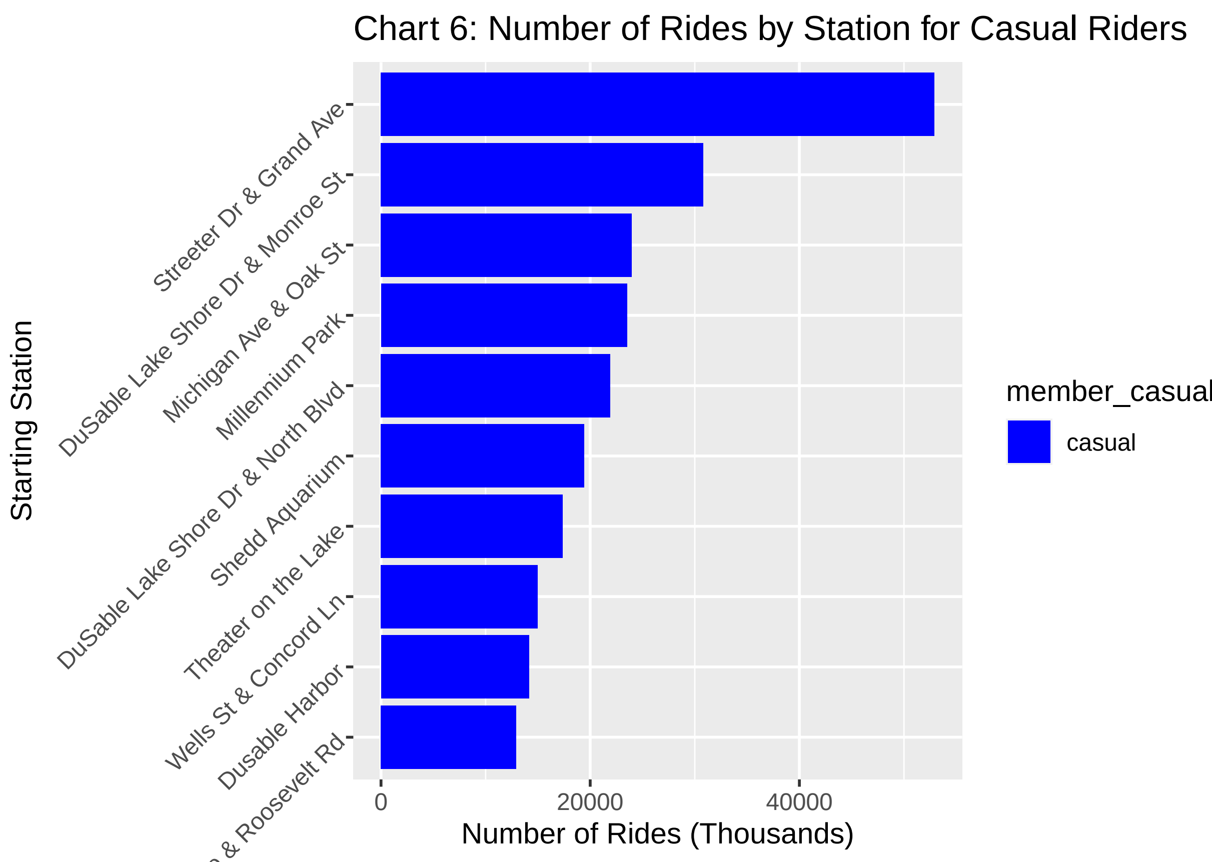
I will now chart these top 10 starting stations for each group.

```
Chart5 <- ggplot(data = Table5, aes(x = number_of_rides,
                            y = reorder(start_station_name, number_of_rides),
                            fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_manual(values = c("orange")) +
  labs(title =
         "Chart 5: Number of Rides by Station for Cyclistic Members") +
  ylab("Starting Station") + xlab("Number of Rides (Thousands)") +
  theme(axis.text.y = element_text(angle=45, hjust=1))
Chart6 <- ggplot(data = Table6, aes(x = number_of_rides,
                            y = reorder(start_station_name, number_of_rides),
                            fill = member_casual)) +
```

```
  geom_col(position = "dodge") + scale_fill_manual(values = c("blue")) +
  labs(title = "Chart 6: Number of Rides by Station for Casual Riders") +
  ylab("Starting Station") + xlab("Number of Rides (Thousands)") +
  theme(axis.text.y = element_text(angle=45, hjust=1))
Chart5
```




```
Chart6
```



There are a few things that can be taken away from these charts. First, it is interesting that only 1 Starting Station (Wells St & Concord Ln) are in both the members and casual riders top 10 stations. Another interesting takeaway is that causal riders most used starting station (Streeter Dr & Grand Ave) far exeed the other stations, however the members top 10 stations are all realitivly close in the number of rides.

ACT

Now that the data has been analyzed and presented with some impactful data visualizations, it is time to define some final conclusions and set out some recommendations for Cyclistic to implement. The first conclusion I would make is that members tend to be more likely to use Cyclistic during the week while casual riders use Cyclistic more on the weekends. This leads me to believe that members most likely use Cyclistic to get to and from work, while casual use the bikes more for leisure or after work events. Cyclistic members tend to spend on average the same duration on rides throughout the year, which would also leads me to conclude that members are often taking the same routes all throughout the year. Casual riders on the other-hand have their duration increase in the warmer months leading me to believe they often ride different routes at different times of the year. Most likely casual riders use Cyclistic bikes more for sightseeing in the warmer months.

Here are my three recommendations for Cyclistic to convert casual riders to members based on my analysis:

1. Offer special weekend memberships that are specifically for Friday, Saturday and Sunday, since these are the days of the week casual riders bike the most.
2. Offer summer memberships that run from April to October, since the part of the year casual riders use Cyclistic the most.

3. Offer discounts to memberships that are limited to the top locations that casual riders use. Like Streeter Dr & Grand Ave, DuSable Lake Shore Dr & Monroe St, Michigan Ave & Oak St, Millennium Park, and DuSable Lake Shore Dr & North Blvd.