Cyclistic Analysis

Austin Carlson

2023-07-04

Introduction

This data analysis is for a fictional bike company, Cyclistic. The scenario and data was provided by Google Data Analytics Capstone program.

Cyclistic is a bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike.

The executive team at Cyclistic want to maximize the number of annual pass members. To achieve this they want you to look at the differences between day pass members and annual members. In addition, they want to use these differences to help market their annual pass more effectively to day pass members.

Phase 1: Ask

1. Identify the business task

- Analyze May 2023 Data from Cyclistic Users.
- Provide quality recommendations to the Cyclistic marketing team.

2. Consider the Key Stakeholders

Primary Stakeholder(s)

- Lily Moreno: Director of Marketing at Cyclistic
- Cyclistic Executive Team

Secondary Stakeholder(s)

- Cyclistic Marketing Analytics Team

Phase 2: Prepare

1.Identify the Data Source

Dataset: Bike Rental Trip Data for May of 2023 (CC0: Public Domain, dataset made available by Motivate International Inc.) Click here to access the data set. This dataset contains trip data for over 600,000 different trips. The data includes start and end times, start and end stations, and member status. While there is data dating back to 2013, 600,000 is more than sufficient as a sample size.

2. Determine the credibility of the data

I will use the **"ROCCC"** system to determine the credibility and integrity of the data.

Reliability: The data is reliable. The sample size is very large and is a first party source.

Originality: The data is original. The data is sourced internally.

Comprehensiveness: The data is **not** comprehensive. Due to data privacy, all PII has been removed. This may limit the final analysis options.

Current: The data is current. The data is from May of 2023. At the time of this analysis the June set was not available, but it is not likely the data from May to June will shift drastically.

Cited: The data is cited. The data comes from an internal source.

Phase 3: Process

Note: All of my analysis will be done in RStudio.

I will start by loading the necessary packages.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
library(skimr)
```

Next, I will import the necessary files onto R.

```
BD_0523 <- read.csv("202305-divvy-tripdata.csv")
##BD = Bike Data, 0523 = Date of Survey
```

Cleaning the dataset

Now, we can take a quick look at the dataset.

```r
head(BD_0523)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 0D9FA920C3062031 electric_bike 2023-05-07 19:53:48 2023-05-07 19:58:32
## 2 92485E5FB5888ACD electric_bike 2023-05-06 18:54:08 2023-05-06 19:03:35
## 3 FB144B3FC8300187 electric_bike 2023-05-21 00:40:21 2023-05-21 00:44:36
## 4 DDEB93BC2CE9AA77  classic_bike 2023-05-10 16:47:01 2023-05-10 16:59:52
## 5 C07B70172FC92F59  classic_bike 2023-05-09 18:30:34 2023-05-09 18:39:28
## 6 2BA66385DF8F815A  classic_bike 2023-05-30 15:01:21 2023-05-30 15:17:00
##            start_station_name start_station_id            end_station_name
## 1 Southport Ave & Belmont Ave            13229
## 2 Southport Ave & Belmont Ave            13229
## 3        Halsted St & 21st St            13162
## 4     Carpenter St & Huron St            13196     Damen Ave & Cortland St
## 5    Southport Ave & Clark St     TA1308000047 Southport Ave & Belmont Ave
## 6     Clinton St & Madison St     TA1305000032        McClurg Ct & Ohio St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1                 41.93941 -87.66383 41.93000 -87.65000        member
## 2                 41.93948 -87.66385 41.94000 -87.69000        member
## 3                 41.85379 -87.64672 41.86000 -87.65000        member
## 4          13133  41.89456 -87.65345 41.91598 -87.67733        member
## 5          13229  41.95708 -87.66420 41.93948 -87.66375        member
## 6   TA1306000029  41.88275 -87.64119 41.89259 -87.61729        member
```

Lets look at the columns names for the data set.

```r
colnames(BD_0523)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

Lets check the amount of distinct rides.

```r
n_distinct(BD_0523)
```

```
## [1] 604827
```

Lets also check the number of observations in the table.

```r
nrow(BD_0523)
```

```
## [1] 604827
```

Since both the distinct rides and observations are equal there should be no duplicates in the data set. Just to be sure, lets run a duplicate check on the dataset.

```r
nrow(BD_0523[duplicated(BD_0523),])
```

```
## [1] 0
```

Since the result is 0 we have now confirmed there are no duplicates in the data set. From this we can conclude there have been 604827 rides with Cyclistic in May of 2023.

Now, we can check if there are any rows with null or NA values.

```r
nrow(BD_0523[is.null(BD_0523),])
```

```
## [1] 0
```

```r
nrow(BD_0523[is.na(BD_0523),])
```

```
## [1] 1420
```

It appears there are 1420 rows with NA values. Lets remove those rows.

```r
BD_0523_RNA <- na.omit(BD_0523)
#RNA = Removed NA
nrow(BD_0523_RNA[is.na(BD_0523_RNA),])
```

```
## [1] 0
```

```r
#checking again for NA Values
```

For future analysis lets find the length of each trip and the day of the week the trip was taken. Also, we can remove unnecessary columns.

```r
BD_0523_Subset <- subset(BD_0523_RNA, select = -c(start_station_id, end_station_id, start_lat, start_lng
BD_0523_Mut1 <- mutate(BD_0523_Subset, day_of_week = wday(started_at, label=TRUE))
BD_0523_Mut2 <- mutate(BD_0523_Mut1, ride_length = difftime(ended_at, started_at))
BD_0523_Mut2$ride_length <- as.numeric(BD_0523_Mut2$ride_length)
```

Now, lets look at the new table

```r
head(BD_0523_Mut2)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 0D9FA920C3062031 electric_bike 2023-05-07 19:53:48 2023-05-07 19:58:32
## 2 92485E5FB5888ACD electric_bike 2023-05-06 18:54:08 2023-05-06 19:03:35
## 3 FB144B3FC8300187 electric_bike 2023-05-21 00:40:21 2023-05-21 00:44:36
## 4 DDEB93BC2CE9AA77  classic_bike 2023-05-10 16:47:01 2023-05-10 16:59:52
## 5 C07B70172FC92F59  classic_bike 2023-05-09 18:30:34 2023-05-09 18:39:28
## 6 2BA66385DF8F815A  classic_bike 2023-05-30 15:01:21 2023-05-30 15:17:00
##            start_station_name            end_station_name member_casual
## 1 Southport Ave & Belmont Ave                                     member
## 2 Southport Ave & Belmont Ave                                     member
```

```
## 3        Halsted St & 21st St                                    member
## 4     Carpenter St & Huron St     Damen Ave & Cortland St         member
## 5    Southport Ave & Clark St Southport Ave & Belmont Ave         member
## 6     Clinton St & Madison St        McClurg Ct & Ohio St         member
##   day_of_week ride_length
## 1         Sun         284
## 2         Sat         567
## 3         Sun         255
## 4         Wed         771
## 5         Tue         534
## 6         Tue         939
```

Now, we can filter out cases that might impact our analysis

```
BD_0523_filtered <- BD_0523_Mut2 %>%
    filter(ride_length > 10 & ended_at > started_at & end_station_name != "" & start_station_name != ""
#Assuming People are not riding past 24 hours and trips are at least 10 minutes
```

Phase 4: Analyze

I will now check the statistical summary for the dataset.

```
summary(BD_0523_filtered)
```

```
##    ride_id           rideable_type       started_at           ended_at
##  Length:343353      Length:343353      Length:343353      Length:343353
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name end_station_name   member_casual      day_of_week
##  Length:343353      Length:343353      Length:343353      Sun:41255
##  Class :character   Class :character   Class :character   Mon:44911
##  Mode  :character   Mode  :character   Mode  :character   Tue:60642
##                                                           Wed:60094
##                                                           Thu:50079
##                                                           Fri:43406
##                                                           Sat:42966
##   ride_length
##  Min.   :  11.0
##  1st Qu.: 299.0
##  Median : 485.0
##  Mean   : 522.2
##  3rd Qu.: 726.0
##  Max.   :1139.0
##
```

```
BD_0523_filtered %>%
  count(start_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##          start_station_name    n
## 1  Streeter Dr & Grand Ave 3070
## 2 Kingsbury St & Kinzie St 2744
## 3    Wells St & Concord Ln 2656
## 4        Clark St & Elm St 2629
## 5 University Ave & 57th St 2488
```

```
BD_0523_filtered %>%
  count(end_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##               end_station_name    n
## 1      Streeter Dr & Grand Ave 2904
## 2     Kingsbury St & Kinzie St 2838
## 3        Wells St & Concord Ln 2723
## 4     University Ave & 57th St 2664
## 5 Clinton St & Washington Blvd 2628
```

```
BD_0523_filtered %>%
  count(rideable_type, sort = TRUE) %>%
  slice(1:5)
```

```
##   rideable_type      n
## 1  classic_bike 192744
## 2 electric_bike 146808
## 3   docked_bike   3801
```

```
BD_0523_filtered %>%
  count(member_casual, sort = TRUE) %>%
  slice(1:5)
```

```
##   member_casual      n
## 1        member 232810
## 2        casual 110543
```

Observations:

- The two most popular days to ride are Tuesday and Wednesday. The two least popular days to ride are Sunday and Saturday.
- The average ride length is 522 Minutes or 8.7 Hours.
- The two most common place to start and end a trip is Streeter Dr & Grand Ave and Kingsbury St & Kinzie St

Deductions:

- Users who use the bikes for work commuting are not using them as often on the weekend.
- A large amount of users will get a bike at the start of their work day and return it after.
- A large amount of users are starting their trip at a very touristy area.

Now, we can look at the specific stats for both Casuals and Members. Lets start with Casuals.

```r
BD_0523_filtered_c <- BD_0523_filtered %>%
  filter(member_casual == 'casual')
summary(BD_0523_filtered_c)
```

```
##    ride_id           rideable_type       started_at          ended_at
##  Length:110543      Length:110543      Length:110543      Length:110543
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name end_station_name   member_casual      day_of_week
##  Length:110543      Length:110543      Length:110543      Sun:17588
##  Class :character   Class :character   Class :character   Mon:13368
##  Mode  :character   Mode  :character   Mode  :character   Tue:16836
##                                                           Wed:16429
##                                                           Thu:14504
##                                                           Fri:14229
##                                                           Sat:17589
##   ride_length
##  Min.   :  11.0
##  1st Qu.: 350.0
##  Median : 550.0
##  Mean   : 570.1
##  3rd Qu.: 786.0
##  Max.   :1139.0
##
```

```r
BD_0523_filtered_c %>%
  count(start_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##                   start_station_name    n
## 1             Streeter Dr & Grand Ave 1942
## 2  DuSable Lake Shore Dr & Monroe St 1323
## 3 DuSable Lake Shore Dr & North Blvd 1080
## 4               Michigan Ave & Oak St  950
## 5               Wells St & Concord Ln  923
```

```r
BD_0523_filtered_c %>%
  count(end_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##                     end_station_name    n
## 1             Streeter Dr & Grand Ave 2032
## 2  DuSable Lake Shore Dr & Monroe St 1176
## 3 DuSable Lake Shore Dr & North Blvd 1125
## 4                     Millennium Park 1066
## 5               Michigan Ave & Oak St  956
```

```
BD_0523_filtered_c %>%
  count(rideable_type, sort = TRUE) %>%
  slice(1:5)
```

```
##   rideable_type     n
## 1  classic_bike 53533
## 2 electric_bike 53209
## 3   docked_bike  3801
```

```
BD_0523_filtered_c %>%
  count(member_casual, sort = TRUE) %>%
  slice(1:5)
```

```
##   member_casual      n
## 1        casual 110543
```

Observations:

- The two most popular days for Casuals to use the bikes are Saturday and Sunday, while the two least popular are Monday and Friday.
- The average ride length is 570 Minutes or 9.5 hours.
- The two most common place to start and end a trip is Streeter Dr & Grand Ave and DuSable Lake Shore Dr & Monroe St

Deductions:

- Casuals are using their bike more often on the weekends.
- Casuals trip length is longer than the population's average.
- Casuals are starting and ending their trips similar to the population.

Now, lets look at Annual Members

```
BD_0523_filtered_m <- BD_0523_filtered %>%
  filter(member_casual == 'member')
summary(BD_0523_filtered_m)
```

```
##    ride_id          rideable_type       started_at          ended_at        
##  Length:232810      Length:232810      Length:232810      Length:232810     
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  start_station_name end_station_name   member_casual      day_of_week 
##  Length:232810      Length:232810      Length:232810      Sun:23667  
##  Class :character   Class :character   Class :character   Mon:31543  
##  Mode  :character   Mode  :character   Mode  :character   Tue:43806  
##                                                           Wed:43665  
##                                                           Thu:35575  
```

```
##                                                            Fri:29177
##                                                            Sat:25377
##    ride_length
##   Min.   :  11.0
##   1st Qu.: 281.0
##   Median : 456.0
##   Mean   : 499.4
##   3rd Qu.: 693.0
##   Max.   :1139.0
##
```

```r
BD_0523_filtered_m %>%
  count(start_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##             start_station_name    n
## 1     Kingsbury St & Kinzie St 2047
## 2 Clinton St & Washington Blvd 1943
## 3     University Ave & 57th St 1863
## 4            Clark St & Elm St 1849
## 5          Ellis Ave & 60th St 1822
```

```r
BD_0523_filtered_m %>%
  count(end_station_name, sort = TRUE) %>%
  slice(1:5)
```

```
##               end_station_name    n
## 1 Clinton St & Washington Blvd 2191
## 2     Kingsbury St & Kinzie St 2135
## 3     University Ave & 57th St 2022
## 4            Clark St & Elm St 1917
## 5        Wells St & Concord Ln 1828
```

```r
BD_0523_filtered_m %>%
  count(rideable_type, sort = TRUE) %>%
  slice(1:5)
```

```
##   rideable_type      n
## 1  classic_bike 139211
## 2 electric_bike  93599
```

```r
BD_0523_filtered_m %>%
  count(member_casual, sort = TRUE) %>%
  slice(1:5)
```

```
##   member_casual      n
## 1        member 232810
```

Observations:

- The two most popular days for Members to use the bikes are Tuesday and Wednesday, while the two least popular are Sunday and Saturday.

- The average ride length is 499 Minutes or 8.3 hours.
- The two most common place to start a trip is Kingsbury St & Kinzie St and Clinton St & Washington Blvd, where that is flipped for ending a trip

Deductions:

- Members are using their bike more often in the middle of the week.
- Members trip length is shorter than the population's average.
- Members are starting and ending their trips commonly at the same place.

Phase 5: Share

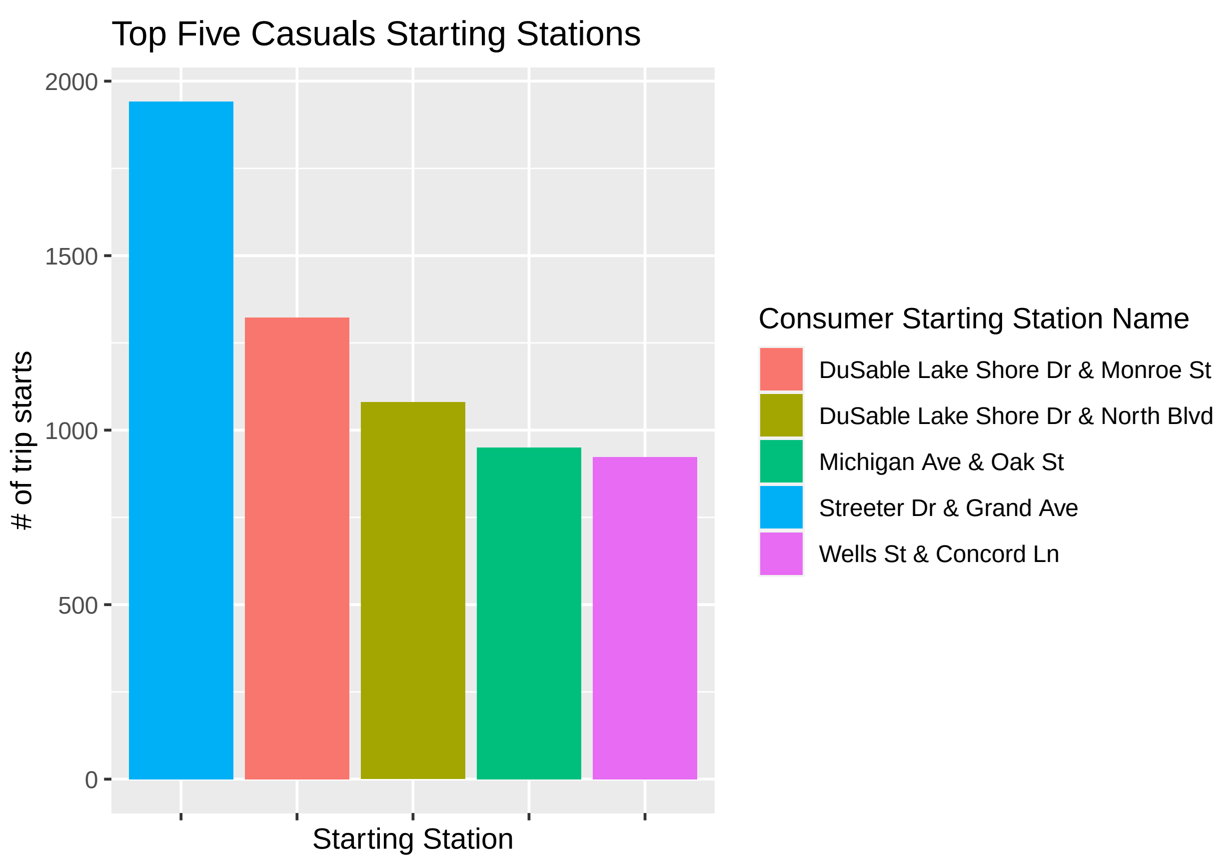
Visualizations*

Here I will create my Vizs to show the relationship between the data.

Fig.1-4 Bar Graphs showing difference between starting locations for Members vs. Casual

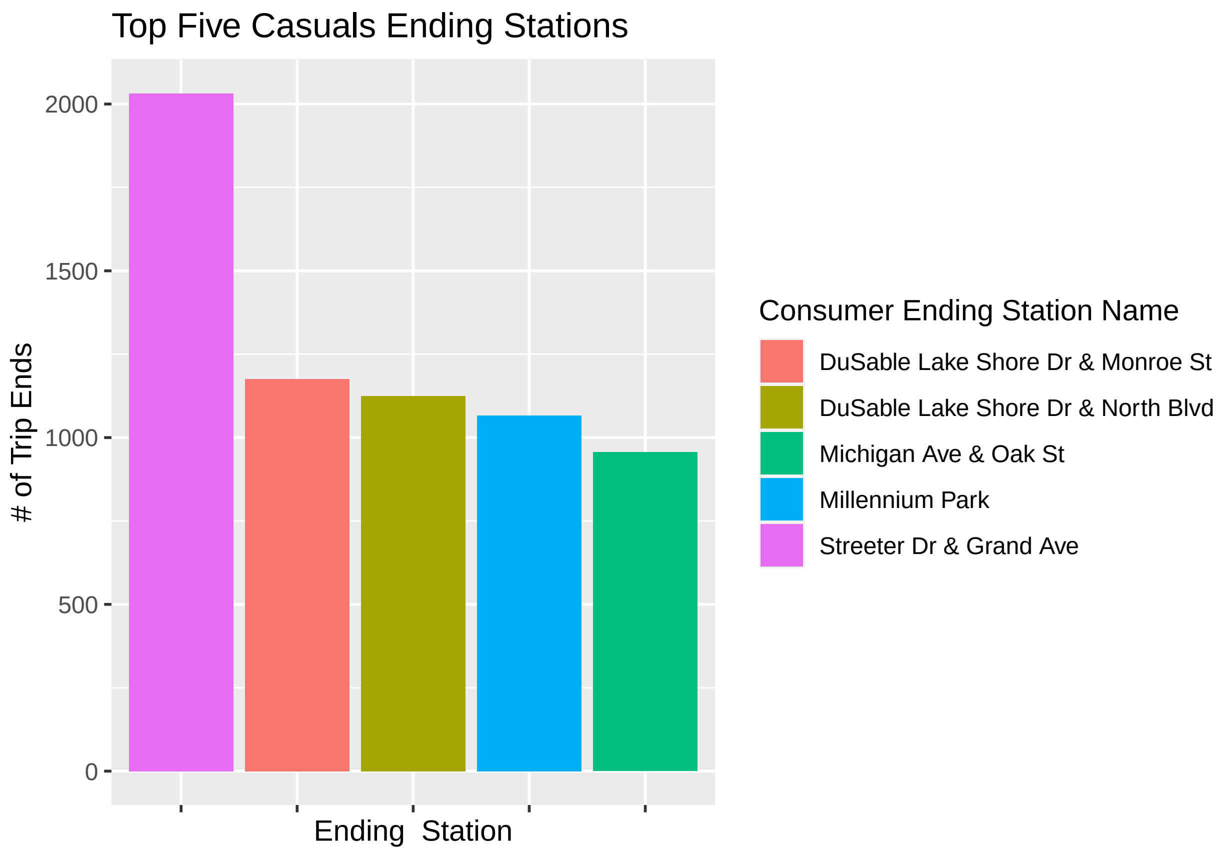
```
ggplot(data = BD_0523_filtered_c, aes(x = start_station_name, fill = start_station_name)) +
  geom_bar() + scale_x_discrete(limits=c("Streeter Dr & Grand Ave", "DuSable Lake Shore Dr & Monroe St"
```

```
## Warning: Removed 104325 rows containing non-finite values (`stat_count()`).
```



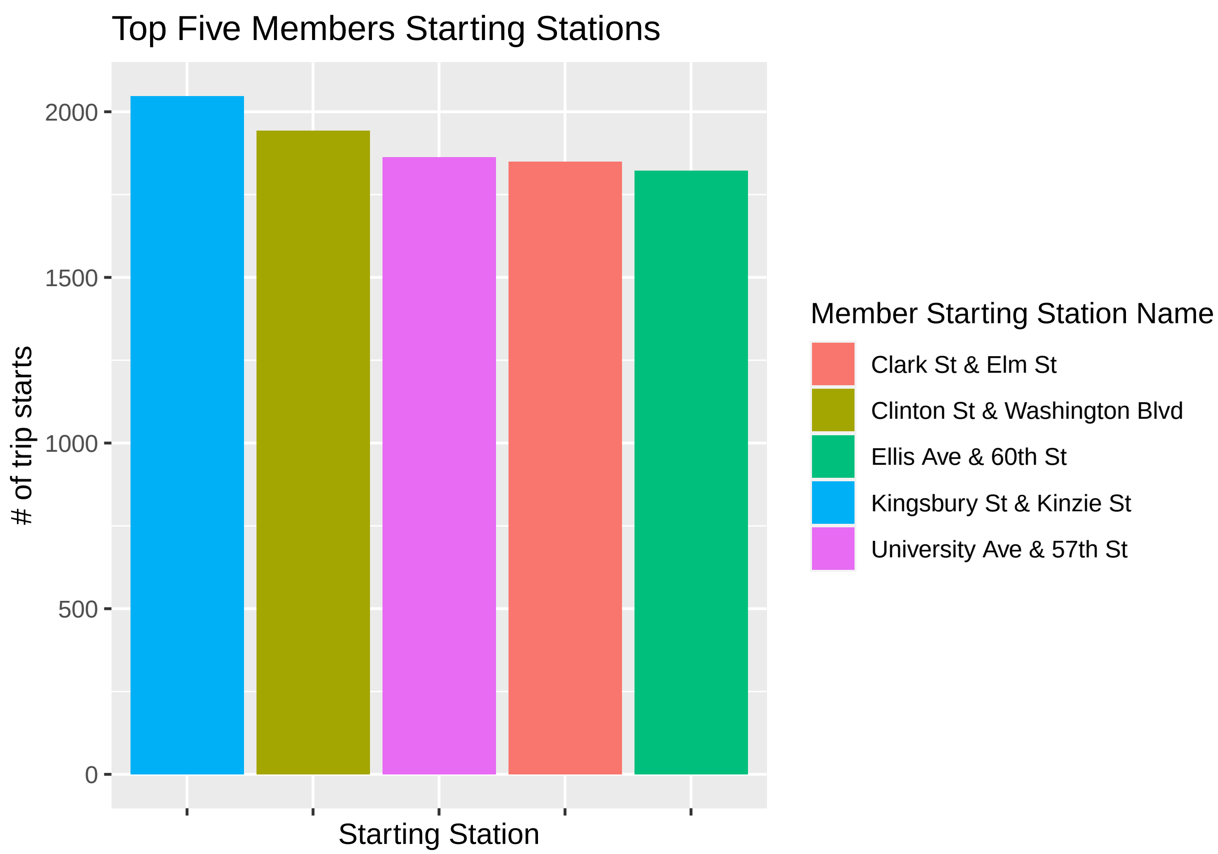
```r
ggplot(data = BD_0523_filtered_c, aes(x = end_station_name, fill = end_station_name)) +
  geom_bar() + scale_x_discrete(limits=c("Streeter Dr & Grand Ave", "DuSable Lake Shore Dr & Monroe St"
```

```
## Warning: Removed 104188 rows containing non-finite values (`stat_count()`).
```




```r
ggplot(data = BD_0523_filtered_m, aes(x = start_station_name, fill = start_station_name)) +
  geom_bar() + scale_x_discrete(limits=c("Kingsbury St & Kinzie St", "Clinton St & Washington Blvd", "U
```

```
## Warning: Removed 223286 rows containing non-finite values (`stat_count()`).
```




```
ggplot(data = BD_0523_filtered_m, aes(x = end_station_name, fill = end_station_name)) +
  geom_bar() + scale_x_discrete(limits=c("Clinton St & Washington Blvd", "Kingsbury St & Kinzie St", "Un
```

```
## Warning: Removed 222717 rows containing non-finite values (`stat_count()`).
```

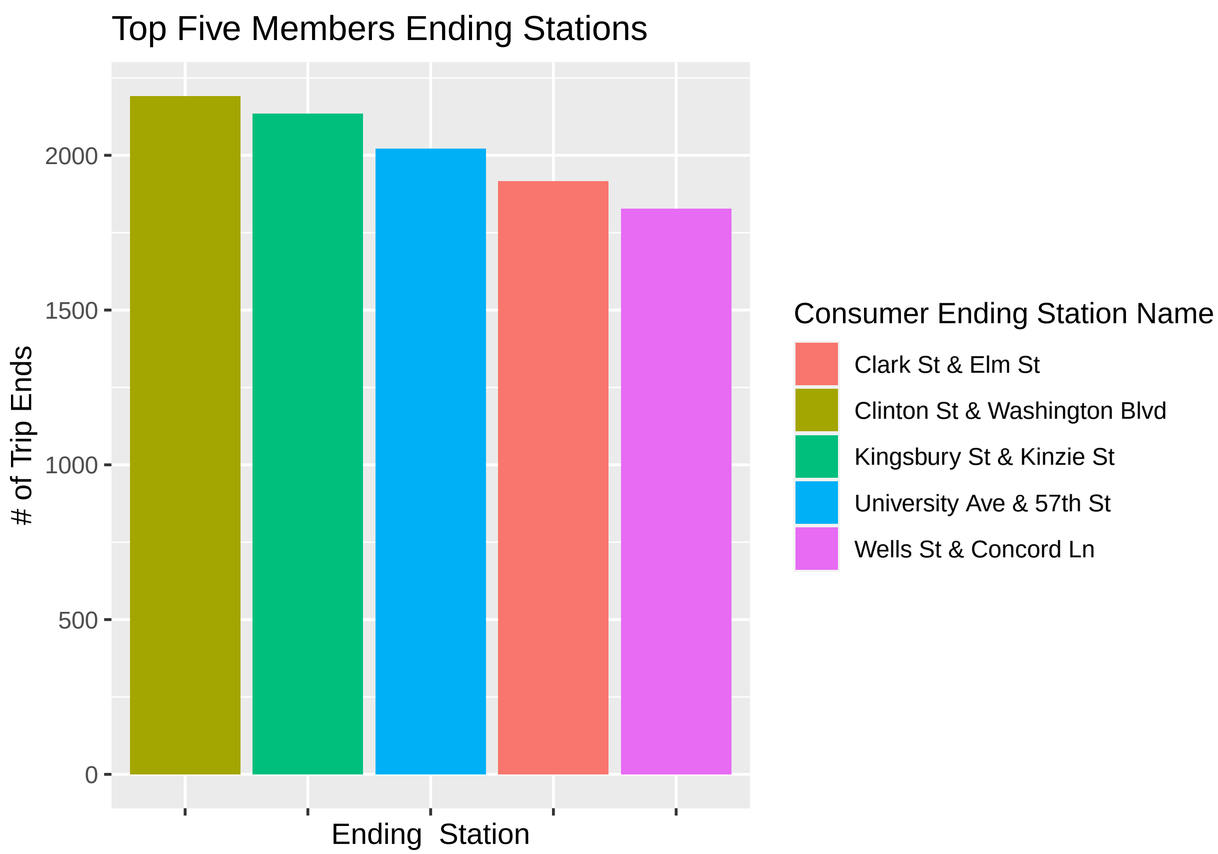
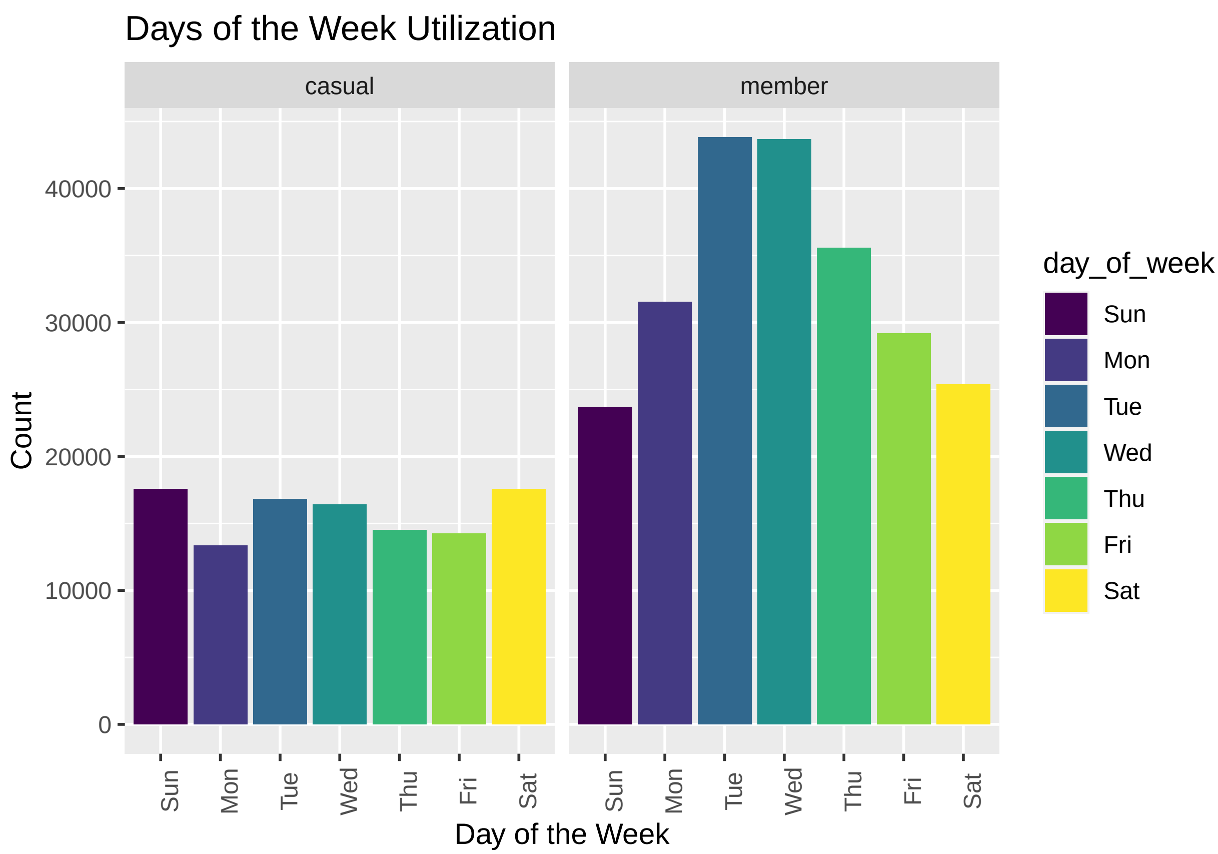


Fig.5 Bar Graph showing Days of the Week Utilization for Members vs. Casual

```
ggplot(data=BD_0523_filtered) +
  geom_bar(mapping=aes(x=day_of_week, fill = day_of_week)) +
  labs(title="Days of the Week Utilization", x="Day of the Week", y="Count") +  theme(axis.text.x = elem
  facet_grid(~member_casual)
```



Phase 6: Act

Recommendations for Cyclistic Marketing Team

1. Based off fig.5, the peak use days and minimal use days are reversed for Members and Casuals. To take advantage of this, the marketing team can run a promotion for annual passes during the weekend at the most popular spots for Casuals.
2. Based off fig.1-4, the starting and ending locations for Casuals are very similar. The marketing team can work to have extra annual memberships promotions posted around those areas.
3. Based off fig.5, and the fact provided by Cyclistic that over 30% of Members use the bikes for commuting purposes, Cyclistic can reach out to that subsection for testimonials about the convenience and cost-savings of commuting to work with Cyclistic bikes.

Recommendations based on limitations of dataset

1. Due to the missing PII, we cannot look at specific individuals. This additional data may show income/location based reasons for membership. It would also help identify average # of rides for Members Vs. Casuals
2. Since there is only quantitative data on the bikes themselves we do not know the human reasons why Casuals are not signing up for memberships.
3. Since we are only checking May, there may be differences in use in colder months.