Google data analytics-Analysis of bike share

Sufyan Shoukat

3/21/2023

Company details

Cyclistic, a bike-sharing company based in Chicago, launched a bike-sharing program in 2016 which has been successful so far. The program includes 5,824 bicycles that are tracked and locked into a network of 692 stations across the city. Riders can unlock bikes from one station and return them to any other station in the system at any time.

The company has two types of riders: members, who have an annual subscription, and casual riders, who use single-ride or full-day passes. Memberships are more profitable for the company than single-ride or full-day passes. Therefore, the director of marketing is looking to increase the number of annual memberships as part of a growth strategy for the company's future.

Steps Ask, Prepare, Process, Analyze, Share, Act

Ask Questions to be answered i-find patterns in the data to differentiate between casual riders and member riders ii-How media can effect the casual riders to become members? iii-why is there difference in bikes between member and casusal riders?

Prepare phase ride__id:identifier for each trip **rideable__type**: type of bike that was used **started__at**: timestamp for when the trip started **ended__at**: timestamp for when the trip ended **stat__station__name**: the name of the station where the trip started **start__station__id**: the id of the station where the trip started **end__station__name**: the name of the station where the trip ended **end__station__id**: the id of the station where the trip ended **start__lat**: latitude value for where the trip started **start__lng**: longitude value for where the trip started **end__lat**: latitude value for where the trip ended **end__lng**: longitude value for where the trip ended **memeber__casual**: type of user I will be using the public dataset located here

Tasks followed

I downloaded the data from jan 2021 to dec 2021(12 csv files) and stored in computer.

Instaled required packages

```r
#install.packages("ggplot2")
#install.packages("dplyr")
#install.packages("tidyverse")
```

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Imported data to RStudio

```r
dt1 <- read.csv("~/Professional career/Bike_share project/202101-divvy-tripdata.csv")
dt2 <- read.csv("~/Professional career/Bike_share project/202102-divvy-tripdata.csv")
dt3 <- read.csv("~/Professional career/Bike_share project/202103-divvy-tripdata.csv")
```

```r
dt4 <- read.csv("~/Professional career/Bike_share project/202104-divvy-tripdata.csv")
dt5 <- read.csv("~/Professional career/Bike_share project/202105-divvy-tripdata.csv")
dt6 <- read.csv("~/Professional career/Bike_share project/202106-divvy-tripdata.csv")
dt7 <- read.csv("~/Professional career/Bike_share project/202107-divvy-tripdata.csv")
dt8 <- read.csv("~/Professional career/Bike_share project/202108-divvy-tripdata.csv")
dt9 <- read.csv("~/Professional career/Bike_share project/202109-divvy-tripdata.csv")
dt10 <- read.csv("~/Professional career/Bike_share project/202110-divvy-tripdata.csv")
dt11 <- read.csv("~/Professional career/Bike_share project/202111-divvy-tripdata.csv")
dt12 <- read.csv("~/Professional career/Bike_share project/202112-divvy-tripdata.csv")
```

View data

```r
view(dt1)             #each file can be viewed by this way to know data

glimpse(dt1)          #view in more detail
```

```
## Rows: 96,834
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2021-01-23 16:14:19", "2021-01-27 18:43:08", "2021~
## $ ended_at           <chr> "2021-01-23 16:24:44", "2021-01-27 18:47:12", "2021~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Wood St & Augu~
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "657", "13258",~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

Merge data

```r
trip_data <- rbind(dt1,dt2,dt3,dt4,dt5,dt6,dt7,dt8,dt9,dt10,dt11,dt12)
```

View types of uers and bikes

```r
unique(trip_data$rideable_type)
```

```
## [1] "electric_bike" "classic_bike"  "docked_bike"
```

```r
unique(trip_data$member_casual)
```

```
## [1] "member" "casual"
```

Process phase

```
colnames(trip_data)                #view the column names (13)
```

```
##  [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"
```

```
nrow(trip_data)    #view no of rows
```

```
## [1] 5595063
```

```
trip_data <- trip_data%>%
  distinct(ride_id,.keep_all = TRUE)    #remove duplicate ride with same id
```

Count NA values

```
sum(is.na(trip_data$start_station_name))
```

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

```
sum(is.na(trip_data$end_station_name))
```

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

```
sum(is.na(trip_data$start_station_id))
```

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

```
sum(is.na(trip_data$end_station_id))
```

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

```
#view first six rows of data frame
head(trip_data)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 E19E6F1B8D4C42ED electric_bike 2021-01-23 16:14:19 2021-01-23 16:24:44
## 2 DC88F20C2C55F27F electric_bike 2021-01-27 18:43:08 2021-01-27 18:47:12
## 3 EC45C94683FE3F27 electric_bike 2021-01-21 22:35:54 2021-01-21 22:37:14
## 4 4FA453A75AE377DB electric_bike 2021-01-07 13:31:13 2021-01-07 13:42:55
## 5 BE5E8EB4E7263A0B electric_bike 2021-01-23 02:24:02 2021-01-23 02:24:45
## 6 5D8969F88C773979 electric_bike 2021-01-09 14:24:07 2021-01-09 15:17:54
##           start_station_name start_station_id end_station_name end_station_id
## 1 California Ave & Cortez St            17660                                
## 2 California Ave & Cortez St            17660                                
## 3 California Ave & Cortez St            17660                                
## 4 California Ave & Cortez St            17660
## 5 California Ave & Cortez St            17660
## 6 California Ave & Cortez St            17660
##   start_lat start_lng end_lat end_lng member_casual
## 1  41.90034 -87.69674   41.89  -87.72        member
## 2  41.90033 -87.69671   41.90  -87.69        member
## 3  41.90031 -87.69664   41.90  -87.70        member
## 4  41.90040 -87.69666   41.92  -87.69        member
## 5  41.90033 -87.69670   41.90  -87.70        casual
## 6  41.90041 -87.69676   41.94  -87.71        casual
```

```
#view columns list and data type
str(trip_data)
```

```
## ’data.frame':    5595063 obs. of  13 variables:
##  $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-0
##  $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-0
##  $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St" "California Ave
##  $ start_station_id  : chr  "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
#summary of data
summary(trip_data)
```

```
##    ride_id          rideable_type       started_at          ended_at        
##  Length:5595063     Length:5595063     Length:5595063     Length:5595063    
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  start_station_name start_station_id   end_station_name   end_station_id    
##  Length:5595063     Length:5595063     Length:5595063     Length:5595063    
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##    start_lat       start_lng         end_lat         end_lng      
##  Min.   :41.64   Min.   :-87.84   Min.   :41.39   Min.   :-88.97  
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66  
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64  
##  Mean   :41.90   Mean   :-87.65   Mean   :41.90   Mean   :-87.65  
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63  
##  Max.   :42.07   Max.   :-87.52   Max.   :42.17   Max.   :-87.49
##                                   NA’s   :4771    NA’s   :4771
##  member_casual
##  Length:5595063
##  Class :character
##  Mode  :character
##
##
##
##
```

Consistency in date format of started_at & ended_at

```r
start<-as.POSIXlt(trip_data$started_at, tz = "","%m/%d/%Y %H:%M")
start<- data.frame(start)
start<- (start[1:nrow(na.omit(start)),])
start<- data.frame(start)

start1<-as.POSIXlt(trip_data$started_at, tz = "","%Y-%m-%d %H:%M")
start1<- data.frame(start1)
start1<- (start1[(nrow(start)+1):nrow(start1),])
start1<-data.frame(start1)

names(start)<- "start"
names(start1)<- "start"
start_time<- (rbind(start,start1))
trip_data$started_at <- NULL
trip_data['started_at']<- start_time



end<-(as.POSIXlt(trip_data$ended_at, tz = "","%m/%d/%Y %H:%M"))
end<- data.frame(end)
end<- (end[1:nrow(na.omit(end)),])
end<-data.frame(end)

end1<-(as.POSIXlt(trip_data$ended_at, tz = "","%Y-%m-%d %H:%M"))
end1<- data.frame(end1)
end1<- (end1[(nrow(end)+1):nrow(end1),])
end1<-data.frame(end1)

names(end)<- "end"
names(end1)<- "end"
end_time<- (rbind(end,end1))
trip_data$ended_at<- NULL
trip_data['ended_at']<- end_time
```

Removing NA values

```r
trip_data <- na.omit(trip_data)
```

Adding weekday column at which each trip started

```
trip_data$started_at <- as.Date(trip_data$started_at, format = "%Y-%m-%d %H:%M:%S")
trip_data$ended_at <- as.Date(trip_data$ended_at, format = "%Y-%m-%d %H:%M:%S")
trip_data$weekday <- weekdays(trip_data$started_at, abbreviate = FALSE)
```

Creating momth column in which each trip started

```
trip_data$month <- format(trip_data$started_at, "%m")
```

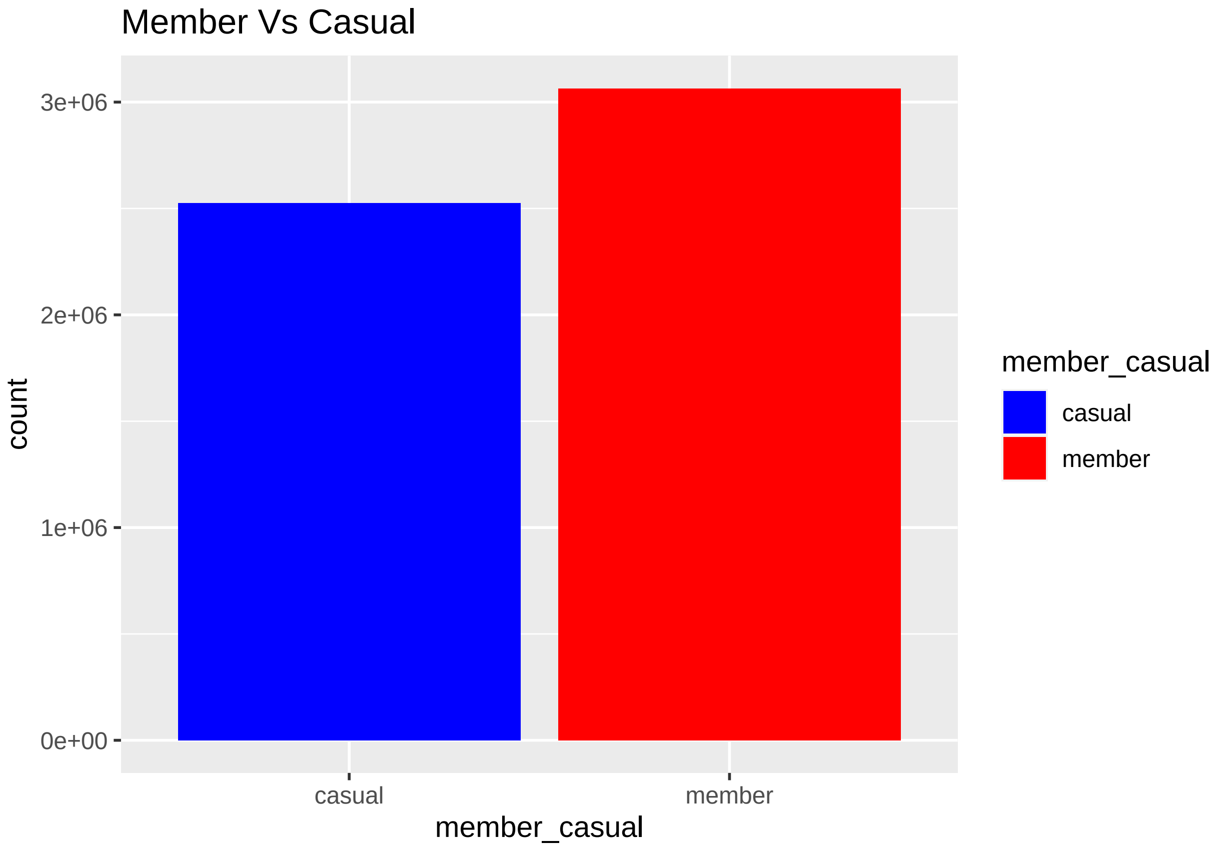
Analyze phase

Important tasks: i- Organize and format data ii- To make data useful and accesible, aggregate it. iii- Identify trends and calculations iv- Calculations **Member and casual users**

```
trip_data%>% group_by(member_casual)%>% summarise(n=n())%>%
  mutate(percent = n*100/sum(n))
```

```
## # A tibble: 2 x 3
##   member_casual       n percent
##   <chr>           <int>   <dbl>
## 1 casual        2525443    45.2
## 2 member        3064735    54.8
```

```
ggplot(data = trip_data, mapping = aes(x = member_casual, fill = member_casual)) +
  geom_bar() +
  scale_fill_manual(values = c("blue", "red")) +
  labs(title = "Member Vs Casual")
```

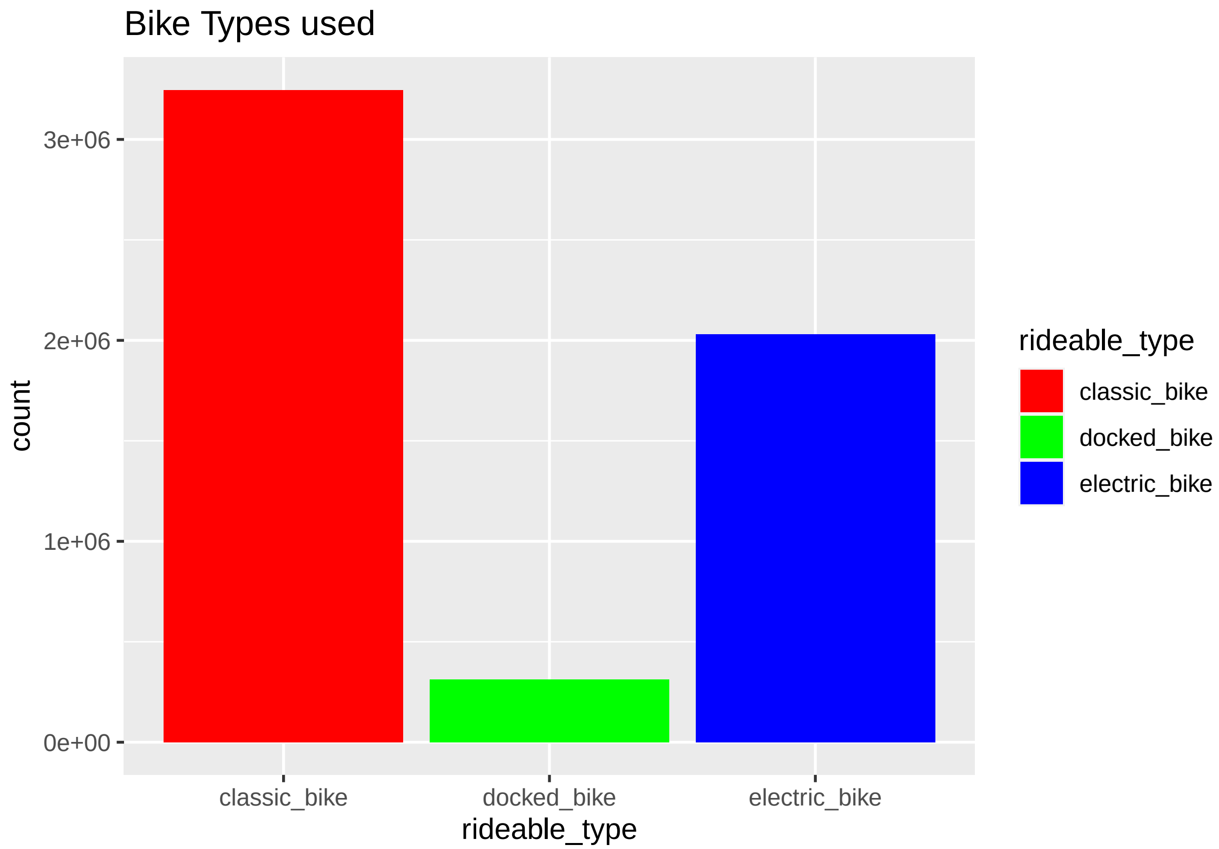


Types of Bikes used

```
trip_data %>%
  group_by(rideable_type) %>%
  summarise(n = n()) %>%
  mutate(percent = n * 100 / sum(n))
```

```
## # A tibble: 3 x 3
##   rideable_type       n percent
##   <chr>           <int>   <dbl>
## 1 classic_bike  3246497   58.1
## 2 docked_bike    312022    5.58
## 3 electric_bike 2031659   36.3
```

```
ggplot(data = trip_data, mapping = aes(x = rideable_type, fill = rideable_type)) +
  geom_bar() +
  scale_fill_manual(values = c("red", "green", "blue", "purple")) +
  labs(title = "Bike Types used")
```



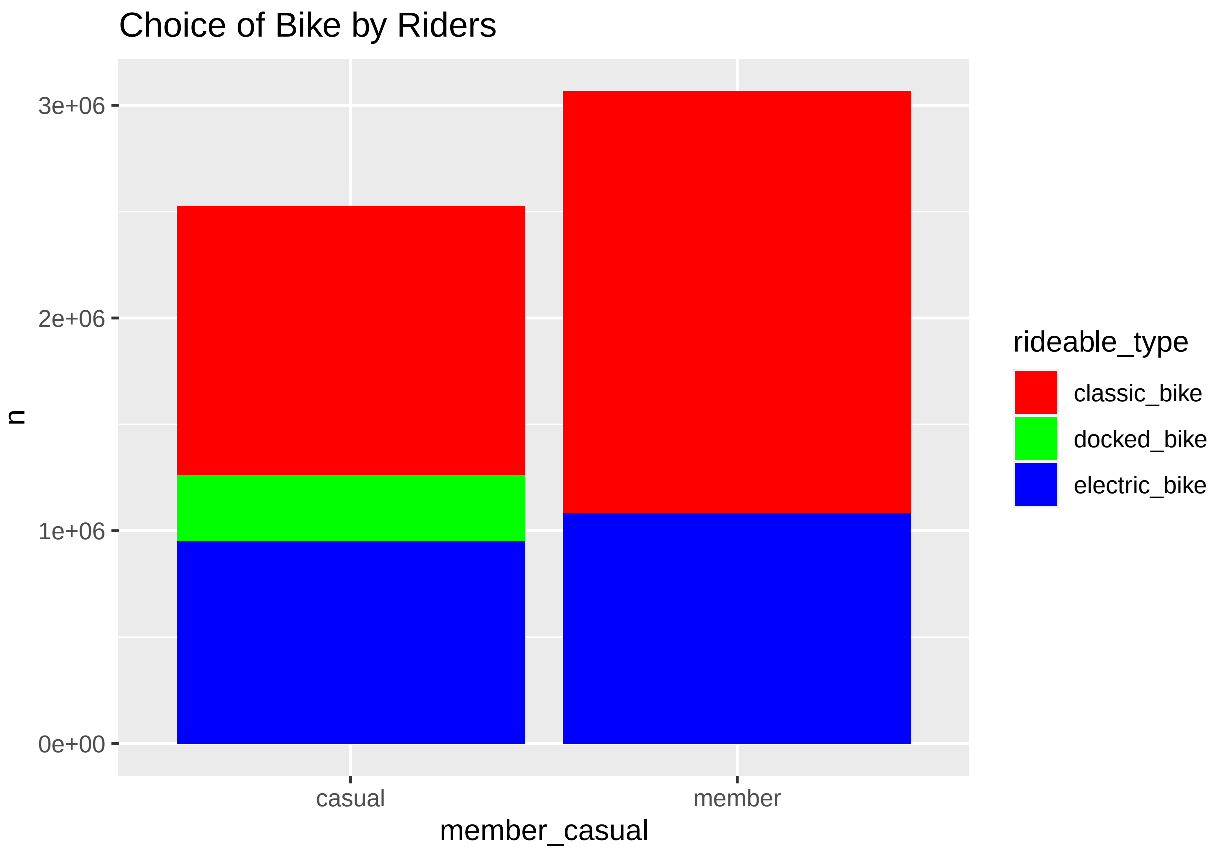
Member type

```
member_type<-trip_data%>% group_by(member_casual,rideable_type) %>% summarise(n=n())%>%  mutate(percent
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

Choice of bike by riders

```
ggplot(data = as.data.frame(member_type),mapping= aes(x= member_casual, y=n, fill =rideable_type)) +geo
```



i- classical bike is most used than electrical

ii- Docked one is least useable

iii- Electrical bikes are almost equally liked by both

Weektable

```
trip_data$weekday <- factor(trip_data$weekday, levels = c("Monday", "Tuesday", "Wednesday", "Thursday",

weektable <- trip_data %>%
  group_by(weekday) %>%
  summarise(n = n()) %>%
  mutate(percent = n * 100 / sum(n))
```

Type of bike with rider

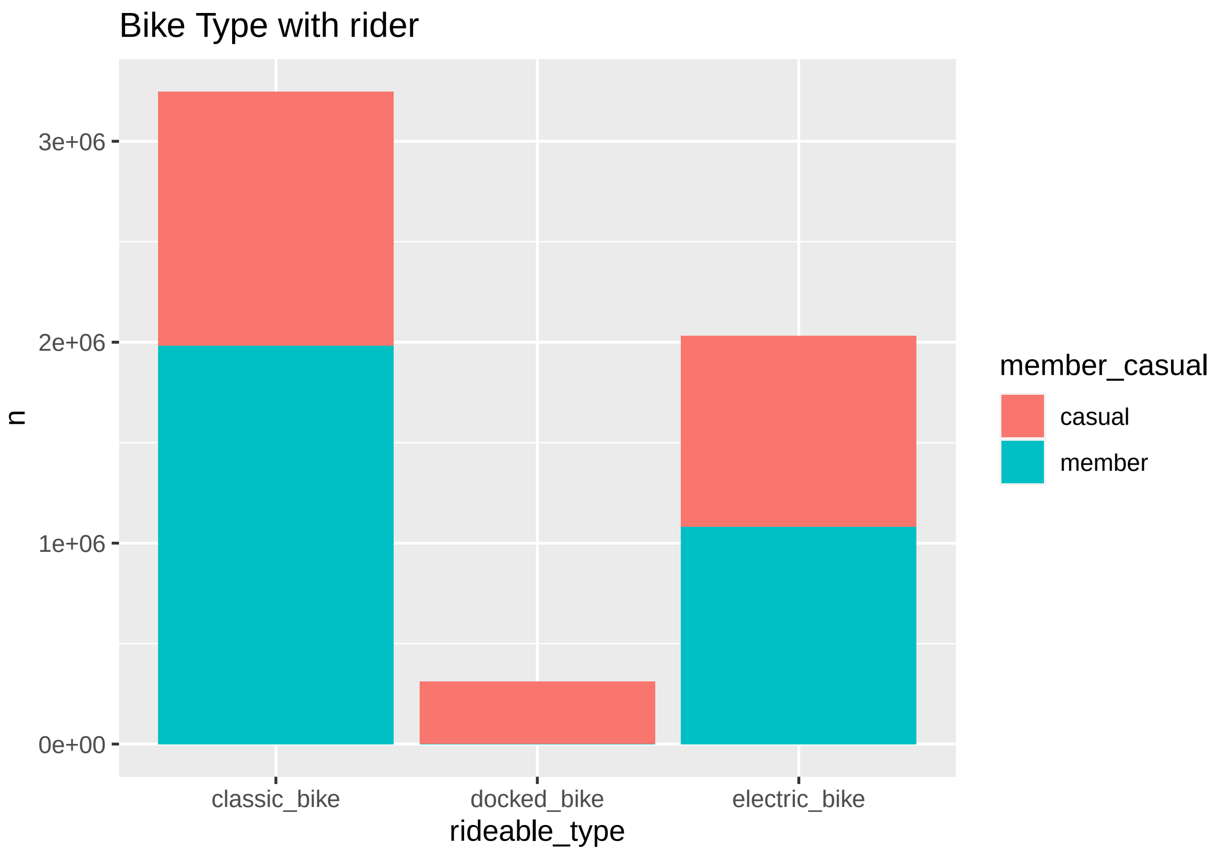
```
rideable_type<-trip_data%>% group_by(rideable_type,member_casual) %>% summarise(n=n())%>%
  mutate(percent = n*100/sum(n))
```

```
## 'summarise()' has grouped output by 'rideable_type'. You can override using the
## '.groups' argument.
```

```
rideable_type
```

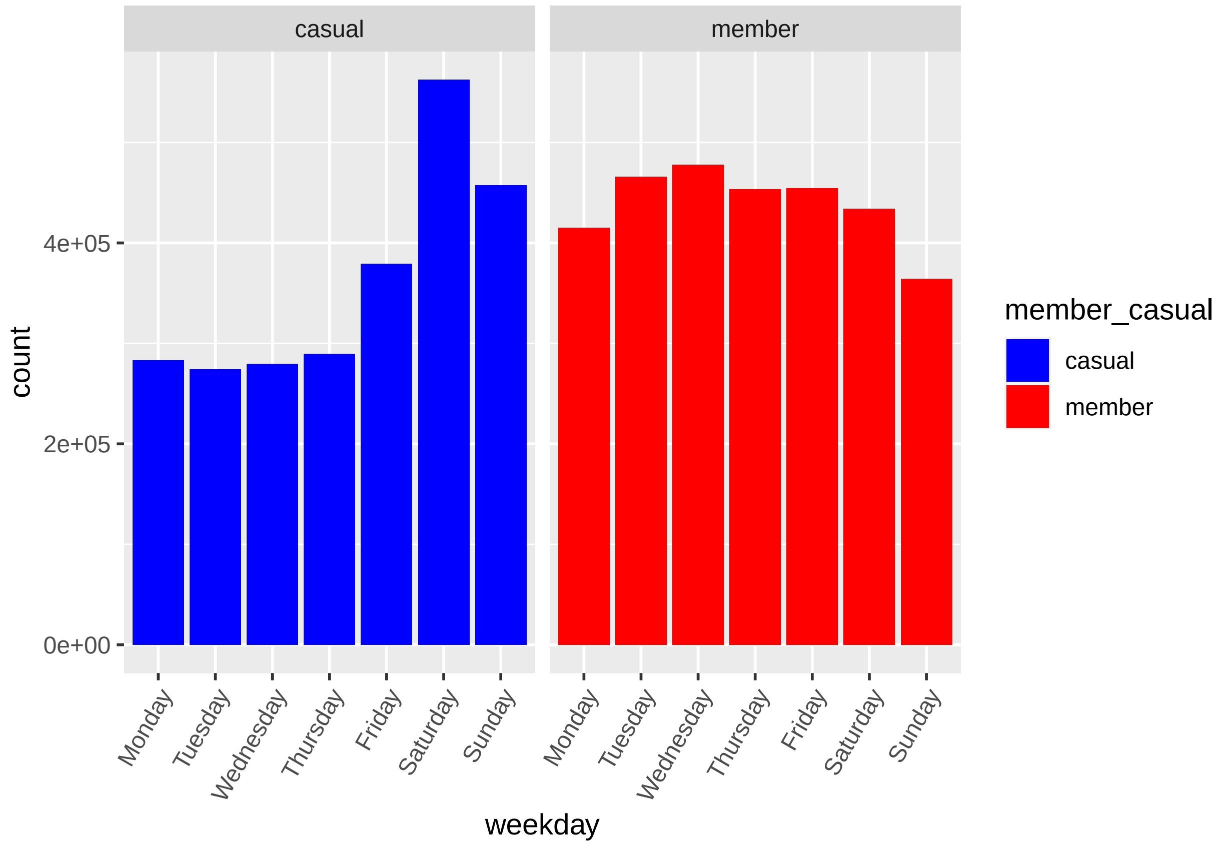
```
## # A tibble: 6 x 4
## # Groups:   rideable_type [3]
##   rideable_type member_casual       n    percent
##   <chr>         <chr>           <int>      <dbl>
## 1 classic_bike  casual        1263430   38.9
## 2 classic_bike  member        1983067   61.1
## 3 docked_bike   casual         312021  100.
## 4 docked_bike   member              1    0.000320
## 5 electric_bike casual         949992   46.8
## 6 electric_bike member        1081667   53.2
```

```
ggplot(data = as.data.frame(rideable_type),mapping= aes(x= rideable_type, y = n, fill =member_casual))
```



Usage per day

```
ggplot(data = trip_data, mapping = aes(x = weekday, fill = member_casual)) +
  geom_bar() +
  facet_wrap(~member_casual) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  scale_fill_manual(values = c("blue", "red"))
```



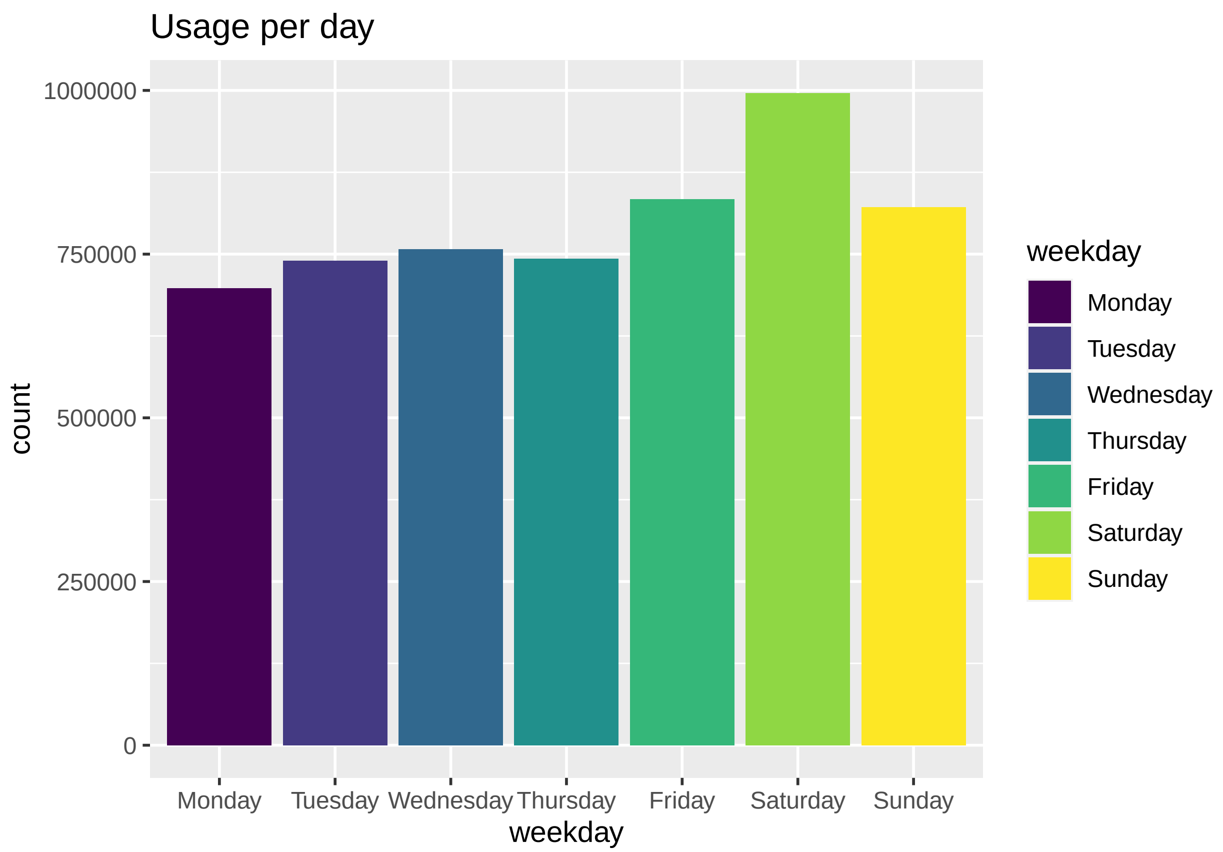
Each day usage of bikes

```
trip_data$weekday<- factor(trip_data$weekday, levels= c("Monday", "Tuesday","Wednesday","Thursday","Fri
 weektable<- trip_data%>% group_by(weekday)%>% summarise(n=n())%>% mutate(percent = n*100/sum(n))

 weektable
```

```
## # A tibble: 7 x 3
##   weekday        n percent
##   <fct>      <int>   <dbl>
## 1 Monday    697820    12.5
## 2 Tuesday   740280    13.2
## 3 Wednesday 757175    13.5
## 4 Thursday  743111    13.3
## 5 Friday    833746    14.9
## 6 Saturday  996231    17.8
## 7 Sunday    821815    14.7
```

```
 ggplot(data = trip_data, aes(x = weekday, fill = weekday)) +
     geom_bar() +
     scale_fill_viridis_d() +
     labs(title = "Usage per day")
```



Summarize

```
trip_data %>%
  group_by(member_casual) %>%
  summarise(n = n()) %>%
  mutate(percent = n * 100 / sum(n))
```

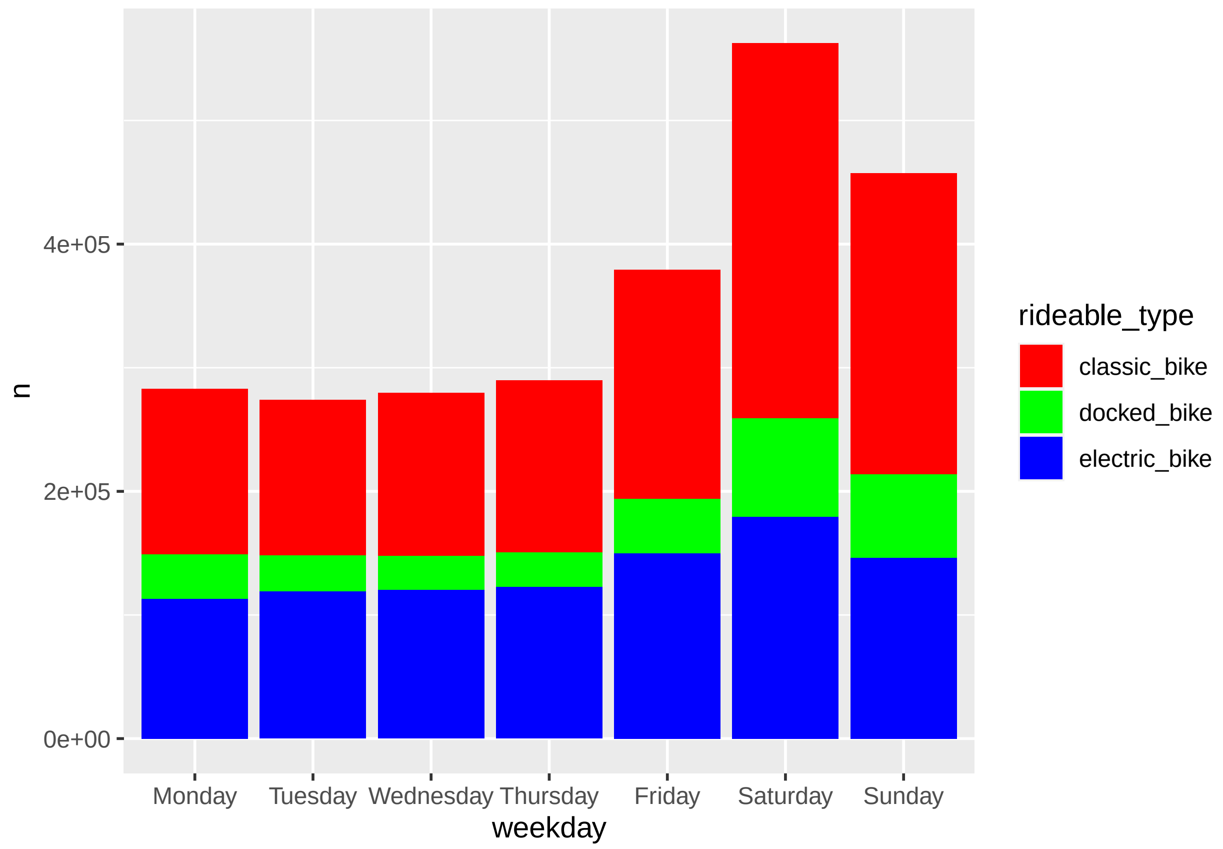
```
## # A tibble: 2 x 3
##   member_casual       n percent
##   <chr>           <int>   <dbl>
## 1 casual        2525443    45.2
## 2 member        3064735    54.8
```

Casual riders trend of using bikes

```
casual_riders<-trip_data%>% filter(member_casual == 'casual')%>%group_by(weekday,rideable_type)%>% summa
```

```
## `summarise()` has grouped output by 'weekday'. You can override using the
## `.groups` argument.
```

```
ggplot(data = casual_riders, aes(x = weekday, y = n, fill = rideable_type)) +
  geom_bar(stat = 'identity') +
  scale_fill_manual(values = c("electric_bike" = "blue", "classic_bike" = "red", "docked_bike" = "green
```

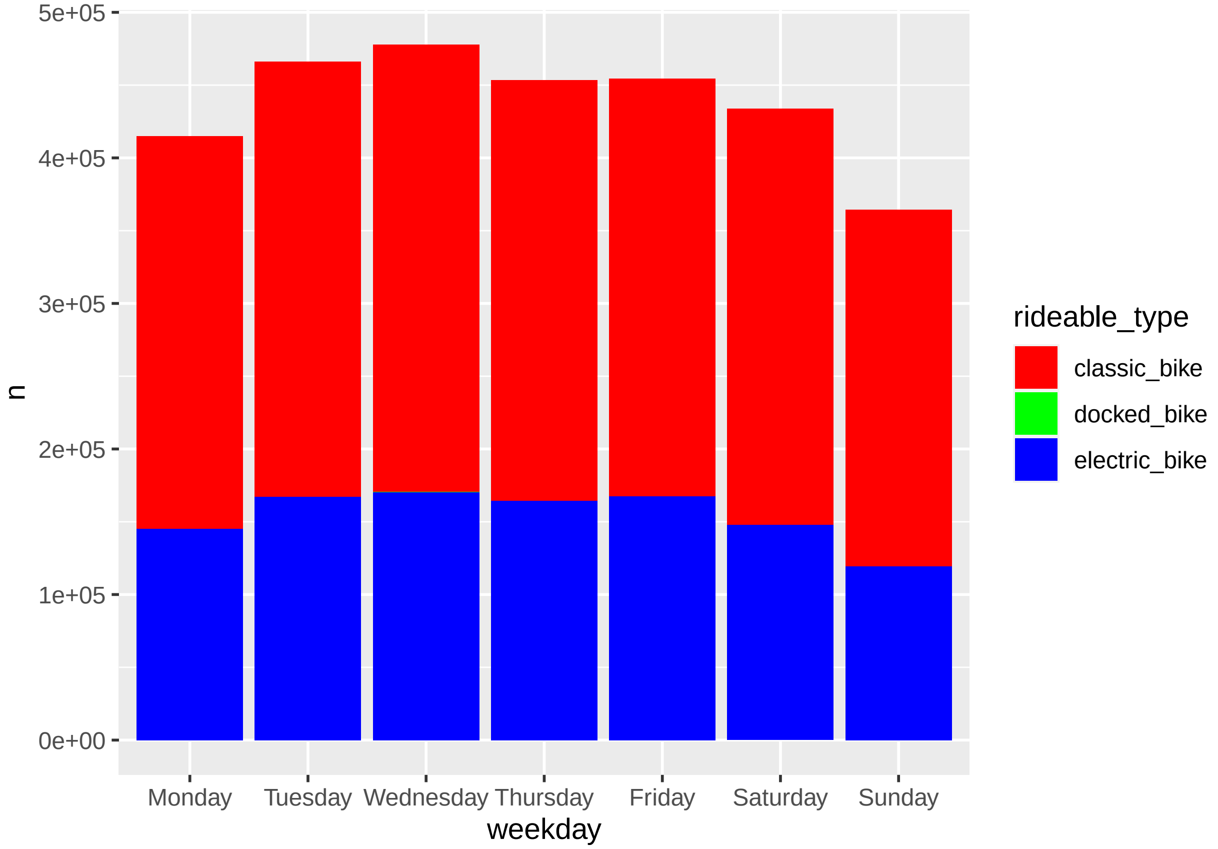


Members trend of using bikes

```
members<-trip_data%>% filter(member_casual == 'member')%>%group_by(weekday,rideable_type)%>% summarise(
```

```
## ‘summarise()‘ has grouped output by ’weekday’. You can override using the
## ‘.groups‘ argument.
```

```
ggplot(data = members, aes(x = weekday, y = n, fill = rideable_type)) +
     geom_bar(stat = 'identity') +
    scale_fill_manual(values = c("electric_bike" = "blue", "classic_bike" = "red", "docked_bike" = "gre
```

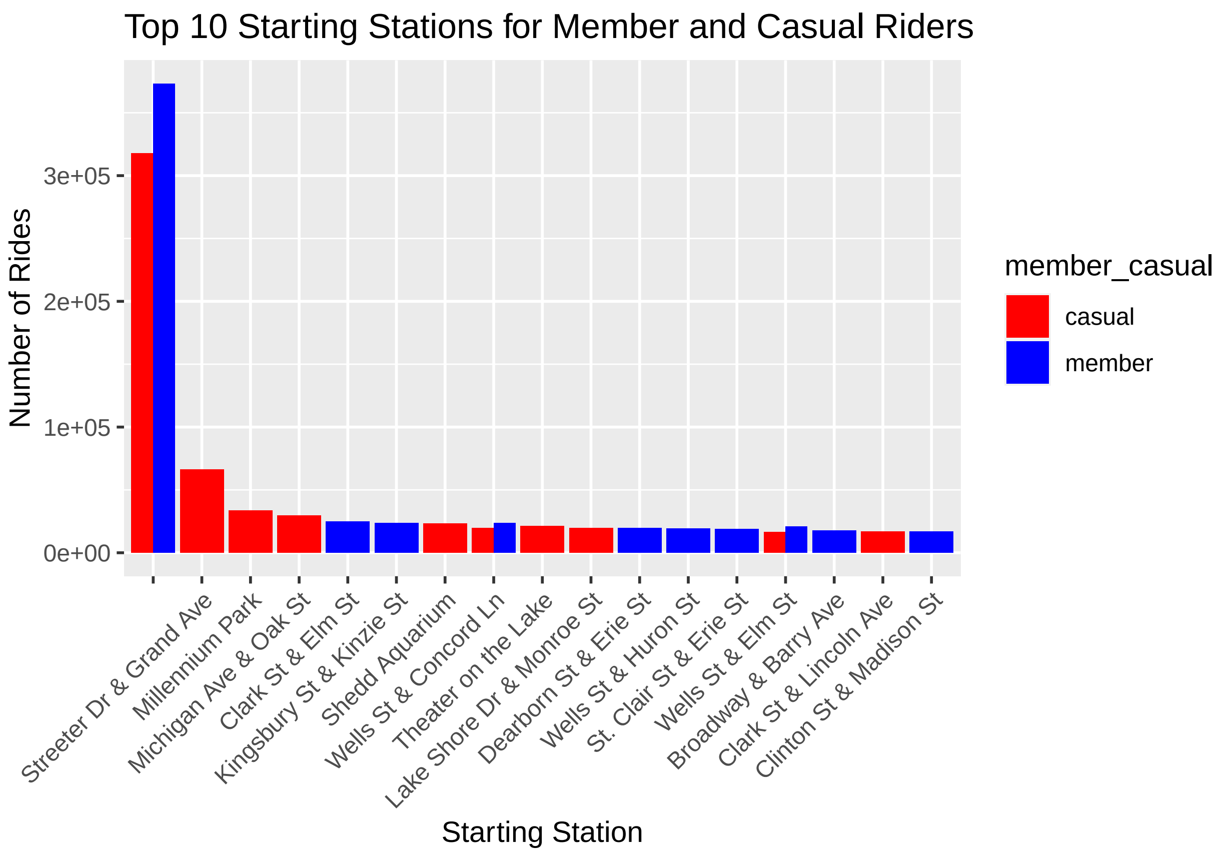


Top starting stations used by member and casual riders

```
top_start_stations <- trip_data %>%
     group_by(start_station_name, member_casual) %>%
     summarise(n = n()) %>%
     arrange(member_casual, desc(n)) %>%
     group_by(member_casual) %>%
     top_n(10, n)
```

```
## `summarise()` has grouped output by 'start_station_name'. You can override
## using the `.groups` argument.
```

```
ggplot(top_start_stations, aes(x = reorder(start_station_name, -n), y = n, fill = member_casual)) +
     geom_col(position = "dodge") +
     scale_fill_manual(values = c("red", "blue")) +
     labs(title = "Top 10 Starting Stations for Member and Casual Riders",
          x = "Starting Station",
          y = "Number of Rides") +
     theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



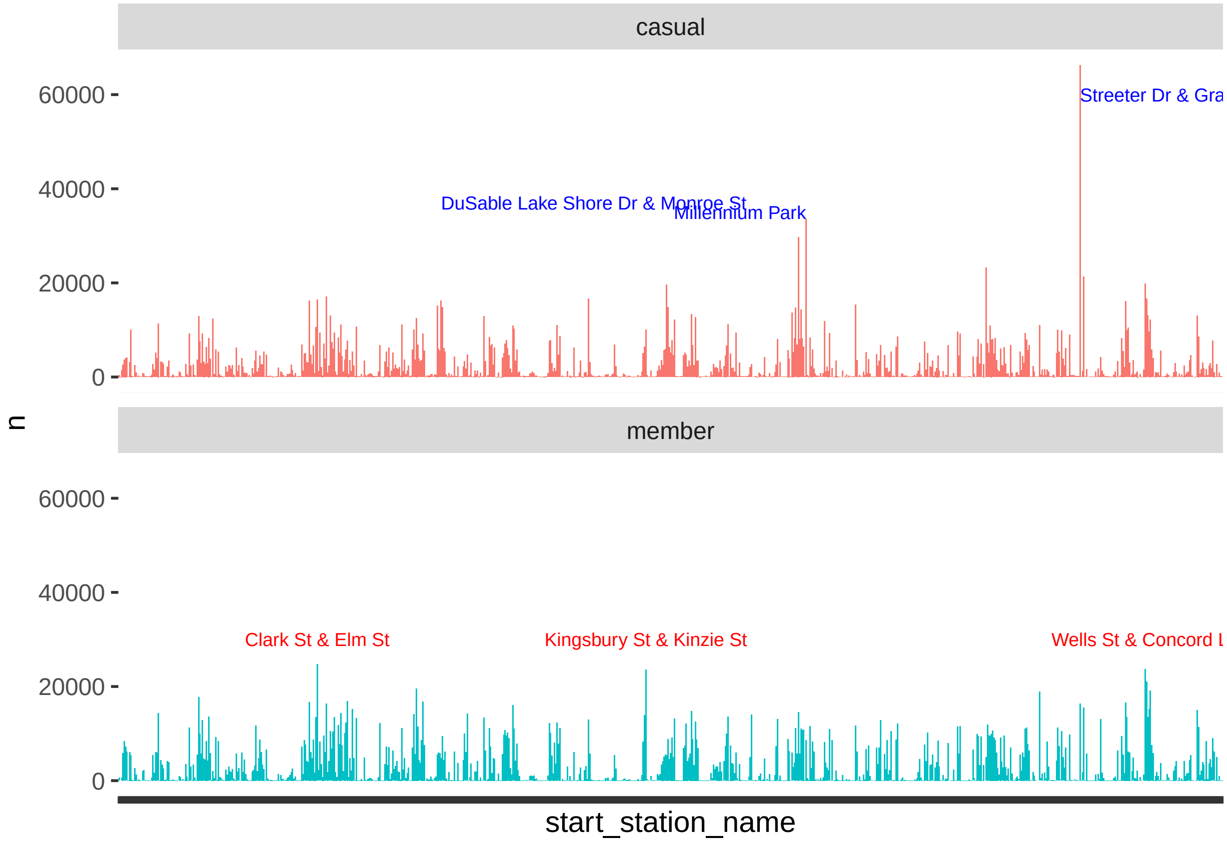
```r
ss <- trip_data %>%
     filter(start_station_name != "") %>%
     group_by(start_station_name, member_casual) %>%
     summarise(n = n()) %>%
     arrange(desc(n))
```

```
## `summarise()` has grouped output by 'start_station_name'. You can override
## using the `.groups` argument.
```

```r
p<- ggplot(data =as.data.frame(ss), aes(x = start_station_name, y=n,fill=member_casual))+
     geom_bar(stat = 'identity')+facet_wrap(~ member_casual, ncol =1)+theme(axis.text.x = element_blank

dat_text <- data.frame(
  label = c("Streeter Dr & Grand Ave", "DuSable Lake Shore Dr & Monroe St", "Millennium Park","Clark St
  member_casual = c("casual", "casual", "casual", "member", "member", "member"),
  x = c("Streeter Dr & Grand Ave", "DuSable Lake Shore Dr & Monroe St", "Millennium Park","Clark St & El
  y = c(60000, 37000, 35000, 30000, 30000, 30000),
  hjust = c(0, 0, 1, 0.5, 0.5, 0.5),
  angle = c(0, 0, 0, 0, 0, 0)
)


p + geom_text(data = dat_text, mapping = aes(x = x, y = y, label = label, hjust = hjust, angle = angle,
     scale_color_manual(values = c("member" = "red", "casual" = "blue"))
```



i- casual users start and end rides from same station.

ii- Lakeshore drive is place where bussiest stations located which is used by casual riders.

Suggestions

There are several marketing strategies that could be implemented to encourage casual riders to become annual members:

i- Offer occasional membership discounts to casual riders, particularly during the summer and on weekends.

ii- Increase the rental price of bikes on weekends, especially for classic and electric bikes, which are preferred more by casual users. This may encourage them to consider purchasing an annual membership instead.

iii- Place banners or advertisements offering special discounts at Lake Shore Drive, specifically targeting casual users, with the hope of encouraging them to become members.

Thanks for spending your time to read, please give your valuable feedback