Cyclist R Notebook

Sakshi Gupta

2024-08-16

Setting up my enviroment

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(RColorBrewer)
library(scales)
```

loading required datasets

```
setwd("C:/Users/HP/Documents/Sakshi/")

df_1<-read.csv("202301-divvy-tripdata.csv")
df_2<-read.csv("202302-divvy-tripdata.csv")
df_3<-read.csv("202303-divvy-tripdata.csv")
df_4<-read.csv("202304-divvy-tripdata.csv")
df_5<-read.csv("202305-divvy-tripdata.csv")
df_6<-read.csv("202306-divvy-tripdata.csv")
df_7<-read.csv("202307-divvy-tripdata.csv")
df_8<-read.csv("202308-divvy-tripdata.csv")
df_9<-read.csv("202309-divvy-tripdata.csv")
df_10<-read.csv("202310-divvy-tripdata.csv")
df_11<-read.csv("202311-divvy-tripdata.csv")
df_12<-read.csv("202312-divvy-tripdata.csv")
```

combining data in one data frame

```
df<-bind_rows(df_1,df_2,df_3,df_4, df_5, df_6 , df_7 , df_8 , df_9 , df_10, df_11, df_12)
```

Exploring the Data

```
head(df)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 F96D5A74A3E41399 electric_bike 2023-01-21 20:05:42 2023-01-21 20:16:33
## 2 13CB7EB698CEDB88  classic_bike 2023-01-10 15:37:36 2023-01-10 15:46:05
## 3 BD88A2E670661CE5 electric_bike 2023-01-02 07:51:57 2023-01-02 08:05:11
## 4 C90792D034FED968  classic_bike 2023-01-22 10:52:58 2023-01-22 11:01:44
## 5 3397017529188E8A  classic_bike 2023-01-12 13:58:01 2023-01-12 14:13:20
## 6 58E68156DAE3E311 electric_bike 2023-01-31 07:18:03 2023-01-31 07:21:16
##              start_station_name start_station_id               end_station_name
## 1   Lincoln Ave & Fullerton Ave     TA1309000058      Hampden Ct & Diversey Ave
## 2         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave & 47th St
## 3        Western Ave & Lunt Ave           RP-005 Valli Produce - Evanston Plaza
## 4         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave & 47th St
## 5         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave & 47th St
## 6 Lakeview Ave & Fullerton Pkwy     TA1309000019      Hampden Ct & Diversey Ave
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1       202480.0  41.92407 -87.64628 41.93000 -87.64000        member
## 2   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 3            599  42.00857 -87.69048 42.03974 -87.69941        casual
## 4   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 5   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 6       202480.0  41.92607 -87.63886 41.93000 -87.64000        member
```

```
str(df)
```

```
## 'data.frame':    5719877 obs. of  13 variables:
##  $ ride_id           : chr  "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C907
92D034FED968" ...
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "electric_bike" "classic_bike"
...
##  $ started_at        : chr  "2023-01-21 20:05:42" "2023-01-10 15:37:36" "2023-01-02 07:51:
57" "2023-01-22 10:52:58" ...
##  $ ended_at          : chr  "2023-01-21 20:16:33" "2023-01-10 15:46:05" "2023-01-02 08:05:
11" "2023-01-22 11:01:44" ...
##  $ start_station_name: chr  "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western
Ave & Lunt Ave" "Kimbark Ave & 53rd St" ...
##  $ start_station_id  : chr  "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
##  $ end_station_name  : chr  "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli P
roduce - Evanston Plaza" "Greenwood Ave & 47th St" ...
##  $ end_station_id    : chr  "202480.0" "TA1308000002" "599" "TA1308000002" ...
##  $ start_lat         : num  41.9 41.8 42 41.8 41.8 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.8 42 41.8 41.8 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "casual" "member" ...
```

```
summary(df)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:5719877     Length:5719877     Length:5719877     Length:5719877
##  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:5719877     Length:5719877     Length:5719877     Length:5719877
##  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.   :41.63   Min.   :-87.94   Min.   : 0.00   Min.   :-88.16
##  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.   :-87.46   Max.   :42.18   Max.   :  0.00
##                                   NA's   :6990    NA's   :6990
##  member_casual
##  Length:5719877
##  Class :character
##  Mode  :character
##
##
##
##
```

Standardizing date

```
df$started_at<-strptime(df$started_at,format="%Y-%m-%d %H:%M:%S")
df$ended_at<-strptime(df$ended_at,format="%Y-%m-%d %H:%M:%S")
```

Converting column type to factor

```
unique(df$rideable_type)
```

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

```
df<-df %>% mutate(rideable_type=factor(rideable_type))
unique(df$member_casual)
```

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

```
df<- df%>% mutate(member_casual=factor(member_casual))
```

checking for Na values

```
summarise(df,across(everything(),~sum(is.na(.))))
```

```
##   ride_id rideable_type started_at ended_at start_station_name start_station_id
## 1       0             0          0        0                  0                0
##   end_station_name end_station_id start_lat start_lng end_lat end_lng
## 1                0              0         0         0    6990    6990
##   member_casual
## 1             0
```

checking distinct values in each column

```
summarise(df, across(everything(),~sum(n_distinct(.))))
```

```
##   ride_id rideable_type started_at ended_at start_station_name start_station_id
## 1 5719877             3    4823909  4835702               1593             1517
##   end_station_name end_station_id start_lat start_lng end_lat end_lng
## 1             1598           1521    789704    748737   13885   14003
##   member_casual
## 1             2
```

adding appropriate columns

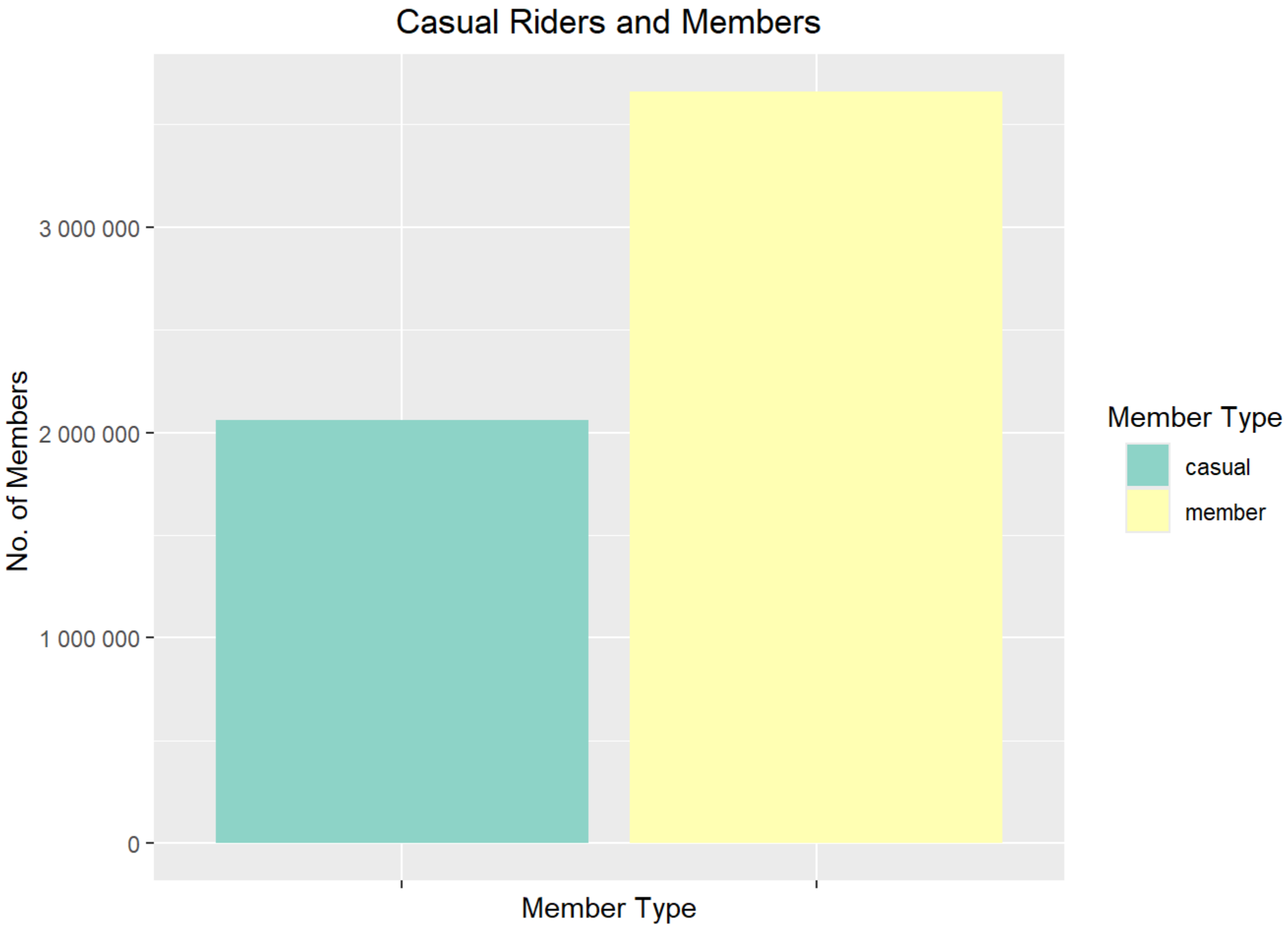
```
df <-df %>% mutate(ride_length=ended_at-started_at, .after = ended_at) %>% mutate(ride_length
=ride_length/60)
df<- df %>% mutate(wkday=weekdays(started_at)) %>% mutate(wkday=factor(wkday,levels=c('Monda
y','Tuesday','Wednesday','Thursday','Friday','Saturday','Sunday')))
df$ride_length<-as.numeric(df$ride_length)
df <- df %>% mutate(mnth=months(started_at)) %>% mutate(mnth = factor(mnth, levels = c('Janua
ry','February',"March"    ,  "April"   ,   "May"     ,   "June"     ,  "July"      , "August" , "S
eptember", "October"  , "November" , "December" )))
```

Filtering out negative ride length

```
df<-filter(df,ride_length>=0)
```

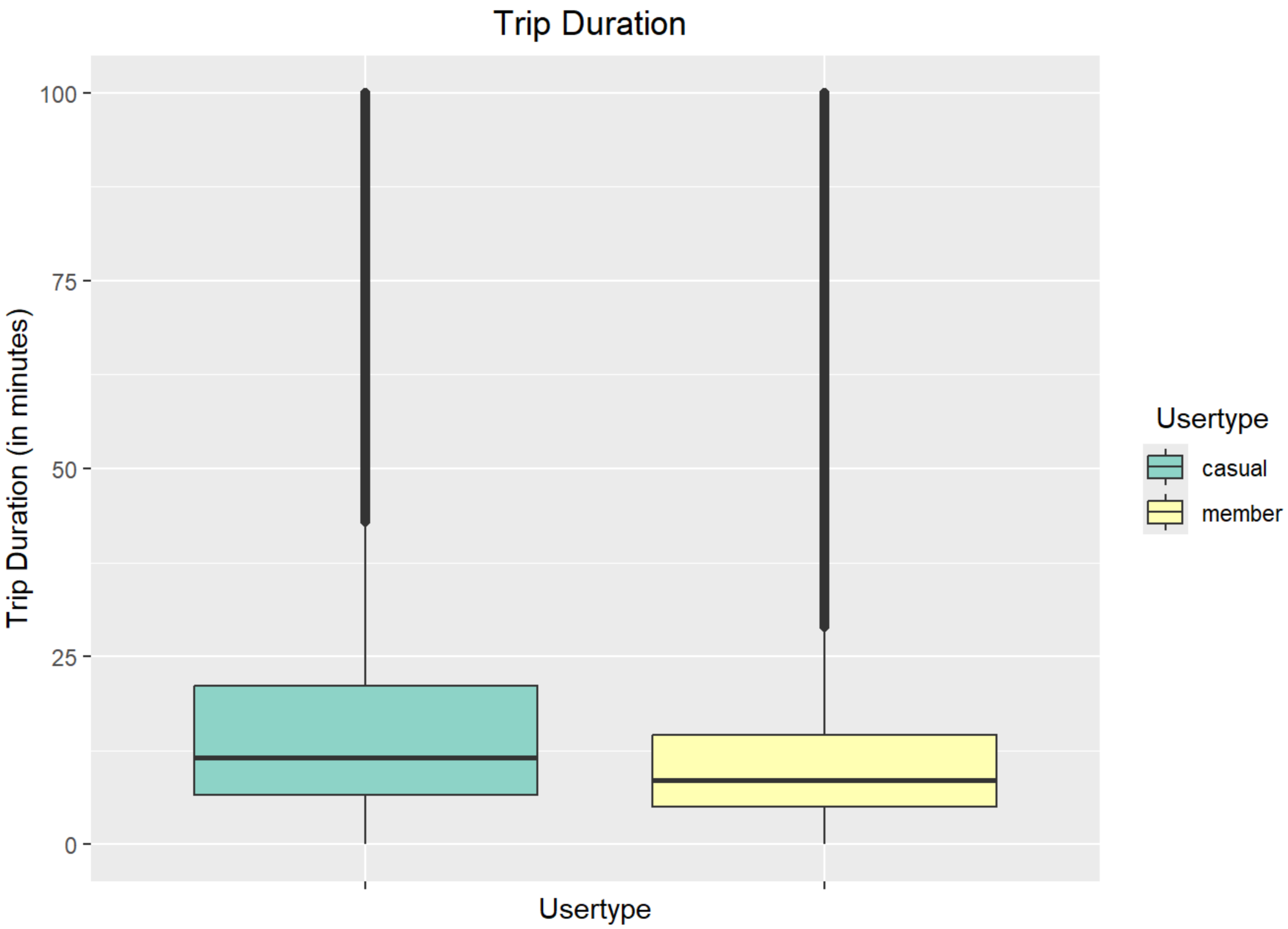
No, of Different Users

```
df%>% count(member_casual) %>% ggplot(aes(x=member_casual , y=n, fill=member_casual))+geom_co
l() +
  theme( axis.text.x = element_blank() ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5, ) +
  labs(  x='Member Type' ,y= "No. of Members " , title= "Casual Riders and Members", fill='Me
mber Type') +
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(labels = label_number())
```



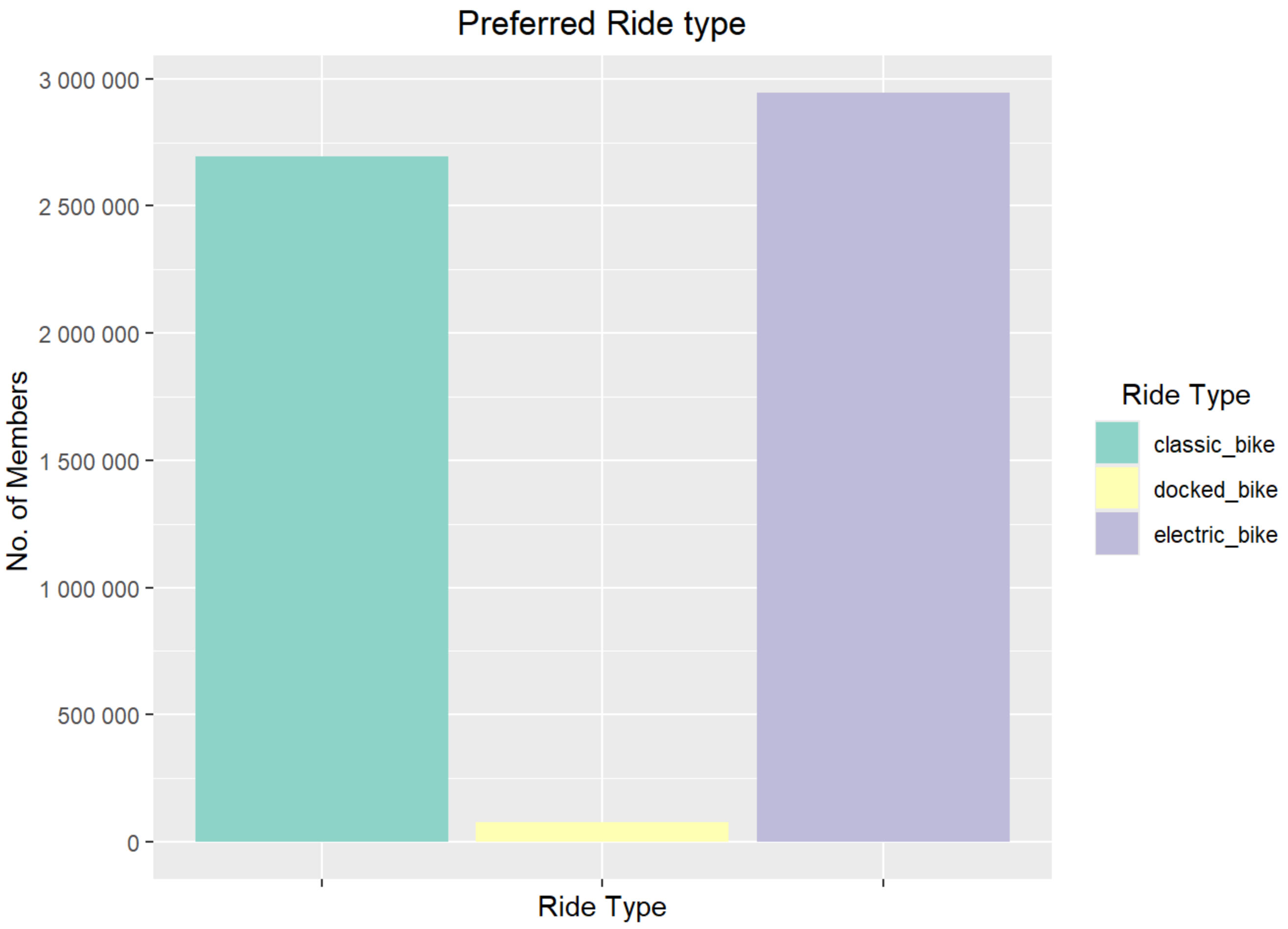
trip duration Boxplot

```
df %>%
  ggplot(aes(y=ride_length, x=member_casual, fill=member_casual))+geom_boxplot()+ ylim(0,100)
+
  theme( axis.text.x = element_blank() ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5) +
  labs(fill=" Usertype ", x="Usertype" , y= "Trip Duration (in minutes)" , title= "Trip Durat
ion ") +
  scale_fill_brewer(palette = "Set3")
```

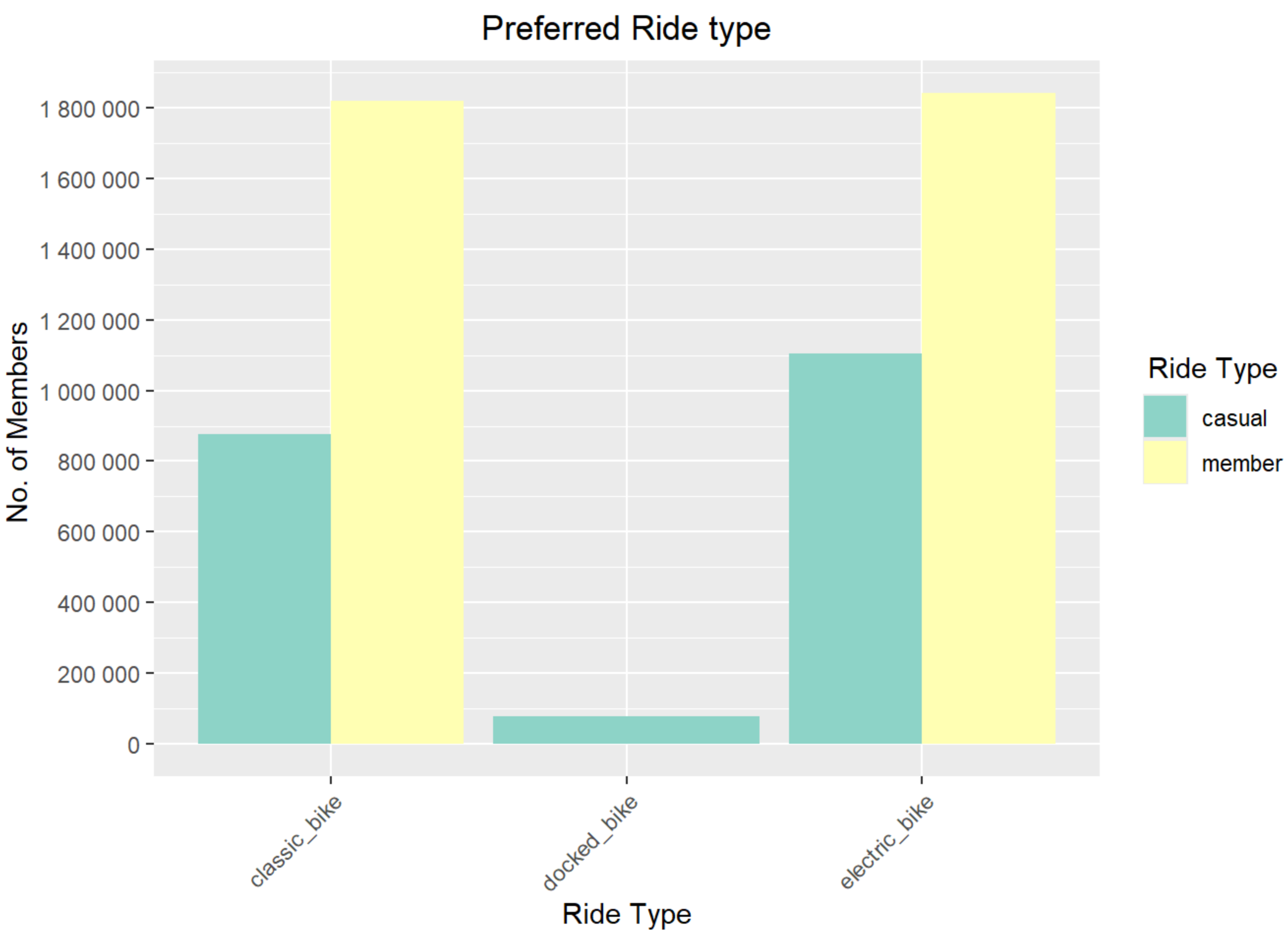


Ridetype

```
df%>% count(rideable_type) %>% ggplot(aes(x=rideable_type , y=n, fill=rideable_type))+geom_co
l() +
  theme( axis.text.x = element_blank() ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5, ) +
  labs(x="Ride Type", y= "No. of Members " , title= "Preferred Ride type", fill='Ride Type')
+
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(labels = label_number() ,breaks = pretty_breaks(10))
```

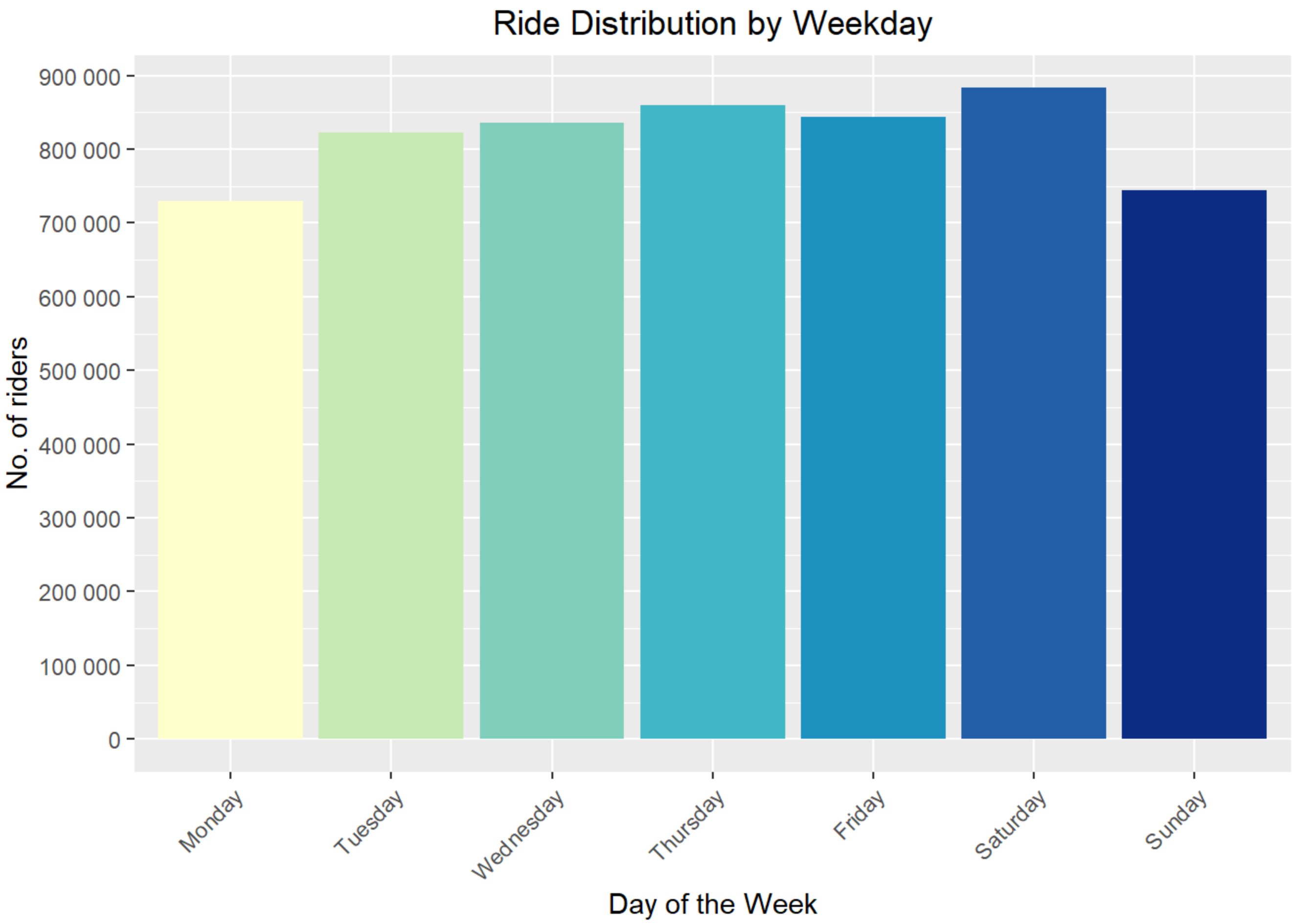


```
df%>% group_by(member_casual) %>% count(rideable_type) %>% ggplot(aes(x=rideable_type , y=n,
fill=member_casual))+geom_bar(stat='identity', position = 'dodge') +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1),
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5, ) +
  labs(x="Ride Type", y= "No. of Members " , title= "Preferred Ride type", fill='Ride Type')
+
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(labels = label_number(), ,breaks = pretty_breaks(10))
```

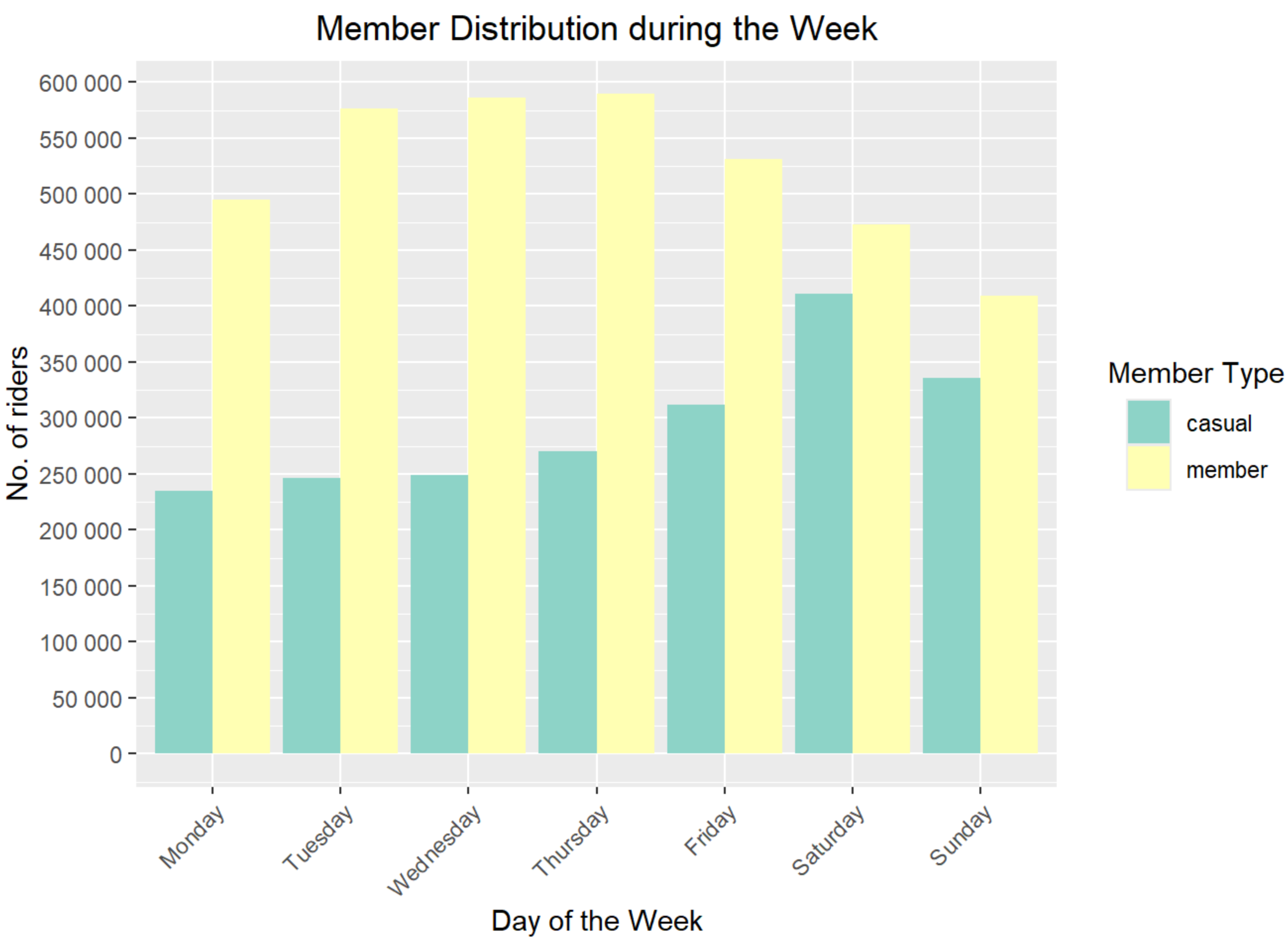


Ridership during the weekday

```
df %>% count(wkday) %>% ggplot(aes(x=wkday , y=n, fill=wkday))+geom_col() +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1) ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5, legend.position =
'none') +
  labs(fill=" Day ", x="Day of the Week" , y= "No. of riders" , title= "Ride Distribution by
Weekday") +
  scale_fill_brewer(palette = "YlGnBu") +
  scale_y_continuous(labels = label_number(),breaks = pretty_breaks(10))
```

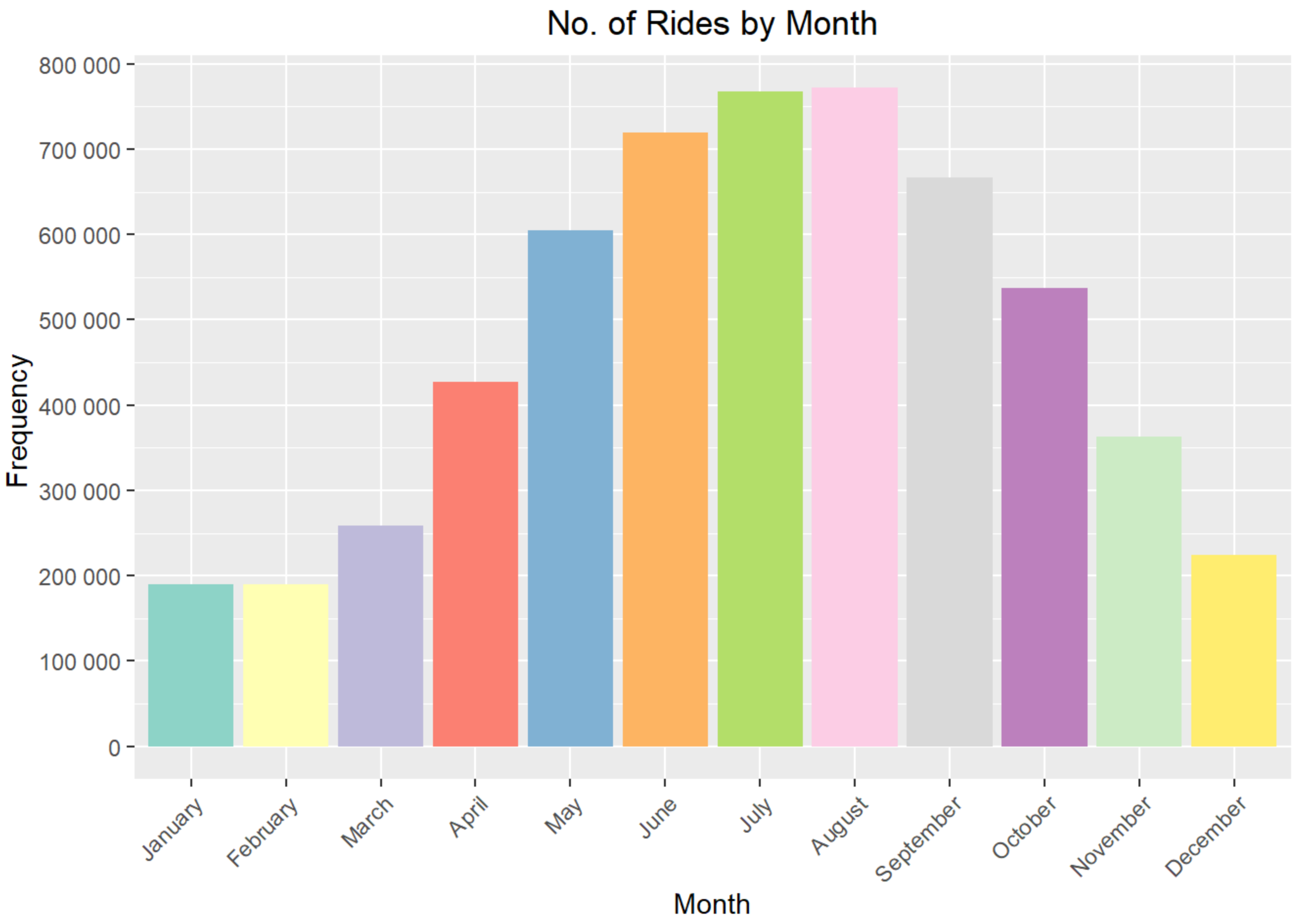


```
df %>% group_by(member_casual)  %>% count(wkday) %>% ggplot(aes(x=wkday , y=n, fill=member_ca
sual))+geom_bar(stat='identity', position = 'dodge') +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1) ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5) +
  labs(fill=" Member Type ", x="Day of the Week" , y= "No. of riders" , title= "Member Distri
bution during the Week") +
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(labels = label_number(), breaks = pretty_breaks(10))
```

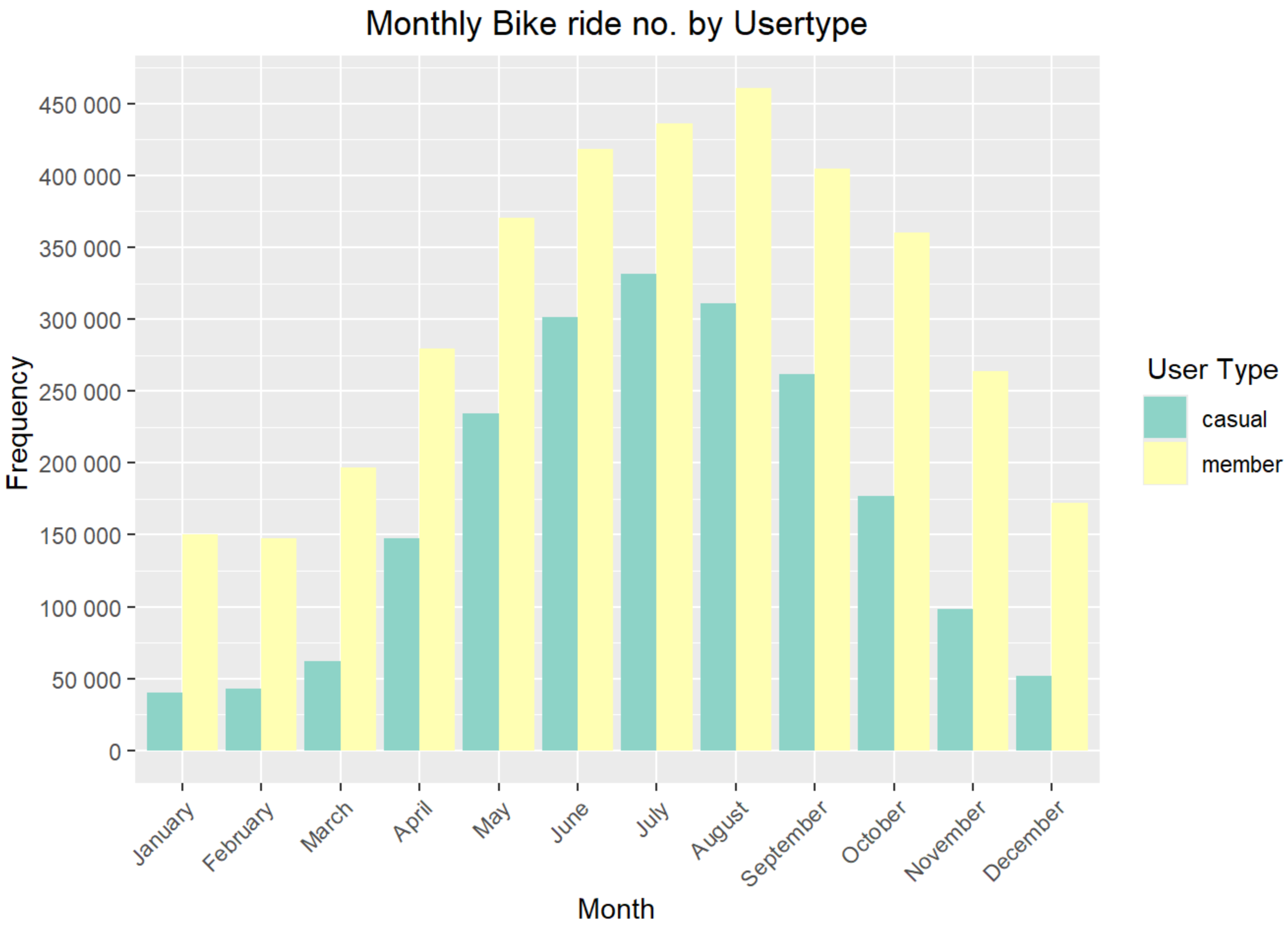


No. of rides by month

```
df %>% count(mnth) %>%  ggplot(aes(x=mnth , y=n, fill=mnth))+geom_col() +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1) ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5, legend.position =
'none') +
  labs( x="Month" , y= "Frequency " , title= "No. of Rides by Month", fill='Month') +
  scale_fill_brewer(palette = 'Set3') +
  scale_y_continuous(labels = label_number(),breaks = pretty_breaks(10))
```



```
df %>% group_by(member_casual) %>%  count(mnth) %>%  ggplot(aes(x=mnth , y=n, fill=member_cas
ual))+geom_bar(position = 'dodge', stat='identity') +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1) ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5) +
  labs( x="Month" , y= "Frequency " , title= "Monthly Bike ride no. by Usertype", fill='User
Type') +
  scale_y_continuous(labels = label_number(), breaks = pretty_breaks(10))+
  scale_fill_brewer(palette = "Set3")
```



Station with most no. rides

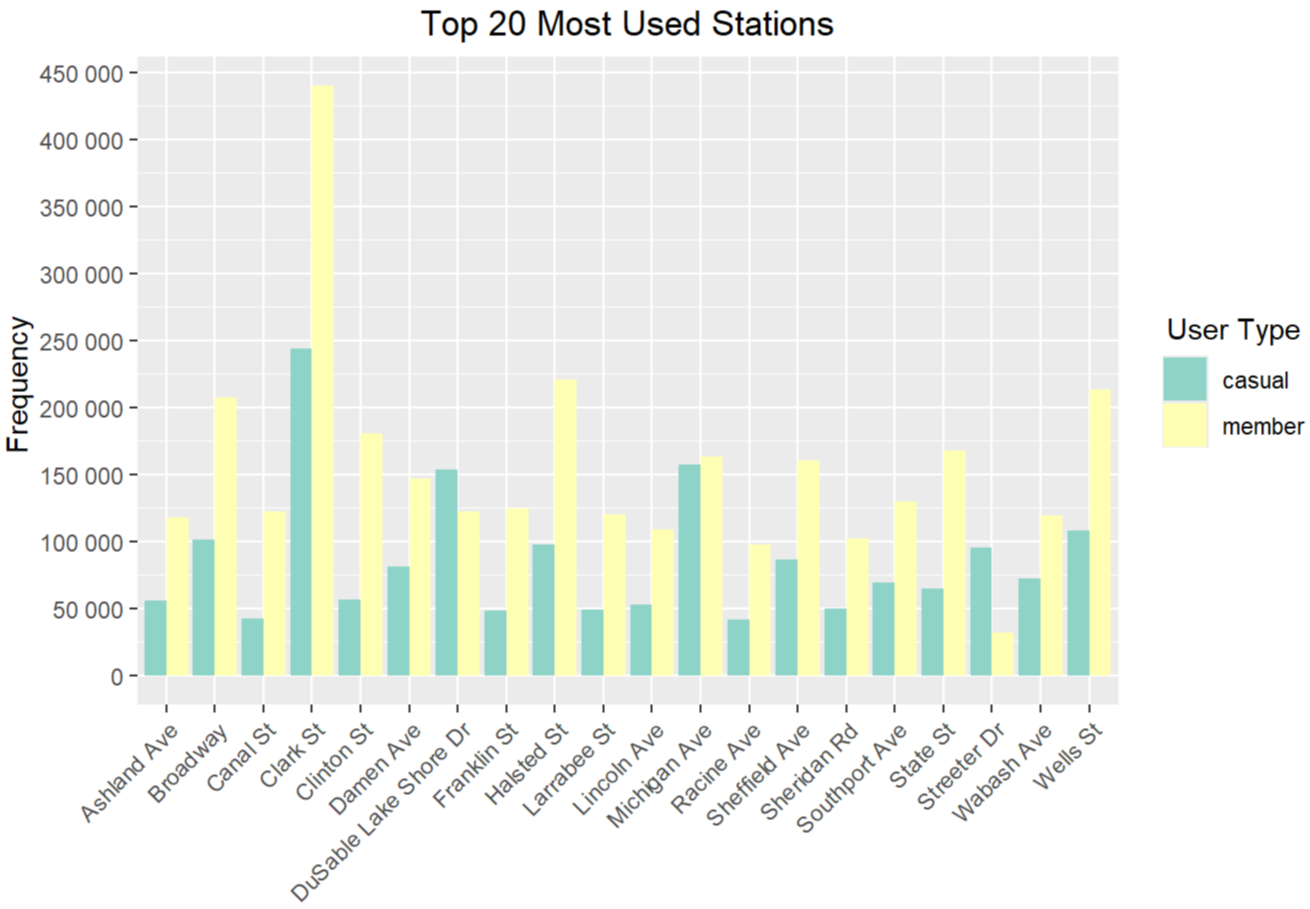
Identifying Area names and Top 20 most used stations

```
df<-df%>% separate(start_station_name, 'from_area', sep=' &', remove=FALSE)
df<-df%>% separate(end_station_name, 'to_area', sep=' &', remove=FALSE)

member_station<-full_join(df %>% filter(member_casual=='member') %>%count(from_area) %>% rename(station_name=from_area),
          df %>%filter(member_casual=='member') %>%  count(to_area) %>% rename(station_name=to_area),
          by='station_name') %>%
  filter(station_name!="") %>%
  mutate(freq=n.x+n.y) %>%
  select(c(station_name, freq))

casual_station<-full_join(df %>% filter(member_casual=='casual') %>%count(from_area) %>% rename(station_name=from_area),
                         df %>%filter(member_casual=='casual') %>%  count(to_area) %>% rename(station_name=to_area),
                         by='station_name') %>%
  filter(station_name!="") %>%
  mutate(freq=n.x+n.y) %>%
  select(c(station_name, freq))

full_join(member_station, casual_station, by="station_name") %>%
  rename(member=freq.x, casual=freq.y) %>%
  mutate(total_freq=(member+casual)) %>%
  arrange(desc(total_freq))%>%
  slice(1:20)%>%
  select(1:3)%>%
  pivot_longer(cols = c('member','casual')) %>%
  group_by(name)%>%
  ggplot(aes(x=station_name, y=value, fill=name))+
  geom_bar(position = 'dodge', stat='identity')  +
  theme( axis.text.x = element_text(angle=45, vjust=1, hjust=1) ,
         plot.title = element_text(hjust = 0.5, ),legend.title.align = 0.5) +
  labs( x="Station Name" , y= "Frequency " , title= "Top 20 Most Used Stations", fill='User Type') +
  scale_y_continuous(labels = label_number(), breaks = pretty_breaks(10))+
  scale_fill_brewer(palette = "Set3")
```


Top 20 Most Used Stations
Frequency
450 000
400 000
350 000
300 000
250 000
200 000
150 000
100 000
50 000
0
Ashland Ave
Broadway
Canal St
Clark St
Clinton St
Damen Ave
DuSable Lake Shore Dr
Franklin St
Halsted St
Larrabee St
Lincoln Ave
Michigan Ave
Racine Ave
Sheffield Ave
Sheridan Rd
Southport Ave
State St
Streeter Dr
Wabash Ave
Wells St
Station Name
User Type
casual
member