Google Data Analytics - Case Study

22/07/2021

Preparation

Set environment

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

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

```
library("dplyr")
```

Import datasets

```
[redacted]
q1_2020 <- read.csv("R/bike_sharing/Resources/Divvy_Trips_2020_Q1.csv")
q4_2019 <- read.csv("R/bike_sharing/Resources/Divvy_Trips_2019_Q4.csv")
q3_2019 <- read.csv("R/bike_sharing/Resources/Divvy_Trips_2019_Q3.csv")
q2_2019 <- read.csv("R/bike_sharing/Resources/Divvy_Trips_2019_Q2.csv")
```

Change value “Subscriber and Customer” to "member and casual

```
q4_2019$member_casual <- gsub("Subscriber", "member",q4_2019$member_casual)
q4_2019$member_casual <- gsub("Customer","casual",q4_2019$member_casual)

q3_2019$member_casual <- gsub("Subscriber", "member",q3_2019$member_casual)
q3_2019$member_casual <- gsub("Customer","casual",q3_2019$member_casual)

q2_2019$member_casual <- gsub("Subscriber", "member",q2_2019$member_casual)
q2_2019$member_casual <- gsub("Customer","casual",q2_2019$member_casual)
```

Change datatype of "ride_id" for 2019 data to be combined

```
q4_2019$ride_id <- as.character(q4_2019$ride_id)
q3_2019$ride_id <- as.character(q3_2019$ride_id)
q2_2019$ride_id <- as.character(q2_2019$ride_id)
```

Combine all dataset into variable "all_data"

```
all_data <- bind_rows(q2_2019,q3_2019,q4_2019,q1_2020)
```

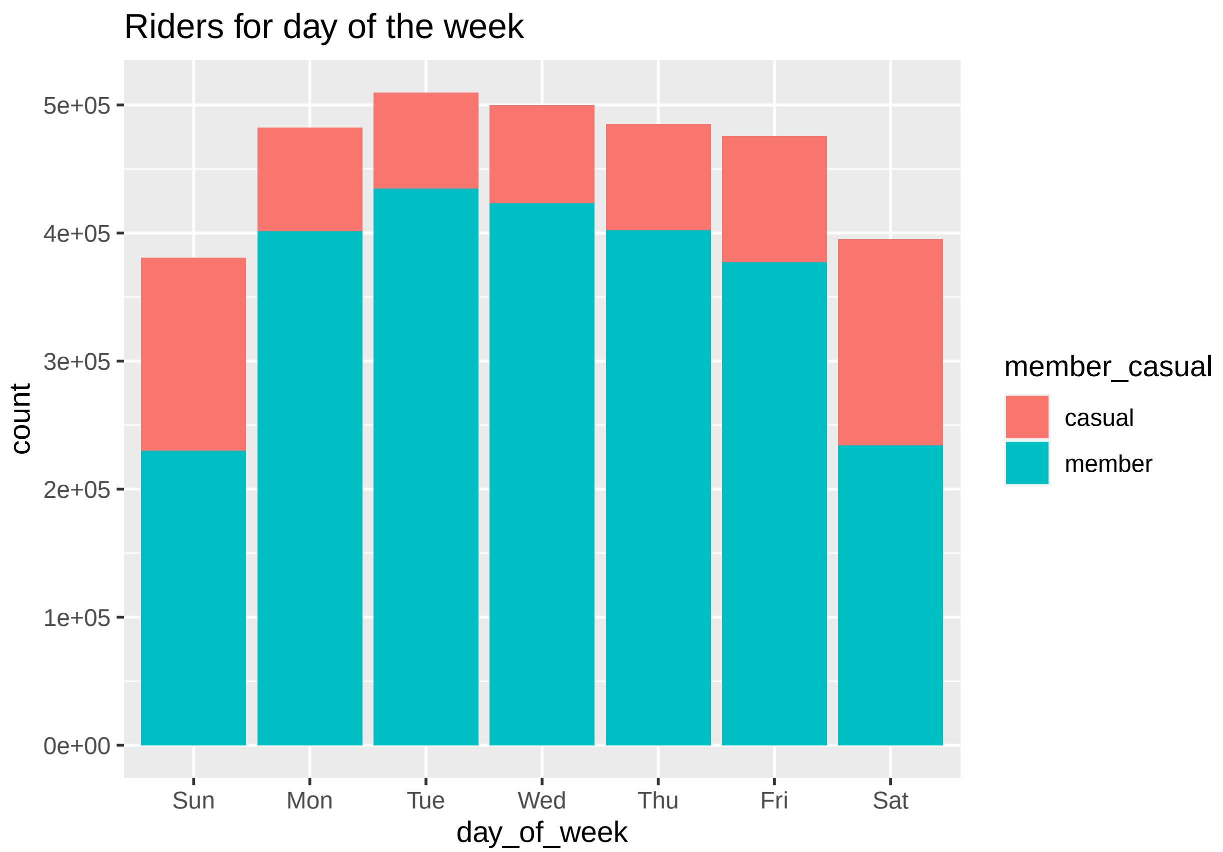
Add columns

```
all_data$ride_length <- difftime(all_data$ended_at,all_data$started_at)/60
all_data$ride_length <- as.integer(all_data$ride_length)
all_data$month <- as.Date(all_data$started_at,format = "%Y-%m-%d")
all_data$month <- format(all_data$month, "%m")
all_data$year <- as.Date(all_data$started_at,format = "%Y-%m-%d")
all_data$year <- format(all_data$year, "%Y")
all_data$year <- as.integer(all_data$year)
all_data$age <- all_data$year - all_data$birthyear
```

Visualizations

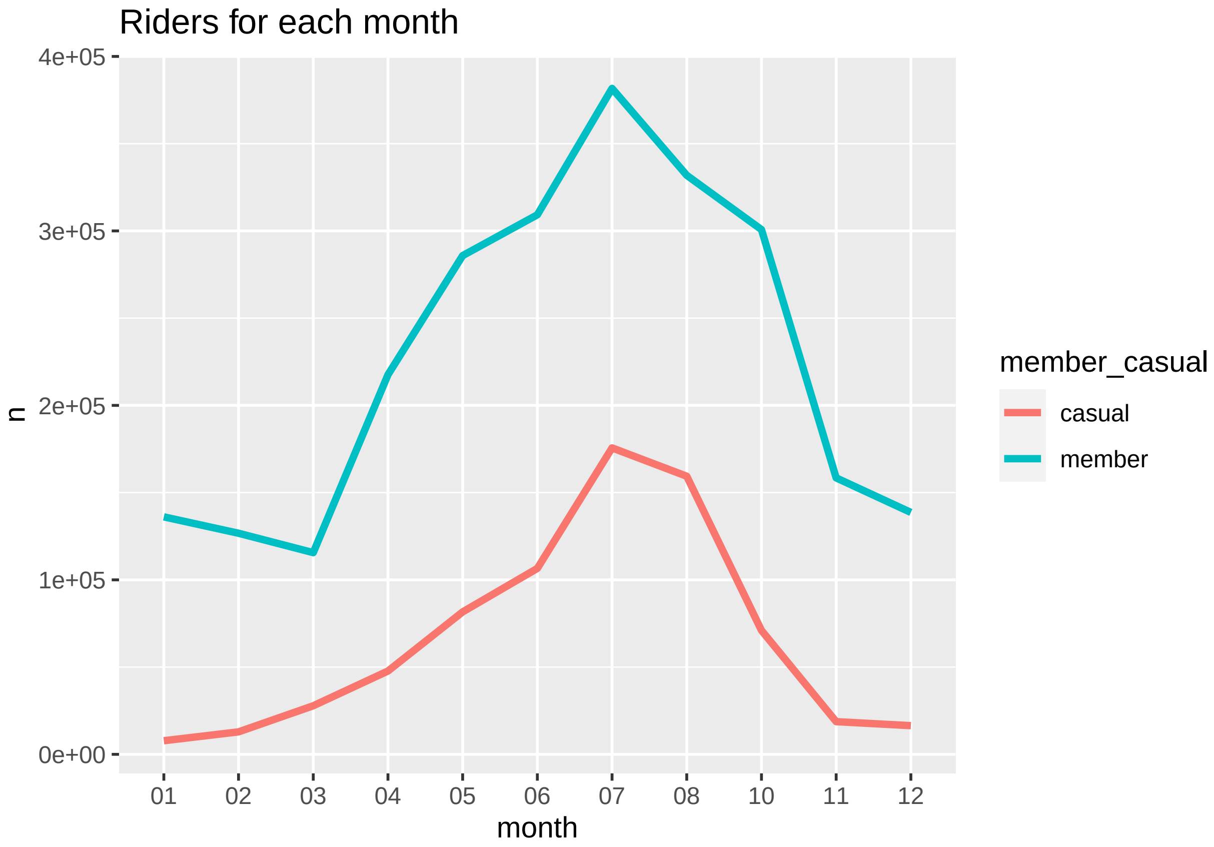
Riders for week of the day grouped by member/casual

```
library(ggplot2)
day_week <- c( 'Sun','Mon', 'Tue', 'Wed', 'Thu', 'Fri',
             'Sat')

ggplot(data=all_data)+
  geom_bar(mapping = aes(x=day_of_week,fill=member_casual))+
  labs(title="Riders for day of the week") +
  xlim(day_week)
```



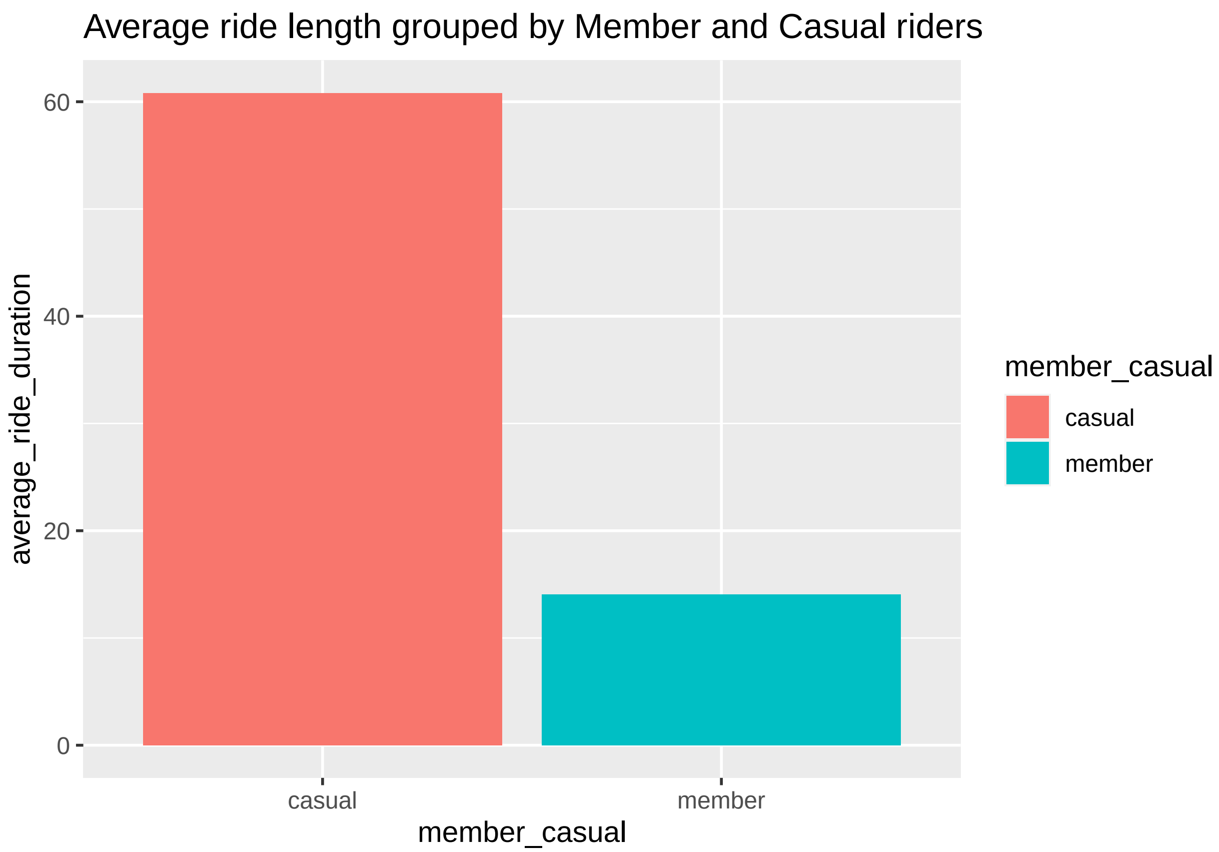
Number of member and casual riders for each month

```r
num_month <- all_data %>%
  group_by(month) %>%
  count(member_casual)

num_month <- data.frame(num_month)
ggplot(data = num_month)+
  geom_line(mapping = aes(x=month, y=n, group=member_casual,color=member_casual), size=1.3)+
  labs(title = "Riders for each month")
```



Average ride length for member and casual riders

```
average_length <-all_data %>%
  group_by(member_casual) %>%
  summarize(average_ride_duration = mean(ride_length))

ggplot(data=average_length)+
  geom_bar(stat = "identity",mapping = aes(x=member_casual,y=average_ride_duration,fill=member_casual))+
  labs(title="Average ride length grouped by Member and Casual riders")
```

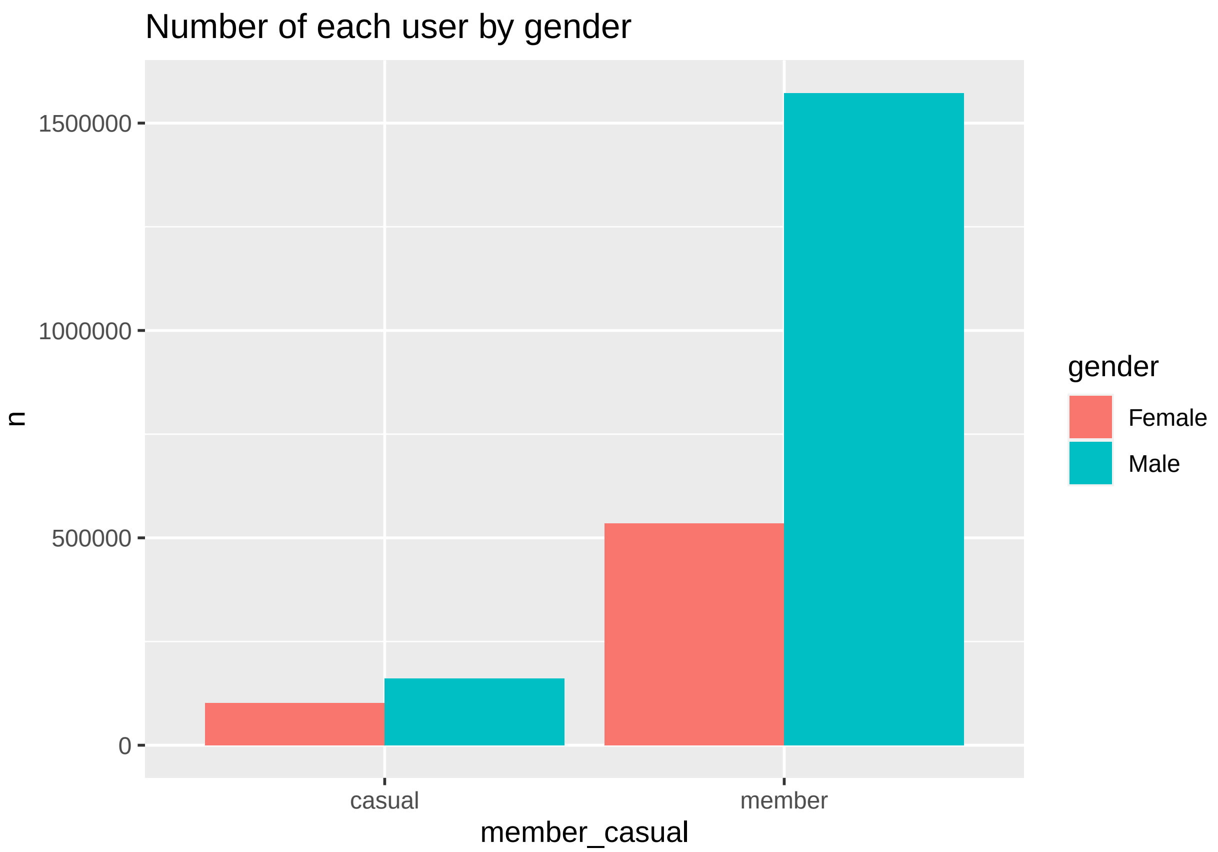


Number of riders by gender

```
p <- all_data %>%
  drop_na(gender) %>%
  group_by(member_casual) %>%
  count(gender)

p<- data.frame(p)
p <- p[p$gender!= "",]

p
```

```
##   member_casual gender       n
## 2        casual Female  101820
## 3        casual   Male  160652
## 5        member Female  534752
## 6        member   Male 1572865
```

```
ggplot(data=p)+
  geom_bar(stat="identity",mapping = aes(x=member_casual,y=n,fill=gender),position="dodge")+
  labs(title = "Number of each user by gender")
```



Number of member/casual riders grouped by end stations

```r
num_station <- all_data %>%
  group_by(end_station_name) %>%
  count(member_casual,sort = (decreasing = TRUE))

num_station <- data.frame(num_station)

top_member <- subset(num_station,member_casual == "member") %>%
  top_n(5)
```

```
## Selecting by n
```

```r
top_casual <- subset(num_station,member_casual == "casual") %>%
  top_n(5)
```

```
## Selecting by n
```

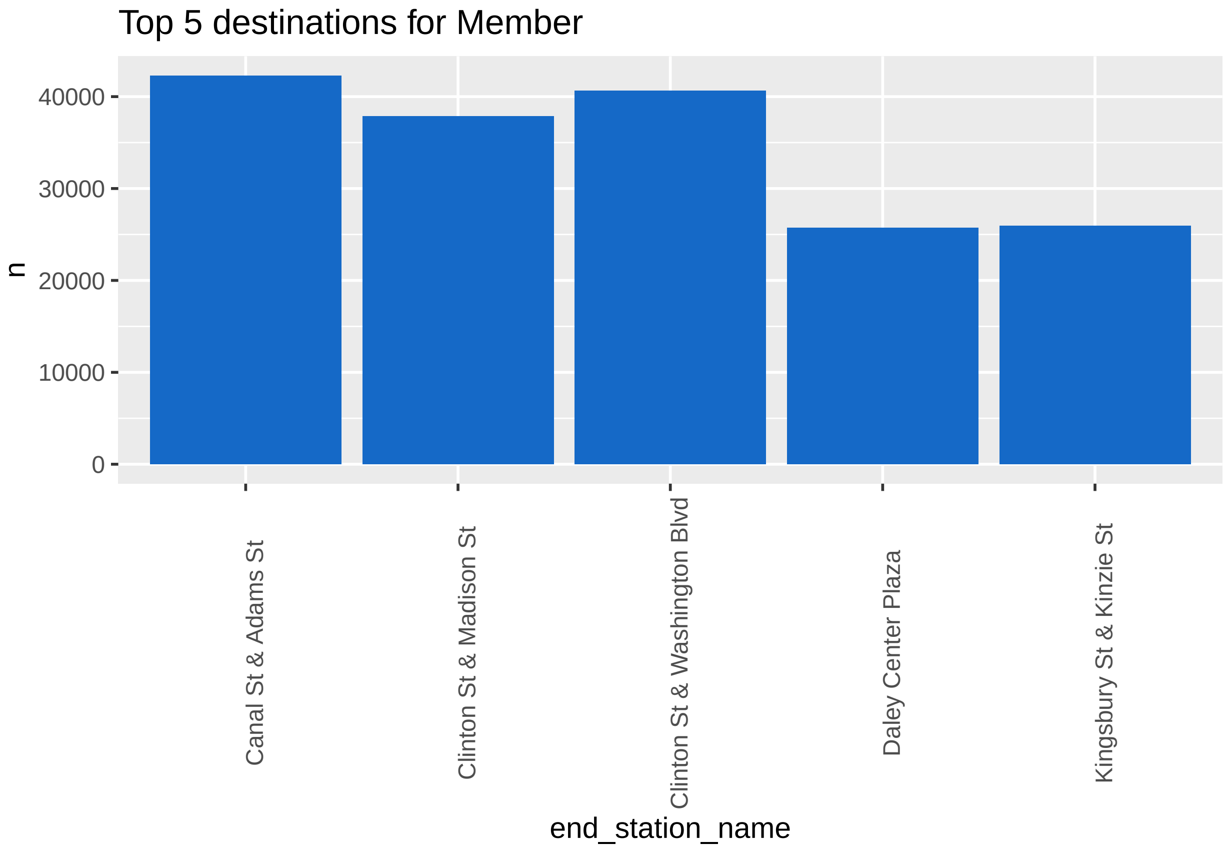
```r
top_member
```

```
##               end_station_name member_casual     n
## 1           Canal St & Adams St        member 42280
## 2 Clinton St & Washington Blvd        member 40654
## 3       Clinton St & Madison St        member 37875
## 4      Kingsbury St & Kinzie St        member 25935
## 5            Daley Center Plaza        member 25729
```

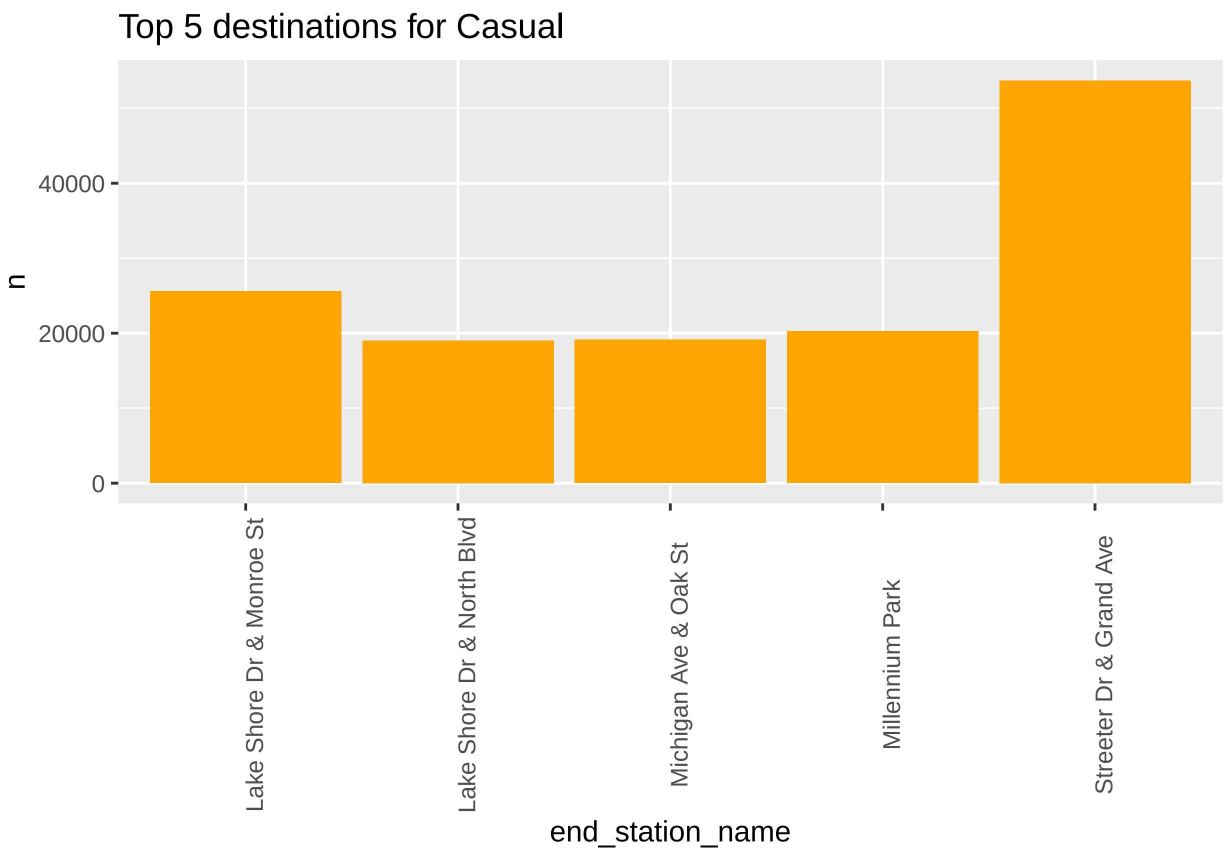
```
top_casual
```

```
##             end_station_name member_casual     n
## 1     Streeter Dr & Grand Ave        casual 53719
## 2   Lake Shore Dr & Monroe St        casual 25596
## 3             Millennium Park        casual 20266
## 4       Michigan Ave & Oak St        casual 19121
## 5 Lake Shore Dr & North Blvd        casual 19008
```

```
ggplot(data=top_member)+
  geom_bar(stat = "identity",mapping=aes(x=end_station_name,y=n),fill='#1569C7')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title="Top 5 destinations for Member")
```



```
ggplot(data=top_casual)+
  geom_bar(stat = "identity",mapping=aes(x=end_station_name,y=n),fill='#FFA500')+
  theme(axis.text.x = element_text(angle = 90))+
  labs(title="Top 5 destinations for Casual")
```



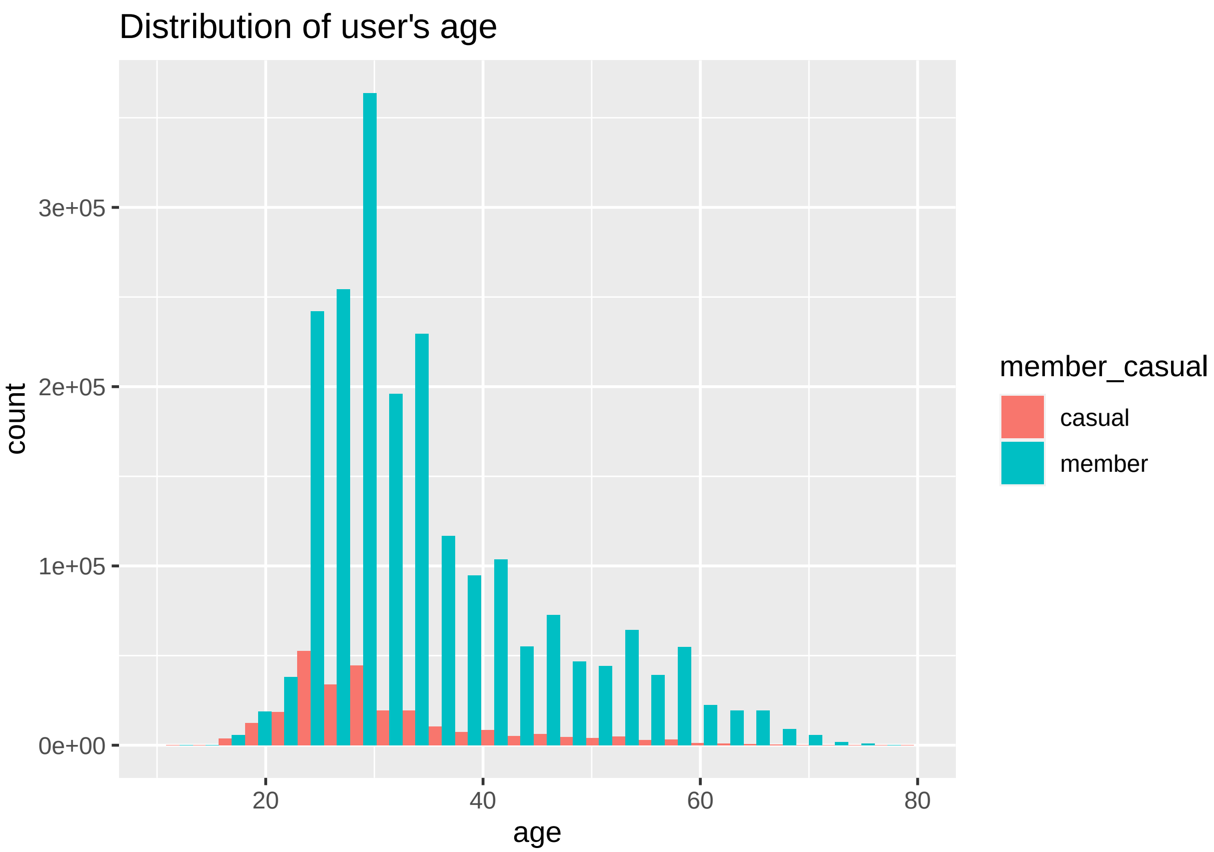
Distribution of age

```
ggplot(data=all_data)+
  geom_histogram(mapping = aes(x=age,fill=member_casual),position = "dodge",bins = 30)+
  scale_x_continuous(limits = c(0, 80))+
  labs(title="Distribution of user's age")+
  xlim(10,80)
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## Warning: Removed 843724 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```


Distribution of user's age
count
3e+05
2e+05
1e+05
0e+00
20
40
60
80
age
member_casual
casual
member