Bike Rideshare Project

8/8/2021

The data downloaded from kaggle website from this link. The bike rideshare company believe that increasing the number of loyal customer is important for the company growth. The purpose o this analysis is explore how causal riders use the bikes differently from loyal (subscriber) riders. Ultimately, the results of the data analysis will be used to help design a plan that encourages causal rider to become subscribers.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

##Loading and isecting datasets

```
Q2_2019 <- read.csv("Divvy_Trips_2019_Q2.csv",as.is=T)
Q3_2019 <- read.csv("Divvy_Trips_2019_Q3.csv",as.is=T)
Q4_2019 <- read.csv("Divvy_Trips_2019_Q4.csv",as.is=T)
Q1_2020 <- read.csv("Divvy_Trips_2020_Q1.csv",as.is=T)
```

There are discrepancies in column names and data type of some columns among the four datasets. Column names and data strucure of Divvy_Trips_2020_Q1 and will be adopted and applied to datasets.

Data cleaning

Renaming some columns in datsets from 2019

```
Q2_2019_new_col_names<- rename(Q2_2019,
                               ride_id = X01...Rental.Details.Rental.ID,
                               rideable_type = X01...Rental.Details.Bike.ID,
                               started_at = X01...Rental.Details.Local.Start.Time,
                               ended_at = X01...Rental.Details.Local.End.Time,
                               start_station_name = X03...Rental.Start.Station.Name,
                               start_station_id = X03...Rental.Start.Station.ID,
                               end_station_name = X02...Rental.End.Station.Name,
                               end_station_id = X02...Rental.End.Station.ID,
                               member_casual = User.Type)

Q3_2019_new_col_names <- rename(Q3_2019,
                                ride_id = trip_id,
                                rideable_type = bikeid,
                                started_at = start_time,
                                ended_at = end_time ,
                                start_station_name = from_station_name ,
                                start_station_id = from_station_id ,
                                end_station_name = to_station_name ,
                                end_station_id = to_station_id ,
                                member_casual = usertype)


Q4_2019_new_col_names <- rename(Q4_2019,
                                ride_id = trip_id,
                                rideable_type = bikeid,
                                started_at = start_time,
                                ended_at = end_time ,
                                start_station_name = from_station_name ,
                                start_station_id = from_station_id ,
                                end_station_name = to_station_name ,
                                end_station_id = to_station_id ,
                                member_casual = usertype)
```

```
glimpse(Q2_2019_new_col_names)
```

```
## Rows: 1,108,163
## Columns: 12
## $ ride_id                                           <int> 22178529, 22178530, ~
## $ started_at                                        <chr> "2019-04-01 00:02:22~
## $ ended_at                                          <chr> "2019-04-01 00:09:48~
## $ rideable_type                                     <int> 6251, 6226, 5649, 41~
## $ X01...Rental.Details.Duration.In.Seconds.Uncapped <chr> "446.0", "1,048.0", ~
## $ start_station_id                                  <int> 81, 317, 283, 26, 20~
## $ start_station_name                                <chr> "Daley Center Plaza"~
## $ end_station_id                                    <int> 56, 59, 174, 133, 12~
## $ end_station_name                                  <chr> "Desplaines St & Kin~
## $ member_casual                                     <chr> "Subscriber", "Subsc~
## $ Member.Gender                                     <chr> "Male", "Female", "M~
## $ X05...Member.Details.Member.Birthday.Year         <int> 1975, 1984, 1990, 19~
```

```
glimpse(Q3_2019_new_col_names)
```

```
## Rows: 1,640,718
## Columns: 12
## $ ride_id            <int> 23479388, 23479389, 23479390, 23479391, 23479392, 2~
## $ started_at         <chr> "2019-07-01 00:00:27", "2019-07-01 00:01:16", "2019~
## $ ended_at           <chr> "2019-07-01 00:20:41", "2019-07-01 00:18:44", "2019~
## $ rideable_type      <int> 3591, 5353, 6180, 5540, 6014, 4941, 3770, 5442, 295~
## $ tripduration       <chr> "1,214.0", "1,048.0", "1,554.0", "1,503.0", "1,213.~
## $ start_station_id   <int> 117, 381, 313, 313, 168, 300, 168, 313, 43, 43, 511~
## $ start_station_name <chr> "Wilton Ave & Belmont Ave", "Western Ave & Monroe S~
## $ end_station_id     <int> 497, 203, 144, 144, 62, 232, 62, 144, 195, 195, 84,~
## $ end_station_name   <chr> "Kimball Ave & Belmont Ave", "Western Ave & 21st St~
## $ member_casual      <chr> "Subscriber", "Customer", "Customer", "Customer", "~
## $ gender             <chr> "Male", "", "", "", "", "Male", "", "", "", "", "",~
## $ birthyear          <int> 1992, NA, NA, NA, NA, 1990, NA, NA, NA, NA, NA, NA,~
```

```
glimpse(Q4_2019_new_col_names)
```

```
## Rows: 704,054
## Columns: 12
## $ ride_id            <int> 25223640, 25223641, 25223642, 25223643, 25223644, 2~
## $ started_at         <chr> "2019-10-01 00:01:39", "2019-10-01 00:02:16", "2019~
## $ ended_at           <chr> "2019-10-01 00:17:20", "2019-10-01 00:06:34", "2019~
## $ rideable_type      <int> 2215, 6328, 3003, 3275, 5294, 1891, 1061, 1274, 601~
## $ tripduration       <chr> "940.0", "258.0", "850.0", "2,350.0", "1,867.0", "3~
## $ start_station_id   <int> 20, 19, 84, 313, 210, 156, 84, 156, 156, 336, 77, 1~
## $ start_station_name <chr> "Sheffield Ave & Kingsbury St", "Throop (Loomis) St~
## $ end_station_id     <int> 309, 241, 199, 290, 382, 226, 142, 463, 463, 336, 5~
## $ end_station_name   <chr> "Leavitt St & Armitage Ave", "Morgan St & Polk St",~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Male", "Female", "Male", "Male", "Female",~
## $ birthyear          <int> 1987, 1998, 1991, 1990, 1987, 1994, 1991, 1995, 199~
```

```
glimpse(Q1_2020)
```

```
## Rows: 426,887
## Columns: 13
## $ ride_id            <chr> "EACB19130B0CDA4A", "8FED874C809DC021", "789F3C21E4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-01-21 20:06:59", "2020-01-30 14:22:39", "2020~
## $ ended_at           <chr> "2020-01-21 20:14:30", "2020-01-30 14:26:22", "2020~
## $ start_station_name <chr> "Western Ave & Leland Ave", "Clark St & Montrose Av~
## $ start_station_id   <int> 239, 234, 296, 51, 66, 212, 96, 96, 212, 38, 117, 1~
## $ end_station_name   <chr> "Clark St & Leland Ave", "Southport Ave & Irving Pa~
## $ end_station_id     <int> 326, 318, 117, 24, 212, 96, 212, 212, 96, 100, 632,~
## $ start_lat          <dbl> 41.9665, 41.9616, 41.9401, 41.8846, 41.8856, 41.889~
## $ start_lng          <dbl> -87.6884, -87.6660, -87.6455, -87.6319, -87.6418, -~
## $ end_lat            <dbl> 41.9671, 41.9542, 41.9402, 41.8918, 41.8899, 41.884~
## $ end_lng            <dbl> -87.6674, -87.6644, -87.6530, -87.6206, -87.6343, -~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

selecting the rquired columns from each dataset

```
Q1_2020_v01 <- Q1_2020 %>%
  select(c("ride_id","started_at","ended_at","rideable_type",
           "start_station_id","start_station_name","end_station_id","end_station_name","member_casual")

Q2_2019_v01 <- Q2_2019_new_col_names %>%
  select(c("ride_id","started_at","ended_at","rideable_type",
           "start_station_id","start_station_name","end_station_id","end_station_name","member_casual")

Q3_2019_v01 <- Q3_2019_new_col_names %>%
  select(c("ride_id","started_at","ended_at","rideable_type",
           "start_station_id","start_station_name","end_station_id","end_station_name","member_casual")

Q4_2019_v01 <- Q4_2019_new_col_names %>%
  select(c("ride_id","started_at","ended_at","rideable_type",
           "start_station_id","start_station_name","end_station_id","end_station_name","member_casual")
```

Now, all columns needed for the analysis from all datasets have identical names. However, “trip_id” and “rideable_type” in 2019 datasets need to be converted to “character datatype to match 2020 dataset.

```
Q2_2019_v02 <- mutate(Q2_2019_v01, ride_id = as.character(ride_id),
                                rideable_type = as.character(rideable_type))

Q3_2019_v02 <- mutate(Q3_2019_v01, ride_id = as.character(ride_id),
                                rideable_type = as.character(rideable_type))
Q4_2019_v02 <- mutate(Q4_2019_v01, ride_id = as.character(ride_id),
                                rideable_type = as.character(rideable_type))


Q1_2020_v02<-Q1_2020_v01
```

Combining all datasets

```
data_v01 <- bind_rows(Q2_2019_v02, Q3_2019_v02, Q4_2019_v02, Q1_2020_v02)

glimpse(data_v01)
```

```
## Rows: 3,879,822
## Columns: 9
## $ ride_id            <chr> "22178529", "22178530", "22178531", "22178532", "22~
## $ started_at         <chr> "2019-04-01 00:02:22", "2019-04-01 00:03:02", "2019~
## $ ended_at           <chr> "2019-04-01 00:09:48", "2019-04-01 00:20:30", "2019~
## $ rideable_type      <chr> "6251", "6226", "5649", "4151", "3270", "3123", "64~
## $ start_station_id   <int> 81, 317, 283, 26, 202, 420, 503, 260, 211, 211, 304~
## $ start_station_name <chr> "Daley Center Plaza", "Wood St & Taylor St", "LaSal~
## $ end_station_id     <int> 56, 59, 174, 133, 129, 426, 500, 499, 211, 211, 232~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "Wabash Ave & Roosevel~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
```

Checking for duplicates

```
length(unique(data_v01$ride_id)) == nrow(data_v01)
```

```
## [1] TRUE
```

Return TRUE means no duplicate

Checking for missing values in the dataset

```
apply(is.na(data_v01), 2, which)
```

```
## $ride_id
## integer(0)
##
## $started_at
## integer(0)
##
## $ended_at
## integer(0)
##
## $rideable_type
## integer(0)
##
## $start_station_id
## integer(0)
##
## $start_station_name
## integer(0)
##
## $end_station_id
## [1] 3867362
##
## $end_station_name
## integer(0)
##
## $member_casual
## integer(0)
```

Columns “started_at” and “ended_at” will be used in following data analysis steps, both do not contain missing data. There are 3867362 missing records in “end_station_id” column, so the “end_station_name” will be used instead if needed.

Creating a new column to calculate trips duration in seconds.

```
data_v02 <- data_v01 %>%
  mutate(ride_duration=difftime(ended_at,started_at, units = "secs"))
```

Inspecting first rows of the dataset

```
head(data_v02)
```

```
##    ride_id          started_at            ended_at rideable_type
## 1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48          6251
## 2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30          6226
## 3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19          5649
## 4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58          4151
## 5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13          3270
## 6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56          3123
##   start_station_id        start_station_name end_station_id
## 1               81        Daley Center Plaza             56
## 2              317       Wood St & Taylor St             59
## 3              283 LaSalle St & Jackson Blvd            174
## 4               26  McClurg Ct & Illinois St            133
## 5              202      Halsted St & 18th St            129
## 6              420       Ellis Ave & 55th St            426
##            end_station_name member_casual ride_duration
## 1 Desplaines St & Kinzie St    Subscriber      446 secs
## 2 Wabash Ave & Roosevelt Rd    Subscriber     1048 secs
## 3     Canal St & Madison St    Subscriber      252 secs
## 4  Kingsbury St & Kinzie St    Subscriber      357 secs
## 5 Blue Island Ave & 18th St    Subscriber     1007 secs
## 6       Ellis Ave & 60th St    Subscriber      257 secs
```

```
min(data_v02$ride_duration)
```

```
## Time difference of -6982 secs
```

There are some non valid time value less than 0. These will dropped

```
data_v03 <- data_v02 %>%filter(ride_duration> 0)

min(data_v03$ride_duration)
```

```
## Time difference of 1 secs
```

Adding new columns for trips starting time in hours, days of the week and months

```
data_v04 <- data_v03 %>%
  mutate(hour= hour(started_at)) %>%
  mutate(days= wday(started_at,lab= T,abbr = F))  %>%
  mutate(month= month(started_at, lab= T,abbr = F))%>%
  mutate(year= year(started_at))
```

```
glimpse(data_v04)
```

```
## Rows: 3,879,599
## Columns: 14
## $ ride_id            <chr> "22178529", "22178530", "22178531", "22178532", "22~
## $ started_at         <chr> "2019-04-01 00:02:22", "2019-04-01 00:03:02", "2019~
## $ ended_at           <chr> "2019-04-01 00:09:48", "2019-04-01 00:20:30", "2019~
## $ rideable_type      <chr> "6251", "6226", "5649", "4151", "3270", "3123", "64~
## $ start_station_id   <int> 81, 317, 283, 26, 202, 420, 503, 260, 211, 211, 304~
## $ start_station_name <chr> "Daley Center Plaza", "Wood St & Taylor St", "LaSal~
## $ end_station_id     <int> 56, 59, 174, 133, 129, 426, 500, 499, 211, 211, 232~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "Wabash Ave & Roosevel~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ ride_duration      <drtn> 446 secs, 1048 secs, 252 secs, 357 secs, 1007 secs~
## $ hour               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, ~
## $ days               <ord> Monday, Monday, Monday, Monday, Monday, Monday, Mon~
## $ month              <ord> April, April, April, April, April, April, April, Ap~
## $ year               <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 201~
```

Finally, "member_casual" categorical variable defined differently in 2019 datasets (Subscriber/ Customer) from 2020 dataset (member/casual).

```
ride_count <- data_v04 %>%group_by(member_casual) %>%
  summarise(number_of_rides = n())

ride_count
```

```
## # A tibble: 4 x 2
##   member_casual number_of_rides
##   <chr>                   <int>
## 1 casual                  48270
## 2 Customer               857468
## 3 member                 378407
## 4 Subscriber            2595454
```

Again will use 2020 format and Subscriber/ Customer to member/casual.

```
data_v04<- data_v04 %>% mutate(member_casual=  case_when(
  member_casual == "member" ~ "member",
  member_casual =="casual" ~ "casual",
  member_casual == "Customer" ~ "casual",  # replace customer with casual
  member_casual == "Subscriber" ~ "member"))    # replace Subscriber with member


ride_count <- data_v04 %>%group_by(member_casual) %>%
  summarise(number_of_rides = n())

ride_count
```

```
## # A tibble: 2 x 2
##   member_casual number_of_rides
##   <chr>                   <int>
## 1 casual                 905738
## 2 member                2973861
```

Data analysis and Visulaization

Some descriptive statistics

Average trip duration based on customer type

```
data_v04 %>%  group_by(member_casual) %>%
  summarize(mean_ride_duration=mean(ride_duration),
            median_ride_duration=median(ride_duration),
            minimum_ride_duration=min(ride_duration),
            maximum_ride_duration=max(ride_duration))
```

```
## # A tibble: 2 x 5
##   member_casual mean_ride_duration median_ride_duration minimum_ride_d~1 maxim~2
##   <chr>         <drtn>             <drtn>               <drtn>           <drtn>
## 1 casual        3538.859 secs      1541 secs            1 secs           938342~
## 2 member         850.078 secs       589 secs            1 secs           905663~
## # ... with abbreviated variable names 1: minimum_ride_duration,
## #   2: maximum_ride_duration
```

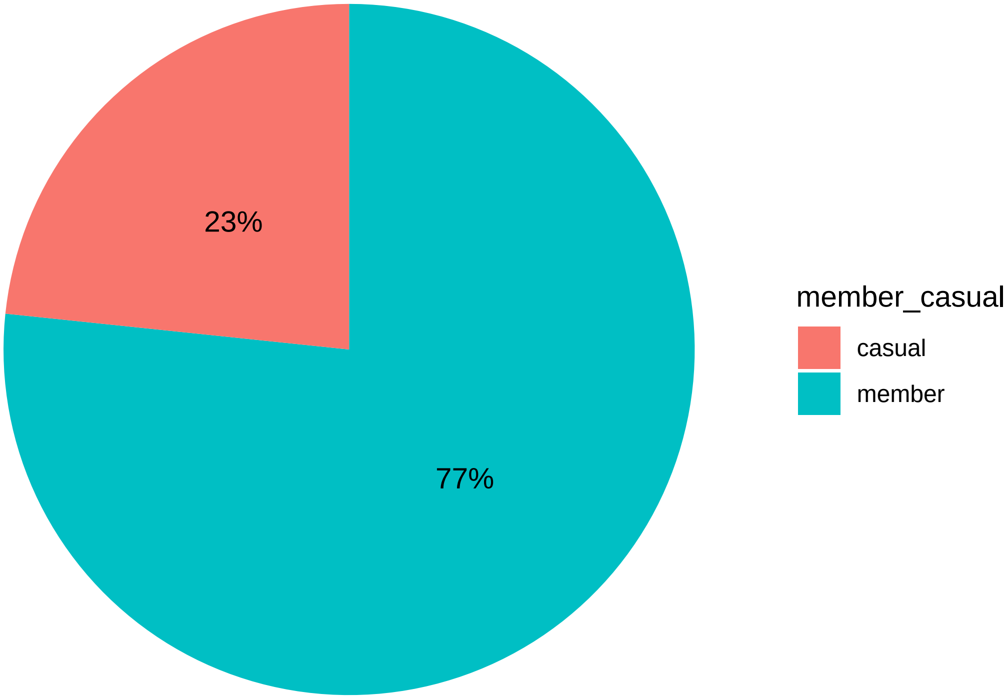
Calculating and plotting number of rides groupped by customer type

```
data_count<- data_v04 %>%
  group_by(member_casual) %>%
  count() %>%
  ungroup() %>%
  mutate(percent=`n`/sum(`n`)) %>%
  arrange(desc(member_casual))

data_count
```

```
## # A tibble: 2 x 3
##   member_casual       n percent
##   <chr>           <int>   <dbl>
## 1 member        2973861   0.767
## 2 casual         905738   0.233
```

```
data_count$label <- scales::percent(data_count$percent)
ggplot(data=data_count)+
  geom_bar(aes(x="", y=percent, fill=member_casual), stat="identity", width = 1)+
  coord_polar("y", start=0)+
  theme_void()+
  geom_text(aes(x=1, y = cumsum(percent) - percent/2, label=label))
```

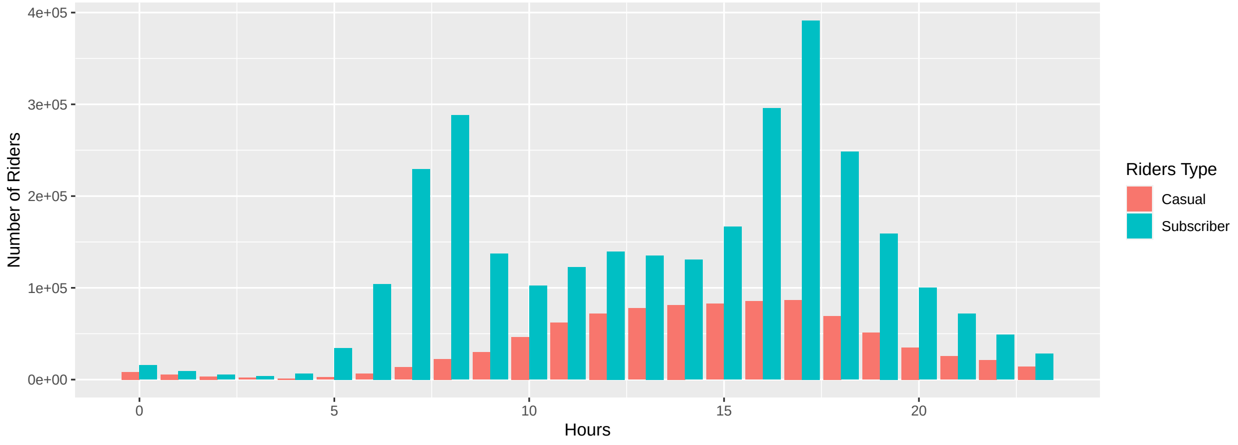


The subscribers (members) represent the majority of riders.

Number of trips grouped by starting time of the trips in hour of the day and customer type (member/casual)

```
data_v04 %>%
  group_by(member_casual, hour) %>%
  summarise(number_of_rides = n()) %>%
  ggplot(aes(x = hour, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete(name = "Riders Type", labels = c("Casual", "Subscri
  xlab("Hours")+ ylab("Number of Riders")
```

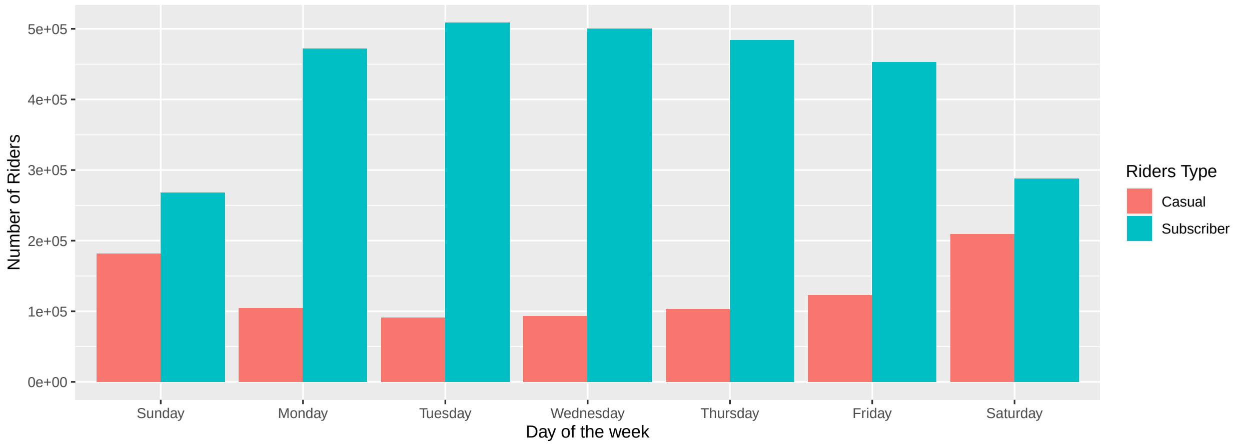
```
## ‘summarise()‘ has grouped output by ’member_casual’. You can override using the
## ‘.groups‘ argument.
```



Number of trips grouped by day of the week and customer type (member/casual)

```
data_v04 %>%
  group_by(member_casual, days) %>%
  summarise(number_of_rides = n()) %>%
  ggplot(aes(x = days, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete (name = "Riders Type", labels = c("Casual", "Subsc
  xlab("Day of the week")+ ylab("Number of Riders")
```

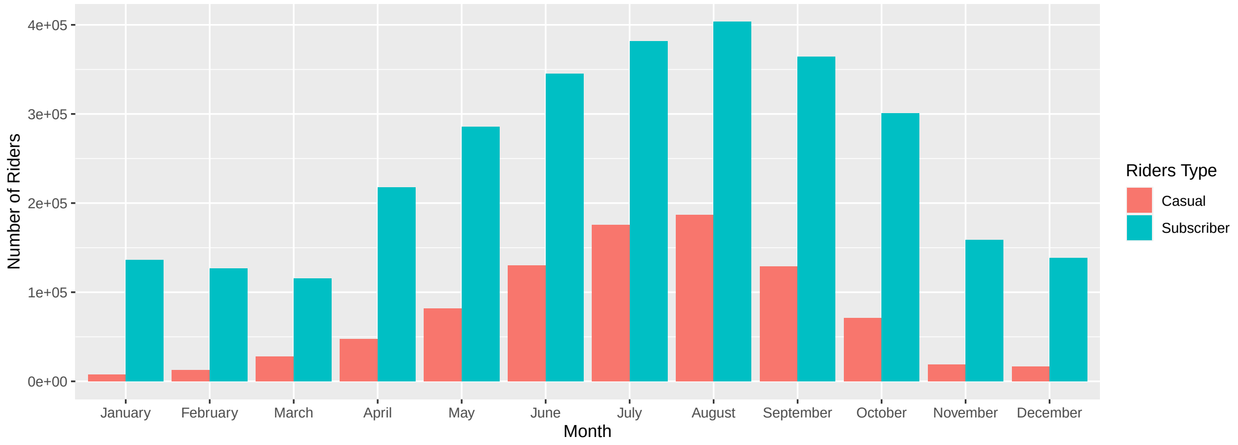
```
## ‘summarise()‘ has grouped output by ’member_casual’. You can override using the
## ‘.groups‘ argument.
```



Number of trips grouped by months and customer type (member/casual)

```
data_v04 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()) %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete (name = "Riders Type", labels = c("Casual", "Subsc
  xlab("Month")+ ylab("Number of Riders")
```

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

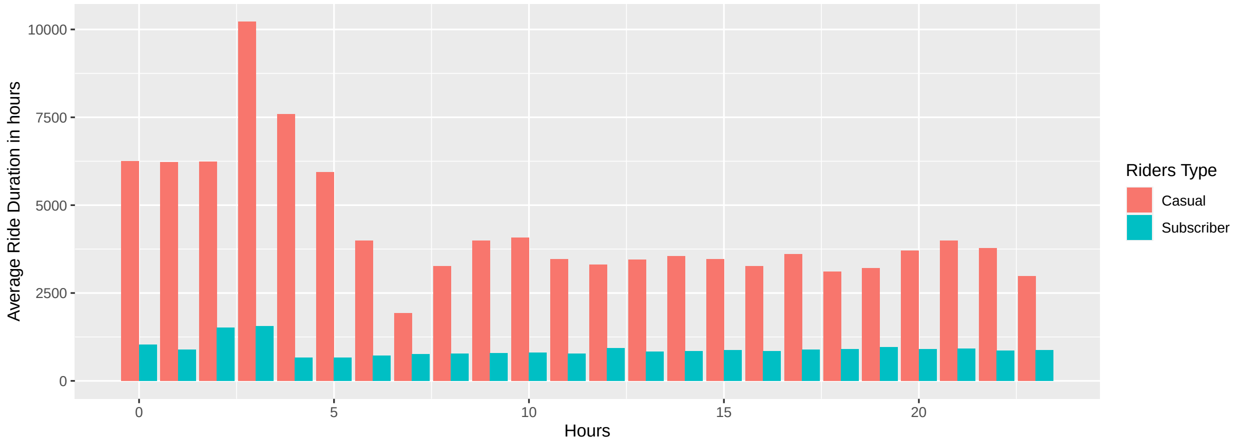


Clearly the number of rides started to increase in in spring and peaked in summer time which parallels with the improvement of weather.The majority of riders are subscribers.

plotting average of rides duration grouped by starting time of the day and customer type (subscriber/casual)

```
data_v04 %>%
  mutate(hour = hour) %>%
  group_by(member_casual, hour) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = hour, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete(name = "Riders Type", labels = c("Casual", "Subscr
  xlab("Hours")+ ylab("Average Ride Duration in hours")
```

```
## ‘summarise()‘ has grouped output by ’member_casual’. You can override using the
## ‘.groups‘ argument.
## Don’t know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```

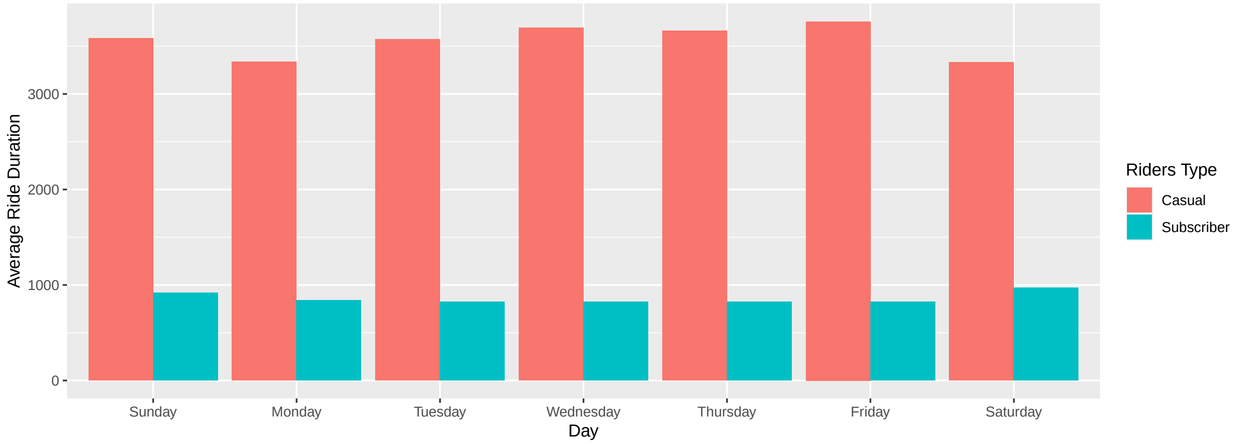


Above figure shows that causal riders tend to ride longer time compared to subscribers throughout the day.

Plotting average of rides duration grouped by the weekdays and customer type (member/casual)

```
data_v04 %>%
  mutate(weekdays = days) %>%
  group_by(member_casual, weekdays) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = weekdays, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete (name = "Riders Type", labels = c("Casual", "Subsc
  xlab("Day")+ ylab("Average Ride Duration")
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```

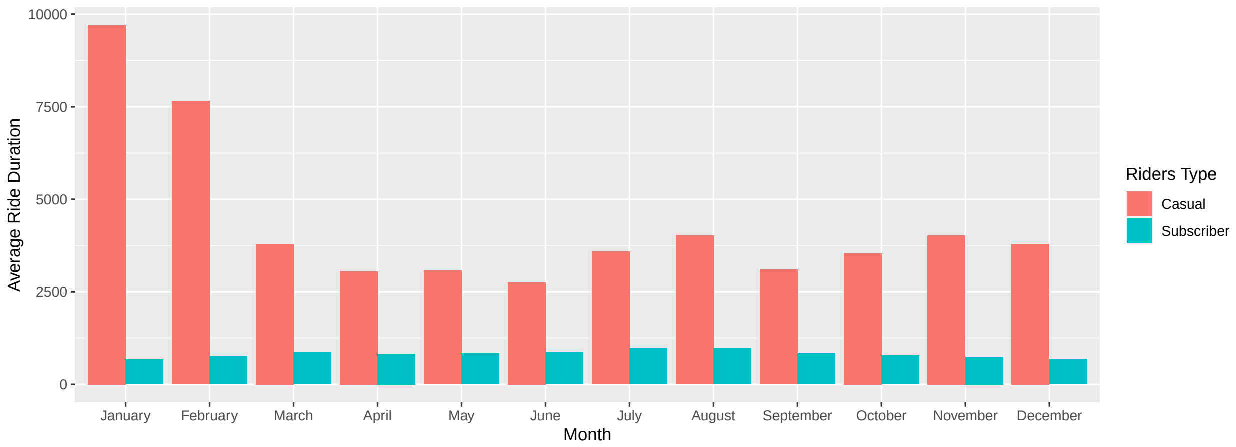


There is no pattern can be seen in average rides duration based on the day of the week.

Plotting average of rides duration grouped months and customer type (member/casual)

```
data_v04 %>%
  group_by(member_casual, month) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + scale_fill_discrete (name = "Riders Type", labels = c("Casual", "Subsc
  xlab("Month")+ ylab("Average Ride Duration")
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```


10000
7500
5000
2500
0
Average Ride Duration
January
February
March
April
May
June
July
August
September
October
November
December
Month
Riders Type
Casual
Subscriber