Cyclistic Case Study Sep21

Hezar K

2022-11-29

This is an analysis for Cyclistic Case Study for Google Data Analytics Course. This is an analysis for September 2021.

STEP ONE: INSTALL REQUIRED PACKAGES AND IMPORT DATA

Install the required packages. **Tidyverse** package to import and wrangling the data and **ggplot2** package for visualization of the data. **Lubridate** package for date parsing and **anytime** package for the datetime conversion.

- install.packages("tidyverse")
- install.packages("ggplot2")
- install.packages("lubridate")
- install.packages("anytime")

```
library(tidyverse)
library(lubridate)
library(data.table)
library(ggplot2)
library(anytime)
```

Import data from local drive.

```
Sep21 <- read_csv("C:/Users/theby/Documents/202109-divvy-tripdata.csv")
```

STEP TWO: EXAMINE THE DATA

Examine the dataframe for an overview of the data. Review column names, **colnames()**, dimensions of the dataframe by row and column, **dim()**, the first, **head()**, and the last, **tail()**, six rows in the dataframe, the summary, **summary()**, statistics on the columns of the dataframe, and review the data type structure of columns, **str()**.

View(Sep21)

```
colnames(Sep21)
```

```
##  [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(Sep21)
```

```
## [1] 756147
```

```
dim(Sep21)
```

```
## [1] 756147     13
```

```
head(Sep21)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ started_at          ended_at            start…² start…³
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 9DC7B962304CB… electr… 2021-09-28 16:07:10 2021-09-28 16:09:54 <NA>    <NA>
## 2 F930E2C6872D6… electr… 2021-09-28 14:24:51 2021-09-28 14:40:05 <NA>    <NA>
## 3 6EF72137900BB… electr… 2021-09-28 00:20:16 2021-09-28 00:23:57 <NA>    <NA>
## 4 78D1DE133B3DB… electr… 2021-09-28 14:51:17 2021-09-28 15:00:06 <NA>    <NA>
## 5 E03D4ACDCAEF6… electr… 2021-09-28 09:53:12 2021-09-28 10:03:44 <NA>    <NA>
## 6 346DE323A2677… electr… 2021-09-28 01:53:18 2021-09-28 02:00:02 <NA>    <NA>
## # … with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²start_station_name, ³start_station_id
```

```
tail(Sep21)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ started_at          ended_at            start…² start…³
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 0A6AA3B1A1EC5… classi… 2021-09-14 23:00:37 2021-09-14 23:10:55 Ellis … KA1503…
## 2 FA66BCAB0D73D… classi… 2021-09-22 15:46:57 2021-09-22 16:01:15 Ellis … 584
## 3 1D44DEFB5D36C… classi… 2021-09-25 16:25:23 2021-09-25 16:40:29 Ellis … KA1503…
## 4 6A346EA57FC23… classi… 2021-09-25 16:26:05 2021-09-25 16:40:30 Ellis … KA1503…
## 5 49360AFD77110… classi… 2021-09-15 17:57:48 2021-09-15 18:24:06 Ellis … KA1503…
## 6 343190A2DC023… electr… 2021-09-11 18:01:06 2021-09-11 18:08:26 Wells … TA1306…
## # … with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²start_station_name, ³start_station_id
```

```
summary(Sep21)
```

```
##    ride_id          rideable_type        started_at
##  Length:756147      Length:756147      Min.   :2021-09-01 00:00:06.00
##  Class :character   Class :character   1st Qu.:2021-09-08 11:14:14.50
##  Mode  :character   Mode  :character   Median :2021-09-15 16:43:37.00
##                                        Mean   :2021-09-15 18:19:01.89
##                                        3rd Qu.:2021-09-23 12:29:54.50
##                                        Max.   :2021-09-30 23:59:48.00
##
##     ended_at                     start_station_name start_station_id
##  Min.   :2021-09-01 00:00:41.00   Length:756147      Length:756147
##  1st Qu.:2021-09-08 11:33:01.00   Class :character   Class :character
##  Median :2021-09-15 17:01:16.00   Mode  :character   Mode  :character
##  Mean   :2021-09-15 18:39:32.52
##  3rd Qu.:2021-09-23 12:44:08.00
##  Max.   :2021-10-01 22:55:35.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:756147      Length:756147      Min.   :41.65   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   :41.57   Min.   :-87.87   Length:756147
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.50
##  NA's   :595     NA's   :595
```

```
str(Sep21)
```

```
## spc_tbl_ [756,147 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1DE133B3DBF
55" ...
##  $ rideable_type     : chr [1:756147] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:756147], format: "2021-09-28 16:07:10" "2021-09-28 14:24:51" ...
##  $ ended_at          : POSIXct[1:756147], format: "2021-09-28 16:09:54" "2021-09-28 14:40:05" ...
##  $ start_station_name: chr [1:756147] NA NA NA NA ...
##  $ start_station_id  : chr [1:756147] NA NA NA NA ...
##  $ end_station_name  : chr [1:756147] NA NA NA NA ...
##  $ end_station_id    : chr [1:756147] NA NA NA NA ...
##  $ start_lat         : num [1:756147] 41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num [1:756147] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:756147] 41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num [1:756147] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:756147] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Create new columns as for *date*, *month*, *day*, *year*, *day_of_week*, and *ride_length* in seconds.

```
Sep21$date <- as.Date(Sep21$started_at)
Sep21$month <- format(as.Date(Sep21$date), "%m")
Sep21$month <- month.name[as.numeric(Sep21$month)]
Sep21$day <- format(as.Date(Sep21$date), "%d")
Sep21$year <- format(as.Date(Sep21$date), "%Y")
Sep21$day_of_week <- format(as.Date(Sep21$date), "%A")
Sep21$ride_length <- difftime(Sep21$ended_at,Sep21$started_at)
```

Convert *ride_length* column to numeric in order to run calculations on the data. First, check to see if the data type is numeric, and then convert if needed.

```
is.numeric(Sep21$ride_length)
```

```
## [1] FALSE
```

Recheck *ride_length* data type.

```
Sep21$ride_length <- as.numeric(as.character(Sep21$ride_length))
is.numeric(Sep21$ride_length)
```

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

STEP THREE: CLEAN DATA

na.omit() will remove all NA from the dataframe.

```
Sep21 <- na.omit(Sep21)
```

Remove rows with the *ride_id* column character length is not 16. This will remove all the scientific ride ids that we noticed while examining the data.

```
Sep21 <- subset(Sep21, nchar(as.character(ride_id)) == 16)
```

Remove rows with the *ride_length* less than 60 seconds or 1 minute.

```
Sep21 <- subset (Sep21, ride_length > 59)
```

STEP FOUR: ANALYZE DATA

Analyze the dataframe by find the **mean**, **median**, **max** (maximum), and **min** (minimum) of *ride_length*.

```
mean(Sep21$ride_length)
```

```
## [1] 1226.085
```

```
median(Sep21$ride_length)
```

```
## [1] 738
```

```
max(Sep21$ride_length)
```

```
## [1] 1971512
```

```
min(Sep21$ride_length)
```

```
## [1] 60
```

Run a statistical summary of the *ride_length*.

```
summary(Sep21$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      60     429     738    1226    1306 1971512
```

Compare the members and casual users

```
aggregate(Sep21$ride_length ~ Sep21$member_casual, FUN = mean)
```

```
##   Sep21$member_casual Sep21$ride_length
## 1              casual         1701.5568
## 2              member          799.4548
```

```
aggregate(Sep21$ride_length ~ Sep21$member_casual, FUN = median)
```

```
##   Sep21$member_casual Sep21$ride_length
## 1              casual               953
## 2              member               593
```

```
aggregate(Sep21$ride_length ~ Sep21$member_casual, FUN = max)
```

```
##   Sep21$member_casual Sep21$ride_length
## 1              casual           1971512
## 2              member             79104
```

```
aggregate(Sep21$ride_length ~ Sep21$member_casual, FUN = min)
```

```
##   Sep21$member_casual Sep21$ride_length
## 1              casual                60
## 2              member                60
```

Aggregate the average ride length by each day of the week for members and users.

```
aggregate(Sep21$ride_length ~ Sep21$member_casual + Sep21$day_of_week, FUN = mean)
```

```
##    Sep21$member_casual Sep21$day_of_week Sep21$ride_length
## 1               casual            Friday         1620.8415
## 2               member            Friday          797.9890
## 3               casual            Monday         1827.1239
## 4               member            Monday          797.4820
## 5               casual          Saturday         1855.2845
## 6               member          Saturday          899.8359
## 7               casual            Sunday         2038.1164
## 8               member            Sunday          938.3356
## 9               casual          Thursday         1421.4412
## 10              member          Thursday          752.0072
## 11              casual           Tuesday         1343.0061
## 12              member           Tuesday          718.4218
## 13              casual         Wednesday         1410.2853
## 14              member         Wednesday          749.6142
```

Sort the days of the week in order.

```
Sep21$day_of_week <- ordered(Sep21$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday"))
```

Assign the aggregate the average ride length by each day of the week for members and users to x.

```
x <- aggregate(Sep21$ride_length ~ Sep21$member_casual + Sep21$day_of_week, FUN = mean)

head(x)
```

```
##   Sep21$member_casual Sep21$day_of_week Sep21$ride_length
## 1              casual            Sunday         2038.1164
## 2              member            Sunday          938.3356
## 3              casual            Monday         1827.1239
## 4              member            Monday          797.4820
## 5              casual           Tuesday         1343.0061
## 6              member           Tuesday          718.4218
```

Find the average ride length of member riders and casual riders per day and assign it to y.

```
y <- Sep21 %>%
  mutate(weekday = wday(started_at)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length), .groups = 'drop') %>%
  arrange(member_casual, weekday)

head(y)
```

```
## # A tibble: 6 × 4
##   member_casual weekday number_of_rides average_duration
##   <chr>           <int>           <int>            <dbl>
## 1 casual              1           56846            2038.
## 2 casual              2           34424            1827.
## 3 casual              3           24102            1343.
## 4 casual              4           34874            1410.
## 5 casual              5           39195            1421.
## 6 casual              6           39014            1621.
```

Analyze the dataframe to find the frequency of member riders, casual riders, classic bikes, docked bikes, and electric bikes.

```
table(Sep21$member_casual)
```

```
##
## casual member
## 290078 323287
```

```
table(Sep21$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        455491         35111        122763
```

```
table(Sep21$day_of_week)
```

```
##
##     Sunday     Monday    Tuesday Wednesday  Thursday     Friday  Saturday
##      94857      74850      67161      93576      98447      82104    102370
```

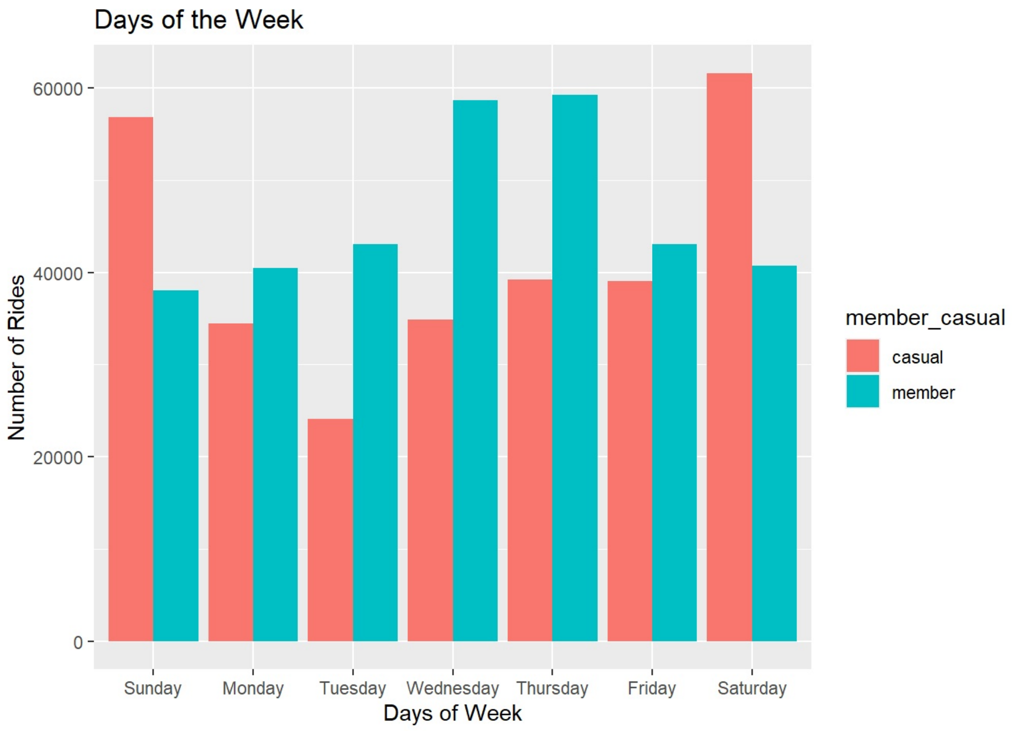
STEP FIVE: VISUALIZATION

Display full digits instead of scientific number.

```
options(scipen=999)
```

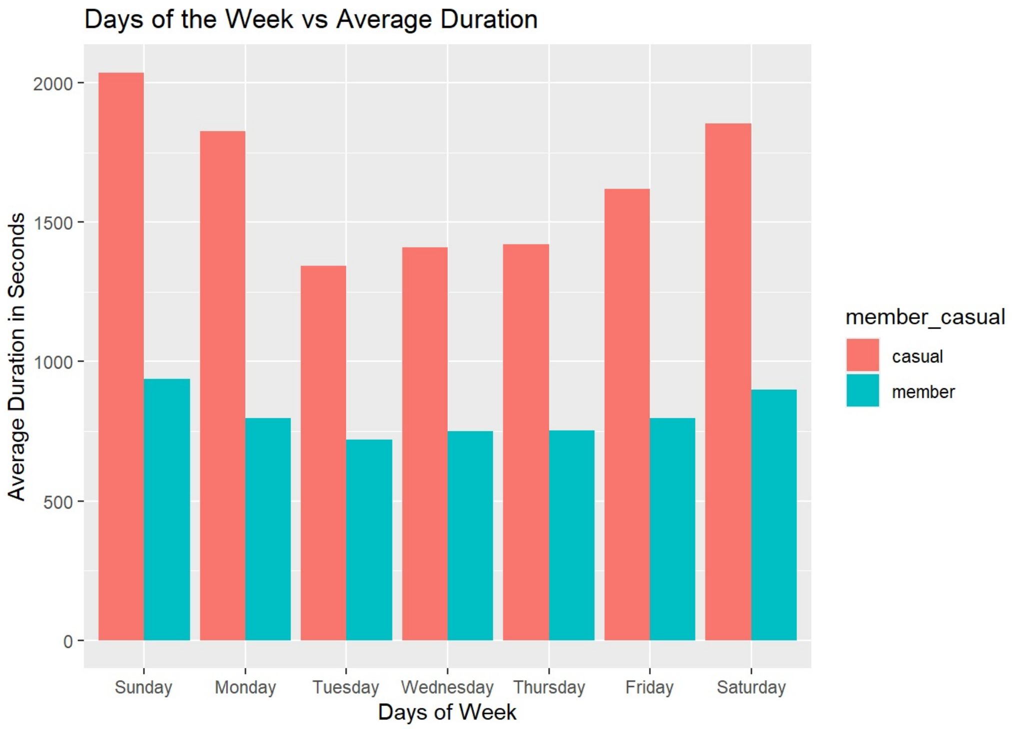
Plot the number of rides by user type during the week.

```
Sep21 %>%
  mutate(day_of_week) %>%
  group_by(member_casual,day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length), .groups = 'drop') %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")+
labs(x = "Days of Week",
     y= "Number of Rides",
     title= "Days of the Week")
```



Plot the duration of the ride by user type during the week.

```
Sep21 %>%
  mutate(day_of_week) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length), .groups = 'drop') %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x = "Days of Week",
       y= "Average Duration in Seconds",
       title= "Days of the Week vs Average Duration")
```



Create new dataframe for plots for weekday trends vs weekend trends.

```
mc<- as.data.frame(table(Sep21$day_of_week,Sep21$member_casual))
```

Rename columns

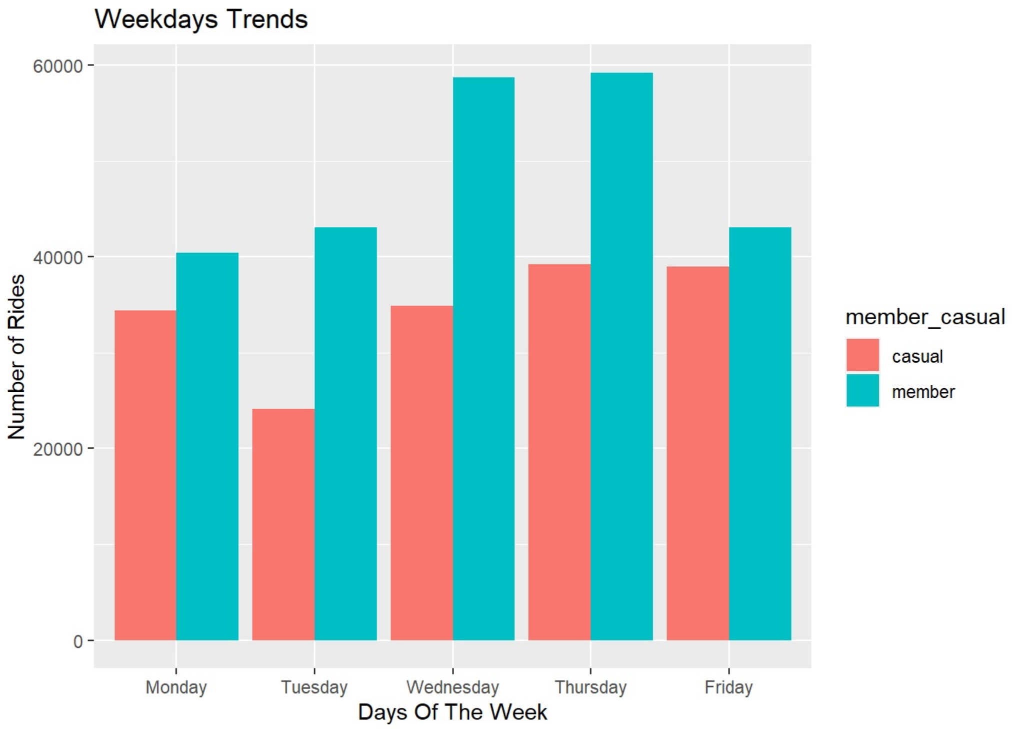
```
mc<-rename(mc, day_of_week = Var1, member_casual = Var2)

head(mc)
```

```
##   day_of_week member_casual  Freq
## 1      Sunday        casual 56846
## 2      Monday        casual 34424
## 3     Tuesday        casual 24102
## 4   Wednesday        casual 34874
## 5    Thursday        casual 39195
## 6      Friday        casual 39014
```

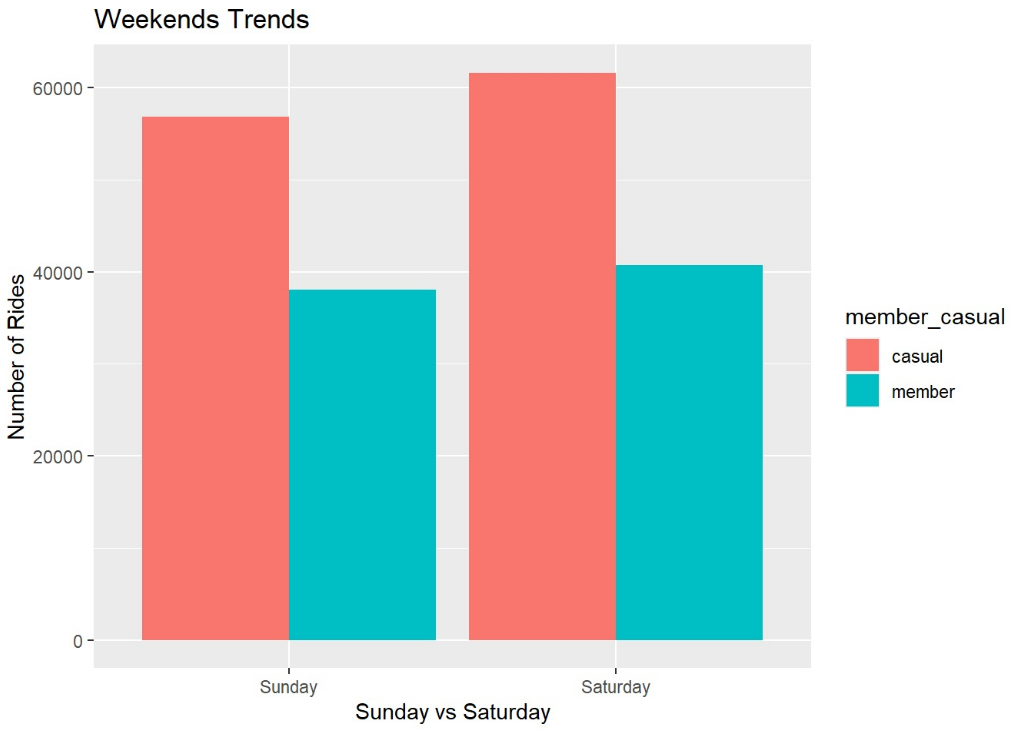
Weekday trends (Monday through Friday).

```
mc %>%
  filter(day_of_week == "Monday" |
           day_of_week == "Tuesday" |
           day_of_week == "Wednesday" |
           day_of_week == "Thursday" |
           day_of_week == "Friday") %>%
  ggplot(aes(x = day_of_week, y = Freq, fill = member_casual))+
  geom_bar(stat = "identity" , position = "dodge") +
  labs(title = "Weekdays Trends",
       x= "Days Of The Week",
       y = "Number of Rides")
```



Weekend trends (Sunday and Saturday).

```
mc %>%
  filter(day_of_week == "Sunday" |
           day_of_week == "Saturday") %>%
  ggplot(aes(x = day_of_week, y = Freq, fill = member_casual))+
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Weekends Trends",
       x= "Sunday vs Saturday",
       y = "Number of Rides")
```



Create dataframe for member and casual riders vs ride type

```
rt<- as.data.frame(table(Sep21$rideable_type,Sep21$member_casual))
```

Rename columns.

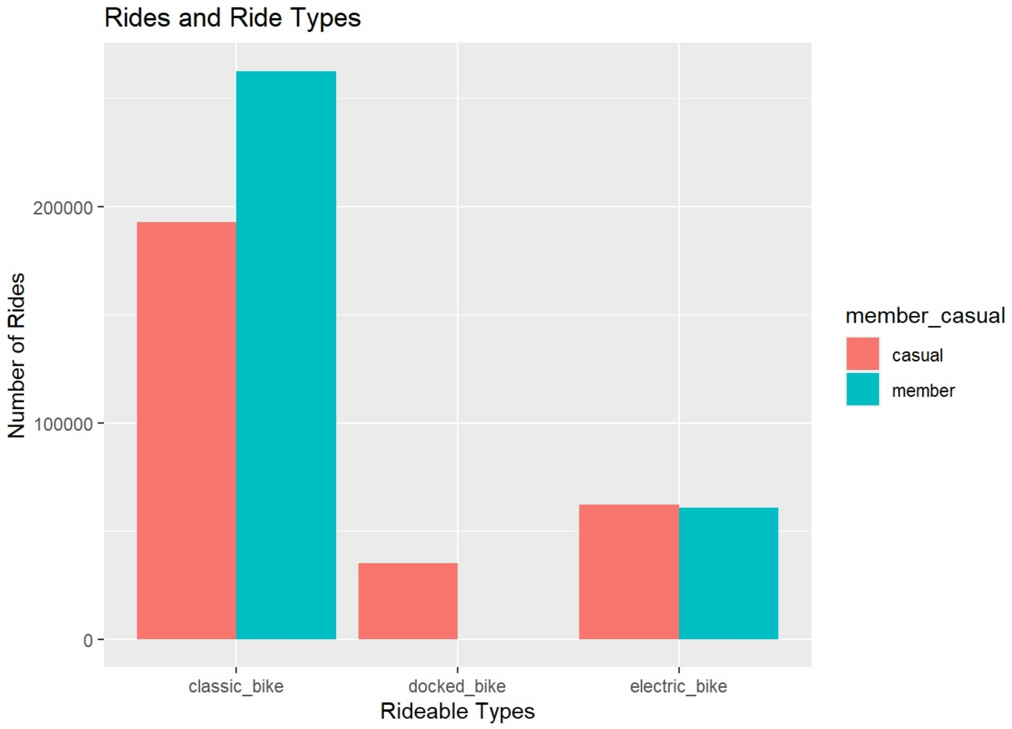
```
rt<-rename(rt, rideable_type = Var1, member_casual = Var2)

head(rt)
```

```
##   rideable_type member_casual   Freq
## 1  classic_bike        casual 192886
## 2   docked_bike        casual  35111
## 3 electric_bike        casual  62081
## 4  classic_bike        member 262605
## 5   docked_bike        member      0
## 6 electric_bike        member  60682
```

Plot for bike user vs bike type.

```
rt %>%
  filter(member_casual == "member" |
           member_casual == "casual") %>%
  ggplot(aes(x = rideable_type, y = Freq, fill = member_casual))+
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Rides and Ride Types",
       x= "Rideable Types",
       y = "Number of Rides")
```



STEP SIX: EXPORT ANALYZED DATA

Save the analyzed data as a new file. fwrite(Sep21, “Sep21.csv”)