Cyclistic Case Study Nov21

Hezar K

2022-11-29

This is an analysis for Cyclistic Case Study for Google Data Analytics Course. This is an analysis for November 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)
```

```
## — Attaching packages ─────────────────────────────── tidyverse 1.3.2 —
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## — Conflicts ────────────────────────────────── tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(ggplot2)
library(anytime)
```

Import data from local drive.

```
Nov21 <- read_csv("C:/Users/theby/Documents/202111-divvy-tripdata.csv")
```

```
## Rows: 359978 Columns: 13
## — Column specification ─────────────────────────────────────
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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(Nov21)

```
colnames(Nov21)
```

```
##  [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(Nov21)
```

```
## [1] 359978
```

```
dim(Nov21)
```

```
## [1] 359978      13
```

```
head(Nov21)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ start…² ended…³ start…⁴ start…⁵ end_s…⁶ end_s…⁷ start…⁸
##   <chr>          <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>
## 1 7C00A93E10556… electr… 11/27/… 11/27/… <NA>    <NA>    <NA>    <NA>       41.9
## 2 90854840DFD50… electr… 11/27/… 11/27/… <NA>    <NA>    <NA>    <NA>       42.0
## 3 0A7D10CDD1440… electr… 11/26/… 11/26/… <NA>    <NA>    <NA>    <NA>       42.0
## 4 2F3BE33085BCF… electr… 11/27/… 11/27/… <NA>    <NA>    <NA>    <NA>       41.9
## 5 D67B4781A1992… electr… 11/26/… 11/26/… <NA>    <NA>    <NA>    <NA>       41.9
## 6 02F85C2C3C5F7… electr… 11/26/… 11/26/… Michig… 13042   <NA>    <NA>       41.9
## # … with 4 more variables: start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²started_at, ³ended_at, ⁴start_station_name, ⁵start_station_id,
## #   ⁶end_station_name, ⁷end_station_id, ⁸start_lat
```

```
tail(Nov21)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ start…² ended…³ start…⁴ start…⁵ end_s…⁶ end_s…⁷ start…⁸
##   <chr>          <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>
## 1 2E383B4D2965B… electr… 11/4/2… 11/4/2… Cityfr… 13427   <NA>    <NA>       41.9
## 2 E00E9F3500D69… electr… 11/29/… 11/29/… Logan … TA1308… <NA>    <NA>       41.9
## 3 8EAA66CE314E5… electr… 11/3/2… 11/3/2… Logan … TA1308… <NA>    <NA>       41.9
## 4 36C2DC8BB1E13… electr… 11/2/2… 11/2/2… Logan … TA1308… <NA>    <NA>       41.9
## 5 8E42FE5C67DF6… electr… 11/10/… 11/10/… Logan … TA1308… <NA>    <NA>       41.9
## 6 4F15069E2D251… electr… 11/30/… 11/30/… Ogden … TA1305… <NA>    <NA>       41.9
## # … with 4 more variables: start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²started_at, ³ended_at, ⁴start_station_name, ⁵start_station_id,
## #   ⁶end_station_name, ⁷end_station_id, ⁸start_lat
```

```
summary(Nov21)
```

```
##    ride_id          rideable_type      started_at         ended_at
## Length:359978      Length:359978      Length:359978      Length:359978
## 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:359978      Length:359978      Length:359978      Length:359978
## 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.   :41.65    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 :41.89    Median :-87.64    Median :41.89    Median :-87.64
## Mean   :41.89    Mean   :-87.65    Mean   :41.89    Mean   :-87.65
## 3rd Qu.:41.93    3rd Qu.:-87.63    3rd Qu.:41.93    3rd Qu.:-87.63
## Max.   :42.07    Max.   :-87.53    Max.   :42.12    Max.   :-87.53
##                                    NA's   :191      NA's   :191
## member_casual
## Length:359978
## Class :character
## Mode  :character
##
##
##
##
```

```
str(Nov21)
```

```
## spc_tbl_ [359,978 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id           : chr [1:359978] "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3BE33085BCFF
02" ...
## $ rideable_type     : chr [1:359978] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
## $ started_at        : chr [1:359978] "11/27/2021 13:27" "11/27/2021 13:38" "11/26/2021 22:03" "11/27/2021 9:5
6" ...
## $ ended_at          : chr [1:359978] "11/27/2021 13:46" "11/27/2021 13:56" "11/26/2021 22:05" "11/27/2021 10:
01" ...
## $ start_station_name: chr [1:359978] NA NA NA NA ...
## $ start_station_id  : chr [1:359978] NA NA NA NA ...
## $ end_station_name  : chr [1:359978] NA NA NA NA ...
## $ end_station_id    : chr [1:359978] NA NA NA NA ...
## $ start_lat         : num [1:359978] 41.9 42 42 41.9 41.9 ...
## $ start_lng         : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ end_lat           : num [1:359978] 42 41.9 42 41.9 41.9 ...
## $ end_lng           : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ member_casual     : chr [1:359978] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   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>
```

Columns *started_at* and *ended_at* need to be convert from character data type to date data type. **Str()** syntax confirms changes.

```
Nov21$started_at <- mdy_hm(Nov21$started_at)
Nov21$ended_at <- mdy_hm(Nov21$ended_at)
str(Nov21)
```

```
## spc_tbl_ [359,978 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id           : chr [1:359978] "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3BE33085BCFF
02" ...
## $ rideable_type     : chr [1:359978] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
## $ started_at        : POSIXct[1:359978], format: "2021-11-27 13:27:00" "2021-11-27 13:38:00" ...
## $ ended_at          : POSIXct[1:359978], format: "2021-11-27 13:46:00" "2021-11-27 13:56:00" ...
## $ start_station_name: chr [1:359978] NA NA NA NA ...
## $ start_station_id  : chr [1:359978] NA NA NA NA ...
## $ end_station_name  : chr [1:359978] NA NA NA NA ...
## $ end_station_id    : chr [1:359978] NA NA NA NA ...
## $ start_lat         : num [1:359978] 41.9 42 42 41.9 41.9 ...
## $ start_lng         : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ end_lat           : num [1:359978] 42 41.9 42 41.9 41.9 ...
## $ end_lng           : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ member_casual     : chr [1:359978] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
##   .. cols(
##   ..     ride_id = col_character(),
##   ..     rideable_type = col_character(),
##   ..     started_at = col_character(),
##   ..     ended_at = col_character(),
##   ..     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.

```
Nov21$date <- as.Date(Nov21$started_at)
Nov21$month <- format(as.Date(Nov21$date), "%m")
Nov21$day <- format(as.Date(Nov21$date), "%d")
Nov21$year <- format(as.Date(Nov21$date), "%Y")
Nov21$day_of_week <- format(as.Date(Nov21$date), "%A")
Nov21$ride_length <- difftime(Nov21$ended_at,Nov21$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(Nov21$ride_length)
```

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

Recheck *ride_length* data type.

```
Nov21$ride_length <- as.numeric(as.character(Nov21$ride_length))
is.numeric(Nov21$ride_length)
```

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

STEP THREE: CLEAN DATA

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

```
Nov21 <- na.omit(Nov21)
```

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.

```
Nov21 <- subset(Nov21, nchar(as.character(ride_id)) == 16)
```

Remove rows with the *ride_length* less than 1 minute.

```
Nov21 <- subset (Nov21, ride_length > "1")
```

STEP FOUR: ANALYZE DATA

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

```
mean(Nov21$ride_length)
```

```
## [1] 853.4811
```

```
median(Nov21$ride_length)
```

```
## [1] 540
```

```
max(Nov21$ride_length)
```

```
## [1] 1336800
```

```
min(Nov21$ride_length)
```

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

Run a statistical summary of the *ride_length*.

```
summary(Nov21$ride_length)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      60.0     300.0     540.0     853.5     900.0 1336800.0
```

Compare the members and casual users

```
aggregate(Nov21$ride_length ~ Nov21$member_casual, FUN = mean)
```

```
##    Nov21$member_casual Nov21$ride_length
## 1               casual         1356.5661
## 2               member          663.4331
```

```
aggregate(Nov21$ride_length ~ Nov21$member_casual, FUN = median)
```

```
##    Nov21$member_casual Nov21$ride_length
## 1               casual               720
## 2               member               480
```

```
aggregate(Nov21$ride_length ~ Nov21$member_casual, FUN = max)
```

```
##    Nov21$member_casual Nov21$ride_length
## 1               casual           1336800
## 2               member             87600
```

```
aggregate(Nov21$ride_length ~ Nov21$member_casual, FUN = min)
```

```
##    Nov21$member_casual Nov21$ride_length
## 1               casual                60
## 2               member                60
```

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

```
aggregate(Nov21$ride_length ~ Nov21$member_casual + Nov21$day_of_week, FUN = mean)
```

```
##      Nov21$member_casual Nov21$day_of_week Nov21$ride_length
## 1                 casual             Friday         1304.6541
## 2                 member             Friday          641.7449
## 3                 casual             Monday         1478.2245
## 4                 member             Monday          652.1121
## 5                 casual           Saturday         1486.5781
## 6                 member           Saturday          733.8206
## 7                 casual             Sunday         1609.6220
## 8                 member             Sunday          738.6727
## 9                 casual           Thursday         1280.3687
## 10                member           Thursday          637.6955
## 11                casual            Tuesday         1065.9940
## 12                member            Tuesday          635.2880
## 13                casual          Wednesday         1102.2266
## 14                member          Wednesday          649.2026
```

Sort the days of the week in order.

```
Nov21$day_of_week <- ordered(Nov21$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(Nov21$ride_length ~ Nov21$member_casual + Nov21$day_of_week, FUN = mean)

head(x)
```

```
##     Nov21$member_casual Nov21$day_of_week Nov21$ride_length
## 1                 casual             Sunday         1609.6220
## 2                 member             Sunday          738.6727
## 3                 casual             Monday         1478.2245
## 4                 member             Monday          652.1121
## 5                 casual            Tuesday         1065.9940
## 6                 member            Tuesday          635.2880
```

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

```
y <- Nov21 %>%
  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           12170            1610.
## 2 casual              2            9327            1478.
## 3 casual              3           10050            1066.
## 4 casual              4            8668            1102.
## 5 casual              5            6890            1280.
## 6 casual              6            8152            1305.
```

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

```
table(Nov21$member_casual)
```

```
##
## casual member
##  69553 184117
```

```
table(Nov21$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        152440          7536         93694
```

STEP FIVE: VISUALIZATION

Display full digits instead of scientific number.

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

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

```
Nov21 %>%
  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 = "Day of Week",
     y= "Number of Rides",
     title= "Days of the Week")
```
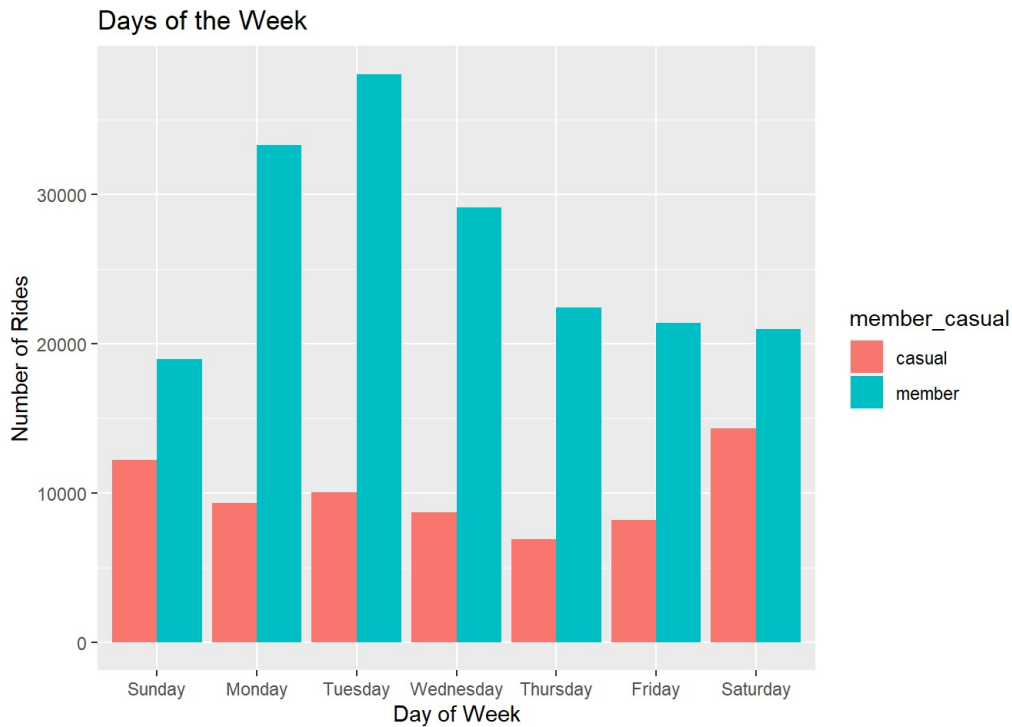


Plot the duration of the ride by user type during the week.

```
Nov21 %>%
  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 = "Day of Week",
       y= "Average Duration in Seconds",
       title= "Days of the Week vs Average Duration")
```

Days of the Week vs Average Duration
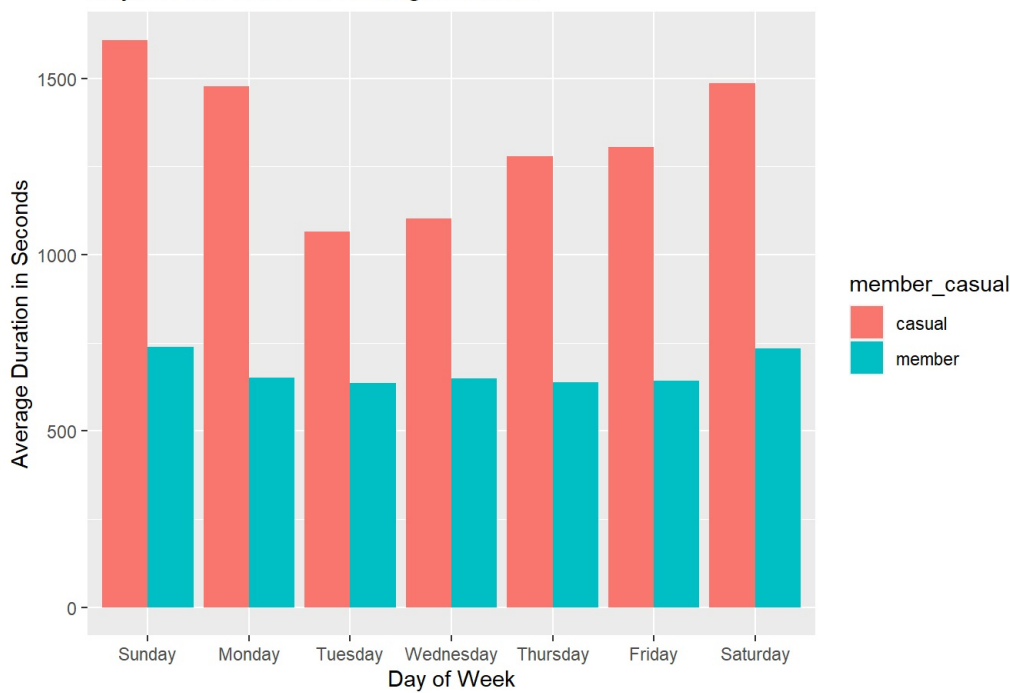
Create new dataframe for plots for weekday trends vs weekend trends.

```
mc<- as.data.frame(table(Nov21$day_of_week,Nov21$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 12170
## 2       Monday        casual  9327
## 3      Tuesday        casual 10050
## 4    Wednesday        casual  8668
## 5     Thursday        casual  6890
## 6       Friday        casual  8152
```

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= "Day Of The Week",
      y = "Rides")
```
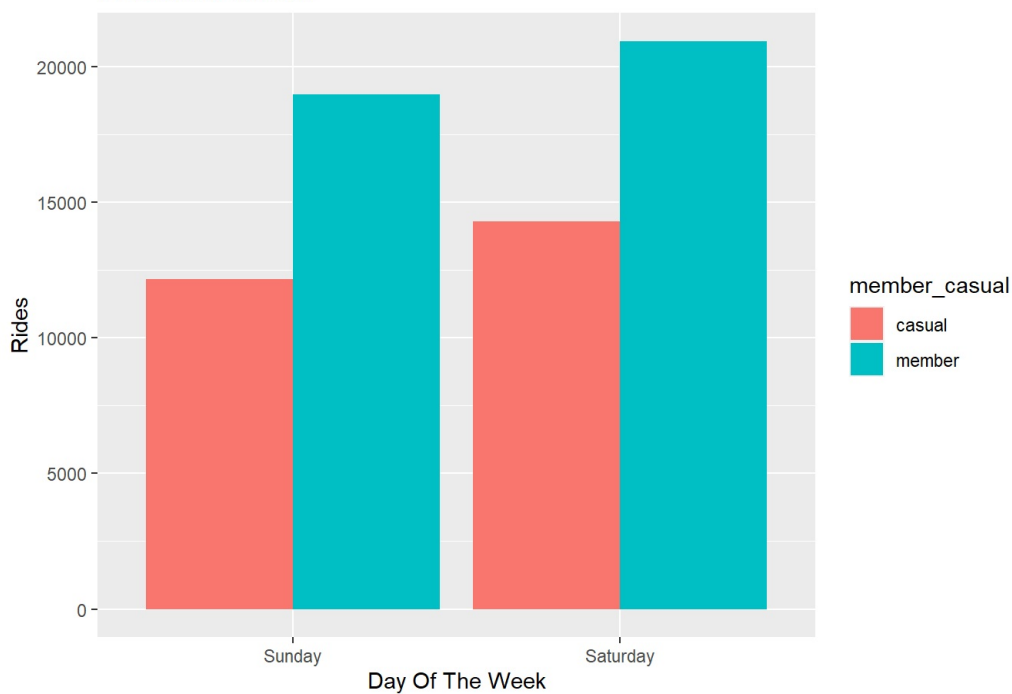
Weekdays Trends
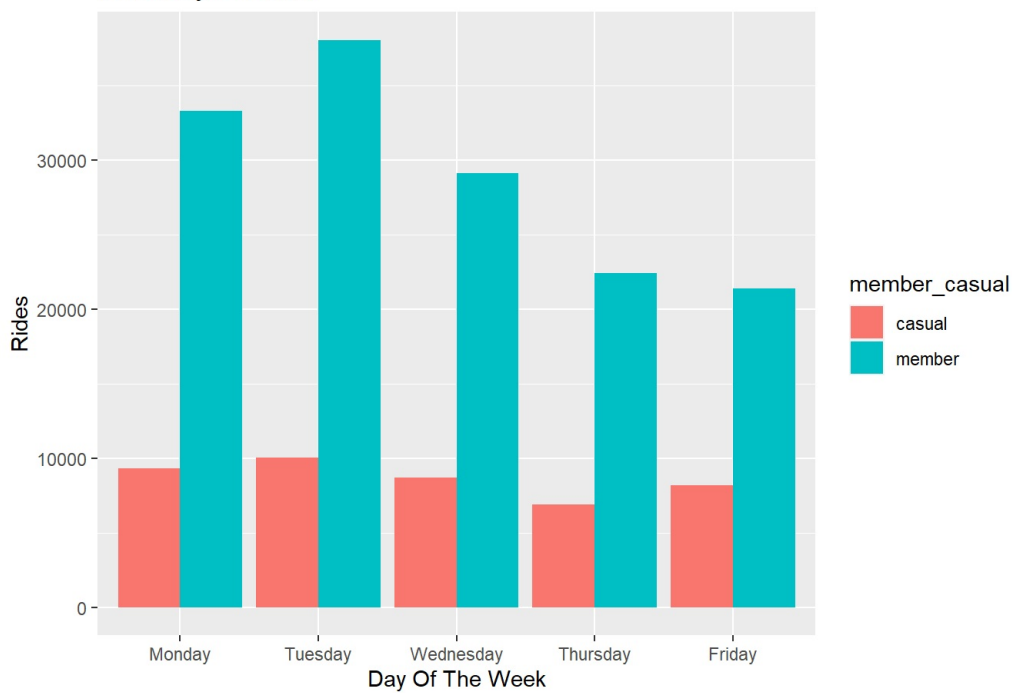
Weekend trends (Sunday and Saturday).

```
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= "Day Of The Week",
       y = "Rides")
```

Weekends Trends

Create dataframe for member and casual riders vs ride type

```
rt<- as.data.frame(table(Nov21$rideable_type,Nov21$member_casual))
```

Rename columns.

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

```
##   rideable_type member_casual   Freq
## 1  classic_bike        casual  31532
## 2   docked_bike        casual   7536
## 3 electric_bike        casual  30485
## 4  classic_bike        member 120908
## 5   docked_bike        member      0
## 6 electric_bike        member  63209
```

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 = "Riders and Ride Types",
       x= "Riders",
       y = "Rides")
```
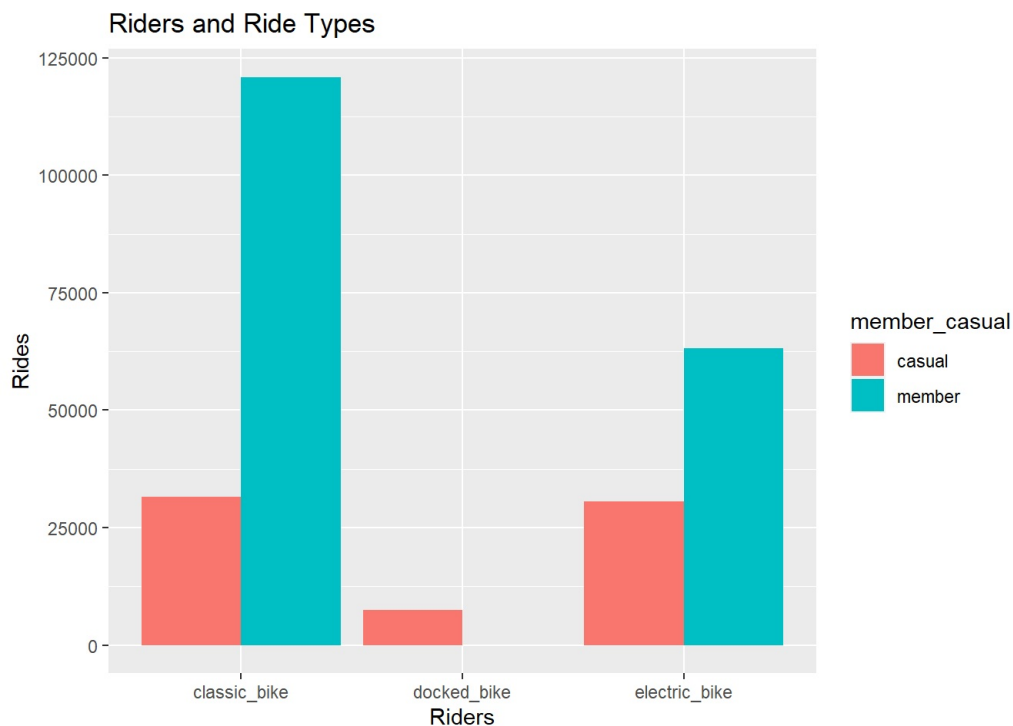

STEP SIX: EXPORT ANALYZED DATA

Save the analyzed data as a new file. fwrite(Nov21, "Nov21.csv")