Cyclistic Case Study Q4_2021

Hezar K

2022-11-29

This is an analysis for Cyclistic Case Study for Google Data Analytics Course. This is an analysis for 2021’s fourth quarter.

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.

```
Oct21 <- read_csv("202110-divvy-tripdata.csv")
```

```
## Rows: 631226 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.
```

```
Nov21 <- read_csv("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.
```

```
Dec21 <- read_csv("202112-divvy-tripdata.csv")
```

```
## Rows: 247540 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()**. Then, we need to combine all data one dataframe. Then we examine dataframes to find dimensions, **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()**.

```
colnames(Oct21)
```

```
##  [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"
```

```
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"
```

```
colnames(Dec21)
```

```
##  [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"
```

Since all column names are the same. We can combine the data for each month into quarters.

```
q4_2021 <- bind_rows(Oct21, Nov21, Dec21)
```

```
View(q4_2021)
```

```
nrow(q4_2021)
```

```
## [1] 1238744
```

```
dim(q4_2021)
```

```
## [1] 1238744      13
```

```
head(q4_2021)
```

```
## # 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 620BC6107255B… electr… 10/22/… 10/22/… Kingsb… KA1503… <NA>    <NA>       41.9
## 2 4471C70731AB2… electr… 10/21/… 10/21/… <NA>    <NA>    <NA>    <NA>       41.9
## 3 26CA69D43D15E… electr… 10/16/… 10/16/… <NA>    <NA>    <NA>    <NA>       41.9
## 4 362947F0437E1… electr… 10/16/… 10/16/… <NA>    <NA>    <NA>    <NA>       41.9
## 5 BB731DE2F2EC5… electr… 10/20/… 10/20/… <NA>    <NA>    <NA>    <NA>       41.9
## 6 7176307BBC097… electr… 10/21/… 10/21/… <NA>    <NA>    <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(q4_2021)
```

```
## # 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 92BBAB97D1683… electr… 12/24/… 12/24/… Canal … 13341   <NA>    <NA>       41.9
## 2 847431F3D5353… electr… 12/12/… 12/12/… Canal … 13341   <NA>    <NA>       41.9
## 3 CF407BBC3B9FA… electr… 12/6/2… 12/6/2… Canal … 13341   Kingsb… KA1503…    41.9
## 4 60BB69EBF5440… electr… 12/2/2… 12/2/2… Canal … 13341   Dearbo… TA1305…    41.9
## 5 C414F654A2863… electr… 12/13/… 12/13/… Lawnda… 362     <NA>    <NA>       41.9
## 6 37AC57E34B2E7… classi… 12/13/… 12/13/… Michig… TA1309… Dearbo… TA1305…    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(q4_2021)
```

```
##    ride_id          rideable_type       started_at          ended_at        
##  Length:1238744     Length:1238744     Length:1238744     Length:1238744    
##  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:1238744     Length:1238744     Length:1238744     Length:1238744    
##  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.   :41.64   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 :41.90   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.   :-87.52   Max.   :42.13   Max.   :-87.52  
##                                   NA's   :819     NA's   :819     
##  member_casual     
##  Length:1238744    
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
## 
```

```
str(q4_2021)
```

```
## spc_tbl_ [1,238,744 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:1238744] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362947F0437E1
514" ...
##  $ rideable_type     : chr [1:1238744] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr [1:1238744] "10/22/2021 12:46" "10/21/2021 9:12" "10/16/2021 16:28" "10/16/2021 16:
17" ...
##  $ ended_at          : chr [1:1238744] "10/22/2021 12:49" "10/21/2021 9:14" "10/16/2021 16:36" "10/16/2021 16:
19" ...
##  $ start_station_name: chr [1:1238744] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:1238744] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:1238744] NA NA NA NA ...
##  $ end_station_id    : chr [1:1238744] NA NA NA NA ...
##  $ start_lat         : num [1:1238744] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:1238744] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:1238744] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:1238744] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:1238744] "member" "member" "member" "member" ...
##  - 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.

```
q4_2021$started_at <- mdy_hm(q4_2021$started_at)
q4_2021$ended_at <- mdy_hm(q4_2021$ended_at)
str(q4_2021)
```

```
## spc_tbl_ [1,238,744 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:1238744] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362947F0437E1
514" ...
##  $ rideable_type     : chr [1:1238744] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:1238744], format: "2021-10-22 12:46:00" "2021-10-21 09:12:00" ...
##  $ ended_at          : POSIXct[1:1238744], format: "2021-10-22 12:49:00" "2021-10-21 09:14:00" ...
##  $ start_station_name: chr [1:1238744] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:1238744] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:1238744] NA NA NA NA ...
##  $ end_station_id    : chr [1:1238744] NA NA NA NA ...
##  $ start_lat         : num [1:1238744] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:1238744] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:1238744] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:1238744] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:1238744] "member" "member" "member" "member" ...
##  - 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.

```
q4_2021$date <- as.Date(q4_2021$started_at)
q4_2021$month <- format(as.Date(q4_2021$date), "%m")
q4_2021$day <- format(as.Date(q4_2021$date), "%d")
q4_2021$year <- format(as.Date(q4_2021$date), "%Y")
q4_2021$day_of_week <- format(as.Date(q4_2021$date), "%A")
q4_2021$ride_length <- difftime(q4_2021$ended_at,q4_2021$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(q4_2021$ride_length)
```

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

Recheck *ride_length* data type.

```
q4_2021$ride_length <- as.numeric(as.character(q4_2021$ride_length))
is.numeric(q4_2021$ride_length)
```

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

STEP THREE: CLEAN DATA

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

```
q4_2021 <- na.omit(q4_2021)
```

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.

```
q4_2021 <- subset(q4_2021, nchar(as.character(ride_id)) == 16)
```

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

```
q4_2021 <- subset (q4_2021, ride_length > "1")
```

STEP FOUR: ANALYZE DATA

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

```
mean(q4_2021$ride_length)
```

```
## [1] 967.9315
```

```
median(q4_2021$ride_length)
```

```
## [1] 600
```

```
max(q4_2021$ride_length)
```

```
## [1] 2442300
```

```
min(q4_2021$ride_length)
```

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

Run a statistical summary of the *ride_length*.

```
summary(q4_2021$ride_length)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      60.0     360.0     600.0     967.9    1020.0 2442300.0
```

Compare the members and casual users

```
aggregate(q4_2021$ride_length ~ q4_2021$member_casual, FUN = mean)
```

```
##   q4_2021$member_casual q4_2021$ride_length
## 1                casual           1520.8113
## 2                member            689.2651
```

```
aggregate(q4_2021$ride_length ~ q4_2021$member_casual, FUN = median)
```

```
##   q4_2021$member_casual q4_2021$ride_length
## 1                casual                 780
## 2                member                 480
```

```
aggregate(q4_2021$ride_length ~ q4_2021$member_casual, FUN = max)
```

```
##   q4_2021$member_casual q4_2021$ride_length
## 1                casual             2442300
## 2                member               87600
```

```
aggregate(q4_2021$ride_length ~ q4_2021$member_casual, FUN = min)
```

```
##   q4_2021$member_casual q4_2021$ride_length
## 1                casual                  60
## 2                member                  60
```

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

```
aggregate(q4_2021$ride_length ~ q4_2021$member_casual + q4_2021$day_of_week, FUN = mean)
```

```
##    q4_2021$member_casual q4_2021$day_of_week q4_2021$ride_length
## 1                 casual              Friday           1422.7698
## 2                 member              Friday            676.6739
## 3                 casual              Monday           1470.2049
## 4                 member              Monday            655.5554
## 5                 casual            Saturday           1678.7859
## 6                 member            Saturday            780.4854
## 7                 casual              Sunday           1843.4167
## 8                 member              Sunday            781.2556
## 9                 casual            Thursday           1262.1898
## 10                member            Thursday            644.8557
## 11                casual             Tuesday           1275.1752
## 12                member             Tuesday            654.4929
## 13                casual           Wednesday           1269.1827
## 14                member           Wednesday            663.8072
```

Sort the days of the week in order.

```
q4_2021$day_of_week <- ordered(q4_2021$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursda
y", "Friday", "Saturday"))
```

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

```
x <- aggregate(q4_2021$ride_length ~ q4_2021$member_casual + q4_2021$day_of_week, FUN = mean)

head(x)
```

```
##   q4_2021$member_casual q4_2021$day_of_week q4_2021$ride_length
## 1                casual              Sunday           1843.4167
## 2                member              Sunday            781.2556
## 3                casual              Monday           1470.2049
## 4                member              Monday            655.5554
## 5                casual             Tuesday           1275.1752
## 6                member             Tuesday            654.4929
```

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

```
y <- q4_2021 %>%
  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           57723            1843.
## 2 casual              2           29725            1470.
## 3 casual              3           32876            1275.
## 4 casual              4           33807            1269.
## 5 casual              5           29976            1262.
## 6 casual              6           44869            1423.
```

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

```
table(q4_2021$member_casual)
```

```
## 
## casual member 
## 302509 600184
```

```
table(q4_2021$rideable_type)
```

```
## 
##  classic_bike   docked_bike electric_bike 
##        564683         34987        303023
```

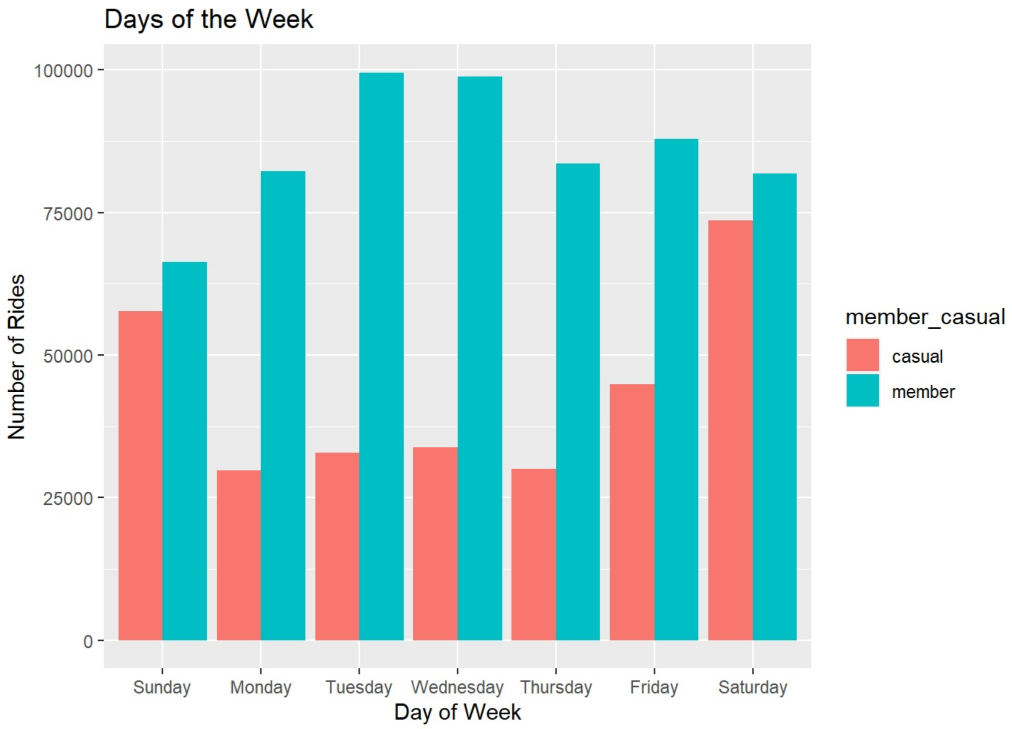
STEP FIVE: VISUALIZATION

Display full digits instead of scientific number.

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

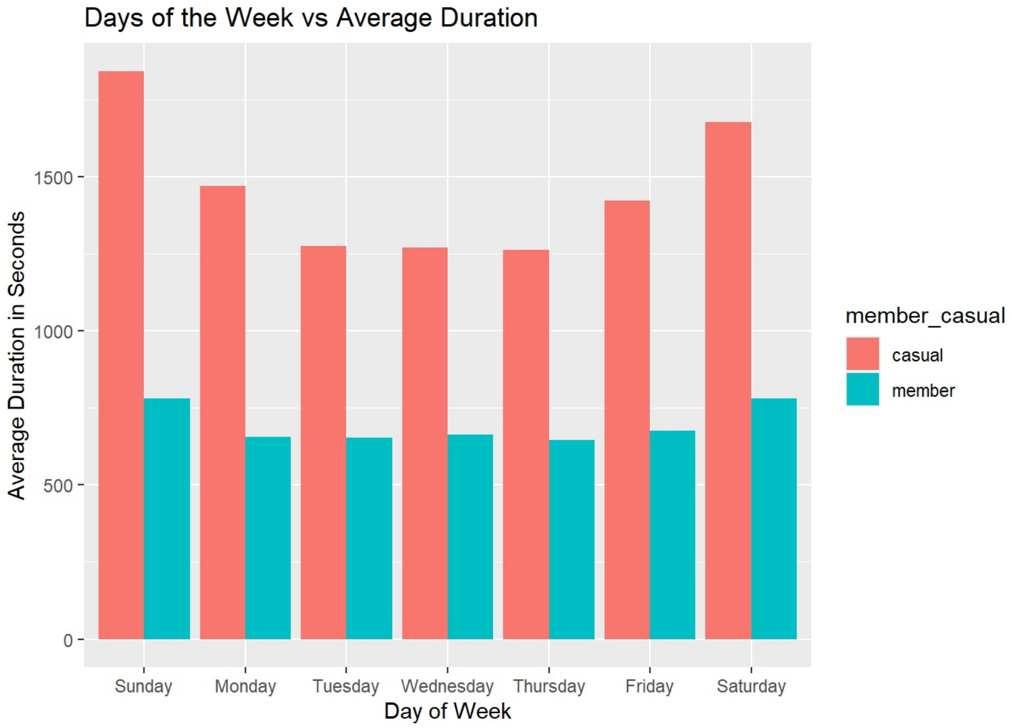
Plot the number of rides by user type during the week.

```
q4_2021 %>%
  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.

```
q4_2021 %>%
  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")
```



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

```
mc<- as.data.frame(table(q4_2021$day_of_week,q4_2021$member_casual))
```

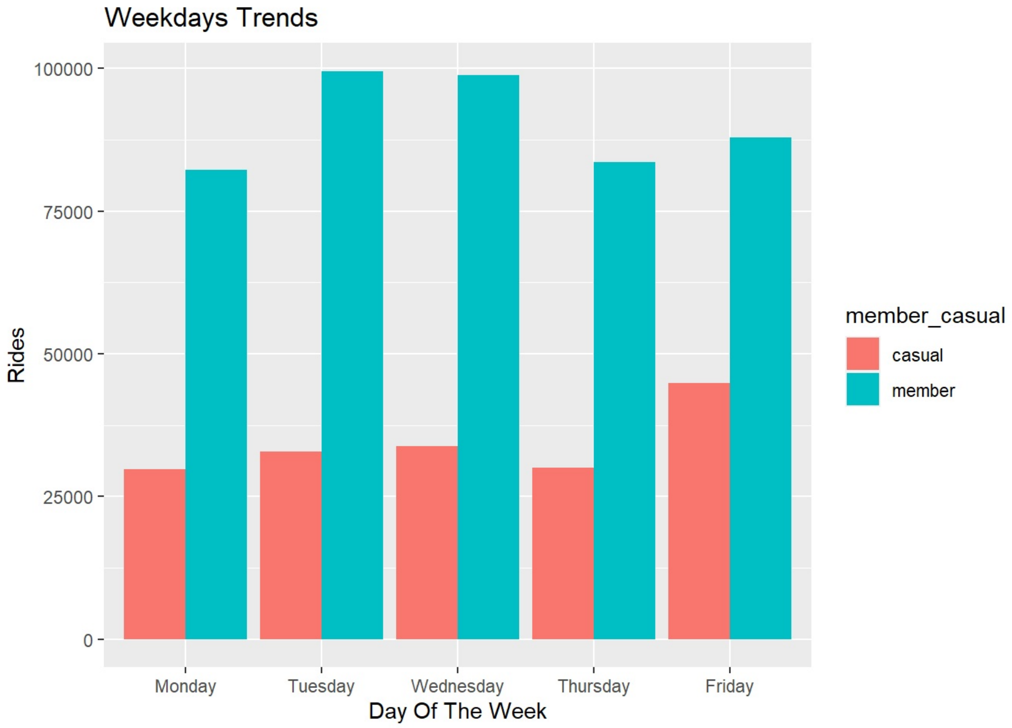
Rename columns

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

```
##   day_of_week member_casual  Freq
## 1      Sunday        casual 57723
## 2      Monday        casual 29725
## 3     Tuesday        casual 32876
## 4   Wednesday        casual 33807
## 5    Thursday        casual 29976
## 6      Friday        casual 44869
```

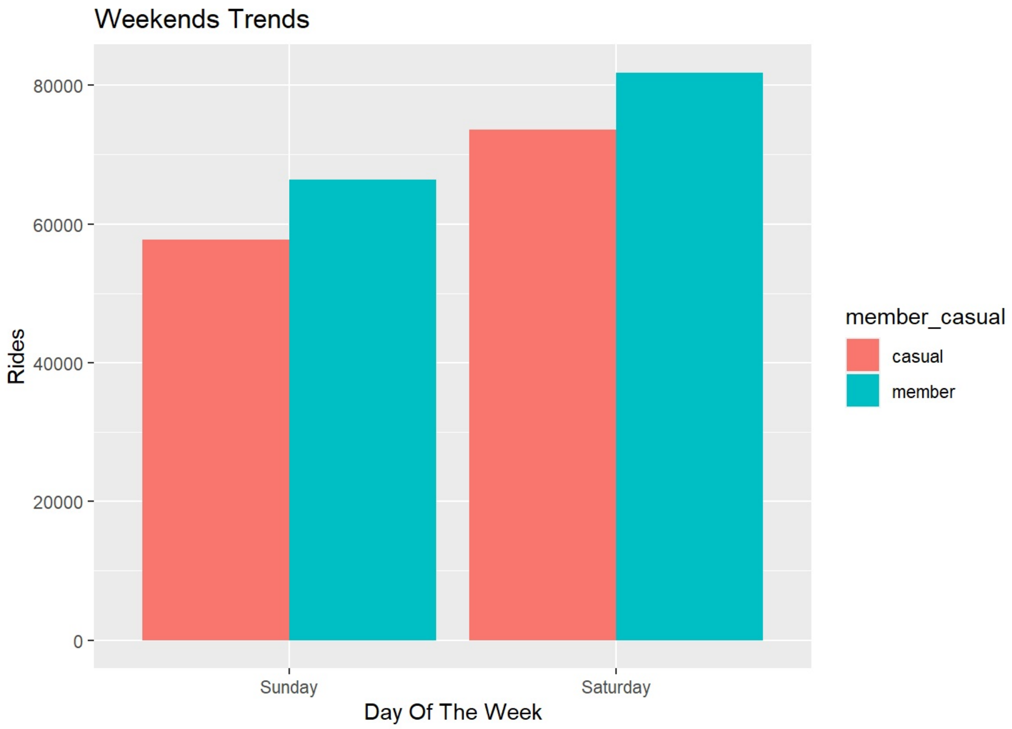
Weekday trends (Monday through Friday).

```
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")
```



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")
```



Create dataframe for member and casual riders vs ride type

```
rt<- as.data.frame(table(q4_2021$rideable_type,q4_2021$member_casual))
```

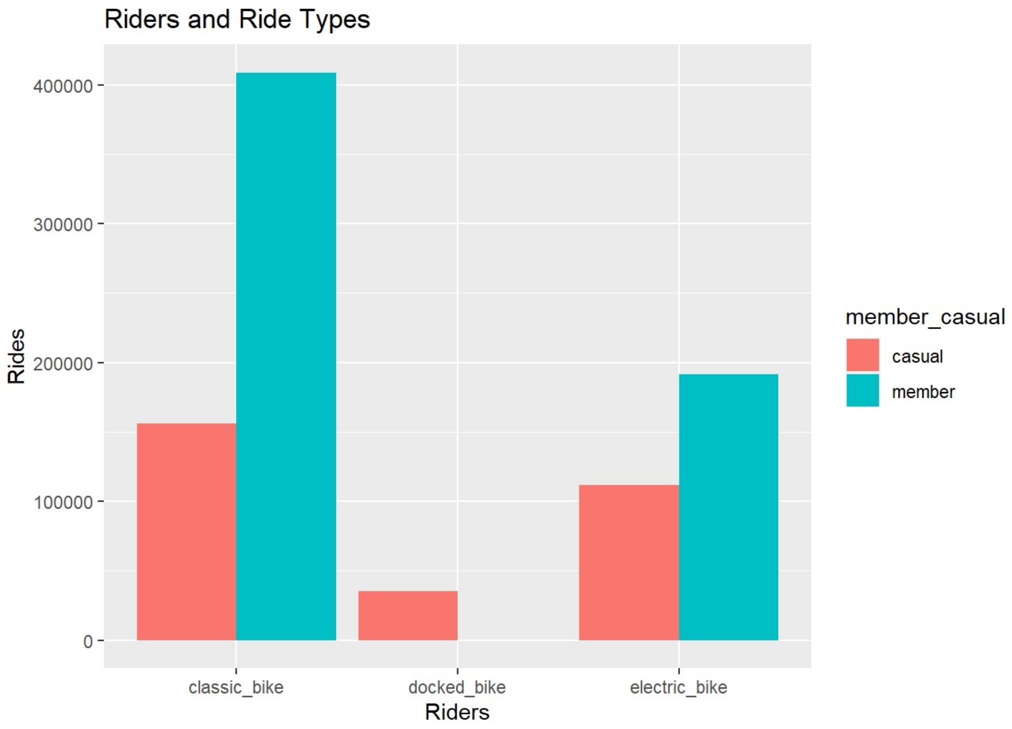
Rename columns.

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

```
##   rideable_type member_casual   Freq
## 1   classic_bike        casual 155745
## 2    docked_bike        casual  34987
## 3 electric_bike        casual 111777
## 4   classic_bike        member 408938
## 5    docked_bike        member      0
## 6 electric_bike        member 191246
```

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(q4_2021, “q4_2021.csv”)