Cyclistic Case Study Aug21

Hezar K

2022-11-29

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

```
Aug21 <- read_csv("C:/Users/theby/Documents/202108-divvy-tripdata.csv")
```

```
## Rows: 804352 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(Aug21)

```
colnames(Aug21)
```

```
##  [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(Aug21)
```

```
## [1] 804352
```

```
dim(Aug21)
```

```
## [1] 804352     13
```

```
head(Aug21)
```

```
## # 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 99103BB87CC6C… electr… 8/10/2… 8/10/2… <NA>    <NA>    <NA>    <NA>       41.8
## 2 EAFCCCFB0A3FC… electr… 8/10/2… 8/10/2… <NA>    <NA>    <NA>    <NA>       41.8
## 3 9EF4F46C57AD2… electr… 8/21/2… 8/21/2… <NA>    <NA>    <NA>    <NA>       42.0
## 4 5834D3208BFAF… electr… 8/21/2… 8/21/2… <NA>    <NA>    <NA>    <NA>       42.0
## 5 CD825CB87ED1D… electr… 8/19/2… 8/19/2… <NA>    <NA>    <NA>    <NA>       41.8
## 6 612F12C94A964… electr… 8/19/2… 8/19/2… <NA>    <NA>    <NA>    <NA>       41.8
## # … 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(Aug21)
```

```
## # 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 2D6861BE1B674… classi… 8/7/20… 8/7/20… Paulin… TA1305… Leavit… TA1308…    41.9
## 2 5E5C9CD681E04… classi… 8/7/20… 8/7/20… Wells … TA1308… Lincol… TA1307…    41.9
## 3 96FB57CF4AA45… electr… 8/9/20… 8/9/20… Broadw… 13323   Clark … 13179      42.0
## 4 226A0910DCCE9… classi… 8/12/2… 8/12/2… Dearbo… TA1305… Clark … 13179      41.9
## 5 1A97D27AE23DE… classi… 8/8/20… 8/8/20… Broadw… 13323   Clark … TA1309…    42.0
## 6 BBC36E4AA3652… electr… 8/27/2… 8/27/2… Paulin… TA1305… Dayton… 13058      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(Aug21)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:804352      Length:804352      Length:804352      Length:804352
##  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:804352      Length:804352      Length:804352      Length:804352
##  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.   :41.58   Min.   :-87.85
##  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.   :42.15   Max.   :-87.51
##                                   NA's   :706     NA's   :706
##  member_casual
##  Length:804352
##  Class :character
##  Mode  :character
##
##
##
##
```

```
str(Aug21)
```

```
## spc_tbl_ [804,352 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFAF1
DA" ...
##  $ rideable_type     : chr [1:804352] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr [1:804352] "8/10/2021 17:15" "8/10/2021 17:23" "8/21/2021 2:34" "8/21/2021 6:52"
...
##  $ ended_at          : chr [1:804352] "8/10/2021 17:22" "8/10/2021 17:39" "8/21/2021 2:50" "8/21/2021 7:08"
...
##  $ start_station_name: chr [1:804352] NA NA NA NA ...
##  $ start_station_id  : chr [1:804352] NA NA NA NA ...
##  $ end_station_name  : chr [1:804352] NA NA NA NA ...
##  $ end_station_id    : chr [1:804352] NA NA NA NA ...
##  $ start_lat         : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num [1:804352] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num [1:804352] -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:804352] "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.

```
Aug21$started_at <- mdy_hm(Aug21$started_at)
Aug21$ended_at <- mdy_hm(Aug21$ended_at)
str(Aug21)
```

```
## spc_tbl_ [804,352 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFAF1
DA" ...
##  $ rideable_type     : chr [1:804352] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:804352], format: "2021-08-10 17:15:00" "2021-08-10 17:23:00" ...
##  $ ended_at          : POSIXct[1:804352], format: "2021-08-10 17:22:00" "2021-08-10 17:39:00" ...
##  $ start_station_name: chr [1:804352] NA NA NA NA ...
##  $ start_station_id  : chr [1:804352] NA NA NA NA ...
##  $ end_station_name  : chr [1:804352] NA NA NA NA ...
##  $ end_station_id    : chr [1:804352] NA NA NA NA ...
##  $ start_lat         : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num [1:804352] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num [1:804352] -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:804352] "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.

```
Aug21$date <- as.Date(Aug21$started_at)
Aug21$month <- format(as.Date(Aug21$date), "%m")
Aug21$day <- format(as.Date(Aug21$date), "%d")
Aug21$year <- format(as.Date(Aug21$date), "%Y")
Aug21$day_of_week <- format(as.Date(Aug21$date), "%A")
Aug21$ride_length <- difftime(Aug21$ended_at,Aug21$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(Aug21$ride_length)
```

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

Recheck *ride_length* data type.

```
Aug21$ride_length <- as.numeric(as.character(Aug21$ride_length))
is.numeric(Aug21$ride_length)
```

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

STEP THREE: CLEAN DATA

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

```
Aug21 <- na.omit(Aug21)
```

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.

```
Aug21 <- subset(Aug21, nchar(as.character(ride_id)) == 16)
```

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

```
Aug21 <- subset (Aug21, ride_length > "1")
```

STEP FOUR: ANALYZE DATA

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

```
mean(Aug21$ride_length)
```

```
## [1] 1278.496
```

```
median(Aug21$ride_length)
```

```
## [1] 780
```

```
max(Aug21$ride_length)
```

```
## [1] 2497740
```

```
min(Aug21$ride_length)
```

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

Run a statistical summary of the *ride_length*.

```
summary(Aug21$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      60     420     780    1278    1380 2497740
```

Compare the members and casual users

```
aggregate(Aug21$ride_length ~ Aug21$member_casual, FUN = mean)
```

```
##   Aug21$member_casual Aug21$ride_length
## 1              casual         1723.3104
## 2              member          820.6077
```

```
aggregate(Aug21$ride_length ~ Aug21$member_casual, FUN = median)
```

```
##   Aug21$member_casual Aug21$ride_length
## 1              casual               960
## 2              member               600
```

```
aggregate(Aug21$ride_length ~ Aug21$member_casual, FUN = max)
```

```
##   Aug21$member_casual Aug21$ride_length
## 1              casual           2497740
## 2              member             89160
```

```
aggregate(Aug21$ride_length ~ Aug21$member_casual, FUN = min)
```

```
##   Aug21$member_casual Aug21$ride_length
## 1              casual                60
## 2              member                60
```

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

```
aggregate(Aug21$ride_length ~ Aug21$member_casual + Aug21$day_of_week, FUN = mean)
```

```
##    Aug21$member_casual Aug21$day_of_week Aug21$ride_length
## 1               casual            Friday         1636.8977
## 2               member            Friday          798.4130
## 3               casual            Monday         1713.1493
## 4               member            Monday          775.5281
## 5               casual          Saturday         1819.5616
## 6               member          Saturday          936.0924
## 7               casual            Sunday         1958.9833
## 8               member            Sunday          944.8154
## 9               casual          Thursday         1548.5790
## 10              member          Thursday          780.2019
## 11              casual           Tuesday         1556.6666
## 12              member           Tuesday          755.0922
## 13              casual         Wednesday         1515.6794
## 14              member         Wednesday          771.2708
```

Sort the days of the week in order.

```
Aug21$day_of_week <- ordered(Aug21$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(Aug21$ride_length ~ Aug21$member_casual + Aug21$day_of_week, FUN = mean)

head(x)
```

```
##   Aug21$member_casual Aug21$day_of_week Aug21$ride_length
## 1              casual            Sunday         1958.9833
## 2              member            Sunday          944.8154
## 3              casual            Monday         1713.1493
## 4              member            Monday          775.5281
## 5              casual           Tuesday         1556.6666
## 6              member           Tuesday          755.0922
```

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

```
y <- Aug21 %>%
  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           72999            1959.
## 2 casual              2           39628            1713.
## 3 casual              3           37439            1557.
## 4 casual              4           32125            1516.
## 5 casual              5           37988            1549.
## 6 casual              6           47861            1637.
```

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

```
table(Aug21$member_casual)
```

```
##
## casual member
## 339534 329840
```

```
table(Aug21$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        497970         44903        126501
```

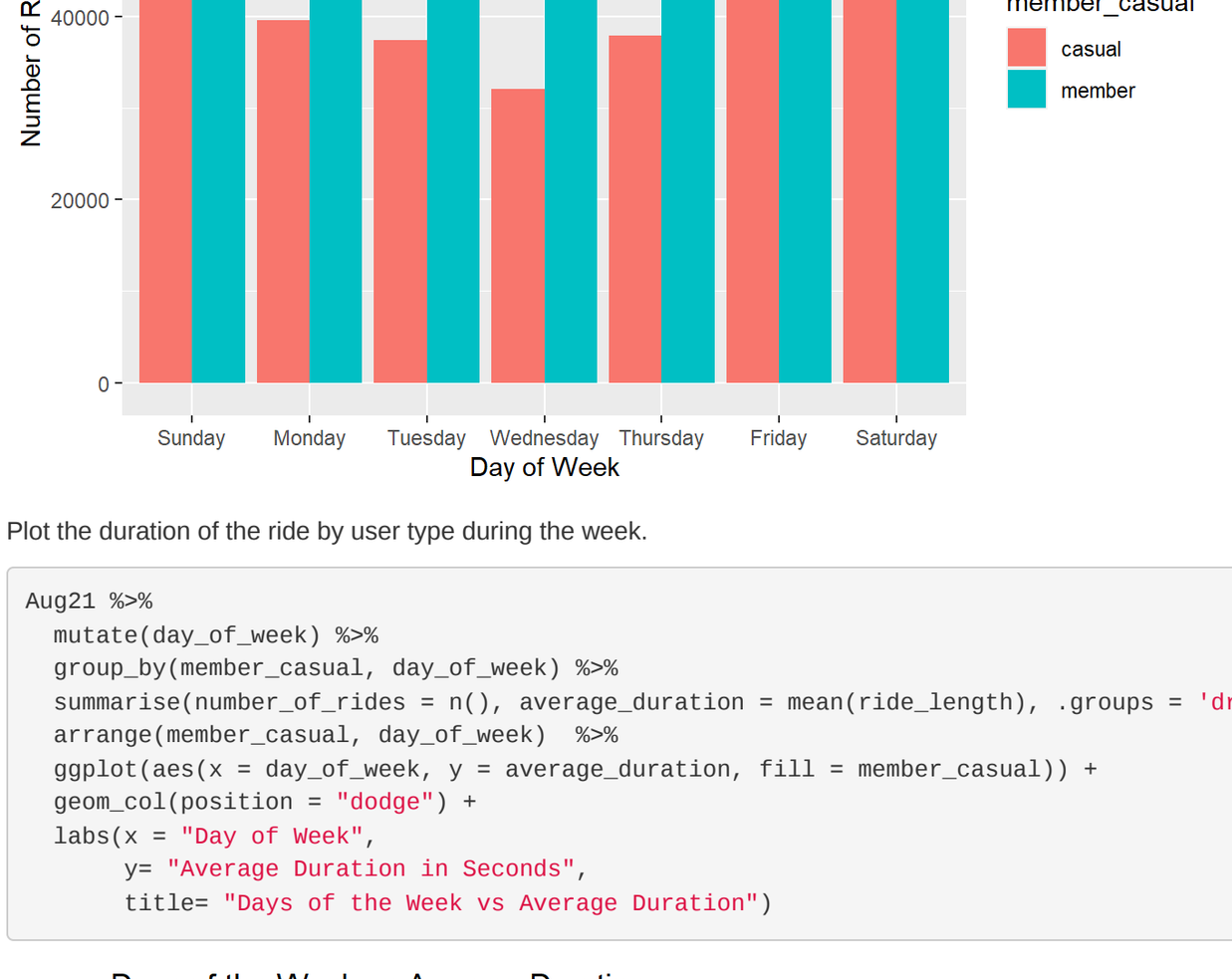
STEP FIVE: VISUALIZATION

Display full digits instead of scientific number.

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

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

```
Aug21 %>%
  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.

```
Aug21 %>%
  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(Aug21$day_of_week,Aug21$member_casual))
```

Rename columns

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

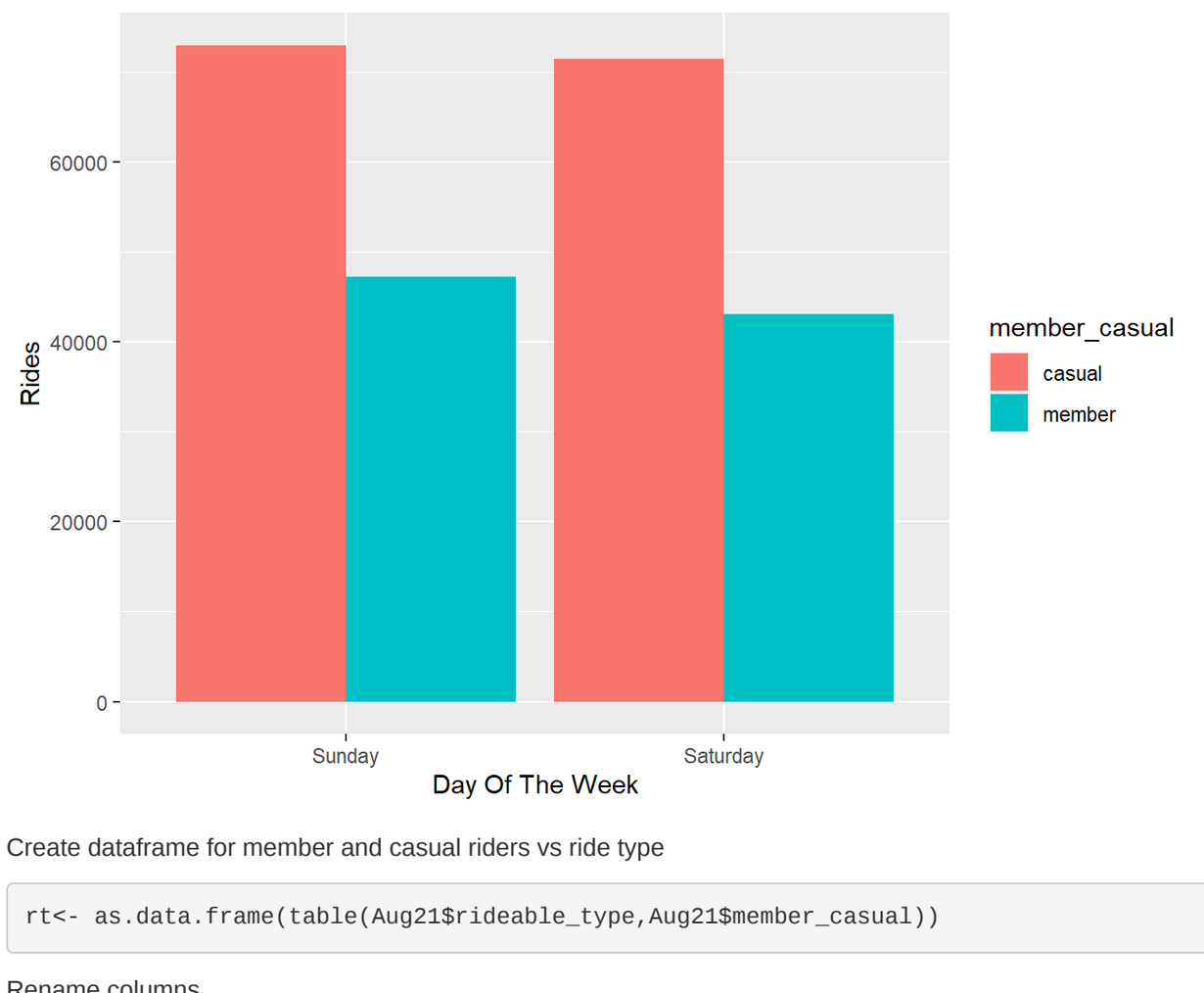
```
##   day_of_week member_casual  Freq
## 1      Sunday        casual 72999
## 2      Monday        casual 39628
## 3     Tuesday        casual 37439
## 4   Wednesday        casual 32125
## 5    Thursday        casual 37988
## 6      Friday        casual 47861
```

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(Aug21$rideable_type,Aug21$member_casual))
```

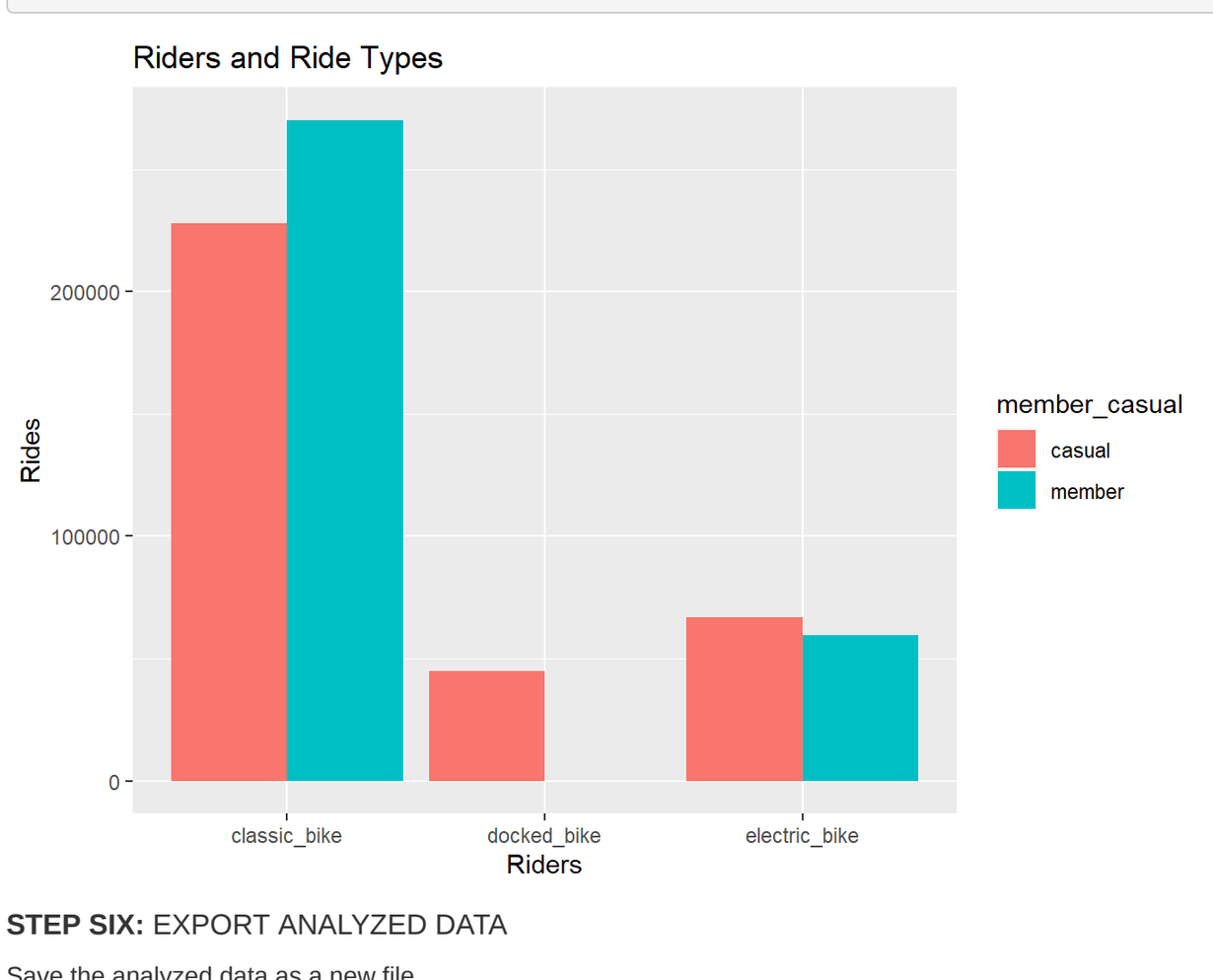
Rename columns.

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

```
##   rideable_type member_casual   Freq
## 1  classic_bike        casual 227807
## 2   docked_bike        casual  44903
## 3 electric_bike        casual  66824
## 4  classic_bike        member 270163
## 5   docked_bike        member      0
## 6 electric_bike        member  59677
```

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(Aug21, “Aug21.csv”)