

Cyclistic Case Study Jul21

Hezar K

2022-11-29

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

```
Jul21 <- read_csv("C:/Users/theby/Documents/202107-divvy-tripdata.csv")
```

```
## Rows: 822410 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(Jul21)

```
colnames(Jul21)
```

```
## [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(Jul21)
```

```
## [1] 822410
```

```
dim(Jul21)
```

```
## [1] 822410      13
```

```
head(Jul21)
```

```
## # 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 0A1B623926EF4... docked... 7/2/20... 7/2/20... Michig... 13001    Halste... KA1504...    41.9
## 2 B2D5583A5A5E7... classi... 7/7/20... 7/7/20... Califo... 17660    Wood S... 13432      41.9
## 3 6F264597DDBF4... classi... 7/25/2... 7/25/2... Wabash... SL-012    Rush S... KA1503...    41.9
## 4 379B58EAB20E8... classi... 7/8/20... 7/8/20... Califo... 17660    Carpen... 13196      41.9
## 5 6615C1E4EB08E... electr... 7/28/2... 7/28/2... Califo... 17660    Elizab... 13197      41.9
## 6 62DC2B32872F9... electr... 7/29/2... 7/29/2... Califo... 17660    Albany... 15655      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(Jul21)
```

```
## # 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 7B47CA3E874D2... electr... 7/4/20... 7/4/20... <NA>     <NA>     <NA>     <NA>     41.7
## 2 1E660BF8DCDA... electr... 7/4/20... 7/4/20... <NA>     <NA>     <NA>     <NA>     41.7
## 3 A2448BDFD9B36... electr... 7/4/20... 7/4/20... <NA>     <NA>     <NA>     <NA>     41.8
## 4 2D612BF853037... electr... 7/3/20... 7/3/20... <NA>     <NA>     <NA>     <NA>     41.7
## 5 6D615D18B765C... electr... 7/3/20... 7/3/20... <NA>     <NA>     <NA>     <NA>     41.7
## 6 0F31D311323F0... electr... 7/4/20... 7/4/20... <NA>     <NA>     <NA>     <NA>     41.7
## # ... 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(Jul21)
```

```
##      ride_id      rideable_type      started_at      ended_at
## Length:822410      Length:822410      Length:822410      Length:822410
## 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:822410      Length:822410      Length:822410      Length:822410
## 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. :41.63      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. : -87.49
## NA's :731      NA's :731
## member_casual
## Length:822410
## Class :character
## Mode :character
##
##
##
##
```

```
str(Jul21)
```

```
## spc_tbl_ [822,410 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id      : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8A
A5" ...
## $ rideable_type : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
## $ started_at   : chr [1:822410] "7/2/2021 14:44" "7/7/2021 16:57" "7/25/2021 11:30" "7/8/2021 22:08" ...
## $ ended_at     : chr [1:822410] "7/2/2021 15:19" "7/7/2021 17:16" "7/25/2021 11:48" "7/8/2021 22:23" ...
## $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Ave
& 16th St" "California Ave & Cortez St" ...
## $ start_station_id : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
## $ end_station_name : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubbard
St" "Carpenter St & Huron St" ...
## $ end_station_id   : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
## $ start_lat        : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng        : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ end_lat          : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng          : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ member_casual    : chr [1:822410] "casual" "casual" "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.

```
Jul21$started_at <- mdy_hm(Jul21$started_at)
Jul21$ended_at <- mdy_hm(Jul21$ended_at)
str(Jul21)
```

```
## spc_tbl_ [822,410 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id      : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8A
A5" ...
## $ rideable_type : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
## $ started_at    : POSIXct[1:822410], format: "2021-07-02 14:44:00" "2021-07-07 16:57:00" ...
## $ ended_at      : POSIXct[1:822410], format: "2021-07-02 15:19:00" "2021-07-07 17:16:00" ...
## $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Ave
& 16th St" "California Ave & Cortez St" ...
## $ start_station_id : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
## $ end_station_name : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubbard
St" "Carpenter St & Huron St" ...
## $ end_station_id   : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
## $ start_lat        : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng        : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ end_lat          : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng          : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ member_casual    : chr [1:822410] "casual" "casual" "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.

```
Jul21$date <- as.Date(Jul21$started_at)
Jul21$month <- format(as.Date(Jul21$date), "%m")
Jul21$day <- format(as.Date(Jul21$date), "%d")
Jul21$year <- format(as.Date(Jul21$date), "%Y")
Jul21$day_of_week <- format(as.Date(Jul21$date), "%A")
Jul21$ride_length <- difftime(Jul21$ended_at, Jul21$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(Jul21$ride_length)
```

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

Recheck *ride_length* data type.

```
Jul21$ride_length <- as.numeric(as.character(Jul21$ride_length))
is.numeric(Jul21$ride_length)
```

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

STEP THREE: CLEAN DATA

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

```
Jul21 <- na.omit(Jul21)
```

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.

```
Jul21 <- subset(Jul21, nchar(as.character(ride_id)) == 16)
```

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

```
Jul21 <- subset (Jul21, ride_length > "1")
```

STEP FOUR: ANALYZE DATA

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

```
mean(Jul21$ride_length)
```

```
## [1] 1462.816
```

```
median(Jul21$ride_length)
```

```
## [1] 840
```

```
max(Jul21$ride_length)
```

```
## [1] 2946420
```

```
min(Jul21$ride_length)
```

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

Run a statistical summary of the *ride_length*.

```
summary(Jul21$ride_length)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       60     480     840    1463    1440 2946420
```

Compare the members and casual users

```
aggregate(Jul21$ride_length ~ Jul21$member_casual, FUN = mean)
```

```
##      Jul21$member_casual Jul21$ride_length
## 1                casual      2008.8268
## 2                member      835.6412
```

```
aggregate(Jul21$ride_length ~ Jul21$member_casual, FUN = median)
```

```
##      Jul21$member_casual Jul21$ride_length
## 1                casual          1020
## 2                member           660
```

```
aggregate(Jul21$ride_length ~ Jul21$member_casual, FUN = max)
```

```
##      Jul21$member_casual Jul21$ride_length
## 1                casual      2946420
## 2                member       75720
```

```
aggregate(Jul21$ride_length ~ Jul21$member_casual, FUN = min)
```

```
##      Jul21$member_casual Jul21$ride_length
## 1                casual           60
## 2                member           60
```

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

```
aggregate(Jul21$ride_length ~ Jul21$member_casual + Jul21$day_of_week, FUN = mean)
```

```
##      Jul21$member_casual Jul21$day_of_week Jul21$ride_length
## 1          casual      Friday      1881.0675
## 2          member      Friday       807.5019
## 3          casual      Monday      2240.9308
## 4          member      Monday       824.2349
## 5          casual      Saturday     2135.3920
## 6          member      Saturday       934.1442
## 7          casual      Sunday      2243.5077
## 8          member      Sunday       949.7615
## 9          casual      Thursday     1890.3745
## 10         member      Thursday       789.7868
## 11         casual      Tuesday     1700.5441
## 12         member      Tuesday       784.6810
## 13         casual      Wednesday    1740.1383
## 14         member      Wednesday     792.6086
```

Sort the days of the week in order.

```
Jul21$day_of_week <- ordered(Jul21$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(Jul21$ride_length ~ Jul21$member_casual + Jul21$day_of_week, FUN = mean)

head(x)
```

```
##      Jul21$member_casual Jul21$day_of_week Jul21$ride_length
## 1          casual      Sunday      2243.5077
## 2          member      Sunday       949.7615
## 3          casual      Monday      2240.9308
## 4          member      Monday       824.2349
## 5          casual      Tuesday     1700.5441
## 6          member      Tuesday       784.6810
```

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

```
y <- Jul21 %>%
  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           59212           2244.
## 2 casual            2           40118           2241.
## 3 casual            3           36498           1701.
## 4 casual            4           37740           1740.
## 5 casual            5           46296           1890.
## 6 casual            6           58902           1881.
```

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

```
table(Jul21$member_casual)
```

```
##
## casual member
## 367208 319687
```

```
table(Jul21$rideable_type)
```

```
##
## classic_bike  docked_bike electric_bike
##      501378      57475      128042
```

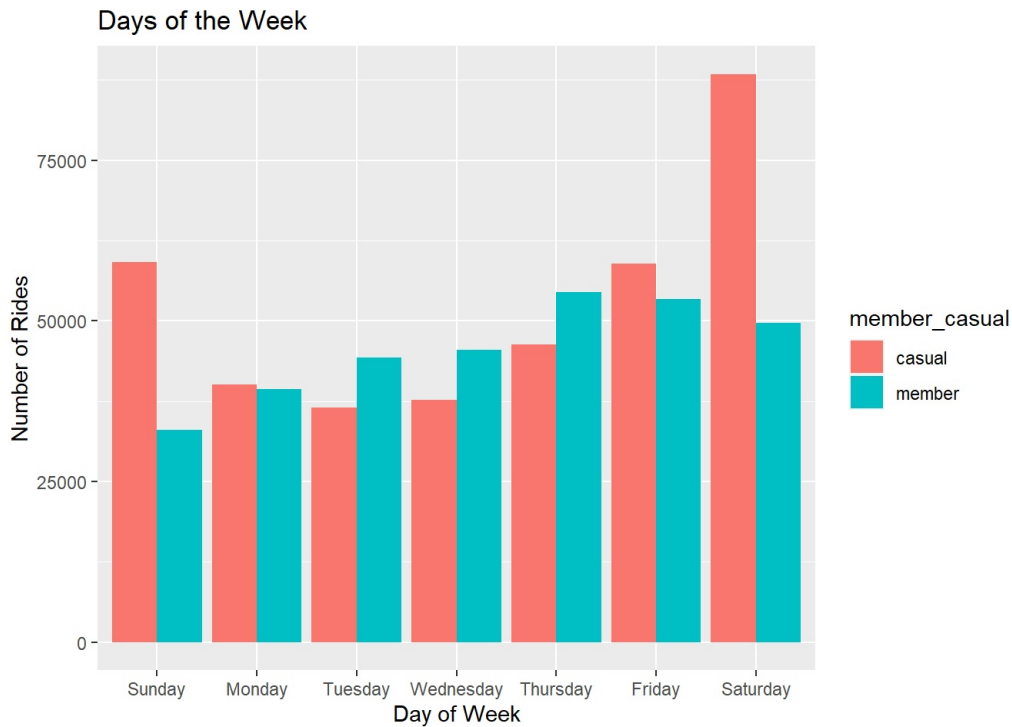
STEP FIVE: VISUALIZATION

Display full digits instead of scientific number.

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

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

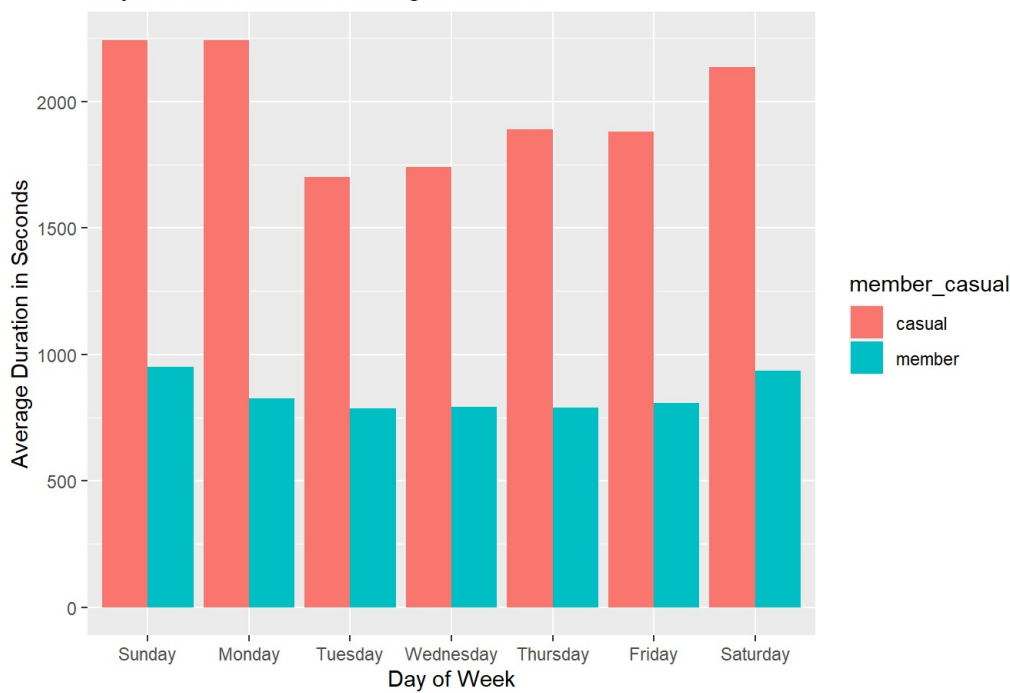
```
Jul21 %>%
  mutate(day_of_week) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length), .groups = 'drop') %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "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.

```
Jul21 %>%
  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(Jul21$day_of_week,Jul21$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 59212
## 2    Monday          casual 40118
## 3   Tuesday          casual 36498
## 4 Wednesday          casual 37740
## 5   Thursday          casual 46296
## 6    Friday          casual 58902
```

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(Jul21$rideable_type,Jul21$member_casual))
```

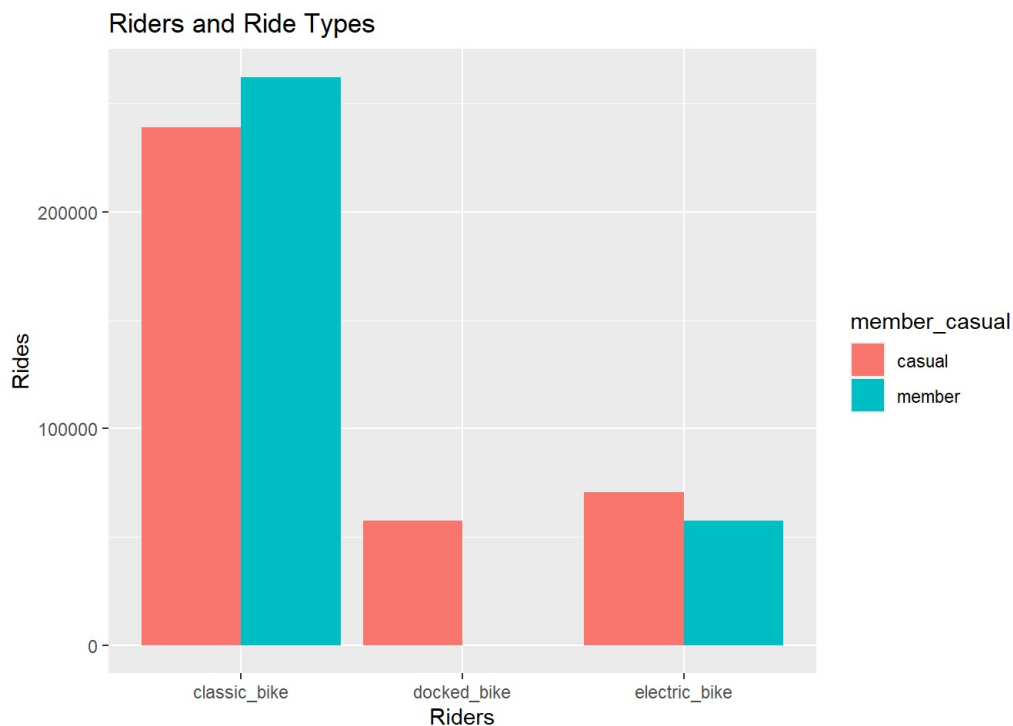
Rename columns.

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

```
##   rideable_type member_casual   Freq
## 1 classic_bike      casual 239043
## 2 docked_bike       casual  57475
## 3 electric_bike     casual  70690
## 4 classic_bike     member 262335
## 5 docked_bike      member    0
## 6 electric_bike    member  57352
```

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(Jul21, "Jul21.csv")`