Cyclistic Bike-share

Ng Min Tri

2022-03-14

R Markdown

#

Install required packages

tidyverse for data import and wrangling

lubridate for date functions

ggplot for visualization

#

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ---------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(lubridate)

## Warning: package 'lubridate' was built under R version 4.1.3

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(ggplot2)
library(dplyr)
setwd("D:\\DH\\3rd - Semester II\\Google Data Analytics\\Case
study\\Data\\CSV")
```

```
#=====================
```

STEP 1: COLLECT DATA

```
 #=====================
```

```
m1 <- read_csv("202101-divvy-tripdata.csv")

## Rows: 96834 Columns: 13
## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id,
end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## 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.

m2 <- read.csv("202102-divvy-tripdata.csv")
m3 <- read.csv("202103-divvy-tripdata.csv")
m4 <- read.csv("202104-divvy-tripdata.csv")
m5 <- read.csv("202105-divvy-tripdata.csv")
m6 <- read.csv("202106-divvy-tripdata.csv")
m7 <- read.csv("202107-divvy-tripdata.csv")
```

```
m8 <- read.csv("202108-divvy-tripdata.csv")
m9 <- read.csv("202109-divvy-tripdata.csv")
m10 <- read.csv("202110-divvy-tripdata.csv")
m11 <- read.csv("202111-divvy-tripdata.csv")
m12 <- read.csv("202112-divvy-tripdata.csv")

#=====================================================
```

STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

```
#=====================================================
```

Compare column names each of the files

```
colnames(m1)

##  [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(m2)

##  [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(m3)

##  [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(m4)

##  [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(m5)

##  [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(m6)

##  [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(m7)

##  [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(m8)

##  [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(m9)

##  [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(m10)

##  [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(m11)

##  [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(m12)
```

```
##  [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"
```

Combine data

```
all_trips <- rbind(m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11, m12)
```

```
#=====================================================
```

STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

```
#=====================================================
```

```
colnames(all_trips)
```

```
##  [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(all_trips)
```

```
## [1] 5595063
```

```
dim(all_trips)
```

```
## [1] 5595063      13
```

```
head(all_trips)
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at
start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 E19E6F~ electric_bike 2021-01-23 16:14:19 2021-01-23 16:24:44 California
Ave ~
## 2 DC88F2~ electric_bike 2021-01-27 18:43:08 2021-01-27 18:47:12 California
Ave ~
## 3 EC45C9~ electric_bike 2021-01-21 22:35:54 2021-01-21 22:37:14 California
Ave ~
## 4 4FA453~ electric_bike 2021-01-07 13:31:13 2021-01-07 13:42:55 California
Ave ~
## 5 BE5E8E~ electric_bike 2021-01-23 02:24:02 2021-01-23 02:24:45 California
Ave ~
## 6 5D8969~ electric_bike 2021-01-09 14:24:07 2021-01-09 15:17:54 California
Ave ~
```

```
## # ... with 8 more variables: start_station_id <chr>, end_station_name
<chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

```
tail(all_trips)
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at
start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 92BBAB~ electric_bike 2021-12-24 08:42:09 2021-12-24 12:29:35 Canal St &
Madi~
## 2 847431~ electric_bike 2021-12-12 06:36:55 2021-12-12 06:56:08 Canal St &
Madi~
## 3 CF407B~ electric_bike 2021-12-06 12:37:50 2021-12-06 12:44:51 Canal St &
Madi~
## 4 60BB69~ electric_bike 2021-12-02 01:57:04 2021-12-02 02:05:21 Canal St &
Madi~
## 5 C414F6~ electric_bike 2021-12-13 02:00:26 2021-12-13 02:14:39 Lawndale
Ave & ~
## 6 37AC57~ classic_bike  2021-12-13 01:45:32 2021-12-13 01:49:09 Michigan
Ave & ~
## # ... with 8 more variables: start_station_id <chr>, end_station_name
<chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

```
str(all_trips)
```

```
## spec_tbl_df [5,595,063 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5595063] "E19E6F1B8D4C42ED"
"DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr [1:5595063] "electric_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:5595063], format: "2021-01-23 16:14:19"
"2021-01-27 18:43:08" ...
##  $ ended_at          : POSIXct[1:5595063], format: "2021-01-23 16:24:44"
"2021-01-27 18:47:12" ...
##  $ start_station_name: chr [1:5595063] "California Ave & Cortez St"
"California Ave & Cortez St" "California Ave & Cortez St" "California Ave &
Cortez St" ...
##  $ start_station_id  : chr [1:5595063] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5595063] NA NA NA NA ...
##  $ end_station_id    : chr [1:5595063] NA NA NA NA ...
##  $ start_lat         : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5595063] "member" "member" "member" "member"
...
```

```
##  - 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>

summary(all_trips)

##    ride_id          rideable_type        started_at
##  Length:5595063     Length:5595063     Min.   :2021-01-01 00:02:05
##  Class :character   Class :character   1st Qu.:2021-06-06 16:52:40
##  Mode  :character   Mode  :character   Median :2021-07-31 18:52:11
##                                        Mean   :2021-07-29 00:48:18
##                                        3rd Qu.:2021-09-24 09:36:16
##                                        Max.   :2021-12-31 16:59:48
##
##     ended_at                    start_station_name start_station_id
##  Min.   :2021-01-01 00:08:39   Length:5595063     Length:5595063
##  1st Qu.:2021-06-06 17:44:21   Class :character   Class :character
##  Median :2021-07-31 19:21:55   Mode  :character   Mode  :character
##  Mean   :2021-07-29 01:10:14
##  3rd Qu.:2021-09-24 09:54:05
##  Max.   :2022-01-03 10:32:18
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5595063     Length:5595063     Min.   :41.64   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.   :41.39   Min.   :-88.97   Length:5595063
##  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.   :-87.49
##  NA's   :4771    NA's   :4771
```

Add columns that list the date, month, day, and year of each ride

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

Add a “ride_length” calculation to all_trips

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
```

Inspect the structure of the columns

```
str(all_trips)
```

```
## spec_tbl_df [5,595,063 x 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5595063] "E19E6F1B8D4C42ED"
"DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr [1:5595063] "electric_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:5595063], format: "2021-01-23 16:14:19"
"2021-01-27 18:43:08" ...
##  $ ended_at          : POSIXct[1:5595063], format: "2021-01-23 16:24:44"
"2021-01-27 18:47:12" ...
##  $ start_station_name: chr [1:5595063] "California Ave & Cortez St"
"California Ave & Cortez St" "California Ave & Cortez St" "California Ave &
Cortez St" ...
##  $ start_station_id  : chr [1:5595063] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5595063] NA NA NA NA ...
##  $ end_station_id    : chr [1:5595063] NA NA NA NA ...
##  $ start_lat         : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5595063] "member" "member" "member" "member"
...
##  $ date              : Date[1:5595063], format: "2021-01-23" "2021-01-27"
...
##  $ month             : chr [1:5595063] "01" "01" "01" "01" ...
##  $ day               : chr [1:5595063] "23" "27" "21" "07" ...
##  $ year              : chr [1:5595063] "2021" "2021" "2021" "2021" ...
##  $ day_of_week       : chr [1:5595063] "Saturday" "Wednesday" "Thursday"
"Thursday" ...
##  $ ride_length       : 'difftime' num [1:5595063] 625 244 80 702 ...
##   ..- attr(*, "units")= chr "secs"
```

```
##  - 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>
```

Convert "ride_length" from Factor to numeric

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

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

Remove "bad" data (ride length < 60 and docked bike)

```
all_trips_v2 <- all_trips[!(all_trips$rideable_type == "docked_bike"
|all_trips$ride_length<60),]
```

```
#=====================================
```

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

```
#=====================================
```

Descriptive analysis on ride_length

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      60     404     700    1122    1237   93596
```

Compare members and casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                1523.4076
## 2                     member                 831.4557
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN =
median)

##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                      900
## 2                     member                      586

aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)

##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                    93596
## 2                     member                    93596

aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)

##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                       60
## 2                     member                       60
```

See the average ride time by each day for members vs casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)

##    all_trips_v2$member_casual all_trips_v2$day_of_week
all_trips_v2$ride_length
## 1                      casual                   Friday
1445.9325
## 2                      member                   Friday
821.6713
## 3                      casual                   Monday
1525.9271
## 4                      member                   Monday
804.4126
## 5                      casual                 Saturday
1665.7982
## 6                      member                 Saturday
934.5948
## 7                      casual                   Sunday
1763.6524
## 8                      member                   Sunday
945.4467
## 9                      casual                 Thursday
1341.5377
## 10                     member                 Thursday
780.6164
## 11                     casual                  Tuesday
1372.2314
## 12                     member                  Tuesday
780.0828
## 13                     casual                Wednesday
```

```
1327.7245
## 14                    member                Wednesday
780.7767
```

#Correct order for day of the week

```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week,
levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
"Saturday"))
```

Run the code again

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)

##    all_trips_v2$member_casual all_trips_v2$day_of_week
all_trips_v2$ride_length
## 1                      casual                  Sunday
1763.6524
## 2                      member                  Sunday
945.4467
## 3                      casual                  Monday
1525.9271
## 4                      member                  Monday
804.4126
## 5                      casual                 Tuesday
1372.2314
## 6                      member                 Tuesday
780.0828
## 7                      casual               Wednesday
1327.7245
## 8                      member               Wednesday
780.7767
## 9                      casual                Thursday
1341.5377
## 10                     member                Thursday
780.6164
## 11                     casual                  Friday
1445.9325
## 12                     member                  Friday
821.6713
## 13                     casual                Saturday
1665.7982
## 14                     member                Saturday
934.5948
```

analyze ridership data by type and weekday

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
```

```
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
  ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, number_of_rides)
```

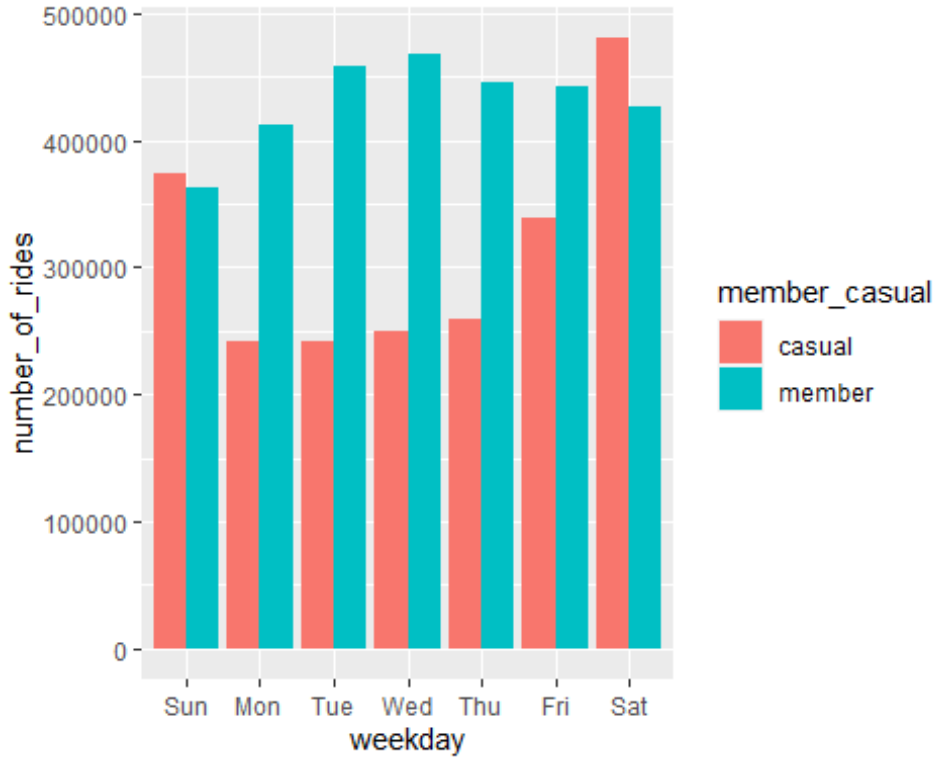
```
## `summarise()` has grouped output by 'member_casual'. You can override
using the
## `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Tue              241495            1372.
##  2 casual        Mon              242240            1526.
##  3 casual        Wed              249076            1328.
##  4 casual        Thu              259671            1342.
##  5 casual        Fri              338314            1446.
##  6 casual        Sun              373162            1764.
##  7 casual        Sat              480544            1666.
##  8 member        Sun              363386             945.
##  9 member        Mon              411969             804.
## 10 member        Sat              426144             935.
## 11 member        Fri              441594             822.
## 12 member        Thu              445121             781.
## 13 member        Tue              458354             780.
## 14 member        Wed              468369             781.
```

Visualize the number of rides by rider type

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  geom_col(position = "dodge")
```

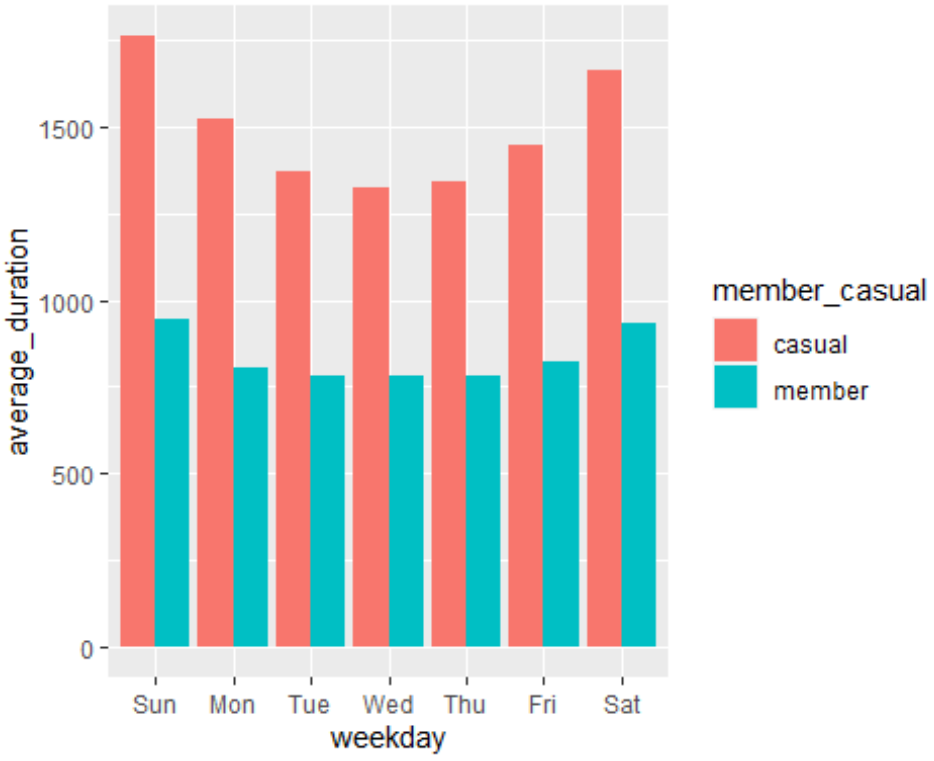
```
## `summarise()` has grouped output by 'member_casual'. You can override
using the
## `.groups` argument.
```



Create a visualization for average duration

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

```
## `summarise()` has grouped output by 'member_casual'. You can override
using the
## `.groups` argument.
```



#==

STEP 5: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

#==

```
counts <- aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)
write.csv(counts, file = 'D:\\DH\\3rd - Semester II\\Google Data
Analytics\\Case study\\Data\\CSV\\avg_ride_length.csv')

ride_counts <- aggregate(all_trips_v2$ride_length ~
all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = length)
write.csv(ride_counts, file = 'D:\\DH\\3rd - Semester II\\Google Data
Analytics\\Case study\\Data\\CSV\\number_of_rides.csv')
```