Bike Sharing

Mizbah

2023-03-13

Ride Sharing Data Analysis

The data is from an open source data platform. It has the bike sharing data of Divvy company for 2nd to 4th quarters of 2019 and the first quarter of 2020.

```
options(warn = -1)
library(tidyverse)
library(lubridate)
library(ggplot2)


setwd("E:/Coursera/Google Data Analytics Professional Certificate/C8 - Google Data Analytics Capstone/Da
getwd()
```

```
## [1] "E:/Coursera/Google Data Analytics Professional Certificate/C8 - Google Data Analytics Capstone/D
```

```
q2_2019 <- read_csv("2019_Q2.csv")
q3_2019 <- read_csv("2019_Q3.csv")
q4_2019 <- read_csv("2019_Q4.csv")
q1_2020 <- read_csv("2020_Q1.csv")
```

Data wrangling and combining data

```
colnames(q3_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q4_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q2_2019)
```

```
##  [1] "01 - Rental Details Rental ID"
##  [2] "01 - Rental Details Local Start Time"
##  [3] "01 - Rental Details Local End Time"
##  [4] "01 - Rental Details Bike ID"
##  [5] "01 - Rental Details Duration In Seconds Uncapped"
##  [6] "03 - Rental Start Station ID"
##  [7] "03 - Rental Start Station Name"
##  [8] "02 - Rental End Station ID"
##  [9] "02 - Rental End Station Name"
## [10] "User Type"
## [11] "Member Gender"
## [12] "05 - Member Details Member Birthday Year"
```

```
colnames(q1_2020)
```

```
##  [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"
```

```
#Let's get the data to uniform format with same column name
q2_2019 <- rename(q2_2019,
                  ride_id="01 - Rental Details Rental ID",
                  rideable_type="01 - Rental Details Bike ID",
                  started_at="01 - Rental Details Local Start Time",
                  ended_at="01 - Rental Details Local End Time",
                  start_station_name = "03 - Rental Start Station Name",
                  start_station_id = "03 - Rental Start Station ID",
                  end_station_name = "02 - Rental End Station Name",
                  end_station_id = "02 - Rental End Station ID",
                  member_casual = "User Type")

q3_2019 <- rename(q3_2019
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = start_time
                  ,ended_at = end_time
                  ,start_station_name = from_station_name
                  ,start_station_id = from_station_id
                  ,end_station_name = to_station_name
                  ,end_station_id = to_station_id
                  ,member_casual = usertype)

q4_2019 <- rename(q4_2019
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = start_time
                  ,ended_at = end_time
                  ,start_station_name = from_station_name
```

```
                  ,start_station_id = from_station_id
                  ,end_station_name = to_station_name
                  ,end_station_id = to_station_id
                  ,member_casual = usertype)

#Inspecting data frames and looking for inconsistencies
str(q1_2020)
```

```
## spc_tbl_ [426,887 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:426887] "EACB19130B0CDA4A" "8FED874C809DC021" "789F3C21E472CA96" "C9A3
##  $ rideable_type     : chr [1:426887] "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:426887], format: "2020-01-21 20:06:59" "2020-01-30 14:22:39" ...
##  $ ended_at          : POSIXct[1:426887], format: "2020-01-21 20:14:30" "2020-01-30 14:26:22" ...
##  $ start_station_name: chr [1:426887] "Western Ave & Leland Ave" "Clark St & Montrose Ave" "Broadway
##  $ start_station_id  : num [1:426887] 239 234 296 51 66 212 96 96 212 38 ...
##  $ end_station_name  : chr [1:426887] "Clark St & Leland Ave" "Southport Ave & Irving Park Rd" "Wilt
##  $ end_station_id    : num [1:426887] 326 318 117 24 212 96 212 212 96 100 ...
##  $ start_lat         : num [1:426887] 42 42 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:426887] -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:426887] 42 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:426887] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:426887] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q4_2019)
```

```
## spc_tbl_ [704,054 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:704054] 25223640 25223641 25223642 25223643 25223644 ...
##  $ started_at        : POSIXct[1:704054], format: "2019-10-01 00:01:39" "2019-10-01 00:02:16" ...
##  $ ended_at          : POSIXct[1:704054], format: "2019-10-01 00:17:20" "2019-10-01 00:06:34" ...
##  $ rideable_type     : num [1:704054] 2215 6328 3003 3275 5294 ...
##  $ tripduration      : num [1:704054] 940 258 850 2350 1867 ...
##  $ start_station_id  : num [1:704054] 20 19 84 313 210 156 84 156 156 336 ...
##  $ start_station_name: chr [1:704054] "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St
##  $ end_station_id    : num [1:704054] 309 241 199 290 382 226 142 463 463 336 ...
##  $ end_station_name  : chr [1:704054] "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave &
```

```
##  $ member_casual     : chr [1:704054] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:704054] "Male" "Male" "Female" "Male" ...
##  $ birthyear         : num [1:704054] 1987 1998 1991 1990 1987 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2019)
```

```
## spc_tbl_ [1,640,718 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:1640718] 23479388 23479389 23479390 23479391 23479392 ...
##  $ started_at        : POSIXct[1:1640718], format: "2019-07-01 00:00:27" "2019-07-01 00:01:16" ...
##  $ ended_at          : POSIXct[1:1640718], format: "2019-07-01 00:20:41" "2019-07-01 00:18:44" ...
##  $ rideable_type     : num [1:1640718] 3591 5353 6180 5540 6014 ...
##  $ tripduration      : num [1:1640718] 1214 1048 1554 1503 1213 ...
##  $ start_station_id  : num [1:1640718] 117 381 313 313 168 300 168 313 43 43 ...
##  $ start_station_name: chr [1:1640718] "Wilton Ave & Belmont Ave" "Western Ave & Monroe St" "Lakevie
##  $ end_station_id    : num [1:1640718] 497 203 144 144 62 232 62 144 195 195 ...
##  $ end_station_name  : chr [1:1640718] "Kimball Ave & Belmont Ave" "Western Ave & 21st St" "Larrabee
##  $ member_casual     : chr [1:1640718] "Subscriber" "Customer" "Customer" "Customer" ...
##  $ gender            : chr [1:1640718] "Male" NA NA NA ...
##  $ birthyear         : num [1:1640718] 1992 NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2019)
```

```
## spc_tbl_ [1,108,163 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id                                          : num [1:1108163] 22178529 22178530 22178531 2217
##  $ started_at                                       : POSIXct[1:1108163], format: "2019-04-01 00:02:2
##  $ ended_at                                         : POSIXct[1:1108163], format: "2019-04-01 00:09:4
##  $ rideable_type                                    : num [1:1108163] 6251 6226 5649 4151 3270 ...
##  $ 01 - Rental Details Duration In Seconds Uncapped: num [1:1108163] 446 1048 252 357 1007 ...
##  $ start_station_id                                 : num [1:1108163] 81 317 283 26 202 420 503 260 2
##  $ start_station_name                               : chr [1:1108163] "Daley Center Plaza" "Wood St &
##  $ end_station_id                                   : num [1:1108163] 56 59 174 133 129 426 500 499 2
##  $ end_station_name                                 : chr [1:1108163] "Desplaines St & Kinzie St" "Wal
##  $ member_casual                                    : chr [1:1108163] "Subscriber" "Subscriber" "Subs
##  $ Member Gender                                    : chr [1:1108163] "Male" "Female" "Male" "Male" .
##  $ 05 - Member Details Member Birthday Year         : num [1:1108163] 1975 1984 1990 1993 1992 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `01 - Rental Details Rental ID` = col_double(),
##   ..   `01 - Rental Details Local Start Time` = col_datetime(format = ""),
##   ..   `01 - Rental Details Local End Time` = col_datetime(format = ""),
##   ..   `01 - Rental Details Bike ID` = col_double(),
##   ..   `01 - Rental Details Duration In Seconds Uncapped` = col_number(),
##   ..   `03 - Rental Start Station ID` = col_double(),
##   ..   `03 - Rental Start Station Name` = col_character(),
##   ..   `02 - Rental End Station ID` = col_double(),
##   ..   `02 - Rental End Station Name` = col_character(),
##   ..   `User Type` = col_character(),
##   ..   `Member Gender` = col_character(),
##   ..   `05 - Member Details Member Birthday Year` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#converting ride_id and rideable_type to character
q4_2019 <-  mutate(q4_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

q3_2019 <- mutate(q3_2019, ride_id = as.character(ride_id),
                  rideable_type=as.character(ride_id))

q2_2019 <-  mutate(q2_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

#Joining data
all_trips <- bind_rows(q2_2019, q3_2019, q4_2019, q1_2020)

#Removing lat, long, birthyear, and gender fields as they were dropped in 2020
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, birthyear, gender, "01 - Rental Details Duration In

#Let's go over the joined data for a bird's eye view
colnames(all_trips)
```

```
## [1] "ride_id"            "started_at"         "ended_at"
```

```
## [4] "rideable_type"      "start_station_id"   "start_station_name"
## [7] "end_station_id"     "end_station_name"   "member_casual"
```

```
nrow(all_trips)
```

```
## [1] 3879822
```

```
head(all_trips)
```

```
## # A tibble: 6 x 9
##   ride_id  started_at          ended_at            rideable_type start~1 start~2
##   <chr>    <dttm>              <dttm>              <chr>           <dbl> <chr>
## 1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48 6251               81 Daley ~
## 2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30 6226              317 Wood S~
## 3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19 5649              283 LaSall~
## 4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58 4151               26 McClur~
## 5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13 3270              202 Halste~
## 6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56 3123              420 Ellis ~
## # ... with 3 more variables: end_station_id <dbl>, end_station_name <chr>,
## #   member_casual <chr>, and abbreviated variable names 1: start_station_id,
## #   2: start_station_name
```

```
str(all_trips)
```

```
## tibble [3,879,822 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" ...
##  $ ended_at          : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" ...
##  $ rideable_type     : chr [1:3879822] "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jacks
##  $ end_station_id    : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal
##  $ member_casual     : chr [1:3879822] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
```

```
summary(all_trips)
```

```
##    ride_id            started_at
##  Length:3879822     Min.   :2019-04-01 00:02:22.00
##  Class :character   1st Qu.:2019-06-23 07:49:09.25
##  Mode  :character   Median :2019-08-14 17:43:38.00
##                     Mean   :2019-08-26 00:49:59.38
##                     3rd Qu.:2019-10-12 12:10:21.00
##                     Max.   :2020-03-31 23:51:34.00
##
##     ended_at                      rideable_type      start_station_id
##  Min.   :2019-04-01 00:09:48.00   Length:3879822     Min.   :  1.0
##  1st Qu.:2019-06-23 08:20:27.75   Class :character   1st Qu.: 77.0
##  Median :2019-08-14 18:02:04.00   Mode  :character   Median :174.0
##  Mean   :2019-08-26 01:14:37.06                      Mean   :202.9
##  3rd Qu.:2019-10-12 12:36:16.75                      3rd Qu.:291.0
```

```
##  Max.   :2020-05-19 20:10:34.00                     Max.   :675.0
##
##  start_station_name end_station_id  end_station_name   member_casual
##  Length:3879822     Min.   :  1.0   Length:3879822     Length:3879822
##  Class :character   1st Qu.: 77.0   Class :character   Class :character
##  Mode  :character   Median :174.0   Mode  :character   Mode  :character
##                     Mean   :203.8
##                     3rd Qu.:291.0
##                     Max.   :675.0
##                     NA's   :1
```

There are a few problems we will need to fix:

- In the "member_casual" column, there are two names for members ("member" and "Subscriber") and two names for casual riders ("Customer" and "casual"). We will need to consolidate that from four to two labels.
- The data can only be aggregated at the ride-level, which is too granular. We will want to add some additional columns of data – such as day, month, year – that provide additional opportunities to aggregate the data.
- We will want to add a calculated field for length of ride since the 2020Q1 data did not have the "tripduration" column. We will add "ride_length" to the entire dataframe for consistency.
- There are some rides where tripduration shows up as negative, including several hundred rides where Divvy took bikes out of circulation for Quality Control reasons. We will want to delete these rides.

```
# In the "member_casual" column, replace "Subscriber" with "member" and "Customer" with "casual"

all_trips <-  all_trips %>%
  mutate(member_casual = recode(member_casual
                                ,"Subscriber" = "member"
                                ,"Customer" = "casual"))

# Add columns that list the date, month, day, and year of each ride
# This will allow us to aggregate ride data for each month, day, or year ... before completing these op

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 (in seconds)
all_trips$ride_length <- difftime(all_trips$ended_at, all_trips$started_at)

#checking column structure
str(all_trips)
```

```
## tibble [3,879,822 x 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" ...
##  $ ended_at          : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" ...
##  $ rideable_type     : chr [1:3879822] "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jacks
```

```
##  $ end_station_id    : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal
##  $ member_casual     : chr [1:3879822] "member" "member" "member" "member" ...
##  $ date              : Date[1:3879822], format: "2019-04-01" "2019-04-01" ...
##  $ month             : chr [1:3879822] "04" "04" "04" "04" ...
##  $ day               : chr [1:3879822] "01" "01" "01" "01" ...
##  $ year              : chr [1:3879822] "2019" "2019" "2019" "2019" ...
##  $ day_of_week       : chr [1:3879822] "Monday" "Monday" "Monday" "Monday" ...
##  $ ride_length       : 'difftime' num [1:3879822] 446 1048 252 357 ...
##   ..- attr(*, "units")= chr "secs"
```

```
# Convert "ride_length" from Factor to numeric so we can run calculations on the data
is.factor(all_trips$ride_length)
```

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

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

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

The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or ride_length was negative. We need to get rid of this records.

```
all_trips_v2<-all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<0),]
head(all_trips_v2)
```

```
## # A tibble: 6 x 15
##   ride_id  started_at          ended_at            rideable_type start~1 start~2
##   <chr>    <dttm>              <dttm>              <chr>           <dbl> <chr>
## 1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48 6251               81 Daley ~
## 2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30 6226              317 Wood S~
## 3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19 5649              283 LaSall~
## 4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58 4151               26 McClur~
## 5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13 3270              202 Halste~
## 6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56 3123              420 Ellis ~
## # ... with 9 more variables: end_station_id <dbl>, end_station_name <chr>,
## #   member_casual <chr>, date <date>, month <chr>, day <chr>, year <chr>,
## #   day_of_week <chr>, ride_length <dbl>, and abbreviated variable names
## #   1: start_station_id, 2: start_station_name
```

Let's run some descriptive anlaysis.

```
#Ride length
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     412     712    1479    1289 9387024
```

```r
#Comparing members and casual riders
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               3552.7502
## 2                     member                850.0662
```

```r
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                     1546
## 2                     member                      589
```

```r
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                  9387024
## 2                     member                  9056634
```

```r
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                        2
## 2                     member                        1
```

```r
# 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                3773.8351
## 2                      member                   Friday                 824.5305
## 3                      casual                   Monday                3372.2869
## 4                      member                   Monday                 842.5726
## 5                      casual                 Saturday                3331.9138
## 6                      member                 Saturday                 968.9337
## 7                      casual                   Sunday                3581.4054
## 8                      member                   Sunday                 919.9746
## 9                      casual                 Thursday                3682.9847
## 10                     member                 Thursday                 823.9278
## 11                     casual                  Tuesday                3596.3599
## 12                     member                  Tuesday                 826.1427
## 13                     casual                Wednesday                3718.6619
## 14                     member                Wednesday                 823.9996
```

```r
# We can see that the days of the week are out of order. Let's fix that.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "

# Now, let's run 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                   Sunday                3581.4054
## 2                      member                   Sunday                 919.9746
## 3                      casual                   Monday                3372.2869
## 4                      member                   Monday                 842.5726
## 5                      casual                  Tuesday                3596.3599
## 6                      member                  Tuesday                 826.1427
## 7                      casual                Wednesday                3718.6619
## 8                      member                Wednesday                 823.9996
## 9                      casual                 Thursday                3682.9847
## 10                     member                 Thursday                 823.9278
## 11                     casual                   Friday                3773.8351
## 12                     member                   Friday                 824.5305
## 13                     casual                 Saturday                3331.9138
## 14                     member                 Saturday                 968.9337
```

```r
# Analyzing 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, weekday)
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              181293            3581.
##  2 casual        Mon              103296            3372.
##  3 casual        Tue               90510            3596.
##  4 casual        Wed               92457            3719.
##  5 casual        Thu              102679            3683.
##  6 casual        Fri              122404            3774.
##  7 casual        Sat              209543            3332.
##  8 member        Sun              267965             920.
##  9 member        Mon              472196             843.
## 10 member        Tue              508445             826.
## 11 member        Wed              500329             824.
## 12 member        Thu              484177             824.
## 13 member        Fri              452790             825.
## 14 member        Sat              287958             969.
```

```r
# Now, let's run the average ride time by each month for members vs casual users.
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$month, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$month all_trips_v2$ride_length
## 1                      casual                 01                9698.9692
## 2                      member                 01                 668.9423
## 3                      casual                 02                7997.1646
## 4                      member                 02                 768.3972
## 5                      casual                 03                4250.2230
## 6                      member                 03                 860.0644
## 7                      casual                 04                3056.5471
```

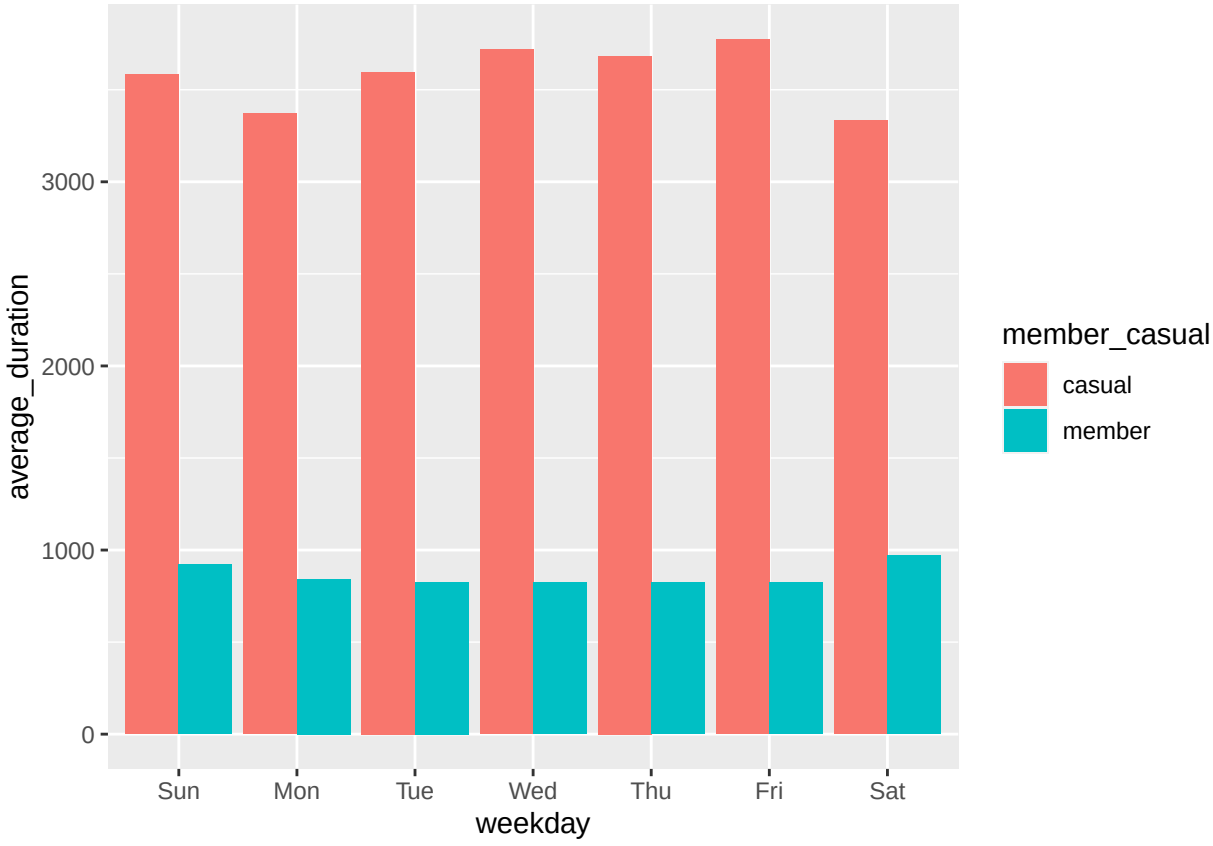
```
## 8                    member              04               810.7675
## 9                    casual              05              3074.3017
## 10                   member              05               830.9938
## 11                   casual              06              2755.2024
## 12                   member              06               872.8465
## 13                   casual              07              3587.0295
## 14                   member              07               986.0000
## 15                   casual              08              4020.4513
## 16                   member              08               971.1324
## 17                   casual              09              3100.0022
## 18                   member              09               847.8389
## 19                   casual              10              3539.5290
## 20                   member              10               781.8533
## 21                   casual              11              4021.9658
## 22                   member              11               745.9243
## 23                   casual              12              3799.9594
## 24                   member              12               684.9098
```

```
# Analyzing ridership data by type and month
all_trips_v2 %>%
  mutate(month=month(started_at, label=TRUE)) %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides=n(), average_duration=mean(ride_length)) %>%
  arrange(member_casual, month)
```

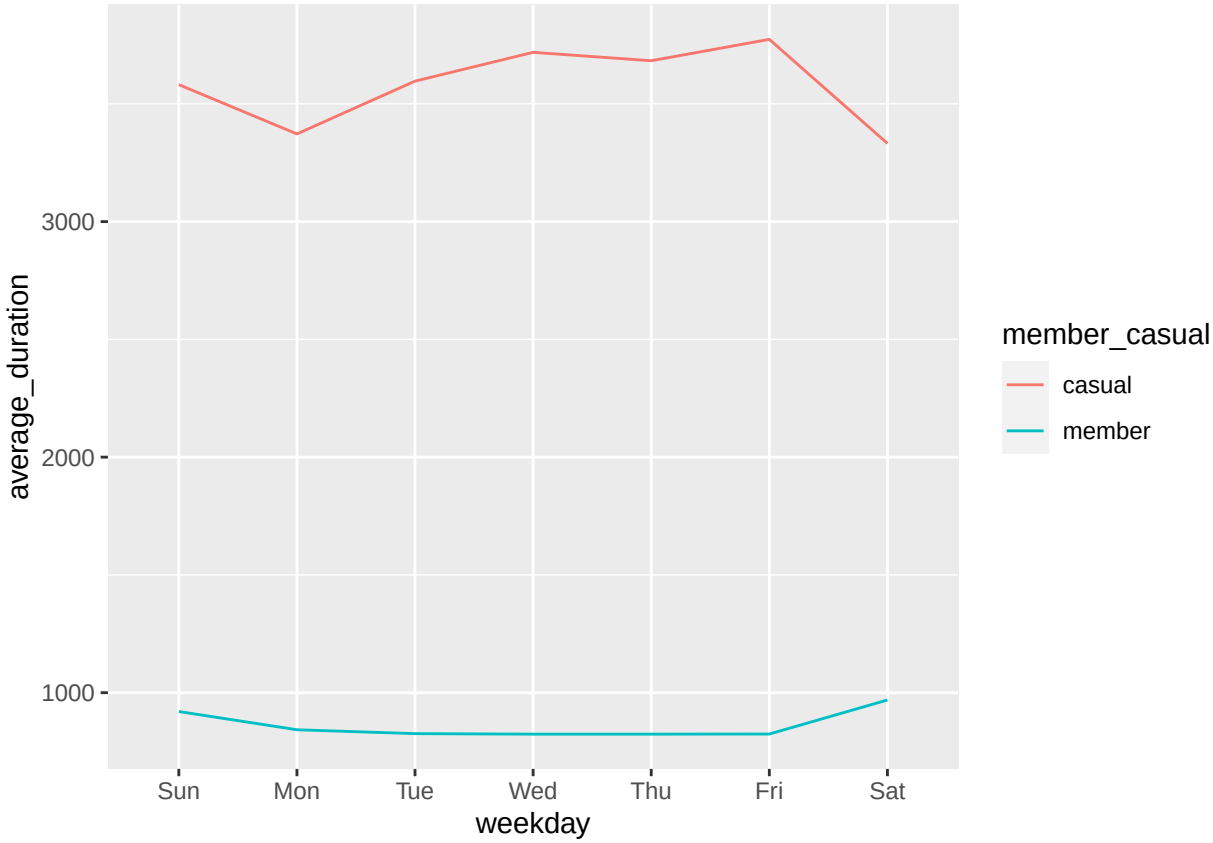
```
## # A tibble: 24 x 4
## # Groups:   member_casual [2]
##    member_casual month number_of_rides average_duration
##    <chr>         <ord>           <int>            <dbl>
##  1 casual        Jan              7785            9699.
##  2 casual        Feb             12314            7997.
##  3 casual        Mar             24615            4250.
##  4 casual        Apr             47744            3057.
##  5 casual        May             81624            3074.
##  6 casual        Jun            130218            2755.
##  7 casual        Jul            175632            3587.
##  8 casual        Aug            186889            4020.
##  9 casual        Sep            129173            3100.
## 10 casual        Oct             71035            3540.
## # ... with 14 more rows
```

Let's do some visualization on our analysis.

```
# Let's visualize the average ride duration by rider type for each day.
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")
```

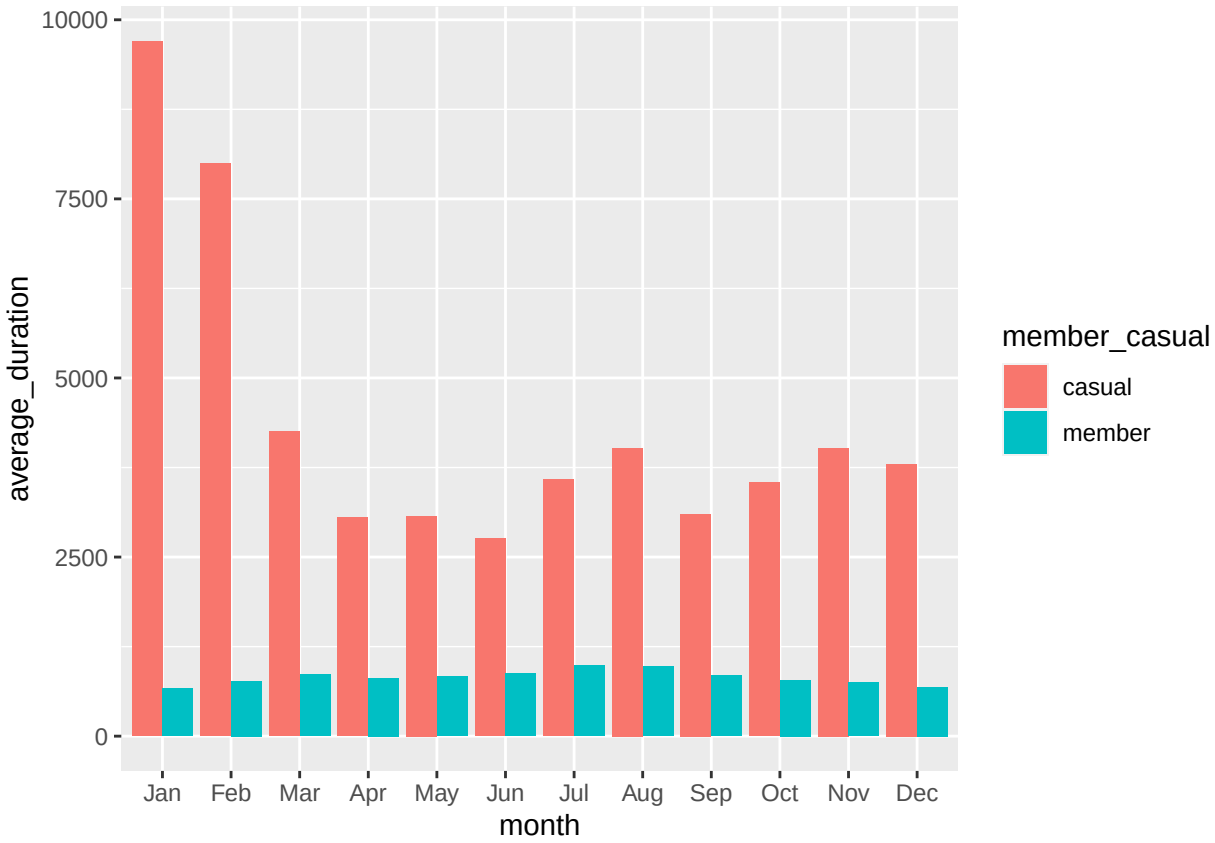



```
#Let's plot it in a line chart.
all_trips_v2 %>%
  mutate(weekday=wday(started_at, label=TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides=n(), average_duration=mean(ride_length)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x=weekday, y=average_duration, color=member_casual, group=member_casual))+
  geom_line()
```

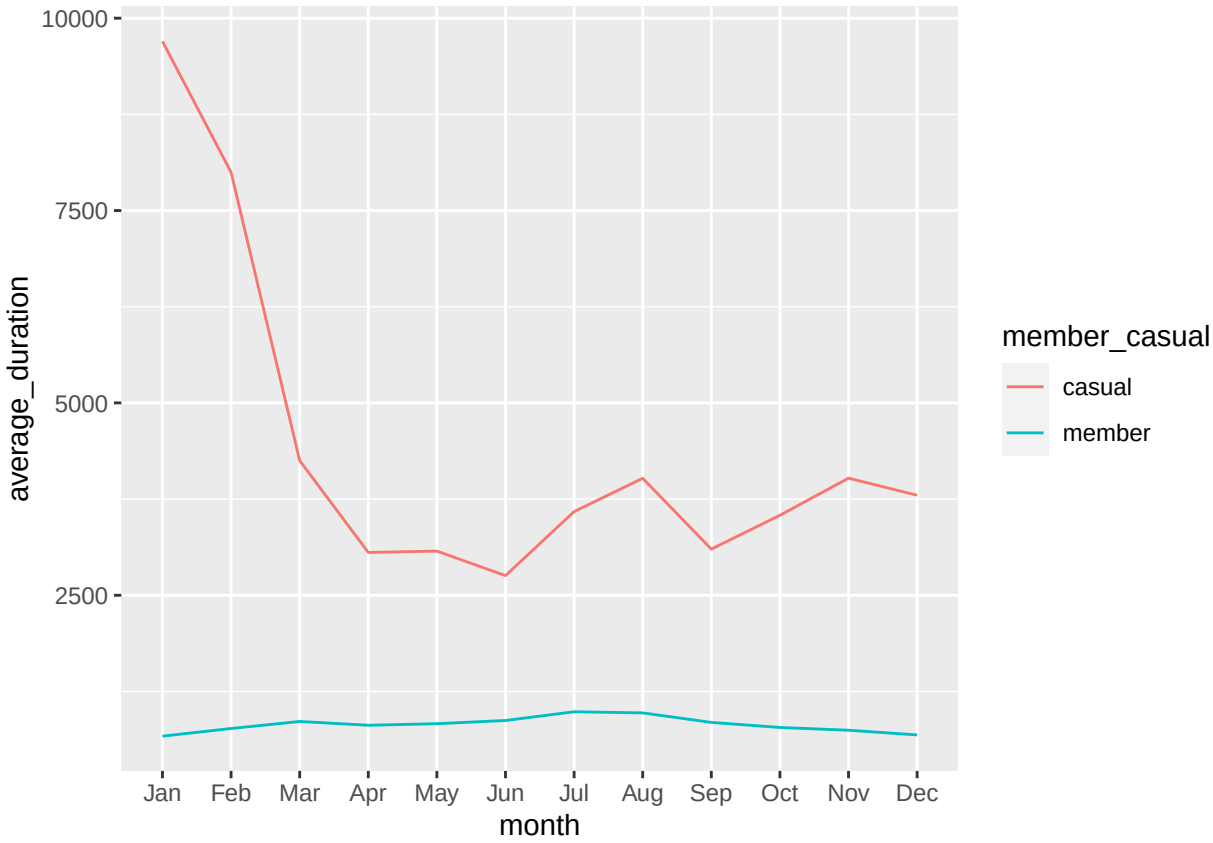



```
# We see a huge difference between the ride duration between casual and subscribed members for each day

# Let's visualize the average ride duration by rider type for each month.
all_trips_v2 %>%
  mutate(month=month(started_at, label=TRUE)) %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides=n(), average_duration=mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=average_duration, fill=member_casual))+
  geom_col(position="dodge")
```




```
#Let's plot it in a line chart as well
all_trips_v2 %>%
  mutate(month=month(started_at, label=TRUE)) %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides=n(), average_duration=mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=average_duration, color=member_casual, group=member_casual))+
  geom_line()
```




```
# A huge difference also exists in monthly ride duration between member types.
```