# Cyclistic

Michelle Salazar

2022-11-17

#====================== # STEP 1: COLLECT DATA #======================

## Install and load packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x 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(ggplot2)
```

#Import .csv files using lapply

```
bike_rides <-list.files(pattern = '*_tripdata.csv')
rides_list <- lapply(bike_rides, read_csv)
```

```
## Rows: 631226 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.
## Rows: 359978 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.
## Rows: 247540 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.
## Rows: 103770 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.
## Rows: 115609 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.
## Rows: 284042 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.
## Rows: 371249 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.
## Rows: 634858 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.
## Rows: 769204 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.
## Rows: 823488 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.
## Rows: 785932 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.
## Rows: 701339 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.
```

```
rides <- bind_rows(rides_list)
```

#Inspect dataframe

```
head(rides)
```

```
## # A tibble: 6 x 13
##   ride_id        ridea~1 started_at          ended_at            start~2 start~3
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 620BC6107255B~ electr~ 2021-10-22 12:46:42 2021-10-22 12:49:50 Kingsb~ KA1503~
## 2 4471C70731AB2~ electr~ 2021-10-21 09:12:37 2021-10-21 09:14:14 <NA>    <NA>
## 3 26CA69D43D15E~ electr~ 2021-10-16 16:28:39 2021-10-16 16:36:26 <NA>    <NA>
## 4 362947F0437E1~ electr~ 2021-10-16 16:17:48 2021-10-16 16:19:03 <NA>    <NA>
## 5 BB731DE2F2EC5~ electr~ 2021-10-20 23:17:54 2021-10-20 23:26:10 <NA>    <NA>
## 6 7176307BBC097~ electr~ 2021-10-21 16:57:37 2021-10-21 17:11:58 <NA>    <NA>
## # ... with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names 1: rideable_type,
## #   2: start_station_name, 3: start_station_id
```

```
colnames(rides)
```

```
##  [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"
```

```
str(rides)
```

```
## spc_tbl_ [5,828,235 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5828235] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362
##  $ rideable_type     : chr [1:5828235] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:5828235], format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct[1:5828235], format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr [1:5828235] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:5828235] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:5828235] NA NA NA NA ...
##  $ end_station_id    : chr [1:5828235] NA NA NA NA ...
##  $ start_lat         : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5828235] "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>
```

#====================== # STEP 2: CLEAN DATA #======================

## Remove columns I do not need

```
rides <- subset(rides, select = -c(start_station_id, end_station_id, start_station_name, end_station_na
```

## Add columns for date, month, day and year of each ride

```
rides$ride_month <- format(as.Date(rides$started_at), "%m")
rides$ride_day <- format(as.Date(rides$started_at), "%d")
rides$ride_year <- format(as.Date(rides$started_at), "%Y")
rides$ride_day_of_week <- format(as.Date(rides$started_at), "%A")
```

## Create ride duration column

```
#rides$ride_length <- difftime(rides$ended_at, rides$started_at, units = 'secs')
rides$ride_length <- as.numeric(difftime(rides$ended_at, rides$started_at))
```

## Remove bad data

```
rides_v2 <- subset(rides, ride_length > 0)
```

#====================== # STEP 3: ANALYSIS #======================

## Get the mean, median, max and min ride times

```
summary(rides_v2$ride_length) #all figures in seconds
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     356     629    1176    1131 2442301
```

## Compare members and casual users

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual, FUN = mean)
```

```
##   rides_v2$member_casual rides_v2$ride_length
## 1                 casual            1761.8174
## 2                 member             766.1685
```

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual, FUN = median)
```

```
##   rides_v2$member_casual rides_v2$ride_length
## 1                 casual                  807
## 2                 member                  533
```

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual, FUN = max)
```

```
##   rides_v2$member_casual rides_v2$ride_length
## 1                 casual              2442301
## 2                 member                93594
```

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual, FUN = min)
```

```
##   rides_v2$member_casual rides_v2$ride_length
## 1                 casual                    1
## 2                 member                    1
```

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

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual + rides_v2$ride_day_of_week, FUN = mean)
```

```
##    rides_v2$member_casual rides_v2$ride_day_of_week rides_v2$ride_length
## 1                  casual                   Friday            1680.8608
## 2                  member                   Friday             751.7498
## 3                  casual                   Monday            1783.8471
## 4                  member                   Monday             739.7066
## 5                  casual                 Saturday            1962.7752
## 6                  member                 Saturday             855.8823
## 7                  casual                   Sunday            2062.0366
## 8                  member                   Sunday             852.9411
## 9                  casual                 Thursday            1540.9259
## 10                 member                 Thursday             737.6950
## 11                 casual                  Tuesday            1548.6993
## 12                 member                  Tuesday             729.8295
## 13                 casual                Wednesday            1502.1538
## 14                 member                Wednesday             727.4074
```

## Fix order of the days of the week

```
rides_v2$ride_day_of_week <- ordered(rides_v2$ride_day_of_week, levels=c("Sunday", "Monday", "Tuesday",
```

## Average ride time by each day for members vs. casual users

```
aggregate(rides_v2$ride_length ~ rides_v2$member_casual + rides_v2$ride_day_of_week, FUN = mean)
```

```
##    rides_v2$member_casual rides_v2$ride_day_of_week rides_v2$ride_length
## 1                  casual                    Sunday            2062.0366
## 2                  member                    Sunday             852.9411
## 3                  casual                    Monday            1783.8471
## 4                  member                    Monday             739.7066
## 5                  casual                   Tuesday            1548.6993
## 6                  member                   Tuesday             729.8295
## 7                  casual                 Wednesday            1502.1538
## 8                  member                 Wednesday             727.4074
## 9                  casual                  Thursday            1540.9259
## 10                 member                  Thursday             737.6950
## 11                 casual                    Friday            1680.8608
## 12                 member                    Friday             751.7498
## 13                 casual                  Saturday            1962.7752
## 14                 member                  Saturday             855.8823
```

## analyze ridership data by type and weekday

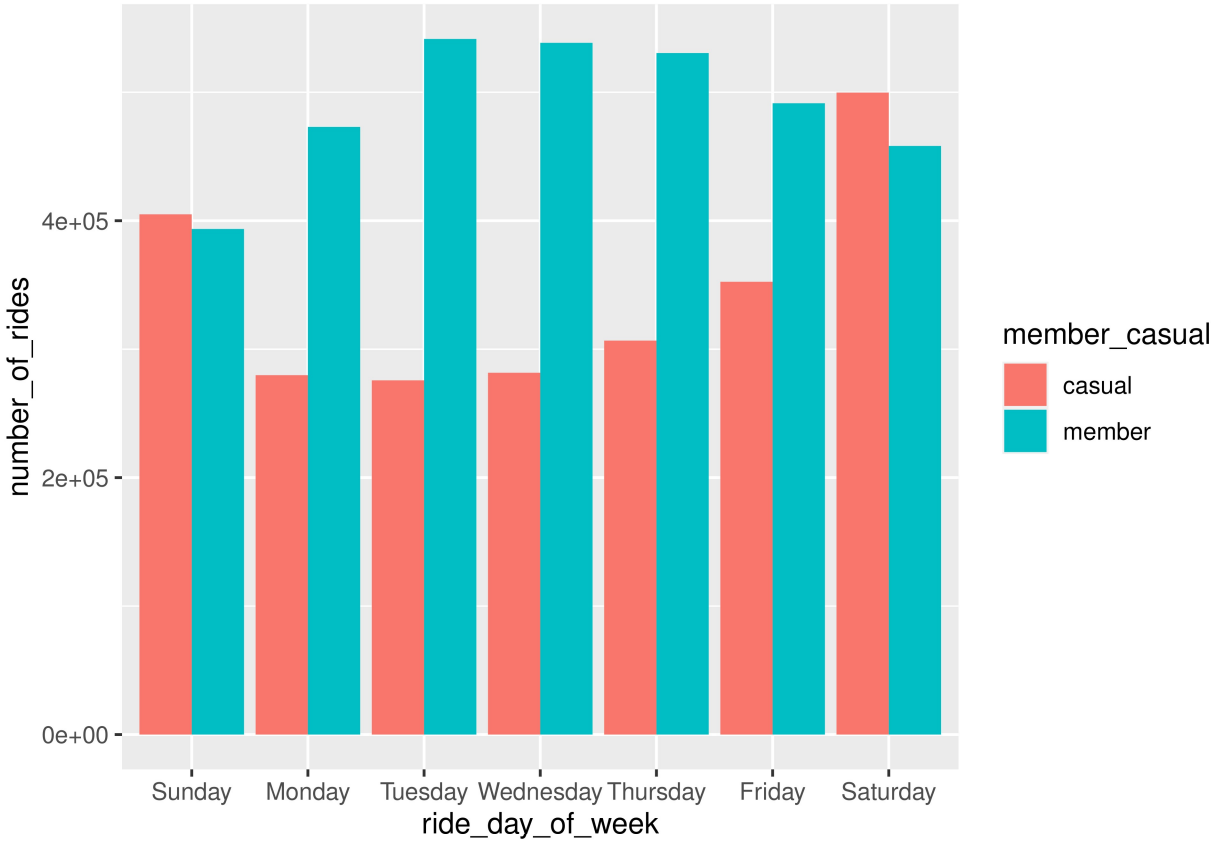
```
rides_v2 %>%
 # mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, ride_day_of_week) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n(), average_duration = mean(ride_length), .groups = 'keep') %>%  #calcul
        # calculates the average duration
  arrange(member_casual, ride_day_of_week)                                  # sorts
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual, ride_day_of_week [14]
##    member_casual ride_day_of_week number_of_rides average_duration
##    <chr>         <ord>                     <int>            <dbl>
##  1 casual        Sunday                   404977            2062.
##  2 casual        Monday                   279762            1784.
##  3 casual        Tuesday                  275745            1549.
##  4 casual        Wednesday                281640            1502.
##  5 casual        Thursday                 306662            1541.
##  6 casual        Friday                   352466            1681.
##  7 casual        Saturday                 499739            1963.
##  8 member        Sunday                   393568             853.
##  9 member        Monday                   473027             740.
```

```
## 10 member        Tuesday                541484            730.
## 11 member        Wednesday              538459            727.
## 12 member        Thursday               530510            738.
## 13 member        Friday                 491436            752.
## 14 member        Saturday               458189            856.
```

## Visualize the number of rides by rider type

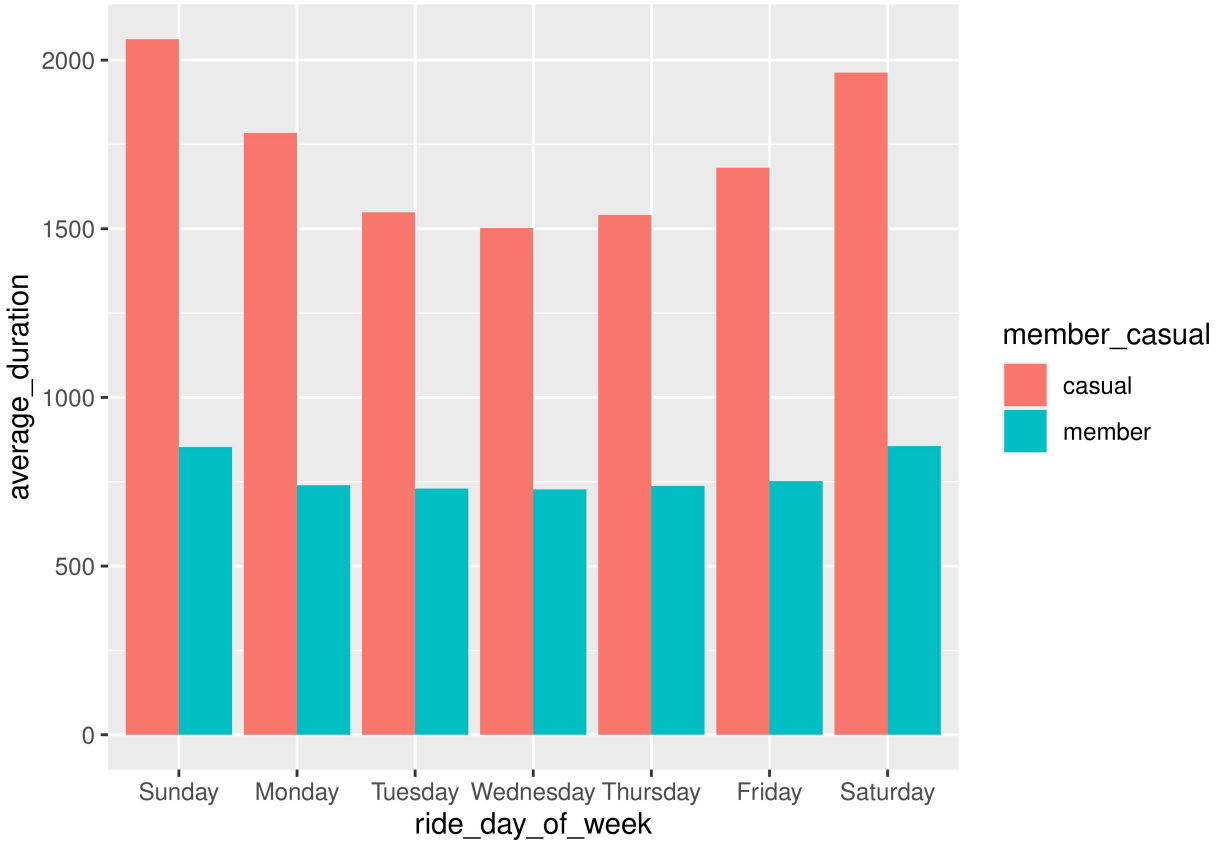
```
rides_v2 %>%
  # mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, ride_day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length), .groups = 'keep') %>%
  arrange(member_casual, ride_day_of_week)  %>%
  ggplot(aes(x = ride_day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```




## Visualization for average duration

```
rides_v2 %>%
  #mutate(weekday = wday(started_at, label = TRUE)) %>%
```

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




#========== # STEP 4: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS #==========

## Create a csv file that we will visualize in Excel or Tableau

```
counts <- aggregate(rides_v2$ride_length ~ rides_v2$member_casual + rides_v2$ride_day_of_week, FUN = mea
write.csv(counts, file = "C:\\Users\\miche\\Dropbox\\PC\\Documents\\R\\Projects\\Cyclistic in R\\avg_ri
```