Cyclist Analysis 2023-02

2023-03-09

Case Study to complete Google Data Analysis Certificate. Investigate how

```r
#import library and files
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors
```

```r
library(lubridate)
library(ggplot2)
getwd()
```

```
## [1] "/Users/chingshawn/Desktop/Data Analysis Project/Google Data analysis Project/DA Project 1"
```

```r
setwd("/Users/chingshawn/Desktop/Data Analysis Project/Google Data analysis Project/DA Project 1")
file202302 <- read_csv("202302-divvy-tripdata.csv")
```

```
## Rows: 190445 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.
```

```r
file202301 <- read_csv("202301-divvy-tripdata.csv")
```

```
## Rows: 190301 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.
```

```r
file202212 <- read_csv("202212-divvy-tripdata.csv")
```

```
## Rows: 181806 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.
```

```r
file202211 <- read_csv("202211-divvy-tripdata.csv")
```

```
## Rows: 337735 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.
```

```
file202210 <- read_csv("202210-divvy-tripdata.csv")
```

```
## Rows: 558685 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.
```

```
file202209 <- read_csv("202209-divvy-tripdata.csv")
```

```
## 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.
```

```
file202208 <- read_csv("202208-divvy-tripdata.csv")
```

```
## 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.
```

```
file202207 <- read_csv("202207-divvy-tripdata.csv")
```

```
## 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.
```

```
file202206 <- read_csv("202206-divvy-tripdata.csv")
```

```
## 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.
```

```
file202205 <- read_csv("202205-divvy-tripdata.csv")
```

```
## 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.
```

```
file202204 <- read_csv("202204-divvy-tripdata.csv")
```

```
## 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.
```

```
file202203 <- read_csv("202203-divvy-tripdata.csv")
```

```
## 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.
```

```
all_trip <- bind_rows(file202302,file202301,file202212,file202211,file202210,file202209,file202208,file202207,file202206,file202205,file202204,file202203)
```

```
colnames(all_trip)
```

```
##  [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"
```

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

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

```
all_trip$ride_length <- difftime(all_trip$ended_at,all_trip$started_at)
all_trip$ride_length <- as.numeric(as.character(all_trip$ride_length))
```

```
all_trip_v2 <- all_trip[!(all_trip$start_station_name == "HQ QR" | all_trip$ride_length < 0),]
#remove start_station_name == "HQ QR" or ride_length < 0 in rows
all_trip_v2 <- na.omit(all_trip_v2)
```

```
summary(all_trip_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     358     628    1013    1128 2061244
```

```
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual, FUN = mean)
```

```
##   all_trip_v2$member_casual all_trip_v2$ride_length
## 1                    casual               1421.7779
## 2                    member                741.6402
```

```
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual, FUN = median)
```

```
##   all_trip_v2$member_casual all_trip_v2$ride_length
## 1                    casual                     821
## 2                    member                     533
```

```
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual, FUN = max)
```

```
##   all_trip_v2$member_casual all_trip_v2$ride_length
## 1                    casual                 2061244
## 2                    member                   89872
```

```
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual, FUN = min)
```

```
##   all_trip_v2$member_casual all_trip_v2$ride_length
## 1                    casual                       0
## 2                    member                       0
```

```
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual + all_trip_v2$day_of_week, FUN = mean)
```

```
##    all_trip_v2$member_casual all_trip_v2$day_of_week all_trip_v2$ride_length
## 1                     casual                  Friday               1329.3054
## 2                     member                  Friday                727.6553
## 3                     casual                  Monday               1466.9029
## 4                     member                  Monday                715.6503
## 5                     casual                Saturday               1593.5945
## 6                     member                Saturday                836.0621
## 7                     casual                  Sunday               1617.6841
## 8                     member                  Sunday                827.1996
## 9                     casual                Thursday               1262.9019
## 10                    member                Thursday                716.3709
## 11                    casual                 Tuesday               1263.8653
## 12                    member                 Tuesday                700.2551
## 13                    casual               Wednesday               1221.2367
## 14                    member               Wednesday                705.9036
```

```
all_trip_v2$day_of_week <- ordered(all_trip_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", 
"Thursday", "Friday", "Saturday"))
aggregate(all_trip_v2$ride_length ~ all_trip_v2$member_casual + all_trip_v2$day_of_week, FUN = mean)
```

```
##    all_trip_v2$member_casual all_trip_v2$day_of_week all_trip_v2$ride_length
## 1                     casual                  Sunday               1617.6841
## 2                     member                  Sunday                827.1996
## 3                     casual                  Monday               1466.9029
## 4                     member                  Monday                715.6503
## 5                     casual                 Tuesday               1263.8653
## 6                     member                 Tuesday                700.2551
## 7                     casual               Wednesday               1221.2367
## 8                     member               Wednesday                705.9036
## 9                     casual                Thursday               1262.9019
## 10                    member                Thursday                716.3709
## 11                    casual                  Friday               1329.3054
## 12                    member                  Friday                727.6553
## 13                    casual                Saturday               1593.5945
## 14                    member                Saturday                836.0621
```

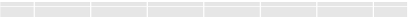
```
all_trip_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()           #calculates the number of rides and average duration
  ,average_duration = mean(ride_length)) %>%        # calculates the average duration
  arrange(member_casual, weekday)
```

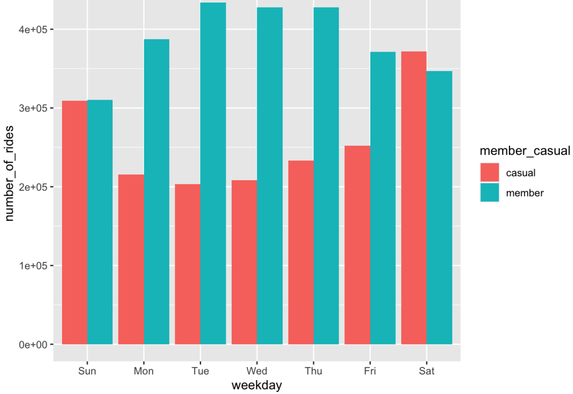
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              309228            1618.
##  2 casual        Mon              215299            1467.
##  3 casual        Tue              203095            1264.
##  4 casual        Wed              208054            1221.
##  5 casual        Thu              233376            1263.
##  6 casual        Fri              251968            1329.
##  7 casual        Sat              371778            1594.
##  8 member        Sun              310117             827.
##  9 member        Mon              387370             716.
## 10 member        Tue              434004             700.
## 11 member        Wed              427602             706.
## 12 member        Thu              427665             716.
## 13 member        Fri              371170             728.
## 14 member        Sat              347103             836.
```

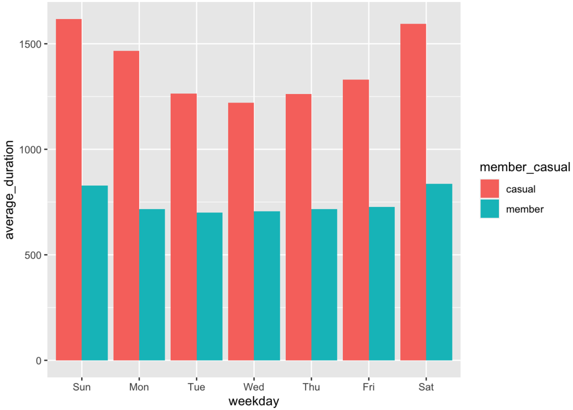
```
all_trip_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n() ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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





```
all_trip_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.
```