capstone cyclistic

Jack Fitzgibbons

3/16/2023

```
##
## Load data sets from cyclistic trip data
##
setwd("C:/Users/Fitzg/OneDrive/Desktop")

df1 <- read.csv("./Google Case Study/202203-tripdata.csv")
df2 <- read.csv("./Google Case Study/202204-tripdata.csv")
df3 <- read.csv("./Google Case Study/202205-tripdata.csv")
df4 <- read.csv("./Google Case Study/202206-tripdata.csv")
df5 <- read.csv("./Google Case Study/202207-tripdata.csv")
df6 <- read.csv("./Google Case Study/202208-tripdata.csv")
df7 <- read.csv("./Google Case Study/202209-tripdata.csv")
df8 <- read.csv("./Google Case Study/202210-tripdata.csv")
df9 <- read.csv("./Google Case Study/202211-tripdata.csv")
df10 <- read.csv("./Google Case Study/202212-tripdata.csv")
df11 <- read.csv("./Google Case Study/202301-tripdata.csv")
df12 <- read.csv("./Google Case Study/202302-tripdata.csv")
```

```
##
## Combine 12 df to 1 df and remove empty cells
##
trip_data <- rbind(df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12)
trip_data <- janitor::remove_empty(trip_data,which = c("cols"))
trip_data <- janitor::remove_empty(trip_data,which = c("rows"))
```

```
##
##inspect the new table
##

colnames(trip_data)  #List of column names
```

```
##  [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(trip_data)  #How many rows are in data frame?
```

```
## [1] 5829084
```

```
dim(trip_data) #Dimensions of the data frame?
```

```
## [1] 5829084      13
```

```
head(trip_data) #See the first 6 rows of data frame
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 47EC0A7F82E65D52  classic_bike 2022-03-21 13:45:01 2022-03-21 13:51:18
## 2 8494861979B0F477 electric_bike 2022-03-16 09:37:16 2022-03-16 09:43:34
## 3 EFE527AF80B66109  classic_bike 2022-03-23 19:52:02 2022-03-23 19:54:48
## 4 9F446FD9DEE3F389  classic_bike 2022-03-01 19:12:26 2022-03-01 19:22:14
## 5 431128AD9AFFEDC0  classic_bike 2022-03-21 18:37:01 2022-03-21 19:19:11
## 6 9AA8A13AF7A85325  classic_bike 2022-03-07 17:10:22 2022-03-07 17:15:04
##                   start_station_name start_station_id
## 1             Wabash Ave & Wacker Pl     TA1307000131
## 2              Michigan Ave & Oak St            13042
## 3              Broadway & Berwyn Ave            13109
## 4             Wabash Ave & Wacker Pl     TA1307000131
## 5 DuSable Lake Shore Dr & North Blvd           LF-005
## 6          Bissell St & Armitage Ave            13059
##                       end_station_name end_station_id start_lat start_lng
## 1             Kingsbury St & Kinzie St   KA1503000043  41.88688 -87.62603
## 2 Orleans St & Chestnut St (NEXT Apts)            620  41.90100 -87.62375
## 3                 Broadway & Ridge Ave          15578  41.97835 -87.65975
## 4           Franklin St & Jackson Blvd   TA1305000025  41.88688 -87.62603
## 5             Loomis St & Jackson Blvd          13206  41.91172 -87.62680
## 6         Southport Ave & Clybourn Ave   TA1309000030  41.91802 -87.65218
##    end_lat   end_lng member_casual
## 1 41.88918 -87.63851        member
## 2 41.89820 -87.63754        member
## 3 41.98404 -87.66027        member
## 4 41.87771 -87.63532        member
## 5 41.87794 -87.66201        member
## 6 41.92077 -87.66371        member
```

```
str(trip_data) #See list of columns and data types
```

```
## ’data.frame’:    5829084 obs. of  13 variables:
##  $ ride_id           : chr  "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F446FD9DEE3F3
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2022-03-21 13:45:01" "2022-03-16 09:37:16" "2022-03-23 19:52:02" "2022-
##  $ ended_at          : chr  "2022-03-21 13:51:18" "2022-03-16 09:43:34" "2022-03-23 19:54:48" "2022-
##  $ start_station_name: chr  "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & Berwyn Ave"
##  $ start_station_id  : chr  "TA1307000131" "13042" "13109" "TA1307000131" ...
##  $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apts)" "Broad
##  $ end_station_id    : chr  "KA1503000043" "620" "15578" "TA1305000025" ...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
##
## add columns that list the date, month, day, and year of each ride
##

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

```
##
##Convert Data/Time stamp to date/time
##

NA_dates1 <- which(is.na(trip_data$started_at))
NA_dates2 <- which(is.na(trip_data$ended_at))
remove(NA_dates1, NA_dates2)

trip_data$started_at <- ymd_hms(trip_data$started_at)
```

```
## Warning: 190445 failed to parse.
```

```
trip_data$ended_at <- ymd_hms(trip_data$ended_at)
```

```
## Warning: 190445 failed to parse.
```

```
##
##add ride_length column in seconds
##

trip_data$ride_length <- difftime(trip_data$ended_at,trip_data$started_at)

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

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

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

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

```
##
##Remove data where ride_length less than 0, and create a new version of the dataframe (v2) since data
##

trip_data_v2 <- trip_data[!(trip_data$start_station_name == "HQ QR" | trip_data$ride_length<0),]
```

```r
# Descriptive analysis on ride_length (all figures in seconds)

summary(trip_data_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##       0     348     615    1165    1105 2483235  190445
```

```r
# Compare members and casual users
aggregate(trip_data_v2$ride_length ~ trip_data_v2$member_casual, FUN = mean)
```

```
##   trip_data_v2$member_casual trip_data_v2$ride_length
## 1                     casual                1743.0612
## 2                     member                 759.8043
```

```r
aggregate(trip_data_v2$ride_length ~ trip_data_v2$member_casual, FUN = median)
```

```
##   trip_data_v2$member_casual trip_data_v2$ride_length
## 1                     casual                      776
## 2                     member                      529
```

```r
aggregate(trip_data_v2$ride_length ~ trip_data_v2$member_casual, FUN = max)
```

```
##   trip_data_v2$member_casual trip_data_v2$ride_length
## 1                     casual                  2483235
## 2                     member                    93594
```

```r
aggregate(trip_data_v2$ride_length ~ trip_data_v2$member_casual, FUN = min)
```

```
##   trip_data_v2$member_casual trip_data_v2$ride_length
## 1                     casual                        0
## 2                     member                        0
```

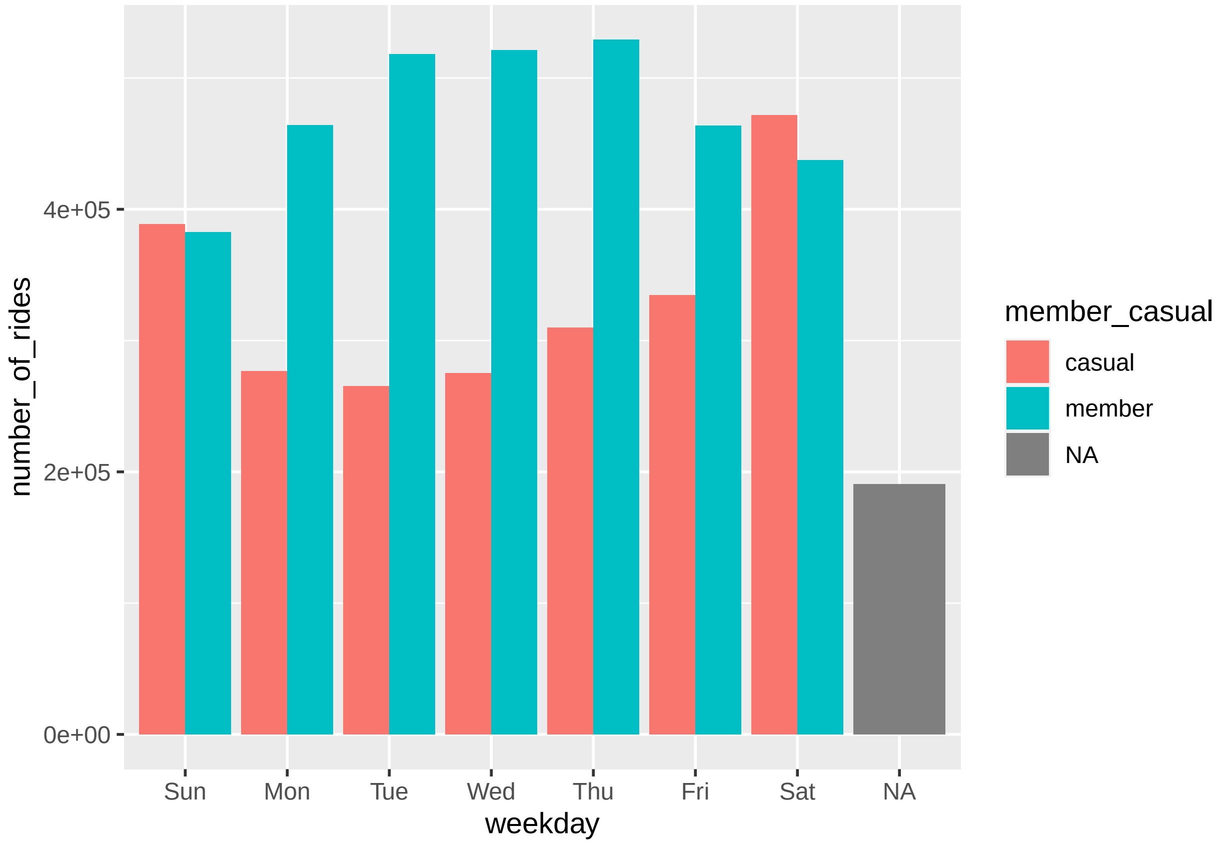
```r
# See the average ride time by each day of the week for members vs casual users
aggregate(trip_data_v2$ride_length ~ trip_data_v2$member_casual + trip_data_v2$day_of_week, FUN = mean)
```

```
##    trip_data_v2$member_casual trip_data_v2$day_of_week trip_data_v2$ride_length
## 1                      casual                   Friday                1680.0409
## 2                      member                   Friday                 748.7728
## 3                      casual                   Monday                1747.1217
## 4                      member                   Monday                 734.0287
## 5                      casual                 Saturday                1951.7751
## 6                      member                 Saturday                 847.5025
## 7                      casual                   Sunday                2046.2966
## 8                      member                   Sunday                 840.2527
## 9                      casual                 Thursday                1523.6812
## 10                     member                 Thursday                 733.9583
## 11                     casual                  Tuesday                1540.2412
## 12                     member                  Tuesday                 722.3249
## 13                     casual                Wednesday                1472.4903
## 14                     member                Wednesday                 723.4344
```

```
#order days of the week
trip_data_v2$day_of_week <- ordered(trip_data_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "
```

```
# visualization for  number of rides by rider type/weekday
trip_data_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")
```

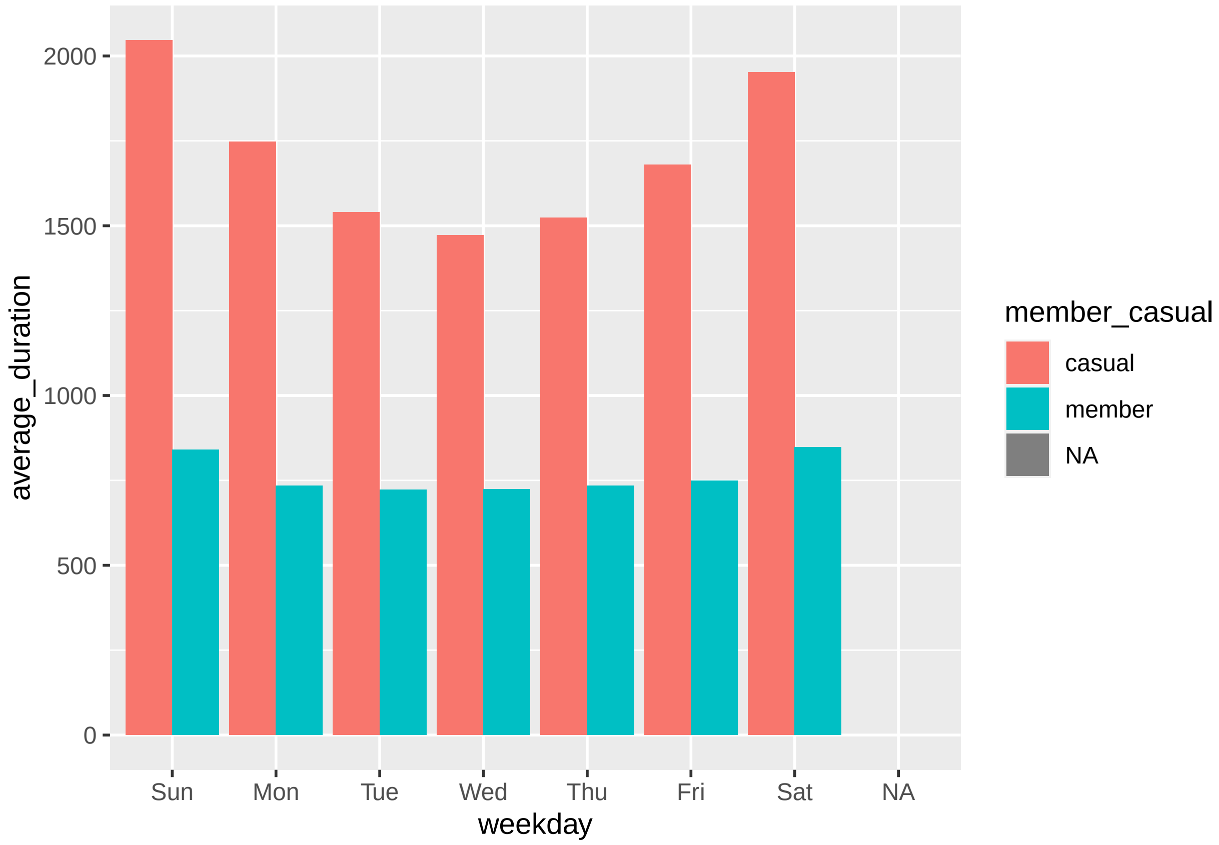
```
## 'summarise()' has grouped output by 'member_casual'. You can override using the
## '.groups' argument.
```



```
#visualization for average duration
trip_data_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.
```

```
## Warning: Removed 1 rows containing missing values (geom_col).
```


average_duration
Sun
Mon
Tue
Wed
Thu
Fri
Sat
NA
weekday
member_casual
casual
member
NA