Cyclistic bike share analysis

Carlos Vasquez

9/20/2020

====================== ### STEP 1: Set up my environment ======================

Notes: setting up my R environment by loading 'tidyverse' and the previous 12 months 'divvy-tripdata' data sets. https://divvy-tripdata.s3.amazonaws.com/index.html

```r
library(tidyverse)
library(janitor)
library(lubridate)
library(scales)
```

```r
q9_2020 <- read_csv("202009-divvy-tripdata.csv")
q10_2020 <- read_csv("202010-divvy-tripdata.csv")
q11_2020 <- read_csv("202011-divvy-tripdata.csv")
q12_2020 <- read_csv("202012-divvy-tripdata.csv")
q1_2021 <- read_csv("202101-divvy-tripdata.csv")
q2_2021 <- read_csv("202102-divvy-tripdata.csv")
q3_2021 <- read_csv("202103-divvy-tripdata.csv")
q4_2021 <- read_csv("202104-divvy-tripdata.csv")
q5_2021 <- read_csv("202105-divvy-tripdata.csv")
q6_2021 <- read_csv("202106-divvy-tripdata.csv")
q7_2021 <- read_csv("202107-divvy-tripdata.csv")
q8_2021 <- read_csv("202108-divvy-tripdata.csv")
```

#====================== ### STEP 2. Make columns consistent and merge them into a single dataframe. #======================

Notes: use colnames function to compare the column names of each data set

```r
colnames(q9_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"
```

```r
colnames(q10_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"
```

```
colnames(q11_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"
```

```
colnames(q12_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"
```

```
colnames(q1_2021)
```

```
##  [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(q2_2021)
```

```
##  [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(q3_2021)
```

```
##  [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(q4_2021)
```

```
##  [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(q5_2021)
```

```
##  [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(q6_2021)
```

```
##  [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(q7_2021)
```

```
##  [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(q8_2021)
```

```
##  [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"
```

Notes: Look for inconsistent data types

```
sapply(q9_2020,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
```

```
##
## $start_station_id
## [1] "numeric"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "numeric"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q10_2020,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "numeric"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "numeric"
##
## $start_lat
## [1] "numeric"
##
```

```
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q11_2020,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "numeric"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "numeric"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q12_2020,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q1_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
```

```
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q2_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q3_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
```

```
##
## $member_casual
## [1] "character"
```

```
sapply(q4_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q5_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
```

```
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q6_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q7_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
```

```
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

```
sapply(q8_2021,class)
```

```
## $ride_id
## [1] "character"
##
## $rideable_type
## [1] "character"
##
## $started_at
## [1] "POSIXct" "POSIXt"
##
## $ended_at
## [1] "POSIXct" "POSIXt"
##
## $start_station_name
## [1] "character"
##
## $start_station_id
## [1] "character"
##
## $end_station_name
## [1] "character"
##
## $end_station_id
## [1] "character"
##
## $start_lat
## [1] "numeric"
##
## $start_lng
## [1] "numeric"
##
## $end_lat
## [1] "numeric"
##
## $end_lng
## [1] "numeric"
##
## $member_casual
## [1] "character"
```

Notes: Mutate data type to make all columns consistent

```
q9_2020 <- mutate(q9_2020, start_station_id = as.character(start_station_id))
q10_2020 <- mutate(q10_2020, start_station_id = as.character(start_station_id))
q11_2020 <- mutate(q11_2020, start_station_id = as.character(start_station_id))
```

```
q9_2020 <- mutate(q9_2020, end_station_id = as.character(end_station_id))
q10_2020 <- mutate(q10_2020, end_station_id = as.character(end_station_id))
q11_2020 <- mutate(q11_2020, end_station_id = as.character(end_station_id))
```

Notes: Merge into one data frame

```
bike_rides <- bind_rows(q9_2020, q10_2020, q11_2020, q12_2020, q1_2021, q2_2021, q3_2021, q4_2021, q5_20
```

#===================== ### STEP 3. Clean up and add data to prepare for analysis #=====================

Notes: Inspect the new data frame

```
dim(bike_rides)
```

```
## [1] 4913072      13
```

```
View(bike_rides)
```

Notes: Remove empty columns and row

```
bike_rides <- janitor::remove_empty(bike_rides, which = c("cols"))
bike_rides <- janitor::remove_empty(bike_rides, which = c("rows"))
dim(bike_rides)
```

```
## [1] 4913072      13
```

Notes: Number of rows remained the same (4,913,072). Preapre data frame for analysis

```
bike_rides$date <- as.Date(bike_rides$started_at)
bike_rides$month <- format(as.Date(bike_rides$date), "%m")
bike_rides$day <- format(as.Date(bike_rides$date), "%d")
bike_rides$year <- format(as.Date(bike_rides$date), "%Y")
bike_rides$day_of_week <- format(as.Date(bike_rides$date), "%A")
bike_rides$minutes <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("min"))
bike_rides$minutes <-  as.numeric(as.character(bike_rides$minutes))
```

Note: Double check newly converted data types

```
is.Date(bike_rides$date)
```

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

```
is.numeric(bike_rides$minutes)
```

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

Notes: Organizing my data frame

```
df <- bike_rides %>%
  filter(minutes>0) %>% drop_na() %>%
  select(-c(ride_id, start_station_name, start_station_id,end_station_name,end_station_id,start_lat,sta
```

Notes: New data frame is 4227857 rows 9 variables

```
View(df)
dim(df)
```

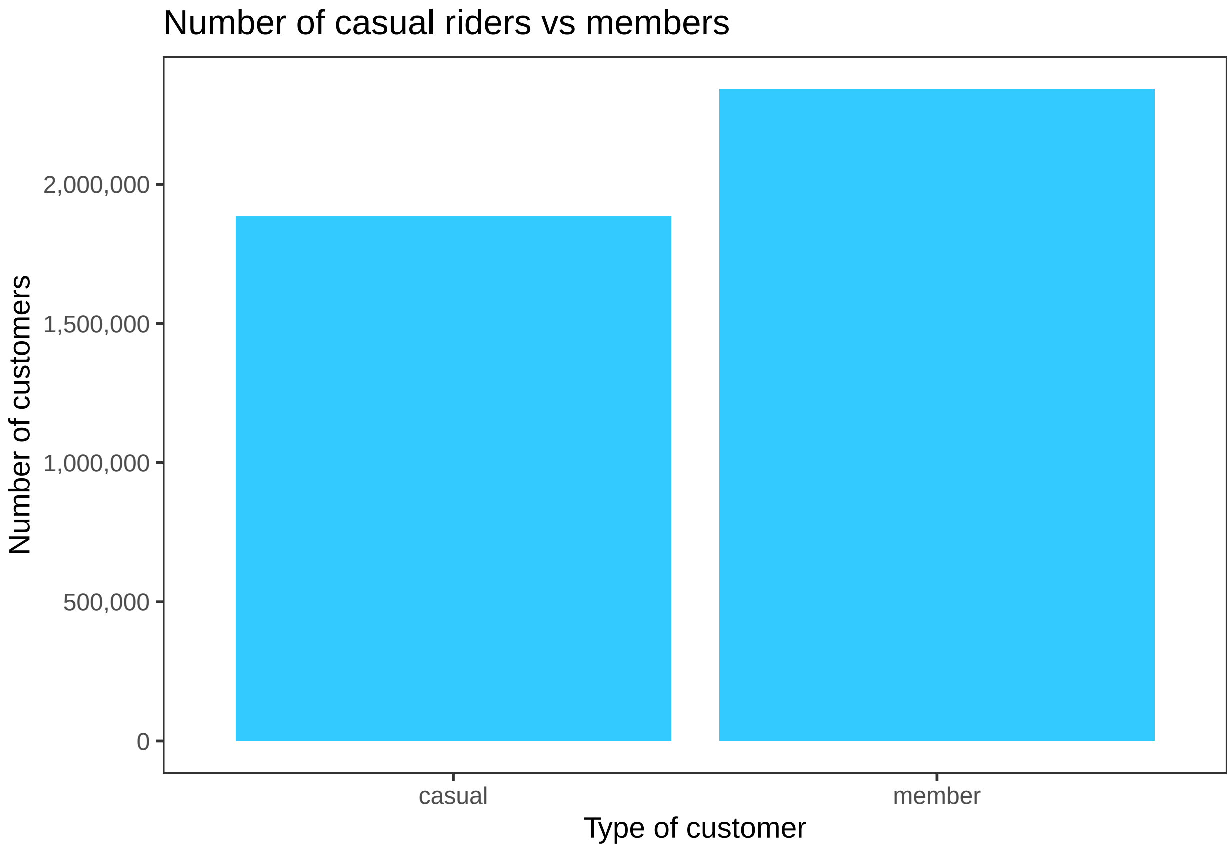
```
## [1] 4227857       9
```

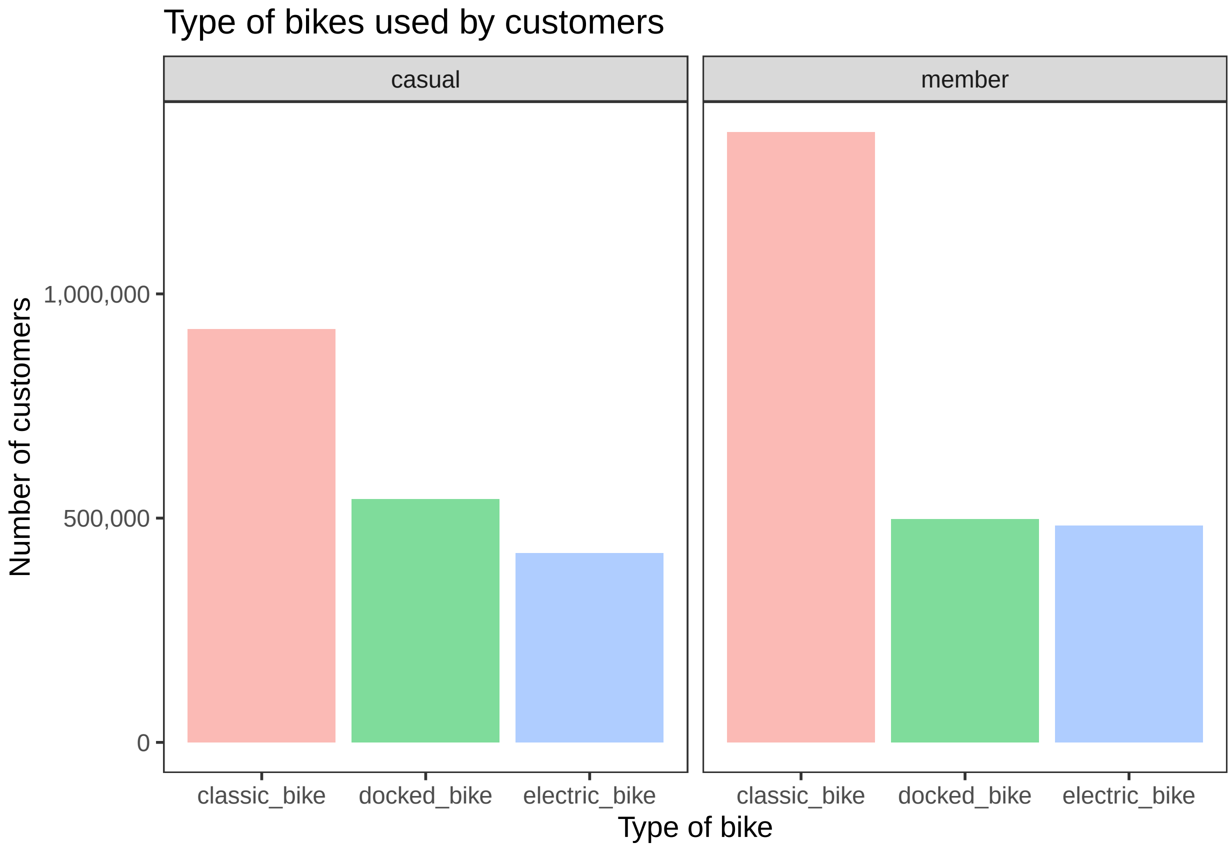
#===================== ### STEP 4. Conduct descriptive analysis #=====================

Business task: How do annual members and casual riders use Cyclistic bikes differently?

Casual = customers who purchase single-ride or full-day passes

Members = customers who purchase annual memberships





Notes: Find the mean, median, max, and min for the ride length (minutes) for customers

```
mean(df$minutes) #average ride (total ride length / rides)
```

```
## [1] 23.32298
```

```
median(df$minutes) #midpoint number in the ascending array of ride lengths
```

```
## [1] 12.93333
```

```
max(df$minutes) #longest ride
```

```
## [1] 55944.15
```

```
min(df$minutes) #shortest ride
```

```
## [1] 0.01666667
```

Notes: Find the mean, median, max, and min for the ride length (minutes) between casual riders and members

```
##   df$member_casual df$minutes
## 1           casual   34.94224
## 2           member   13.96941
##   df$member_casual df$minutes
## 1           casual   17.70000
## 2           member   10.28333
```

```
##   df$member_casual df$minutes
## 1           casual   55944.15
## 2           member   31169.60
```

```
##   df$member_casual df$minutes
## 1           casual 0.01666667
## 2           member 0.01666667
```

Notes: Find the average minutes spend riding bikes by day of the week between casual riders and members

```
##    df$member_casual df$day_of_week df$minutes
## 1            casual         Sunday   40.20812
## 2            member         Sunday   16.04180
## 3            casual         Monday   34.46998
## 4            member         Monday   13.34791
## 5            casual        Tuesday   30.97017
## 6            member        Tuesday   13.17192
## 7            casual      Wednesday   31.18743
## 8            member      Wednesday   13.22661
## 9            casual       Thursday   30.04033
## 10           member       Thursday   13.08012
## 11           casual         Friday   33.46439
## 12           member         Friday   13.68663
## 13           casual       Saturday   37.54052
## 14           member       Saturday   15.55784
```

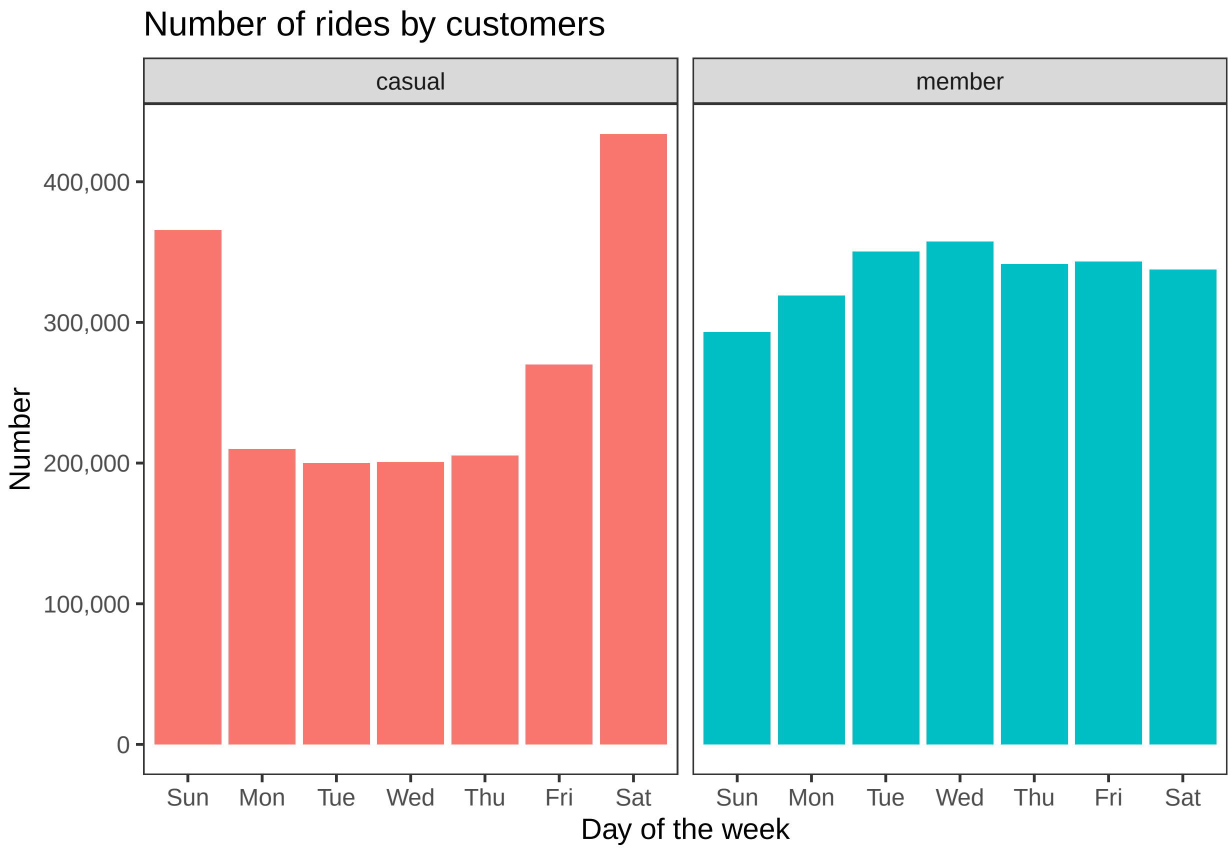
Notes: Find the number of rides per day of the week between casual riders and members

```
## ‘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        Sun              365657             40.2
##  2 casual        Mon              210055             34.5
##  3 casual        Tue              200089             31.0
##  4 casual        Wed              200821             31.2
##  5 casual        Thu              205179             30.0
##  6 casual        Fri              269935             33.5
##  7 casual        Sat              433825             37.5
##  8 member        Sun              293164             16.0
##  9 member        Mon              318952             13.3
## 10 member        Tue              350384             13.2
## 11 member        Wed              357524             13.2
## 12 member        Thu              341329             13.1
## 13 member        Fri              343308             13.7
## 14 member        Sat              337635             15.6
```

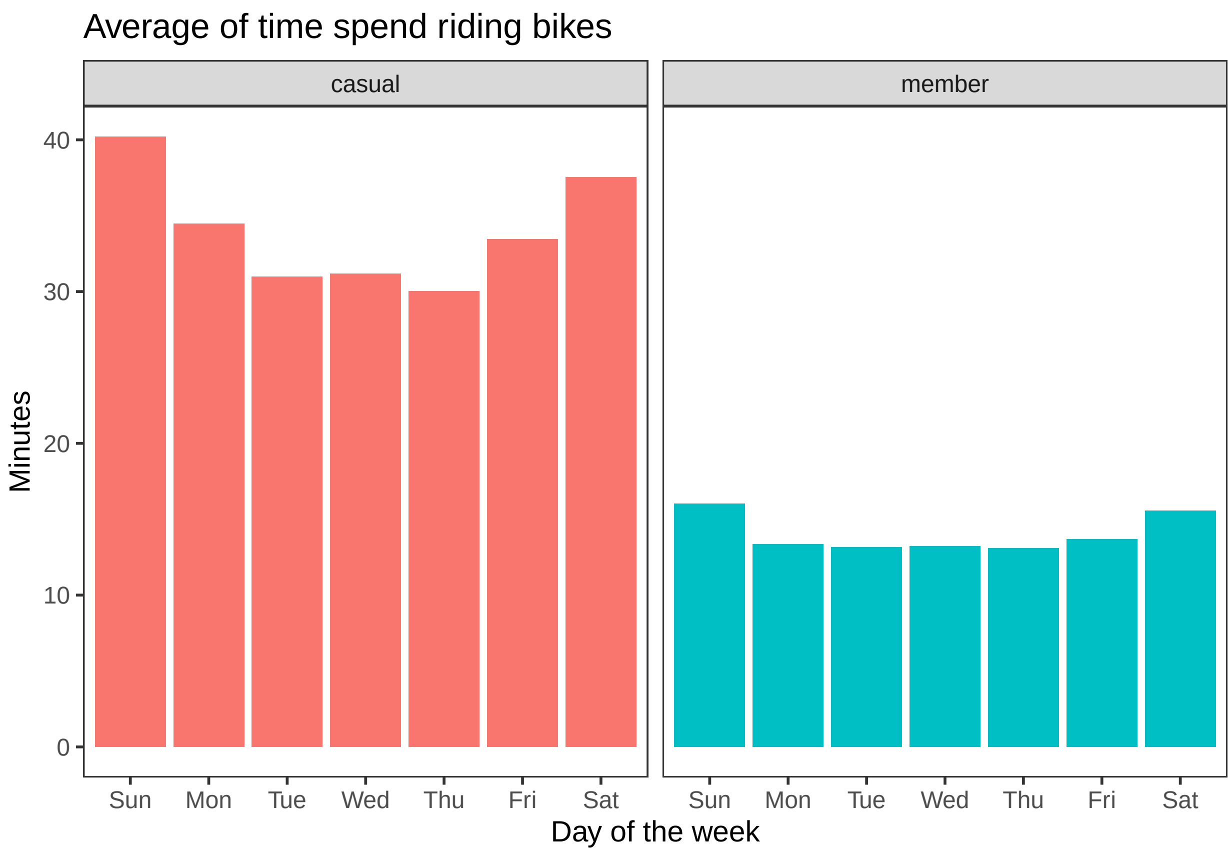
Notes: Visualize the number of rides by rider type

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



Notes: Visualize the average of minutes spend riding bikes

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


Average of time spend riding bikes
casual
member
40
30
20
10
0
Minutes
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day of the week

End