

Cyclistic Case Study Data Prep

Dillon Ngoh

2023-05-16

Packages and setup

I utilized the tidyverse and ggplot2 packages for cleaning, prepping, and visualizing the data. The work directory where the data is located can be set manually in RStudio or using the setwd() function.

```
library(tidyverse)
library(ggplot2)
```

```
#setwd("C:/Users/....")
```

```
#Importing and merging
```

The most recent 12 months data was downloaded from here. They were stored and the number of rows, columns, and column specifications for the first three are shown.

```
may2022 <- 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
## dtm  (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.
```

```
june2022 <- 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
## dtm  (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.
```

```
july2022 <- 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
```

```
## dtm (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.
```

```
aug2022 <- read_csv("202208-divvy-tripdata.csv")
sep2022 <- read_csv("202209-divvy-tripdata.csv")
oct2022 <- read_csv("202210-divvy-tripdata.csv")
nov2022 <- read_csv("202211-divvy-tripdata.csv")
dec2022 <- read_csv("202212-divvy-tripdata.csv")
jan2023 <- read_csv("202301-divvy-tripdata.csv")
feb2023 <- read_csv("202302-divvy-tripdata.csv")
mar2023 <- read_csv("202303-divvy-tripdata.csv")
apr2023 <- read_csv("202304-divvy-tripdata.csv")
```

I used the str() function to check the other tables for variation before merging the tables together. They all had the same column names and data types which made merging them simple.

```
str(may2022)
```

```
## spc_tbl_ [634,858 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id      : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF5
## $ rideable_type : chr [1:634858] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
## $ started_at   : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
## $ ended_at     : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
## $ start_station_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St" "O
## $ start_station_id : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
## $ end_station_name : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood S
## $ end_station_id   : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...
## $ start_lat       : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng       : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...
## $ end_lat         : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng         : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...
## $ member_casual   : chr [1:634858] "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>
```

```
str(jun2022)
```

```
## spc_tbl_ [769,204 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id      : chr [1:769204] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD771D2" "C9C3
```

```
## $ rideable_type      : chr [1:769204] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
## $ started_at        : POSIXct[1:769204], format: "2022-06-30 17:27:53" "2022-06-30 18:39:52" ...
## $ ended_at          : POSIXct[1:769204], format: "2022-06-30 17:35:15" "2022-06-30 18:47:28" ...
## $ start_station_name: chr [1:769204] NA NA NA NA ...
## $ start_station_id  : chr [1:769204] NA NA NA NA ...
## $ end_station_name  : chr [1:769204] NA NA NA NA ...
## $ end_station_id    : chr [1:769204] NA NA NA NA ...
## $ start_lat         : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
## $ start_lng         : num [1:769204] -87.6 -87.6 -87.7 -87.7 -87.6 ...
## $ end_lat           : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
## $ end_lng           : num [1:769204] -87.6 -87.6 -87.6 -87.7 -87.6 ...
## $ member_casual     : chr [1:769204] "casual" "casual" "casual" "casual" ...
## - 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>
```

```
str(jan2023)
```

```
## spc_tbl_ [190,301 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id           : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C907
## $ rideable_type     : chr [1:190301] "electric_bike" "classic_bike" "electric_bike" "classic_bike"
## $ started_at        : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" ...
## $ ended_at          : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" ...
## $ start_station_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western
## $ start_station_id  : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
## $ end_station_name  : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli P
## $ end_station_id    : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...
## $ start_lat         : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
## $ start_lng         : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
## $ end_lat           : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
## $ end_lng           : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
## $ member_casual     : chr [1:190301] "member" "member" "casual" "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>

str(apr2023)

## spc_tbl_ [426,590 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id : chr [1:426590] "8FE8F7D9C10E88C7" "34E4ED3ADF1D821B" "5296BF07A2F77CB5" "4075
## $ rideable_type : chr [1:426590] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
## $ started_at : POSIXct[1:426590], format: "2023-04-02 08:37:28" "2023-04-19 11:29:02" ...
## $ ended_at : POSIXct[1:426590], format: "2023-04-02 08:41:37" "2023-04-19 11:52:12" ...
## $ start_station_name: chr [1:426590] NA NA NA NA ...
## $ start_station_id : chr [1:426590] NA NA NA NA ...
## $ end_station_name : chr [1:426590] NA NA NA NA ...
## $ end_station_id : chr [1:426590] NA NA NA NA ...
## $ start_lat : num [1:426590] 41.8 41.9 41.9 41.9 41.9 ...
## $ start_lng : num [1:426590] -87.6 -87.7 -87.7 -87.7 -87.7 ...
## $ end_lat : num [1:426590] 41.8 41.9 41.9 41.9 41.9 ...
## $ end_lng : num [1:426590] -87.6 -87.7 -87.7 -87.7 -87.6 ...
## $ member_casual : chr [1:426590] "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>

```

After merging all the data together, I checked the table to make sure the number of rows and columns match up and the total number of members vs casuals for reference after manipulating the data.

```

trips <- bind_rows(may2022, jun2022, jul2022, aug2022, sep2022, oct2022, nov2022, dec2022, jan2023, feb2023)
str(trips)

```

```

## spc_tbl_ [5,859,061 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id : chr [1:5859061] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF
## $ rideable_type : chr [1:5859061] "classic_bike" "classic_bike" "classic_bike" "classic_bike" .
## $ started_at : POSIXct[1:5859061], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
## $ ended_at : POSIXct[1:5859061], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
## $ start_station_name: chr [1:5859061] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St"
## $ start_station_id : chr [1:5859061] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...

```

```
## $ end_station_name : chr [1:5859061] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood St" ...
## $ end_station_id   : chr [1:5859061] "TA1309000025" "15534" "13221" "TA1305000030" ...
## $ start_lat        : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng        : num [1:5859061] -87.6 -87.6 -87.6 -87.6 -87.6 ...
## $ end_lat          : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng          : num [1:5859061] -87.6 -87.6 -87.7 -87.6 -87.7 ...
## $ member_casual    : chr [1:5859061] "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>

table(trips$member_casual)

##
##   casual   member
## 2358307 3500754
```

Cleaning and preparing the data

I removed `ride_id`, latitude, and longitude as I was not going to utilize that information in my visualizations. It also saves some time and space when importing and downloading data. If I ever needed or thought it could be important to my analysis I could always return and add them back.

```
trips <- trips %>%
  select(-c(ride_id, start_lat, start_lng, end_lat, end_lng))
```

I wanted to be able to visualize casual and member cyclists differences by day of the week, so I added a day of the week column using the column containing the date of the trip. Using the `table()` function allowed me to quickly see the most rides occurred on Saturdays.

For the sake of visualizing riders by month in RStudio, I also created a column for that.

```
trips$date <- as.Date(trips$started_at)
trips$day_of_week <- format(as.Date(trips$date), "%A")
trips$month <- format(as.Date(trips$date), "%B")

table(trips$day_of_week)
```

```
##
##   Friday   Monday  Saturday   Sunday  Thursday  Tuesday Wednesday
##   847560   760316   922357   790917   879355   817065   841491
```

```
table(trips$month)
```

```
##
##      April      August  December  February  January      July      June      March
##    426590    785932    181806    190445    190301    823488    769204    258678
##      May  November   October  September
##    634858    337735    558685    701339
```

Using the start and end times, I made a column for the length of the trip in minutes rounded to 2 decimals places. The str() function shows the changes made to the dataframe.

```
trips$ride_length <- round(difftime(trips$ended_at, trips$started_at, units='mins'), 2)
trips$ride_length <- as.numeric(as.character(trips$ride_length))
```

```
str(trips)
```

```
## tibble [5,859,061 x 12] (S3: tbl_df/tbl/data.frame)
## $ rideable_type      : chr [1:5859061] "classic_bike" "classic_bike" "classic_bike" "classic_bike" .
## $ started_at         : POSIXct[1:5859061], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
## $ ended_at           : POSIXct[1:5859061], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
## $ start_station_name: chr [1:5859061] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St"
## $ start_station_id   : chr [1:5859061] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
## $ end_station_name   : chr [1:5859061] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood
## $ end_station_id     : chr [1:5859061] "TA1309000025" "15534" "13221" "TA1305000030" ...
## $ member_casual      : chr [1:5859061] "member" "member" "member" "member" ...
## $ date               : Date[1:5859061], format: "2022-05-23" "2022-05-11" ...
## $ day_of_week        : chr [1:5859061] "Monday" "Wednesday" "Thursday" "Tuesday" ...
## $ month              : chr [1:5859061] "May" "May" "May" "May" ...
## $ ride_length        : num [1:5859061] 33.35 37.9 21.83 8.7 5.02 ...
```

Here I wanted to check whether there were trips that had no time, were at 0 or negative minutes, or were greater than 24 hours.

```
filter(trips, is.null(ride_length))
```

```
## # A tibble: 0 x 12
## #   i 12 variables: rideable_type <chr>, started_at <dtm>, ended_at <dtm>,
## #     start_station_name <chr>, start_station_id <chr>, end_station_name <chr>,
## #     end_station_id <chr>, member_casual <chr>, date <date>, day_of_week <chr>,
## #     month <chr>, ride_length <dbl>
```

```
filter(trips, ride_length <= 0)
```

```
## # A tibble: 544 x 12
##   rideable_type started_at      ended_at      start_station_name
##   <chr>         <dtm>         <dtm>         <chr>
## 1 electric_bike 2022-05-18 19:56:48 2022-05-18 19:56:48 Orleans St & Merchandi-
## 2 classic_bike 2022-05-26 16:07:29 2022-05-26 16:07:29 University Ave & 57th ~
## 3 classic_bike 2022-05-25 18:04:53 2022-05-25 18:04:53 Clark St & Leland Ave
## 4 classic_bike 2022-05-29 20:40:52 2022-05-29 20:40:52 Greenview Ave & Divers-
## 5 classic_bike 2022-05-08 17:07:10 2022-05-08 17:07:10 Halsted St & Polk St
## 6 electric_bike 2022-05-29 07:38:27 2022-05-29 07:38:27 Kedzie Ave & Milwaukee-
## 7 electric_bike 2022-05-12 18:32:16 2022-05-12 18:32:16 Damen Ave & Wellington~
## 8 classic_bike 2022-05-14 13:19:59 2022-05-14 13:19:59 Southport Ave & Roscoe-
## 9 electric_bike 2022-05-16 19:06:08 2022-05-16 19:06:08 Wells St & Hubbard St
## 10 electric_bike 2022-05-05 06:02:08 2022-05-05 06:02:08 Clark St & Bryn Mawr A-
## # i 534 more rows
```

```
## # i 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, member_casual <chr>, date <date>, day_of_week <chr>,
## #   month <chr>, ride_length <dbl>
```

```
filter(trips, ride_length >= 1440)
```

```
## # A tibble: 5,345 x 12
##   rideable_type started_at      ended_at      start_station_name
##   <chr>          <dtm>          <dtm>          <chr>
## 1 docked_bike   2022-05-16 02:28:43 2022-05-26 10:37:35 Halsted & 63rd - Kenne-
## 2 docked_bike   2022-05-29 12:19:02 2022-05-30 15:22:48 DuSable Lake Shore Dr ~
## 3 docked_bike   2022-05-30 04:52:35 2022-06-01 04:50:30 Wabash Ave & 16th St
## 4 classic_bike  2022-05-13 19:35:55 2022-05-14 20:35:50 State St & 123rd St
## 5 classic_bike  2022-05-24 21:10:50 2022-05-25 22:10:44 Daley Center Plaza
## 6 classic_bike  2022-05-25 06:28:29 2022-05-26 07:28:20 Southport Ave & Wavela-
## 7 docked_bike   2022-05-05 08:48:19 2022-05-06 09:48:19 Prairie Ave & 43rd St
## 8 classic_bike  2022-05-13 21:40:17 2022-05-14 22:40:13 California Ave & Fletc-
## 9 docked_bike   2022-05-15 04:33:48 2022-05-17 04:38:30 Michigan Ave & Washing-
## 10 classic_bike 2022-05-12 00:13:43 2022-05-13 01:13:39 Dearborn St & Monroe St
## # i 5,335 more rows
## # i 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, member_casual <chr>, date <date>, day_of_week <chr>,
## #   month <chr>, ride_length <dbl>
```

The above showed 544 trips with 0 or negative minutes and 5345 trips longer than 24 hours so those were removed.

```
trips_v2 <- trips[!(trips$ride_length <= 0 | trips$ride_length > 1440),]
str(trips_v2)
```

```
## tibble [5,853,172 x 12] (S3: tbl_df/tbl/data.frame)
## $ rideable_type      : chr [1:5853172] "classic_bike" "classic_bike" "classic_bike" "classic_bike" .
## $ started_at        : POSIXct[1:5853172], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
## $ ended_at          : POSIXct[1:5853172], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
## $ start_station_name: chr [1:5853172] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St"
## $ start_station_id  : chr [1:5853172] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
## $ end_station_name  : chr [1:5853172] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood
## $ end_station_id    : chr [1:5853172] "TA1309000025" "15534" "13221" "TA1305000030" ...
## $ member_casual     : chr [1:5853172] "member" "member" "member" "member" ...
## $ date              : Date[1:5853172], format: "2022-05-23" "2022-05-11" ...
## $ day_of_week       : chr [1:5853172] "Monday" "Wednesday" "Thursday" "Tuesday" ...
## $ month             : chr [1:5853172] "May" "May" "May" "May" ...
## $ ride_length       : num [1:5853172] 33.35 37.9 21.83 8.7 5.02 ...
```

```
table(trips_v2$member_casual)
```

```
##
## casual member
## 2353445 3499727
```

Preliminary analysis and visualization

I wanted to organize the days of the week and month in chronological order so that it appears as Sunday through Saturday for the week. For the month it should appear as January through December. Then I would be able to see the average ride length for casual vs member cyclists.

```
trips_v2$day_of_week <- ordered(trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday",
                                                             "Thursday", "Friday", "Saturday"))

trips_v2$month <- ordered(trips_v2$month, levels=c("January", "February", "March", "April", "May", "June",
                                                  "July", "August", "September", "October"))

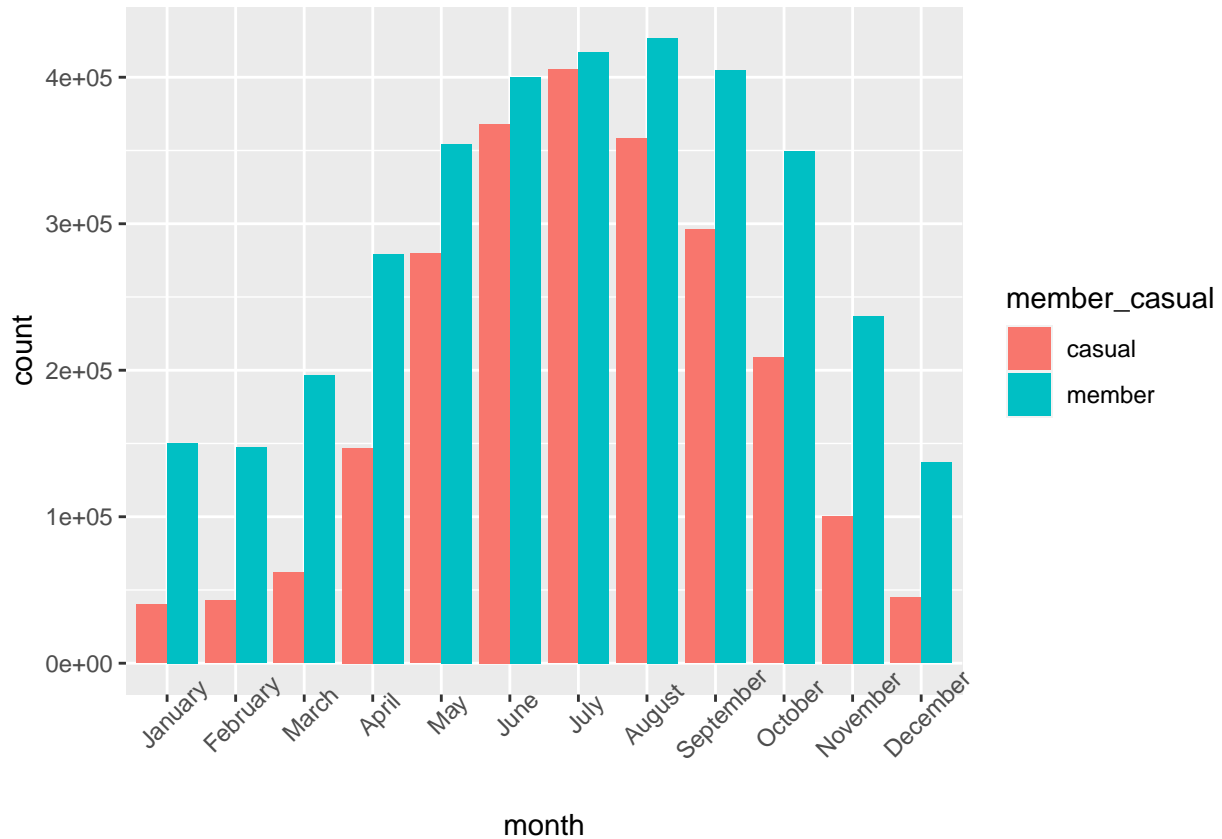
aggregate(trips_v2$ride_length ~ trips_v2$member_casual + trips_v2$day_of_week, FUN = function(x) {round(
  ##      trips_v2$member_casual trips_v2$day_of_week trips_v2$ride_length
  ## 1          casual          Sunday          24.49
  ## 2          member          Sunday          13.46
  ## 3          casual          Monday          21.50
  ## 4          member          Monday          11.67
  ## 5          casual          Tuesday          19.09
  ## 6          member          Tuesday          11.65
  ## 7          casual        Wednesday          18.27
  ## 8          member        Wednesday          11.64
  ## 9          casual        Thursday          19.00
  ## 10         member        Thursday          11.81
  ## 11         casual          Friday          20.26
  ## 12         member          Friday          12.02
  ## 13         casual        Saturday          23.80
  ## 14         member        Saturday          13.58

```

Here you can easily see the number of rides peaks in the summertime where member riders peak in August and casual riders peak in July.

```
ggplot(data = trips_v2, aes(x = month, fill = member_casual))+
  geom_bar(stat="count", position = "dodge") +
  theme(axis.text.x = element_text(angle = 45))

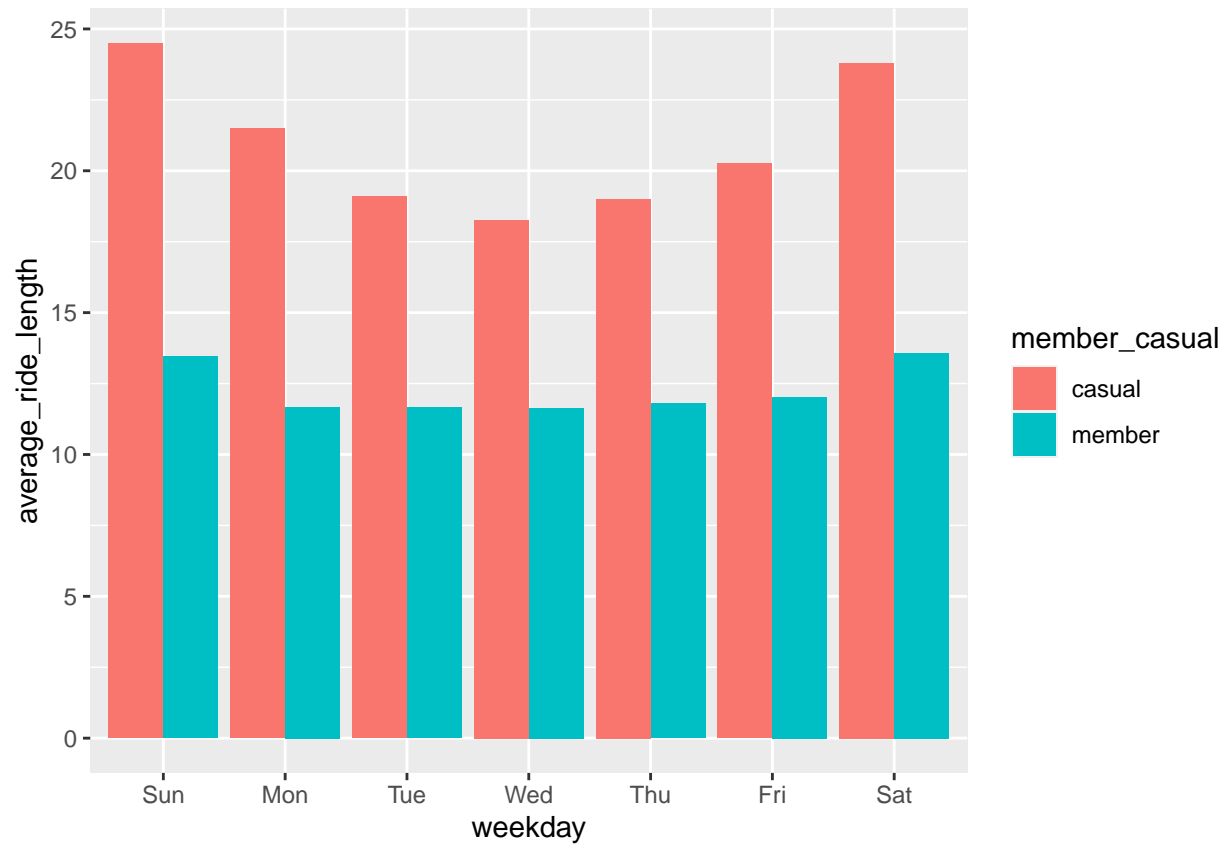
```

You can see average ride length is greater on the weekends for casual riders while it is a couple minutes greater on the weekends for member riders, but otherwise fairly consistent throughout the weekdays for members.

```
trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(), average Ride Length = mean(ride_length)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x = weekday, y = average_ride_length, fill = member_casual)) +
    geom_col(position = "dodge")
```

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



After cleaning and prepping the data, the data is ready to be exported as a csv where I utilized Tableau Public for further analysis.

```
write.csv(trips_v2, 'trips_dataset.csv')
```