Cyclistic

Susan

6/29/2021

Install Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
```

Set Working Directory and Load Data

```
setwd("/Users/Suze/Cyclistic Data <20MB")
q4_2019 <- read_csv("Divvy_Trips/Divvy_Trips_2019_Q4.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   trip_id = col_double(),
##   start_time = col_datetime(format = ""),
##   end_time = col_datetime(format = ""),
##   bikeid = col_double(),
##   tripduration = col_number(),
##   from_station_id = col_double(),
##   from_station_name = col_character(),
##   to_station_id = col_double(),
##   to_station_name = col_character(),
##   usertype = col_character(),
##   gender = col_character(),
##   birthyear = col_double()
## )
```

```
q1_2018 <- read_csv("Divvy_Trips/Divvy_Trips_2018_Q1.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `01 - Rental Details Rental ID` = col_double(),
##   `01 - Rental Details Local Start Time` = col_datetime(format = ""),
##   `01 - Rental Details Local End Time` = col_datetime(format = ""),
##   `01 - Rental Details Bike ID` = col_double(),
##   `01 - Rental Details Duration In Seconds Uncapped` = col_number(),
##   `03 - Rental Start Station ID` = col_double(),
##   `03 - Rental Start Station Name` = col_character(),
##   `02 - Rental End Station ID` = col_double(),
##   `02 - Rental End Station Name` = col_character(),
##   `User Type` = col_character(),
##   `Member Gender` = col_character(),
##   `05 - Member Details Member Birthday Year` = col_double()
## )
```

```
q4_2018 <- read_csv("Divvy_Trips/Divvy_Trips_2018_Q4.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   trip_id = col_double(),
##   start_time = col_datetime(format = ""),
##   end_time = col_datetime(format = ""),
##   bikeid = col_double(),
##   tripduration = col_number(),
##   from_station_id = col_double(),
##   from_station_name = col_character(),
##   to_station_id = col_double(),
##   to_station_name = col_character(),
##   usertype = col_character(),
##   gender = col_character(),
##   birthyear = col_double()
## )
```

```
q1_2020 <- read_csv("Divvy_Trips/Divvy_Trips_2020_Q1.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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_double(),
##   end_station_name = col_character(),
##   end_station_id = col_double(),
##   start_lat = col_double(),
##   start_lng = col_double(),
##   end_lat = col_double(),
##   end_lng = col_double(),
##   member_casual = col_character()
## )
```

Combine Data

```
## Compare the column names of each file
colnames(q4_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q1_2018)
```

```
##  [1] "01 - Rental Details Rental ID"
##  [2] "01 - Rental Details Local Start Time"
##  [3] "01 - Rental Details Local End Time"
##  [4] "01 - Rental Details Bike ID"
##  [5] "01 - Rental Details Duration In Seconds Uncapped"
##  [6] "03 - Rental Start Station ID"
##  [7] "03 - Rental Start Station Name"
##  [8] "02 - Rental End Station ID"
##  [9] "02 - Rental End Station Name"
## [10] "User Type"
## [11] "Member Gender"
## [12] "05 - Member Details Member Birthday Year"
```

```
colnames(q4_2018)
```

```
##  [1] "trip_id"           "start_time"        "end_time"
```

```
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q1_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"
```

```
## Rename the columns to align with most recent naming system
(q4_2019 <- rename(q4_2019
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype))
```

```
## # A tibble: 704,054 x 12
##     ride_id started_at          ended_at            rideable_type tripduration
##       <dbl> <dttm>              <dttm>                      <dbl>        <dbl>
##  1 25223640 2019-10-01 00:01:39 2019-10-01 00:17:20          2215          940
##  2 25223641 2019-10-01 00:02:16 2019-10-01 00:06:34          6328          258
##  3 25223642 2019-10-01 00:04:32 2019-10-01 00:18:43          3003          850
##  4 25223643 2019-10-01 00:04:32 2019-10-01 00:43:43          3275         2350
##  5 25223644 2019-10-01 00:04:34 2019-10-01 00:35:42          5294         1867
##  6 25223645 2019-10-01 00:04:38 2019-10-01 00:10:51          1891          373
##  7 25223646 2019-10-01 00:04:52 2019-10-01 00:22:45          1061         1072
##  8 25223647 2019-10-01 00:04:57 2019-10-01 00:29:16          1274         1458
##  9 25223648 2019-10-01 00:05:20 2019-10-01 00:29:18          6011         1437
## 10 25223649 2019-10-01 00:05:20 2019-10-01 02:23:46          2957         8306
## # ... with 704,044 more rows, and 7 more variables: start_station_id <dbl>,
## #   start_station_name <chr>, end_station_id <dbl>, end_station_name <chr>,
## #   member_casual <chr>, gender <chr>, birthyear <dbl>
```

```
(q1_2018 <- rename(q1_2018
                   ,ride_id = '01 - Rental Details Rental ID'
                   ,rideable_type = '01 - Rental Details Bike ID'
                   ,started_at = '01 - Rental Details Local Start Time'
                   ,ended_at = '01 - Rental Details Local End Time'
                   ,start_station_name = '03 - Rental Start Station Name'
                   ,start_station_id = '03 - Rental Start Station ID'
                   ,end_station_name = '02 - Rental End Station Name'
                   ,end_station_id = '02 - Rental End Station ID'
                   ,member_casual = 'User Type'
                   ))
```

```
## # A tibble: 387,145 x 12
##     ride_id started_at          ended_at            rideable_type
##       <dbl> <dttm>              <dttm>                      <dbl>
##  1 17536702 2018-01-01 00:12:00 2018-01-01 00:17:23          3304
##  2 17536703 2018-01-01 00:41:35 2018-01-01 00:47:52          5367
##  3 17536704 2018-01-01 00:44:46 2018-01-01 01:33:10          4599
##  4 17536705 2018-01-01 00:53:10 2018-01-01 01:05:37          2302
##  5 17536706 2018-01-01 00:53:37 2018-01-01 00:56:40          3696
##  6 17536707 2018-01-01 00:56:15 2018-01-01 01:00:41          6298
##  7 17536708 2018-01-01 00:57:26 2018-01-01 01:02:40          1169
##  8 17536709 2018-01-01 01:00:29 2018-01-01 01:13:43          6351
##  9 17536710 2018-01-01 01:07:12 2018-01-01 01:31:53          1920
## 10 17536711 2018-01-01 01:07:54 2018-01-06 10:04:02          4783
## # ... with 387,135 more rows, and 8 more variables:
## #   01 - Rental Details Duration In Seconds Uncapped <dbl>,
## #   start_station_id <dbl>, start_station_name <chr>, end_station_id <dbl>,
## #   end_station_name <chr>, member_casual <chr>, Member Gender <chr>,
## #   05 - Member Details Member Birthday Year <dbl>
```

```
(q4_2018 <- rename(q4_2018
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype))
```

```
## # A tibble: 642,686 x 12
##     ride_id started_at          ended_at            rideable_type tripduration
##       <dbl> <dttm>              <dttm>                      <dbl>        <dbl>
##  1 20983530 2018-10-01 00:01:17 2018-10-01 00:29:35          4551         1698
##  2 20983531 2018-10-01 00:03:59 2018-10-01 00:10:55           847          416
##  3 20983532 2018-10-01 00:05:14 2018-10-01 00:14:08          6188          534
##  4 20983533 2018-10-01 00:05:48 2018-10-01 00:18:46          6372          778
##  5 20983534 2018-10-01 00:07:29 2018-10-01 00:25:51          1927         1102
##  6 20983535 2018-10-01 00:07:36 2018-10-01 00:11:25          2392          229
##  7 20983536 2018-10-01 00:08:09 2018-10-01 00:58:48           308         3039
##  8 20983537 2018-10-01 00:09:29 2018-10-01 00:15:23          1187          354
##  9 20983538 2018-10-01 00:09:33 2018-10-01 00:12:27          6247          174
## 10 20983539 2018-10-01 00:09:44 2018-10-01 00:21:06          3083          682
## # ... with 642,676 more rows, and 7 more variables: start_station_id <dbl>,
## #   start_station_name <chr>, end_station_id <dbl>, end_station_name <chr>,
## #   member_casual <chr>, gender <chr>, birthyear <dbl>
```

```
## Inspect dataframes

str(q1_2018)
```

```
## spec_tbl_df [387,145 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id                                          : num [1:387145] 17536702 17536703 17536704 17536
```

```
##  $ started_at                                        : POSIXct[1:387145], format: "2018-01-01 00:12:00
##  $ ended_at                                          : POSIXct[1:387145], format: "2018-01-01 00:17:23
##  $ rideable_type                                     : num [1:387145] 3304 5367 4599 2302 3696 ...
##  $ 01 - Rental Details Duration In Seconds Uncapped: num [1:387145] 323 377 2904 747 183 ...
##  $ start_station_id                                  : num [1:387145] 69 253 98 125 129 304 164 182 99
##  $ start_station_name                                : chr [1:387145] "Damen Ave & Pierce Ave" "Winthro
##  $ end_station_id                                    : num [1:387145] 159 325 509 364 205 299 174 142 9
##  $ end_station_name                                  : chr [1:387145] "Claremont Ave & Hirsch St" "Cla
##  $ member_casual                                     : chr [1:387145] "Subscriber" "Subscriber" "Subsc
##  $ Member Gender                                     : chr [1:387145] "Male" "Male" "Male" "Male" ...
##  $ 05 - Member Details Member Birthday Year          : num [1:387145] 1988 1984 1989 1983 1989 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `01 - Rental Details Rental ID` = col_double(),
##   ..   `01 - Rental Details Local Start Time` = col_datetime(format = ""),
##   ..   `01 - Rental Details Local End Time` = col_datetime(format = ""),
##   ..   `01 - Rental Details Bike ID` = col_double(),
##   ..   `01 - Rental Details Duration In Seconds Uncapped` = col_number(),
##   ..   `03 - Rental Start Station ID` = col_double(),
##   ..   `03 - Rental Start Station Name` = col_character(),
##   ..   `02 - Rental End Station ID` = col_double(),
##   ..   `02 - Rental End Station Name` = col_character(),
##   ..   `User Type` = col_character(),
##   ..   `Member Gender` = col_character(),
##   ..   `05 - Member Details Member Birthday Year` = col_double()
##   .. )
```

```
str(q1_2020)
```

```
## spec_tbl_df [426,887 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:426887] "EACB19130B0CDA4A" "8FED874C809DC021" "789F3C21E472CA96" "C9A38
##  $ rideable_type     : chr [1:426887] "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:426887], format: "2020-01-21 20:06:59" "2020-01-30 14:22:39" ...
##  $ ended_at          : POSIXct[1:426887], format: "2020-01-21 20:14:30" "2020-01-30 14:26:22" ...
##  $ start_station_name: chr [1:426887] "Western Ave & Leland Ave" "Clark St & Montrose Ave" "Broadway
##  $ start_station_id  : num [1:426887] 239 234 296 51 66 212 96 96 212 38 ...
##  $ end_station_name  : chr [1:426887] "Clark St & Leland Ave" "Southport Ave & Irving Park Rd" "Wilto
##  $ end_station_id    : num [1:426887] 326 318 117 24 212 96 212 212 96 100 ...
##  $ start_lat         : num [1:426887] 42 42 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:426887] -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:426887] 42 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:426887] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:426887] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
```

```
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
```

```
str(q4_2018)
```

```
## spec_tbl_df [642,686 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:642686] 2.1e+07 2.1e+07 2.1e+07 2.1e+07 2.1e+07 ...
##  $ started_at        : POSIXct[1:642686], format: "2018-10-01 00:01:17" "2018-10-01 00:03:59" ...
##  $ ended_at          : POSIXct[1:642686], format: "2018-10-01 00:29:35" "2018-10-01 00:10:55" ...
##  $ rideable_type     : num [1:642686] 4551 847 6188 6372 1927 ...
##  $ tripduration      : num [1:642686] 1698 416 534 778 1102 ...
##  $ start_station_id  : num [1:642686] 85 13 59 328 93 229 148 374 268 125 ...
##  $ start_station_name: chr [1:642686] "Michigan Ave & Oak St" "Wilton Ave & Diversey Pkwy" "Wabash Av
##  $ end_station_id    : num [1:642686] 166 144 197 419 159 318 11 130 289 175 ...
##  $ end_station_name  : chr [1:642686] "Ashland Ave & Wrightwood Ave" "Larrabee St & Webster Ave" "Mi
##  $ member_casual     : chr [1:642686] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:642686] "Male" "Female" "Male" "Female" ...
##  $ birthyear         : num [1:642686] 1992 1982 1986 1960 1993 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
```

```
str(q4_2019)
```

```
## spec_tbl_df [704,054 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:704054] 25223640 25223641 25223642 25223643 25223644 ...
##  $ started_at        : POSIXct[1:704054], format: "2019-10-01 00:01:39" "2019-10-01 00:02:16" ...
##  $ ended_at          : POSIXct[1:704054], format: "2019-10-01 00:17:20" "2019-10-01 00:06:34" ...
##  $ rideable_type     : num [1:704054] 2215 6328 3003 3275 5294 ...
##  $ tripduration      : num [1:704054] 940 258 850 2350 1867 ...
##  $ start_station_id  : num [1:704054] 20 19 84 313 210 156 84 156 156 336 ...
##  $ start_station_name: chr [1:704054] "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St
##  $ end_station_id    : num [1:704054] 309 241 199 290 382 226 142 463 463 336 ...
##  $ end_station_name  : chr [1:704054] "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave &
##  $ member_casual     : chr [1:704054] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:704054] "Male" "Male" "Female" "Male" ...
##  $ birthyear         : num [1:704054] 1987 1998 1991 1990 1987 ...
##  - attr(*, "spec")=
```

```
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
```

```
## Change ride_id and rideable_type to chr

q1_2018 <-  mutate(q1_2018, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q4_2018 <-  mutate(q4_2018, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q4_2019 <-  mutate(q4_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

## Combine datasets into one dataframe

all_trips <- bind_rows(q1_2018, q1_2020, q4_2018, q4_2019)

## Remove unnecessary columns

all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, birthyear, gender, "01 - Rental Details Duration In
```

Clean & Add Data to Prepare for Analysis

```
## Look at new table
colnames(all_trips)
```

```
## [1] "ride_id"            "started_at"         "ended_at"
## [4] "rideable_type"      "start_station_id"   "start_station_name"
## [7] "end_station_id"     "end_station_name"   "member_casual"
```

```
nrow(all_trips)
```

```
## [1] 2160772
```

```
dim(all_trips)
```

```
## [1] 2160772       9
```

```
head(all_trips)
```

```
## # A tibble: 6 x 9
##   ride_id started_at          ended_at            rideable_type start_station_id
##   <chr>   <dttm>              <dttm>              <chr>                        <dbl>
## 1 175367~ 2018-01-01 00:12:00 2018-01-01 00:17:23 3304                            69
## 2 175367~ 2018-01-01 00:41:35 2018-01-01 00:47:52 5367                           253
## 3 175367~ 2018-01-01 00:44:46 2018-01-01 01:33:10 4599                            98
## 4 175367~ 2018-01-01 00:53:10 2018-01-01 01:05:37 2302                           125
## 5 175367~ 2018-01-01 00:53:37 2018-01-01 00:56:40 3696                           129
## 6 175367~ 2018-01-01 00:56:15 2018-01-01 01:00:41 6298                           304
## # ... with 4 more variables: start_station_name <chr>, end_station_id <dbl>,
## #   end_station_name <chr>, member_casual <chr>
```

```
str(all_trips)
```

```
## tibble [2,160,772 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:2160772] "17536702" "17536703" "17536704" "17536705" ...
##  $ started_at        : POSIXct[1:2160772], format: "2018-01-01 00:12:00" "2018-01-01 00:41:35" ...
##  $ ended_at          : POSIXct[1:2160772], format: "2018-01-01 00:17:23" "2018-01-01 00:47:52" ...
##  $ rideable_type     : chr [1:2160772] "3304" "5367" "4599" "2302" ...
##  $ start_station_id  : num [1:2160772] 69 253 98 125 129 304 164 182 99 99 ...
##  $ start_station_name: chr [1:2160772] "Damen Ave & Pierce Ave" "Winthrop Ave & Lawrence Ave" "LaSal
##  $ end_station_id    : num [1:2160772] 159 325 509 364 205 299 174 142 99 99 ...
##  $ end_station_name  : chr [1:2160772] "Claremont Ave & Hirsch St" "Clark St & Winnemac Ave (Temp)"
##  $ member_casual     : chr [1:2160772] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
```

```
summary(all_trips)
```

```
##    ride_id            started_at                    ended_at
##  Length:2160772     Min.   :2018-01-01 00:12:00   Min.   :2018-01-01 00:17:23
##  Class :character   1st Qu.:2018-10-12 20:20:06   1st Qu.:2018-10-12 20:41:10
##  Mode  :character   Median :2019-10-04 13:50:49   Median :2019-10-04 14:11:49
##                     Mean   :2019-04-18 21:26:27   Mean   :2019-04-18 21:45:32
##                     3rd Qu.:2019-12-07 11:03:28   3rd Qu.:2019-12-07 11:21:31
##                     Max.   :2020-03-31 23:51:34   Max.   :2020-05-19 20:10:34
##
##  rideable_type      start_station_id start_station_name end_station_id
##  Length:2160772     Min.   :  2.0    Length:2160772     Min.   :  2.0
##  Class :character   1st Qu.: 77.0    Class :character   1st Qu.: 77.0
##  Mode  :character   Median :173.0    Mode  :character   Median :172.0
##                     Mean   :199.2                       Mean   :199.1
##                     3rd Qu.:288.0                       3rd Qu.:288.0
##                     Max.   :675.0                       Max.   :675.0
##                                                         NA's   :1
##  end_station_name   member_casual
##  Length:2160772     Length:2160772
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
## Inspect labels for 'member_casual', consolidate to two
unique(all_trips[c("member_casual")])
```

```
## # A tibble: 4 x 1
##   member_casual
##   <chr>
## 1 Subscriber
## 2 Customer
## 3 member
## 4 casual
```

```
table(all_trips$member_casual)
```

```
##
##     casual   Customer     member Subscriber
##      48480     190873     378407    1543012
```

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

```
##
##  casual  member
##  239353 1921419
```

```
## Add descriptive columns for each ride (date, month, day, year) so we can
## aggregate in a more helpful way

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

```
## # A tibble: 6 x 14
##   ride_id started_at          ended_at            rideable_type start_station_id
##   <chr>   <dttm>              <dttm>              <chr>                      <dbl>
## 1 175367~ 2018-01-01 00:12:00 2018-01-01 00:17:23 3304                          69
## 2 175367~ 2018-01-01 00:41:35 2018-01-01 00:47:52 5367                         253
## 3 175367~ 2018-01-01 00:44:46 2018-01-01 01:33:10 4599                          98
## 4 175367~ 2018-01-01 00:53:10 2018-01-01 01:05:37 2302                         125
## 5 175367~ 2018-01-01 00:53:37 2018-01-01 00:56:40 3696                         129
## 6 175367~ 2018-01-01 00:56:15 2018-01-01 01:00:41 6298                         304
## # ... with 9 more variables: start_station_name <chr>, end_station_id <dbl>,
## #   end_station_name <chr>, member_casual <chr>, date <date>, month <chr>,
## #   day <chr>, year <chr>, day_of_week <chr>
```

```
## Add a column for ride_length (in seconds)
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
str(all_trips)
```

```
## tibble [2,160,772 x 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:2160772] "17536702" "17536703" "17536704" "17536705" ...
##  $ started_at        : POSIXct[1:2160772], format: "2018-01-01 00:12:00" "2018-01-01 00:41:35" ...
##  $ ended_at          : POSIXct[1:2160772], format: "2018-01-01 00:17:23" "2018-01-01 00:47:52" ...
##  $ rideable_type     : chr [1:2160772] "3304" "5367" "4599" "2302" ...
##  $ start_station_id  : num [1:2160772] 69 253 98 125 129 304 164 182 99 99 ...
##  $ start_station_name: chr [1:2160772] "Damen Ave & Pierce Ave" "Winthrop Ave & Lawrence Ave" "LaSal
##  $ end_station_id    : num [1:2160772] 159 325 509 364 205 299 174 142 99 99 ...
##  $ end_station_name  : chr [1:2160772] "Claremont Ave & Hirsch St" "Clark St & Winnemac Ave (Temp)"
##  $ member_casual     : chr [1:2160772] "member" "member" "member" "member" ...
##  $ date              : Date[1:2160772], format: "2018-01-01" "2018-01-01" ...
##  $ month             : chr [1:2160772] "01" "01" "01" "01" ...
##  $ day               : chr [1:2160772] "01" "01" "01" "01" ...
##  $ year              : chr [1:2160772] "2018" "2018" "2018" "2018" ...
##  $ day_of_week       : chr [1:2160772] "Monday" "Monday" "Monday" "Monday" ...
##  $ ride_length       : 'difftime' num [1:2160772] 323 377 2904 747 ...
##   ..- attr(*, "units")= chr "secs"
```

```
is.numeric(all_trips$ride_length)
```

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

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

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

```
## Remove bad data (bikes being checked for quality or negative ride time).
### Create V2 version of dataframe
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<0),]
```

Descriptive Analysis

```
## Look at ride_length
summary(all_trips_v2$ride_length)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        1      342      557     1147      943 14340041
```

```
## Compare members to casual riders
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                4268.7437
## 2                     member                 764.3251
```

```r
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                     1366
## 2                     member                      516
```

```r
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                 14340041
## 2                     member                 13561217
```

```r
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                        2
## 2                     member                        1
```

```r
## Average ride time by day for members vs casual riders
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Sunday                4168.7688
## 2                      member                   Sunday                 838.0304
## 3                      casual                   Monday                3565.7127
## 4                      member                   Monday                 751.7841
## 5                      casual                  Tuesday                4420.5960
## 6                      member                  Tuesday                 743.3060
## 7                      casual                Wednesday                3958.1885
## 8                      member                Wednesday                 736.7186
## 9                      casual                 Thursday                4223.2260
## 10                     member                 Thursday                 764.4709
## 11                     casual                   Friday                5233.4108
## 12                     member                   Friday                 733.5178
## 13                     casual                 Saturday                4345.7333
## 14                     member                 Saturday                 873.7174
```

```r
## Analyze by type and weekday
all_trips_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)
```

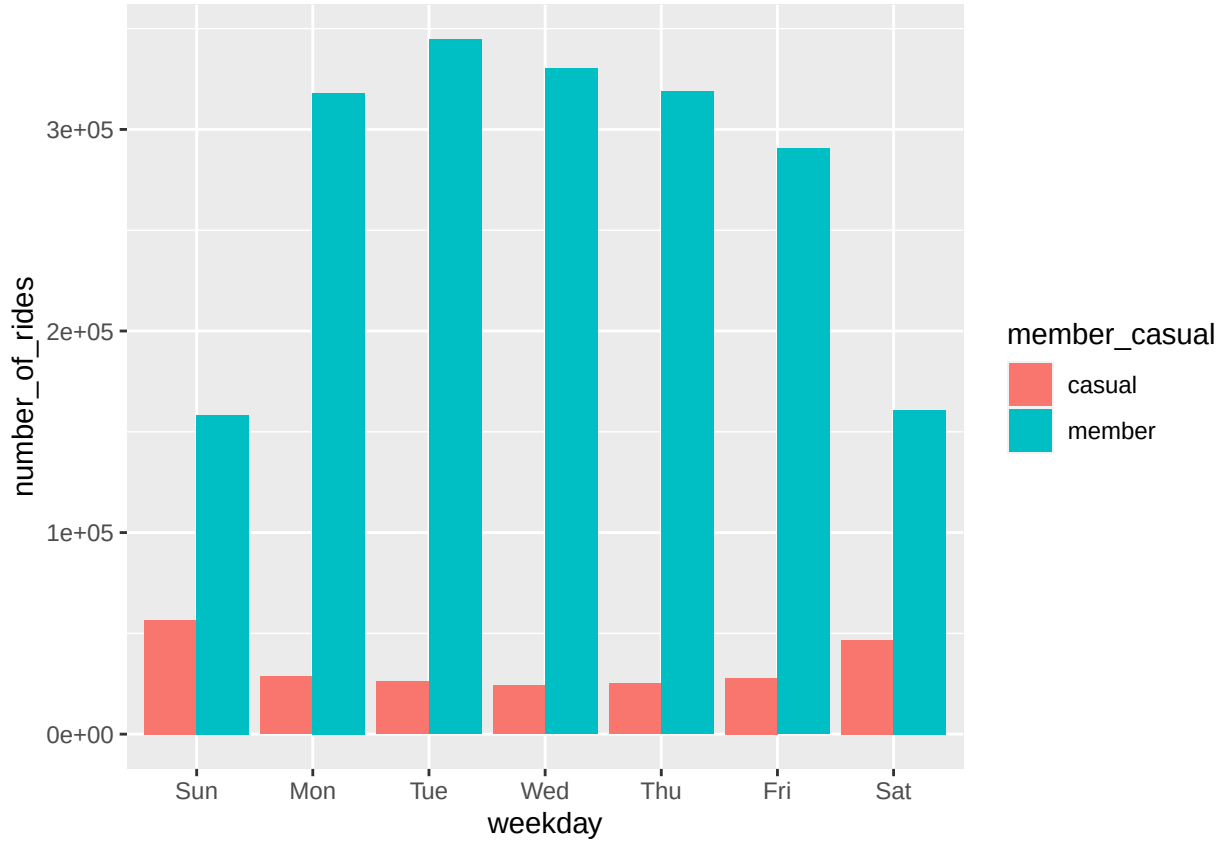
```
## `summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
```

```
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun               56687            4169.
##  2 casual        Mon               28593            3566.
##  3 casual        Tue               26096            4421.
##  4 casual        Wed               24303            3958.
##  5 casual        Thu               25256            4223.
##  6 casual        Fri               27889            5233.
##  7 casual        Sat               46756            4346.
##  8 member        Sun              158290             838.
##  9 member        Mon              318083             752.
## 10 member        Tue              344697             743.
## 11 member        Wed              330243             737.
## 12 member        Thu              318969             764.
## 13 member        Fri              290383             734.
## 14 member        Sat              160740             874.
```

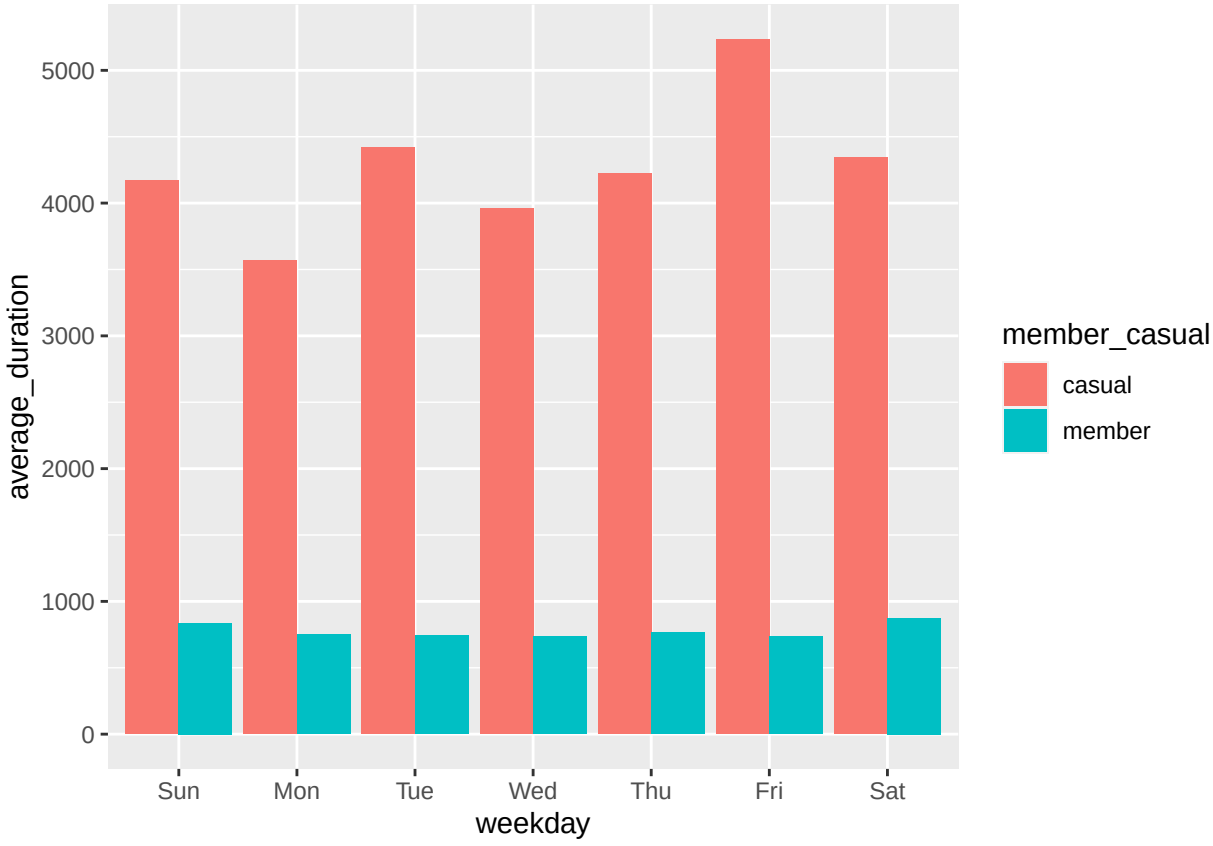
```r
## Viz - number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, 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.
```




```
## Viz - average duration
all_trips_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.
```




```
## Export summary file
counts <- all_trips_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)
```

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

```
write.csv(counts, file = '~/number_of_rides_avg_ride_length.csv')
```