Case Study on How Annual Subscribers differ from Casual Subscribers

Esther Anthony Oyeniyi

2022-11-21

STEP 1: UPLOADING THE DATA

In this step, the data was uploaded using a code chunk that automatically sets the data structure to what it is required to be. for example, column types were set to datetime(dttm), double (dbl), strings, etc.This was done for the month in the year 2020 and 2021 by changing the figures as they had similar format.Meanwhile, the data for year 2019 were in quarters and so were uploaded using a different code. The embedded R code chunk is like this:

```
library(tidyverse)
Q1_2020<- read_csv("cyclist/2020 DATA/Divvy_Trips_2020_Q1.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202004<- read_csv("cyclist/2020 DATA/202004-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202005<- read_csv("cyclist/2020 DATA/202005-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202006<- read_csv("cyclist/2020 DATA/202006-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202007<- read_csv("cyclist/2020 DATA/202007-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202008<- read_csv("cyclist/2020 DATA/202008-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202009<- read_csv("cyclist/2020 DATA/202009-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202010<- read_csv("cyclist/2020 DATA/202010-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
```

```
X202011<- read_csv("cyclist/2020 DATA/202011-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202012<- read_csv("cyclist/2020 DATA/202012-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202101<- read_csv("cyclist/2021 DATA/202101-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202102<- read_csv("cyclist/2021 DATA/202101-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202103<- read_csv("cyclist/2021 DATA/202103-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202104<- read_csv("cyclist/2021 DATA/202104-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202105<- read_csv("cyclist/2021 DATA/202105-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202106<- read_csv("cyclist/2021 DATA/202106-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202107<- read_csv("cyclist/2021 DATA/202108-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202108<- read_csv("cyclist/2021 DATA/202108-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202109<- read_csv("cyclist/2021 DATA/202109-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202110<- read_csv("cyclist/2021 DATA/202110-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202111<- read_csv("cyclist/2021 DATA/202111-divvy-tripdata.csv",
                   col_types = cols(ride_id = col_character(),
                                    rideable_type = col_character(),
                                    started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col
X202112<- read_csv("cyclist/2021 DATA/202112-divvy-tripdata.csv",
```

```
                  col_types = cols(ride_id = col_character(),
                                   rideable_type = col_character(),
                                   started_at = col_datetime(format = "%d/%m/%Y %H:%M"), ended_at = col

q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
q1_2019 <- read_csv("Divvy_Trips_2019_Q1.csv")
## Check the structures to ensure coltypes are in order.
str(q1_2019)
```

```
## spec_tbl_df [365,069 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ trip_id          : num [1:365069] 21742443 21742444 21742445 21742446 21742447 ...
##  $ start_time       : POSIXct[1:365069], format: "2019-01-01 00:04:37" "2019-01-01 00:08:13" ...
##  $ end_time         : POSIXct[1:365069], format: "2019-01-01 00:11:07" "2019-01-01 00:15:34" ...
##  $ bikeid           : num [1:365069] 2167 4386 1524 252 1170 ...
##  $ tripduration     : num [1:365069] 390 441 829 1783 364 ...
##  $ from_station_id  : num [1:365069] 199 44 15 123 173 98 98 211 150 268 ...
##  $ from_station_name: chr [1:365069] "Wabash Ave & Grand Ave" "State St & Randolph St" "Racine Ave &
##  $ to_station_id    : num [1:365069] 84 624 644 176 35 49 49 142 148 141 ...
##  $ to_station_name  : chr [1:365069] "Milwaukee Ave & Grand Ave" "Dearborn St & Van Buren St (*)" "W
##  $ usertype         : chr [1:365069] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender           : chr [1:365069] "Male" "Female" "Female" "Male" ...
##  $ birthyear        : num [1:365069] 1989 1990 1994 1993 1994 ...
##  - 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2019)
```

```
## spec_tbl_df [1,108,163 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ 01 - Rental Details Rental ID                   : num [1:1108163] 22178529 22178530 22178531 2217
##  $ 01 - Rental Details Local Start Time            : POSIXct[1:1108163], format: "2019-04-01 00:02:2
##  $ 01 - Rental Details Local End Time              : POSIXct[1:1108163], format: "2019-04-01 00:09:4
##  $ 01 - Rental Details Bike ID                     : num [1:1108163] 6251 6226 5649 4151 3270 ...
##  $ 01 - Rental Details Duration In Seconds Uncapped: num [1:1108163] 446 1048 252 357 1007 ...
##  $ 03 - Rental Start Station ID                    : num [1:1108163] 81 317 283 26 202 420 503 260 2
##  $ 03 - Rental Start Station Name                  : chr [1:1108163] "Daley Center Plaza" "Wood St &
##  $ 02 - Rental End Station ID                      : num [1:1108163] 56 59 174 133 129 426 500 499 2
##  $ 02 - Rental End Station Name                    : chr [1:1108163] "Desplaines St & Kinzie St" "Wal
##  $ User Type                                       : chr [1:1108163] "Subscriber" "Subscriber" "Subs
```

```
##  $ Member Gender                                    : chr [1:1108163] "Male" "Female" "Male" "Male" .
##  $ 05 - Member Details Member Birthday Year         : num [1:1108163] 1975 1984 1990 1993 1992 ...
##  - 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2019)
```

```
## spec_tbl_df [1,640,718 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ trip_id          : num [1:1640718] 23479388 23479389 23479390 23479391 23479392 ...
##  $ start_time       : POSIXct[1:1640718], format: "2019-07-01 00:00:27" "2019-07-01 00:01:16" ...
##  $ end_time         : POSIXct[1:1640718], format: "2019-07-01 00:20:41" "2019-07-01 00:18:44" ...
##  $ bikeid           : num [1:1640718] 3591 5353 6180 5540 6014 ...
##  $ tripduration     : num [1:1640718] 1214 1048 1554 1503 1213 ...
##  $ from_station_id  : num [1:1640718] 117 381 313 313 168 300 168 313 43 43 ...
##  $ from_station_name: chr [1:1640718] "Wilton Ave & Belmont Ave" "Western Ave & Monroe St" "Lakeview
##  $ to_station_id    : num [1:1640718] 497 203 144 144 62 232 62 144 195 195 ...
##  $ to_station_name  : chr [1:1640718] "Kimball Ave & Belmont Ave" "Western Ave & 21st St" "Larrabee S
##  $ usertype         : chr [1:1640718] "Subscriber" "Customer" "Customer" "Customer" ...
##  $ gender           : chr [1:1640718] "Male" NA NA NA ...
##  $ birthyear        : num [1:1640718] 1992 NA NA NA NA ...
##  - 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q4_2019)
```

```
## spec_tbl_df [704,054 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ trip_id          : num [1:704054] 25223640 25223641 25223642 25223643 25223644 ...
```

```
##  $ start_time       : POSIXct[1:704054], format: "2019-10-01 00:01:39" "2019-10-01 00:02:16" ...
##  $ end_time         : POSIXct[1:704054], format: "2019-10-01 00:17:20" "2019-10-01 00:06:34" ...
##  $ bikeid           : num [1:704054] 2215 6328 3003 3275 5294 ...
##  $ tripduration     : num [1:704054] 940 258 850 2350 1867 ...
##  $ from_station_id  : num [1:704054] 20 19 84 313 210 156 84 156 156 336 ...
##  $ from_station_name: chr [1:704054] "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St"
##  $ to_station_id    : num [1:704054] 309 241 199 290 382 226 142 463 463 336 ...
##  $ to_station_name  : chr [1:704054] "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave &
##  $ usertype         : 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Cleaning the Data

The data in 2019 was renamed.

```
(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_t~1 tripd~2 start~3
##       <dbl> <dttm>              <dttm>                     <dbl>   <dbl>   <dbl>
##  1 25223640 2019-10-01 00:01:39 2019-10-01 00:17:20         2215     940      20
##  2 25223641 2019-10-01 00:02:16 2019-10-01 00:06:34         6328     258      19
##  3 25223642 2019-10-01 00:04:32 2019-10-01 00:18:43         3003     850      84
##  4 25223643 2019-10-01 00:04:32 2019-10-01 00:43:43         3275    2350     313
##  5 25223644 2019-10-01 00:04:34 2019-10-01 00:35:42         5294    1867     210
##  6 25223645 2019-10-01 00:04:38 2019-10-01 00:10:51         1891     373     156
##  7 25223646 2019-10-01 00:04:52 2019-10-01 00:22:45         1061    1072      84
##  8 25223647 2019-10-01 00:04:57 2019-10-01 00:29:16         1274    1458     156
##  9 25223648 2019-10-01 00:05:20 2019-10-01 00:29:18         6011    1437     156
```

```
## 10 25223649 2019-10-01 00:05:20 2019-10-01 02:23:46         2957    8306     336
## # ... with 704,044 more rows, 6 more variables: start_station_name <chr>,
## #   end_station_id <dbl>, end_station_name <chr>, member_casual <chr>,
## #   gender <chr>, birthyear <dbl>, and abbreviated variable names
## #   1: rideable_type, 2: tripduration, 3: start_station_id
```

```
(Q3_2019 <- rename(q3_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: 1,640,718 x 12
##     ride_id started_at          ended_at            rideable_t~1 tripd~2 start~3
##       <dbl> <dttm>              <dttm>                     <dbl>   <dbl>   <dbl>
##  1 23479388 2019-07-01 00:00:27 2019-07-01 00:20:41         3591    1214     117
##  2 23479389 2019-07-01 00:01:16 2019-07-01 00:18:44         5353    1048     381
##  3 23479390 2019-07-01 00:01:48 2019-07-01 00:27:42         6180    1554     313
##  4 23479391 2019-07-01 00:02:07 2019-07-01 00:27:10         5540    1503     313
##  5 23479392 2019-07-01 00:02:13 2019-07-01 00:22:26         6014    1213     168
##  6 23479393 2019-07-01 00:02:21 2019-07-01 00:07:31         4941     310     300
##  7 23479394 2019-07-01 00:02:24 2019-07-01 00:23:12         3770    1248     168
##  8 23479395 2019-07-01 00:02:26 2019-07-01 00:28:16         5442    1550     313
##  9 23479396 2019-07-01 00:02:34 2019-07-01 00:28:57         2957    1583      43
## 10 23479397 2019-07-01 00:02:45 2019-07-01 00:29:14         6091    1589      43
## # ... with 1,640,708 more rows, 6 more variables: start_station_name <chr>,
## #   end_station_id <dbl>, end_station_name <chr>, member_casual <chr>,
## #   gender <chr>, birthyear <dbl>, and abbreviated variable names
## #   1: rideable_type, 2: tripduration, 3: start_station_id
```

```
(Q2_2019 <- rename(q2_2019
                  ,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: 1,108,163 x 12
##     ride_id started_at          ended_at            rideable_t~1 01 - ~2 start~3
##       <dbl> <dttm>              <dttm>                     <dbl>   <dbl>   <dbl>
##  1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48         6251     446      81
##  2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30         6226    1048     317
##  3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19         5649     252     283
##  4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58         4151     357      26
##  5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13         3270    1007     202
##  6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56         3123     257     420
```

```
##  7 22178535 2019-04-01 00:26:33 2019-04-01 00:35:41         6418     548     503
##  8 22178536 2019-04-01 00:29:48 2019-04-01 00:36:11         4513     383     260
##  9 22178537 2019-04-01 00:32:07 2019-04-01 01:07:44         3280    2137     211
## 10 22178538 2019-04-01 00:32:19 2019-04-01 01:07:39         5534    2120     211
## # ... with 1,108,153 more rows, 6 more variables: 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>,
## #   and abbreviated variable names 1: rideable_type,
## #   2: `01 - Rental Details Duration In Seconds Uncapped`, 3: start_station_id
```

```
(Q1_2019 <- rename(q1_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: 365,069 x 12
##     ride_id started_at          ended_at            rideable_t~1 tripd~2 start~3
##       <dbl> <dttm>              <dttm>                     <dbl>   <dbl>   <dbl>
##  1 21742443 2019-01-01 00:04:37 2019-01-01 00:11:07         2167     390     199
##  2 21742444 2019-01-01 00:08:13 2019-01-01 00:15:34         4386     441      44
##  3 21742445 2019-01-01 00:13:23 2019-01-01 00:27:12         1524     829      15
##  4 21742446 2019-01-01 00:13:45 2019-01-01 00:43:28          252    1783     123
##  5 21742447 2019-01-01 00:14:52 2019-01-01 00:20:56         1170     364     173
##  6 21742448 2019-01-01 00:15:33 2019-01-01 00:19:09         2437     216      98
##  7 21742449 2019-01-01 00:16:06 2019-01-01 00:19:03         2708     177      98
##  8 21742450 2019-01-01 00:18:41 2019-01-01 00:20:21         2796     100     211
##  9 21742451 2019-01-01 00:18:43 2019-01-01 00:47:30         6205    1727     150
## 10 21742452 2019-01-01 00:19:18 2019-01-01 00:24:54         3939     336     268
## # ... with 365,059 more rows, 6 more variables: start_station_name <chr>,
## #   end_station_id <dbl>, end_station_name <chr>, member_casual <chr>,
## #   gender <chr>, birthyear <dbl>, and abbreviated variable names
## #   1: rideable_type, 2: tripduration, 3: start_station_id
```

DATA MERGING

Data of all the month in 2020 and 2021 will be merged into one as all_trips_2020 and all_trips_2021 respectively and all the quarters in 2019 as all_trip_2019.

```
all_trips_2019 <- bind_rows(Q1_2019,Q2_2019,Q3_2019,Q4_2019)
all_trips_2020 <- bind_rows(Q1_2020, X202004, X202005, X202006,
                            X202007, X202008, X202009, X202010,
                            X202011, X202012)
all_trips_2021 <- bind_rows(X202101, X202102,
                        X202103, X202104, X202105, X202106,
                        X202107, X202108, X202109,
                        X202110, X202111, X202112)
```

DATA CLEANING 2

The merged data, all_trips_2019,all_trips_2020,and all_trips_2021 will be cleaned using these code chunk:

```
library(lubridate)
## change the string in rideable_type of 2019 to suit rideable_type_2020 and 2021)
mutate(all_trips_2019, ride_id = as.character(ride_id, rideable_type = as.character(rideable_type)))
```

```
## # A tibble: 3,818,004 x 15
##    ride_id  started_at          ended_at            rideable_t~1 tripd~2 start~3
##    <chr>    <dttm>              <dttm>                     <dbl>   <dbl>   <dbl>
##  1 21742443 2019-01-01 00:04:37 2019-01-01 00:11:07         2167     390     199
##  2 21742444 2019-01-01 00:08:13 2019-01-01 00:15:34         4386     441      44
##  3 21742445 2019-01-01 00:13:23 2019-01-01 00:27:12         1524     829      15
##  4 21742446 2019-01-01 00:13:45 2019-01-01 00:43:28          252    1783     123
##  5 21742447 2019-01-01 00:14:52 2019-01-01 00:20:56         1170     364     173
##  6 21742448 2019-01-01 00:15:33 2019-01-01 00:19:09         2437     216      98
##  7 21742449 2019-01-01 00:16:06 2019-01-01 00:19:03         2708     177      98
##  8 21742450 2019-01-01 00:18:41 2019-01-01 00:20:21         2796     100     211
##  9 21742451 2019-01-01 00:18:43 2019-01-01 00:47:30         6205    1727     150
## 10 21742452 2019-01-01 00:19:18 2019-01-01 00:24:54         3939     336     268
## # ... with 3,817,994 more rows, 9 more variables: start_station_name <chr>,
## #   end_station_id <dbl>, end_station_name <chr>, member_casual <chr>,
## #   gender <chr>, birthyear <dbl>,
## #   `01 - Rental Details Duration In Seconds Uncapped` <dbl>,
## #   `Member Gender` <chr>, `05 - Member Details Member Birthday Year` <dbl>,
## #   and abbreviated variable names 1: rideable_type, 2: tripduration,
## #   3: start_station_id
```

```
all_trips_2019 <- all_trips_2019 %>%
  mutate(member_casual = recode(member_casual
                                ,"Subscriber" = "member"
                                ,"Customer" = "casual"))

## CREATE NEW COLS SUCH AS DATE,DAY, MONTH, YEAR AND DAY OF WEEK
all_trips_2019$date <- as.Date(all_trips_2019$started_at) #The default format is yyyy-mm-dd
all_trips_2019$month <- format(as.Date(all_trips_2019$date), "%m")
all_trips_2019$day <- format(as.Date(all_trips_2019$date), "%d")
all_trips_2019$year <- format(as.Date(all_trips_2019$date), "%Y")
all_trips_2019$day_of_week <- format(as.Date(all_trips_2019$date), "%A")


all_trips_2019$ride_length <- difftime(all_trips_2019$ended_at,all_trips_2019$started_at)
is.factor(all_trips_2019$ride_length)
```

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

```
all_trips_2019$ride_length <- as.numeric(as.character(all_trips_2019$ride_length))
is.numeric(all_trips_2019$ride_length)
```

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

```
all_trips_2020$ride_length <- difftime(all_trips_2020$ended_at,all_trips_2020$started_at)
is.factor(all_trips_2020$ride_length)
```

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

```
all_trips_2020$ride_length <- as.numeric(as.character(all_trips_2020$ride_length))
is.numeric(all_trips_2020$ride_length)
```

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

```
all_trips_2021$ride_length <- difftime(all_trips_2021$ended_at,all_trips_2021$started_at)
is.factor(all_trips_2021$ride_length)
```

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

```
all_trips_2021$ride_length <- as.numeric(as.character(all_trips_2021$ride_length))
is.numeric(all_trips_2021$ride_length)
```

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

```
## remove bad data
all_trips_2019_v2 <- all_trips_2019[!(all_trips_2019$start_station_name == "HQ QR" | all_trips_2019$ride
all_trips_2020_v2 <- all_trips_2020[!(all_trips_2020$start_station_name == "HQ QR" | all_trips_2020$ride
all_trips_2021_v2 <- all_trips_2021[!(all_trips_2021$start_station_name == "HQ QR" | all_trips_2021$ride
## CREATE NEW COLS SUCH AS DATE,DAY, MONTH, YEAR AND DAY OF WEEK
all_trips_2020_v2$date <- as.Date(all_trips_2020_v2$started_at)
#The default format is yyyy-mm-dd
all_trips_2020_v2$month <- format(as.Date(all_trips_2020_v2$date), "%m")
all_trips_2020_v2$day <- format(as.Date(all_trips_2020_v2$date), "%d")
all_trips_2020_v2$year <- format(as.Date(all_trips_2020_v2$date), "%Y")
all_trips_2020_v2$day_of_week <- format(as.Date(all_trips_2020_v2$date), "%A")

all_trips_2021_v2$date <- as.Date(all_trips_2021_v2$started_at)
#The default format is yyyy-mm-dd
all_trips_2021_v2$month <- format(as.Date(all_trips_2021_v2$date), "%m")
all_trips_2021_v2$day <- format(as.Date(all_trips_2021_v2$date), "%d")
all_trips_2021_v2$year <- format(as.Date(all_trips_2021_v2$date), "%Y")
all_trips_2021_v2$day_of_week <- format(as.Date(all_trips_2021_v2$date), "%A")

## select only the needed by deselecting unwanted data
all_trips_2019_v3 <- all_trips_2019_v2 %>%
  select(-c("birthyear", "gender", "01 - Rental Details Duration In Seconds Uncapped", "05 - Member Deta

all_trips_2020_v3 <- all_trips_2020_v2 %>%
  select(-c("start_lat", "start_lng", "end_lat", "end_lng", "weekday", "Weekday"))

all_trips_2021_v3 <- all_trips_2021_v2 %>%
  select(-c("start_lat", "start_lng", "end_lat", "end_lng", "weekday"))
```

DATA ANALYSIS

This will be done separately for each of the year

```
#YEAR 2019
summary(all_trips_2019_v3$ride_length)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      1.02      6.85     11.82     24.17     21.40 177200.37
```

```
# Compare members and casual users
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual, FUN = mean)
```

```
##   all_trips_2019_v3$member_casual all_trips_2019_v3$ride_length
## 1                          casual                      57.01802
## 2                          member                      14.32780
```

```
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual, FUN = median)
```

```
##   all_trips_2019_v3$member_casual all_trips_2019_v3$ride_length
## 1                          casual                      25.83333
## 2                          member                       9.80000
```

```
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual, FUN = max)
```

```
##   all_trips_2019_v3$member_casual all_trips_2019_v3$ride_length
## 1                          casual                      177200.4
## 2                          member                      150943.9
```

```
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual, FUN = min)
```

```
##   all_trips_2019_v3$member_casual all_trips_2019_v3$ride_length
## 1                          casual                      1.016667
## 2                          member                      1.016667
```

```
# The average ride time by each day for members vs casual users
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual + all_trips_2019_v3$day_of_we
```

```
##    all_trips_2019_v3$member_casual all_trips_2019_v3$day_of_week
## 1                           casual                         Friday
## 2                           member                         Friday
## 3                           casual                         Monday
## 4                           member                         Monday
## 5                           casual                       Saturday
## 6                           member                       Saturday
## 7                           casual                         Sunday
## 8                           member                         Sunday
## 9                           casual                       Thursday
## 10                          member                       Thursday
## 11                          casual                        Tuesday
## 12                          member                        Tuesday
## 13                          casual                      Wednesday
## 14                          member                      Wednesday
##    all_trips_2019_v3$ride_length
## 1                       60.17561
## 2                       13.89748
## 3                       54.49989
## 4                       14.24928
## 5                       54.06111
## 6                       16.30271
## 7                       56.18519
## 8                       15.40290
## 9                       59.95112
## 10                      13.77979
## 11                      57.41328
## 12                      14.15259
## 13                      60.33407
## 14                      13.80984
```

```r
# Notice that the days of the week are out of order. I'll fix that.
all_trips_2019_v3$day_of_week <- ordered(all_trips_2019_v3$day_of_week, levels=c("Sunday", "Monday",
"Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
# The average ride time by each day for members vs casual users
aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual + all_trips_2019_v3$day_of_we
```

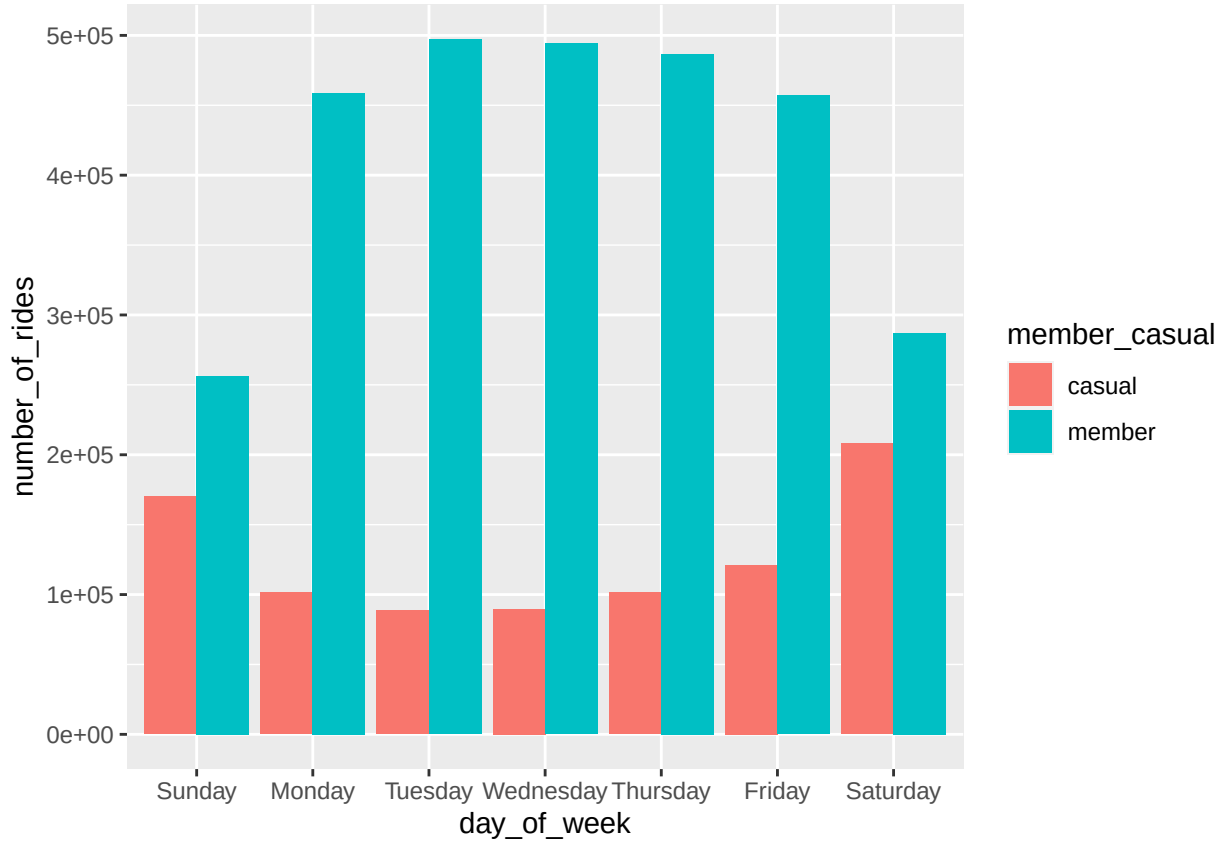
```
##    all_trips_2019_v3$member_casual all_trips_2019_v3$day_of_week
## 1                           casual                         Sunday
## 2                           member                         Sunday
## 3                           casual                         Monday
## 4                           member                         Monday
## 5                           casual                        Tuesday
## 6                           member                        Tuesday
## 7                           casual                      Wednesday
## 8                           member                      Wednesday
## 9                           casual                       Thursday
## 10                          member                       Thursday
## 11                          casual                         Friday
## 12                          member                         Friday
## 13                          casual                       Saturday
## 14                          member                       Saturday
##    all_trips_2019_v3$ride_length
## 1                       56.18519
## 2                       15.40290
## 3                       54.49989
## 4                       14.24928
## 5                       57.41328
## 6                       14.15259
## 7                       60.33407
## 8                       13.80984
## 9                       59.95112
## 10                      13.77979
## 11                      60.17561
## 12                      13.89748
## 13                      54.06111
## 14                      16.30271
```

```r
# analyze ridership data by type and day_of_week
all_trips_2019_v3 %>%
  group_by(member_casual, day_of_week) %>% #groups by user type and day_of_week
  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, day_of_week) # sorts
```

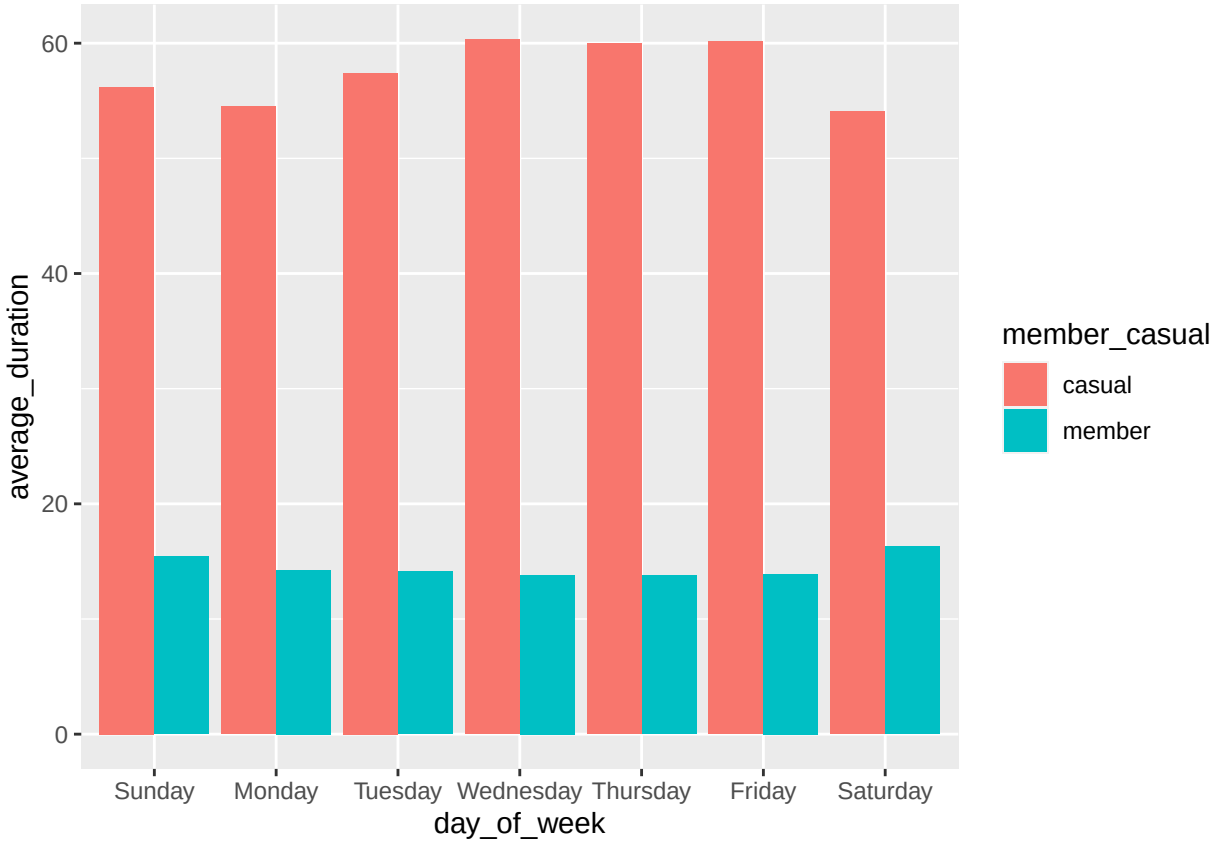
```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                <int>            <dbl>
##  1 casual        Sunday              170173             56.2
##  2 casual        Monday              101489             54.5
##  3 casual        Tuesday              88655             57.4
##  4 casual        Wednesday            89745             60.3
##  5 casual        Thursday            101372             60.0
##  6 casual        Friday              121141             60.2
```

```
##  7 casual        Saturday            208056            54.1
##  8 member        Sunday              256234            15.4
##  9 member        Monday              458780            14.2
## 10 member        Tuesday             497025            14.2
## 11 member        Wednesday           494277            13.8
## 12 member        Thursday            486915            13.8
## 13 member        Friday              456966            13.9
## 14 member        Saturday            287163            16.3
```

```r
# visualize the number of rides by rider type
all_trips_2019_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week, na = TRUE) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```




```r
# Create a visualization for average duration
all_trips_2019_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```




```
counts_2019 <- aggregate(all_trips_2019_v3$ride_length ~ all_trips_2019_v3$member_casual + all_trips_201


#YEAR 2020
summary(all_trips_2020_v3$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     480     840    1697    1560 9387000   94609
```

```
# Compare members and casual users
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual, FUN = mean)
```

```
##   all_trips_2020_v3$member_casual all_trips_2020_v3$ride_length
## 1                          casual                     2895.8621
## 2                          member                      946.5658
```

```
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual, FUN = median)
```

```
##   all_trips_2020_v3$member_casual all_trips_2020_v3$ride_length
## 1                          casual                          1320
## 2                          member                           660
```

```
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual, FUN = max)
```

```
##   all_trips_2020_v3$member_casual all_trips_2020_v3$ride_length
## 1                          casual                       9387000
## 2                          member                       5627640
```

```
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual, FUN = min)
```

```
##   all_trips_2020_v3$member_casual all_trips_2020_v3$ride_length
## 1                          casual                             0
## 2                          member                             0
```

```
# The average ride time by each day for members vs casual users
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual + all_trips_2020_v3$day_of_we
          FUN = mean)
```

```
##    all_trips_2020_v3$member_casual all_trips_2020_v3$day_of_week
## 1                          casual                         Friday
## 2                          member                         Friday
## 3                          casual                         Monday
## 4                          member                         Monday
## 5                          casual                       Saturday
## 6                          member                       Saturday
## 7                          casual                         Sunday
## 8                          member                         Sunday
## 9                          casual                       Thursday
## 10                         member                       Thursday
## 11                         casual                        Tuesday
## 12                         member                        Tuesday
## 13                         casual                      Wednesday
## 14                         member                      Wednesday
##    all_trips_2020_v3$ride_length
## 1                      2771.1856
## 2                       934.1218
## 3                      2857.2656
## 4                       899.2084
## 5                      2949.8484
## 6                      1074.8281
## 7                      3295.1075
## 8                      1093.6767
## 9                      2842.8887
## 10                      888.3074
## 11                     2618.9917
## 12                      872.6844
## 13                     2611.5219
## 14                      887.9519
```

```
# Notice that the days of the week are out of order. Let's fix that.
all_trips_2020_v3$day_of_week <- ordered(all_trips_2020_v3$day_of_week, levels=c("Sunday", "Monday", "Tu
# Find the average ride time by each day for members vs casual users
aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual + all_trips_2020_v3$day_of_we
          FUN = mean)
```

```
##    all_trips_2020_v3$member_casual all_trips_2020_v3$day_of_week
## 1                          casual                         Sunday
## 2                          member                         Sunday
## 3                          casual                         Monday
## 4                          member                         Monday
## 5                          casual                        Tuesday
## 6                          member                        Tuesday
## 7                          casual                      Wednesday
```

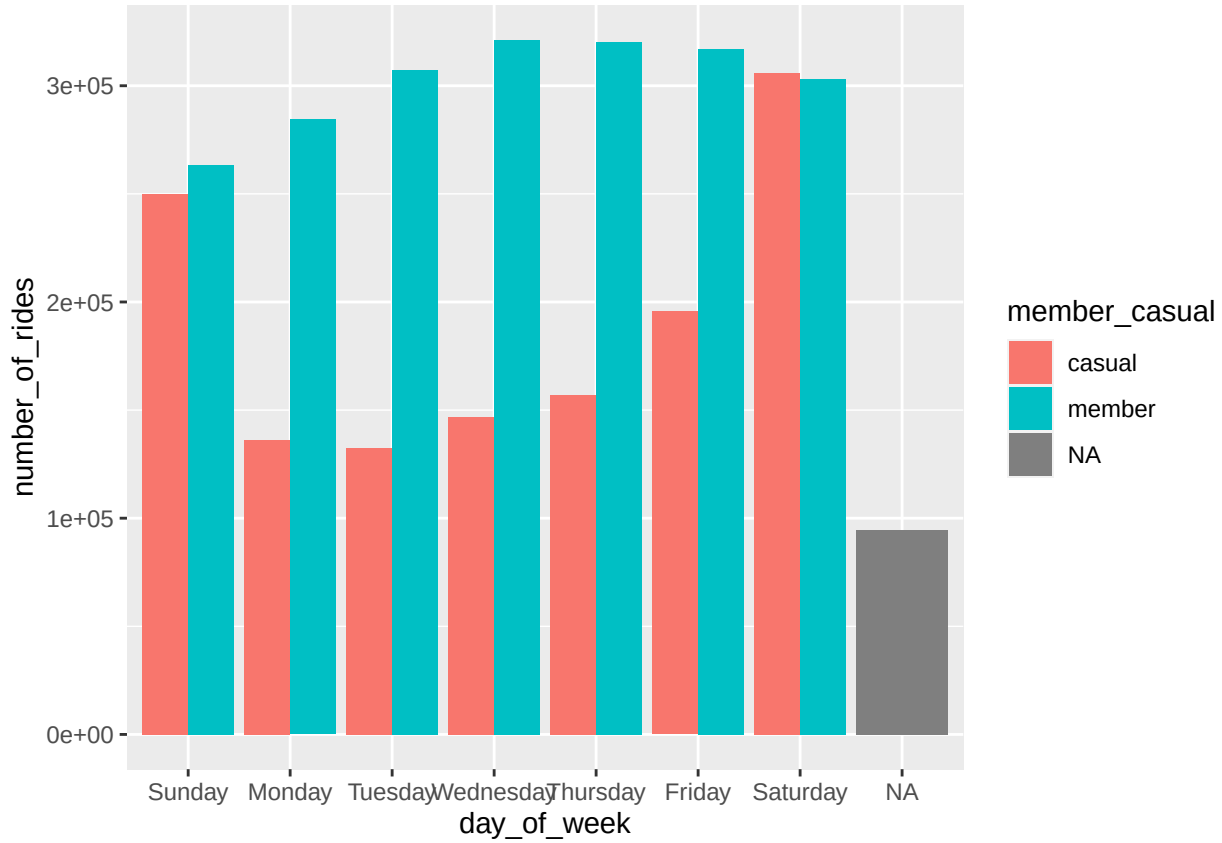
```
## 8                          member                    Wednesday
## 9                          casual                     Thursday
## 10                         member                     Thursday
## 11                         casual                       Friday
## 12                         member                       Friday
## 13                         casual                     Saturday
## 14                         member                     Saturday
##    all_trips_2020_v3$ride_length
## 1                      3295.1075
## 2                      1093.6767
## 3                      2857.2656
## 4                       899.2084
## 5                      2618.9917
## 6                       872.6844
## 7                      2611.5219
## 8                       887.9519
## 9                      2842.8887
## 10                      888.3074
## 11                     2771.1856
## 12                      934.1218
## 13                     2949.8484
## 14                     1074.8281
```

```
# analyze ridership data by type and day_of_week
all_trips_2020_v3 %>%
group_by(member_casual, day_of_week) %>% #groups by user type and day_of_week
  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, day_of_week) # sorts
```

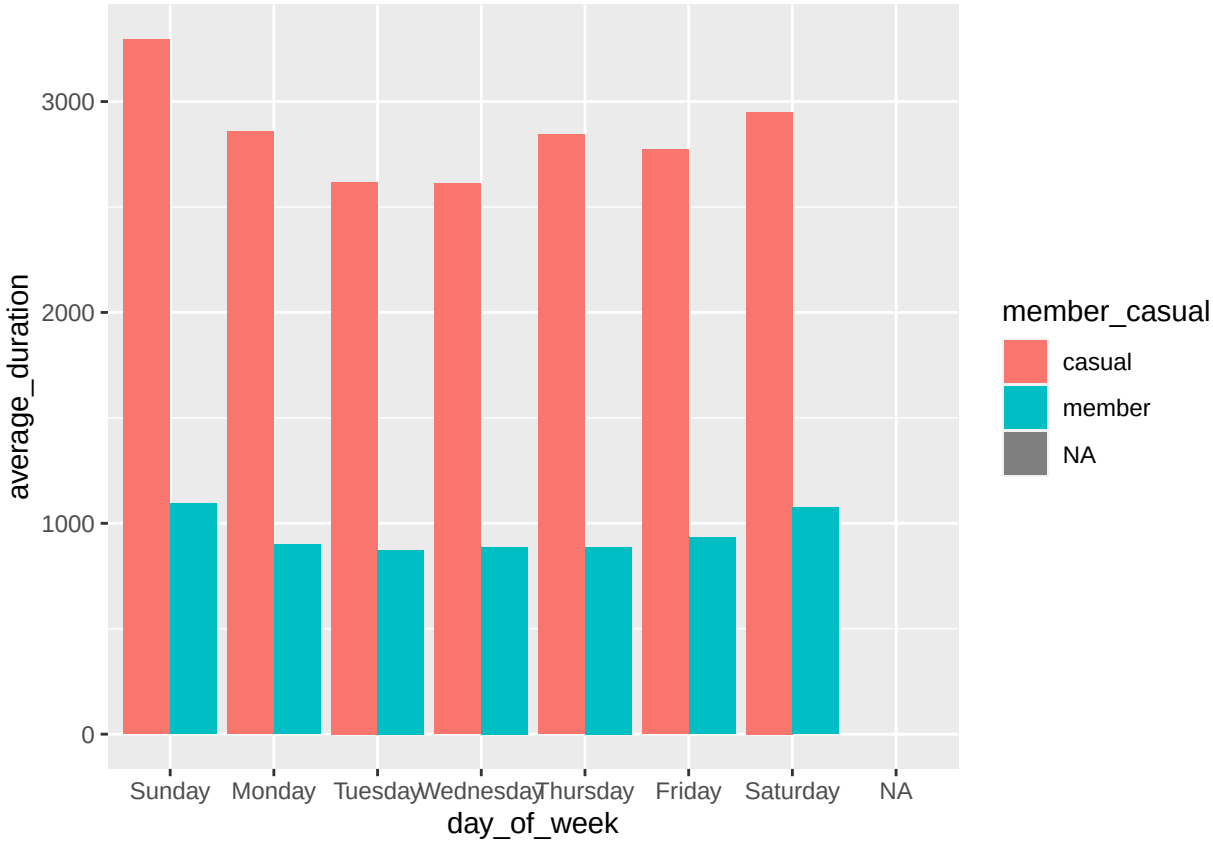
```
## # A tibble: 15 x 4
## # Groups:   member_casual [3]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                 <int>            <dbl>
##  1 casual        Sunday               250044            3295.
##  2 casual        Monday               136060            2857.
##  3 casual        Tuesday              132365            2619.
##  4 casual        Wednesday            146854            2612.
##  5 casual        Thursday             156806            2843.
##  6 casual        Friday               195578            2771.
##  7 casual        Saturday             305957            2950.
##  8 member        Sunday               263221            1094.
##  9 member        Monday               284370             899.
## 10 member        Tuesday              307181             873.
## 11 member        Wednesday            321323             888.
## 12 member        Thursday             320079             888.
## 13 member        Friday               317113             934.
## 14 member        Saturday             303025            1075.
## 15 <NA>          <NA>                  94609              NA
```

```
# Visualize the number of rides by rider type
all_trips_2020_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
```

```
  arrange(member_casual, day_of_week, na = TRUE) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```




```
# Create a visualization for average duration
all_trips_2020_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```




```
counts_2020 <- aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual + all_trips_202

counts_2020_1 <- aggregate(all_trips_2020_v3$ride_length ~ all_trips_2020_v3$member_casual + all_trips_1

#YEAR 2021
summary(all_trips_2021_v3$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     420     720    1334    1320 3356640  696565
```

```
# Compare members and casual users
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual, FUN = mean)
```

```
##   all_trips_2021_v3$member_casual all_trips_2021_v3$ride_length
## 1                          casual                     1980.9887
## 2                          member                      816.3187
```

```
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual, FUN = median)
```

```
##   all_trips_2021_v3$member_casual all_trips_2021_v3$ride_length
## 1                          casual                           960
## 2                          member                           600
```

```
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual, FUN = max)
```

```
##   all_trips_2021_v3$member_casual all_trips_2021_v3$ride_length
## 1                          casual                        3356640
## 2                          member                          93600
```

```
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual, FUN = min)
```

```
##   all_trips_2021_v3$member_casual all_trips_2021_v3$ride_length
## 1                          casual                             0
## 2                          member                             0
```

```
# The average ride time by each day for members vs casual users
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual + all_trips_2021_v3$day_of_we
          FUN = mean)
```

```
##    all_trips_2021_v3$member_casual all_trips_2021_v3$day_of_week
## 1                          casual                         Friday
## 2                          member                         Friday
## 3                          casual                         Monday
## 4                          member                         Monday
## 5                          casual                       Saturday
## 6                          member                       Saturday
## 7                          casual                         Sunday
## 8                          member                         Sunday
## 9                          casual                       Thursday
## 10                         member                       Thursday
## 11                         casual                        Tuesday
## 12                         member                        Tuesday
## 13                         casual                      Wednesday
## 14                         member                      Wednesday
##    all_trips_2021_v3$ride_length
## 1                      1883.5499
## 2                       794.0223
## 3                      1928.8068
## 4                       790.6884
## 5                      2131.5252
## 6                       918.5661
## 7                      2318.1110
## 8                       942.0297
## 9                      1705.6112
## 10                      765.1741
## 11                     1745.5962
## 12                      762.3587
## 13                     1727.0750
## 14                      766.6581
```

```
# Notice that the days of the week are out of order. Let's fix that.
all_trips_2021_v3$day_of_week <- ordered(all_trips_2021_v3$day_of_week, levels=c("Sunday", "Monday", "Tu
# Find the average ride time by each day for members vs casual users
aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual + all_trips_2021_v3$day_of_we
          FUN = mean)
```

```
##    all_trips_2021_v3$member_casual all_trips_2021_v3$day_of_week
## 1                          casual                         Sunday
## 2                          member                         Sunday
## 3                          casual                         Monday
## 4                          member                         Monday
## 5                          casual                        Tuesday
## 6                          member                        Tuesday
## 7                          casual                      Wednesday
```

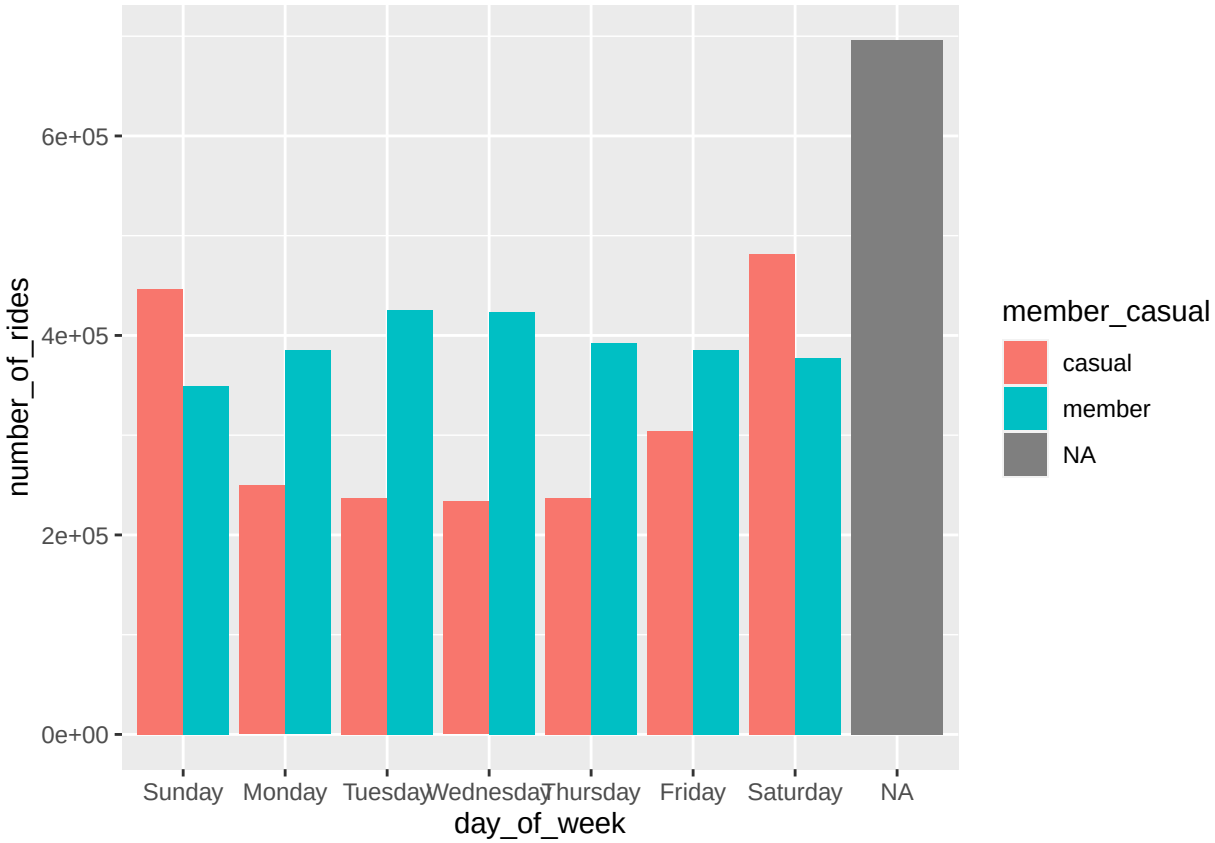
```
## 8                          member                    Wednesday
## 9                          casual                     Thursday
## 10                         member                     Thursday
## 11                         casual                       Friday
## 12                         member                       Friday
## 13                         casual                     Saturday
## 14                         member                     Saturday
##    all_trips_2021_v3$ride_length
## 1                      2318.1110
## 2                       942.0297
## 3                      1928.8068
## 4                       790.6884
## 5                      1745.5962
## 6                       762.3587
## 7                      1727.0750
## 8                       766.6581
## 9                      1705.6112
## 10                      765.1741
## 11                     1883.5499
## 12                      794.0223
## 13                     2131.5252
## 14                      918.5661
```

```
# Analyze ridership data by type and day_of_week
all_trips_2021_v3 %>%
  group_by(member_casual, day_of_week) %>% #groups by user type and day_of_week
  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, day_of_week) # sorts
```

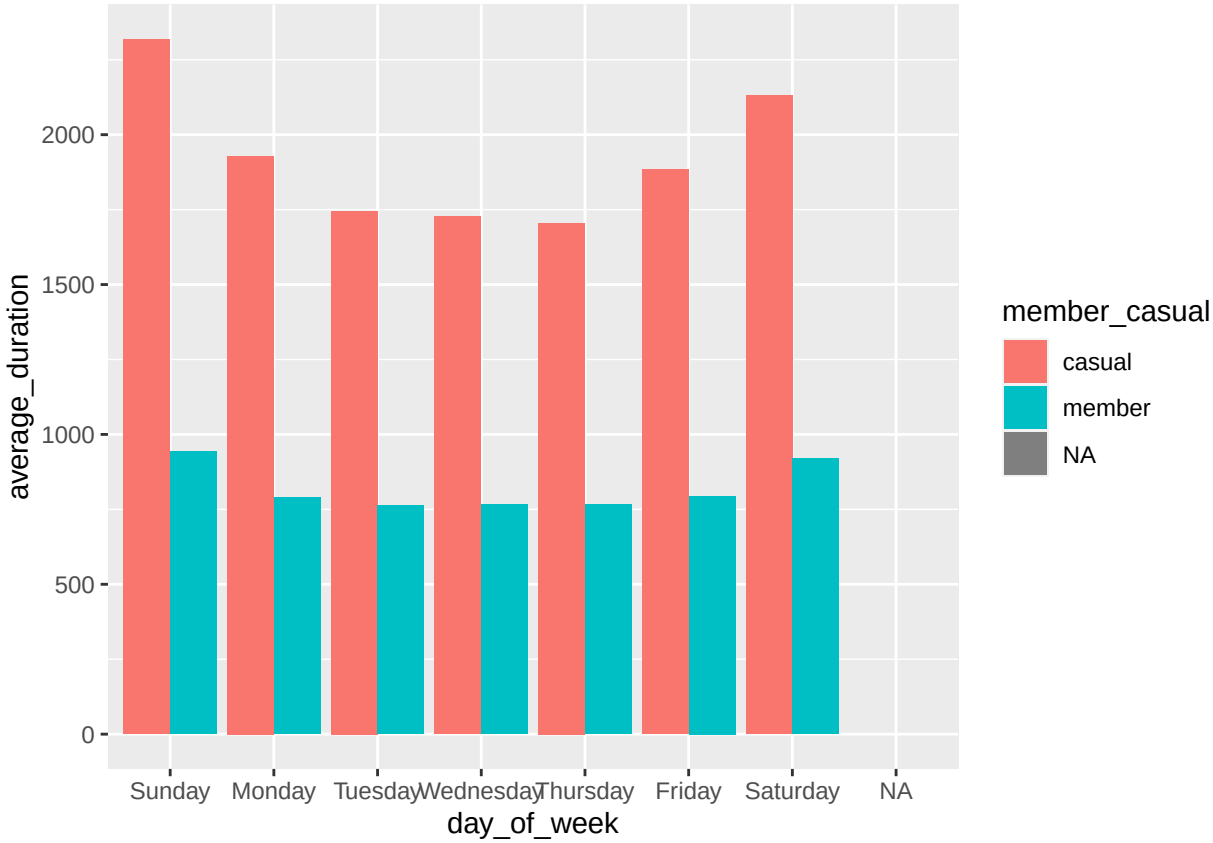
```
## # A tibble: 15 x 4
## # Groups:   member_casual [3]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                 <int>            <dbl>
##  1 casual        Sunday               446418            2318.
##  2 casual        Monday               249544            1929.
##  3 casual        Tuesday              236647            1746.
##  4 casual        Wednesday            233528            1727.
##  5 casual        Thursday             237290            1706.
##  6 casual        Friday               303948            1884.
##  7 casual        Saturday             481517            2132.
##  8 member        Sunday               349337             942.
##  9 member        Monday               384895             791.
## 10 member        Tuesday              425114             762.
## 11 member        Wednesday            423854             767.
## 12 member        Thursday             392780             765.
## 13 member        Friday               385054             794.
## 14 member        Saturday             377627             919.
## 15 <NA>          <NA>                 696565              NA
```

```
# Visualize the number of rides by rider type
all_trips_2021_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
```

```
  arrange(member_casual, day_of_week, na = TRUE) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```




```
# Create a visualization for average duration
all_trips_2021_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```




```
counts_2021 <- aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual + all_trips_202

counts_2021_1 <- aggregate(all_trips_2021_v3$ride_length ~ all_trips_2021_v3$member_casual + all_trips_1


## EXPORT THE DATA
write.csv(counts_2019,"C://Users//User//Desktop//GOOGLE DATA ANALYTICS RESOURCES//.csv folder//avg_ride_
write.csv(counts_2020,"C://Users//User//Desktop//GOOGLE DATA ANALYTICS RESOURCES//.csv folder//avg_ride_
write.csv(counts_2021,"C://Users//User//Desktop//GOOGLE DATA ANALYTICS RESOURCES//.csv folder//avg_ride_
write.csv(counts_2020_1,"C://Users//User//Desktop//GOOGLE DATA ANALYTICS RESOURCES//.csv folder//avg_ri
write.csv(counts_2021_1,"C://Users//User//Desktop//GOOGLE DATA ANALYTICS RESOURCES//.csv folder//avg_ri
```