Divvy DataSet

#Install Packages for working

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.0
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the  ]8;;http://conflicted.r-lib.org/ conflicted package ]8;;  to force all conflict
s to become errors
```

```
library(lubridate)
library(ggplot2)
```

Data Collection

```
q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
```

```
## Rows: 1108163 Columns: 12
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): 03 - Rental Start Station Name, 02 - Rental End Station Name, User...
## dbl  (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID, 03 - R...
## num  (1): 01 - Rental Details Duration In Seconds Uncapped
## dttm (2): 01 - Rental Details Local Start Time, 01 - Rental Details Local En...
##
## 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.
```

```
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
```

```
## Rows: 1640718 Columns: 12
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## 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.
```

```
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
```

```
## Rows: 704054 Columns: 12
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## 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.
```

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

```
## Rows: 426887 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## dttm (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.
```

Checking Structure of data

```
print('Q2 2019 Data')
```

```
## [1] "Q2 2019 Data"
```

```
colnames(q2_2019)
```

```
##  [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"
```

```
print('Q3 2019 Data')
```

```
## [1] "Q3 2019 Data"
```

```
colnames(q3_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"
```

```
print('Q4 2019 Data')
```

```
## [1] "Q4 2019 Data"
```

```
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"
```

```
print('Q1 2020 Data')
```

```
## [1] "Q1 2020 Data"
```

```
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"
```

Fixing Columns inconsistencies, taking reference lastest edition as new format.

```
(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))
```

ride_id <dbl>	started_at <dttm>	ended_at <dttm>	rideable_type <dbl>	tripduration <dbl>
25223640	2019-10-01 00:01:39	2019-10-01 00:17:20	2215	940
25223641	2019-10-01 00:02:16	2019-10-01 00:06:34	6328	258
25223642	2019-10-01 00:04:32	2019-10-01 00:18:43	3003	850
25223643	2019-10-01 00:04:32	2019-10-01 00:43:43	3275	2350
25223644	2019-10-01 00:04:34	2019-10-01 00:35:42	5294	1867
25223645	2019-10-01 00:04:38	2019-10-01 00:10:51	1891	373
25223646	2019-10-01 00:04:52	2019-10-01 00:22:45	1061	1072
25223647	2019-10-01 00:04:57	2019-10-01 00:29:16	1274	1458
25223648	2019-10-01 00:05:20	2019-10-01 00:29:18	6011	1437
25223649	2019-10-01 00:05:20	2019-10-01 02:23:46	2957	8306

1-10 of 10,000 rows | 1-6 of 12 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
(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))
```

ride_id <dbl>	started_at <dttm>	ended_at <dttm>	rideable_type <dbl>	tripduration <dbl>
23479388	2019-07-01 00:00:27	2019-07-01 00:20:41	3591	1214
23479389	2019-07-01 00:01:16	2019-07-01 00:18:44	5353	1048
23479390	2019-07-01 00:01:48	2019-07-01 00:27:42	6180	1554
23479391	2019-07-01 00:02:07	2019-07-01 00:27:10	5540	1503
23479392	2019-07-01 00:02:13	2019-07-01 00:22:26	6014	1213
23479393	2019-07-01 00:02:21	2019-07-01 00:07:31	4941	310
23479394	2019-07-01 00:02:24	2019-07-01 00:23:12	3770	1248
23479395	2019-07-01 00:02:26	2019-07-01 00:28:16	5442	1550
23479396	2019-07-01 00:02:34	2019-07-01 00:28:57	2957	1583
23479397	2019-07-01 00:02:45	2019-07-01 00:29:14	6091	1589

1-10 of 10,000 rows | 1-6 of 12 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
(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"))
```

ride_id <dbl>	**started_at** <dttm>	**ended_at** <dttm>	**rideable_type** <dbl>
22178529	2019-04-01 00:02:22	2019-04-01 00:09:48	6251
22178530	2019-04-01 00:03:02	2019-04-01 00:20:30	6226
22178531	2019-04-01 00:11:07	2019-04-01 00:15:19	5649
22178532	2019-04-01 00:13:01	2019-04-01 00:18:58	4151
22178533	2019-04-01 00:19:26	2019-04-01 00:36:13	3270
22178534	2019-04-01 00:19:39	2019-04-01 00:23:56	3123
22178535	2019-04-01 00:26:33	2019-04-01 00:35:41	6418
22178536	2019-04-01 00:29:48	2019-04-01 00:36:11	4513
22178537	2019-04-01 00:32:07	2019-04-01 01:07:44	3280
22178538	2019-04-01 00:32:19	2019-04-01 01:07:39	5534

1-10 of 10,000 rows | 1-4 of 12 columns Previous **1** 2 3 4 5 6 ... 1000 Next

Checking if dtypes are correct to join them together.

```
str(q1_2020)
```

```
## spc_tbl_ [426,887 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:426887] "EACB19130B0CDA4A" "8FED874C809DC021" "789F3C21E472C
A96" "C9A388DAC6ABF313" ...
##  $ rideable_type     : chr [1:426887] "docked_bike" "docked_bike" "docked_bike" "docked_bi
ke" ...
##  $ 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 & Belmont Ave" "Clark St & Randolph St" ...
##  $ 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" "Wilton Ave & Belmont Ave" "Fairbanks Ct & Grand Ave" ...
##  $ 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q4_2019)
```

```
## spc_tbl_ [704,054 × 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" "Milwaukee Ave & Grand Ave" "Lakeview Ave & Fullerton Pkwy" ...
##  $ 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" "W
abash Ave & Grand Ave" "Kedzie Ave & Palmer Ct" ...
##  $ 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2019)
```

```
## spc_tbl_ [1,640,718 × 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id          : num [1:1640718] 23479388 23479389 23479390 23479391 23479392 ...
##  $ started_at       : POSIXct[1:1640718], format: "2019-07-01 00:00:27" "2019-07-01 00:0
1:16" ...
##  $ ended_at         : POSIXct[1:1640718], format: "2019-07-01 00:20:41" "2019-07-01 00:1
8:44" ...
##  $ rideable_type    : num [1:1640718] 3591 5353 6180 5540 6014 ...
##  $ tripduration     : num [1:1640718] 1214 1048 1554 1503 1213 ...
##  $ start_station_id  : num [1:1640718] 117 381 313 313 168 300 168 313 43 43 ...
##  $ start_station_name: chr [1:1640718] "Wilton Ave & Belmont Ave" "Western Ave & Monroe S
t" "Lakeview Ave & Fullerton Pkwy" "Lakeview Ave & Fullerton Pkwy" ...
##  $ end_station_id   : num [1:1640718] 497 203 144 144 62 232 62 144 195 195 ...
##  $ end_station_name  : chr [1:1640718] "Kimball Ave & Belmont Ave" "Western Ave & 21st St"
"Larrabee St & Webster Ave" "Larrabee St & Webster Ave" ...
##  $ member_casual    : 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(q2_2019)
```

```
## spc_tbl_ [1,108,163 × 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id                                         : num [1:1108163] 22178529 22178530 221
78531 22178532 22178533 ...
##  $ started_at                                      : POSIXct[1:1108163], format: "2019-04-
01 00:02:22" "2019-04-01 00:03:02" ...
##  $ ended_at                                        : POSIXct[1:1108163], format: "2019-04-
01 00:09:48" "2019-04-01 00:20:30" ...
##  $ rideable_type                                   : num [1:1108163] 6251 6226 5649 4151 3
270 ...
##  $ 01 - Rental Details Duration In Seconds Uncapped: num [1:1108163] 446 1048 252 357 1007
...
##  $ start_station_id                                : num [1:1108163] 81 317 283 26 202 420
503 260 211 211 ...
##  $ start_station_name                              : chr [1:1108163] "Daley Center Plaza"
"Wood St & Taylor St" "LaSalle St & Jackson Blvd" "McClurg Ct & Illinois St" ...
##  $ end_station_id                                  : num [1:1108163] 56 59 174 133 129 426
500 499 211 211 ...
##  $ end_station_name                                : chr [1:1108163] "Desplaines St & Kinz
ie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...
##  $ member_casual                                   : chr [1:1108163] "Subscriber" "Subscri
ber" "Subscriber" "Subscriber" ...
##  $ Member Gender                                   : chr [1:1108163] "Male" "Female" "Mal
e" "Male" ...
##  $ 05 - Member Details Member Birthday Year        : num [1:1108163] 1975 1984 1990 1993 1
992 ...
##  - 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>
```

Converting numbers to characters, to merge together.

```
q4_2019 <-  mutate(q4_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q3_2019 <-  mutate(q3_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q2_2019 <-  mutate(q2_2019, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
```

Stacking together

```
all_trips <- bind_rows(q2_2019, q3_2019, q4_2019, q1_2020)
```

Removing redundancies from older formats (Using -c to keep all others columns.)

```
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, birthyear, gender, "01 - Rental Details D
uration In Seconds Uncapped", "05 - Member Details Member Birthday Year", "Member Gender", "t
ripduration"))
```

Inspecting health of the new dataset

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

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

```
## [1] 3879822
```

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

```
## [1] 3879822       9
```

```
head(all_trips)  #See the first 6 rows of data frame.  Also tail(all_trips)
```

ride_id	started_at	ended_at	rideable_type	start_station_id
<chr>	<dttm>	<dttm>	<chr>	<dbl>
22178529	2019-04-01 00:02:22	2019-04-01 00:09:48	6251	81
22178530	2019-04-01 00:03:02	2019-04-01 00:20:30	6226	317
22178531	2019-04-01 00:11:07	2019-04-01 00:15:19	5649	283
22178532	2019-04-01 00:13:01	2019-04-01 00:18:58	4151	26
22178533	2019-04-01 00:19:26	2019-04-01 00:36:13	3270	202
22178534	2019-04-01 00:19:39	2019-04-01 00:23:56	3123	420

6 rows | 1-5 of 9 columns

```
str(all_trips)  #See list of columns and data types (numeric, character, etc)
```

```
## tibble [3,879,822 × 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:0
3:02" ...
##  $ ended_at          : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:2
0:30" ...
##  $ rideable_type     : chr [1:3879822] "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle
St & Jackson Blvd" "McClurg Ct & Illinois St" ...
##  $ end_station_id    : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt
Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...
##  $ member_casual     : chr [1:3879822] "Subscriber" "Subscriber" "Subscriber" "Subscriber"
...
```

```
summary(all_trips)  #Statistical summary of data. Mainly for numerics
```

```
##    ride_id            started_at                   
##  Length:3879822     Min.   :2019-04-01 00:02:22.00  
##  Class :character   1st Qu.:2019-06-23 07:49:09.25  
##  Mode  :character   Median :2019-08-14 17:43:38.00  
##                     Mean   :2019-08-26 00:49:59.38  
##                     3rd Qu.:2019-10-12 12:10:21.00  
##                     Max.   :2020-03-31 23:51:34.00  
##                                                     
##     ended_at                      rideable_type      start_station_id
##  Min.   :2019-04-01 00:09:48.00   Length:3879822     Min.   :  1.0   
##  1st Qu.:2019-06-23 08:20:27.75   Class :character   1st Qu.: 77.0   
##  Median :2019-08-14 18:02:04.00   Mode  :character   Median :174.0   
##  Mean   :2019-08-26 01:14:37.06                      Mean   :202.9   
##  3rd Qu.:2019-10-12 12:36:16.75                      3rd Qu.:291.0   
##  Max.   :2020-05-19 20:10:34.00                      Max.   :675.0   
##                                                                      
##  start_station_name end_station_id  end_station_name   member_casual     
##  Length:3879822     Min.   :  1.0   Length:3879822     Length:3879822    
##  Class :character   1st Qu.: 77.0   Class :character   Class :character  
##  Mode  :character   Median :174.0   Mode  :character   Mode  :character  
##                     Mean   :203.8                                        
##                     3rd Qu.:291.0                                        
##                     Max.   :675.0                                        
##                     NA's   :1
```

Fixing names of members types to work with new format, the error has been seen while inspecting the data with view()

```
unique(all_trips$member_casual)
```

```
## [1] "Subscriber" "Customer"   "member"     "casual"
```

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

```
##
##     casual   Customer     member Subscriber
##      48480     857474     378407    2595461
```

Fixing names

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

Checking if fixed

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

```
##
##  casual  member
##  905954 2973868
```

Adding columns that will help us in the analysis of data

Helpful resource to do this (https://www.statmethods.net/input/dates.html (https://www.statmethods.net/input/dates.html))

```
all_trips$date <- as.Date(all_trips$started_at) #The default format is yyyy-mm-dd
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")
```

Adding calculated column (Ride_Length)

Resource to do this(https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html (https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html))

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
```

Checking for the result

```
str(all_trips)
```

```
## tibble [3,879,822 × 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:0
3:02" ...
##  $ ended_at          : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:2
0:30" ...
##  $ rideable_type     : chr [1:3879822] "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle
St & Jackson Blvd" "McClurg Ct & Illinois St" ...
##  $ end_station_id    : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt
Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...
##  $ member_casual     : chr [1:3879822] "member" "member" "member" "member" ...
##  $ date              : Date[1:3879822], format: "2019-04-01" "2019-04-01" ...
##  $ month             : chr [1:3879822] "04" "04" "04" "04" ...
##  $ day               : chr [1:3879822] "01" "01" "01" "01" ...
##  $ year              : chr [1:3879822] "2019" "2019" "2019" "2019" ...
##  $ day_of_week       : chr [1:3879822] "Monday" "Monday" "Monday" "Monday" ...
##  $ ride_length       : 'difftime' num [1:3879822] 446 1048 252 357 ...
##   ..- attr(*, "units")= chr "secs"
```

Ride_length has a format that I do not need.

```
is.factor(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
```

Final checking before cleaning

```
str(all_trips)
```

```
## tibble [3,879,822 × 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:0
3:02" ...
##  $ ended_at          : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:2
0:30" ...
##  $ rideable_type     : chr [1:3879822] "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle
St & Jackson Blvd" "McClurg Ct & Illinois St" ...
##  $ end_station_id    : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt
Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...
##  $ member_casual     : chr [1:3879822] "member" "member" "member" "member" ...
##  $ date              : Date[1:3879822], format: "2019-04-01" "2019-04-01" ...
##  $ month             : chr [1:3879822] "04" "04" "04" "04" ...
##  $ day               : chr [1:3879822] "01" "01" "01" "01" ...
##  $ year              : chr [1:3879822] "2019" "2019" "2019" "2019" ...
##  $ day_of_week       : chr [1:3879822] "Monday" "Monday" "Monday" "Monday" ...
##  $ ride_length       : num [1:3879822] 446 1048 252 357 1007 ...
```

Cleaning from bad data with emptys and negative values, seen by inspecting the data

Good practice is to do not replace the database and create a new version.

Useful Resources (https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/ (https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/))

```
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<
0),]
summary(all_trips_v2)
```

```
##    ride_id            started_at                    
##  Length:3876042     Min.   :2019-04-01 00:02:22.00  
##  Class :character   1st Qu.:2019-06-22 23:44:33.25  
##  Mode  :character   Median :2019-08-14 16:56:35.00  
##                     Mean   :2019-08-25 20:15:33.77  
##                     3rd Qu.:2019-10-11 23:23:20.75  
##                     Max.   :2020-03-31 23:51:34.00  
##     ended_at                     rideable_type      start_station_id
##  Min.   :2019-04-01 00:09:48.00   Length:3876042     Min.   :  1.0   
##  1st Qu.:2019-06-23 00:16:46.00   Class :character   1st Qu.: 77.0   
##  Median :2019-08-14 17:15:04.00   Mode  :character   Median :174.0   
##  Mean   :2019-08-25 20:40:12.92                      Mean   :202.4   
##  3rd Qu.:2019-10-12 00:26:13.50                      3rd Qu.:290.0   
##  Max.   :2020-05-19 20:10:34.00                      Max.   :673.0   
##  start_station_name end_station_id  end_station_name   member_casual     
##  Length:3876042     Min.   :  1.0   Length:3876042     Length:3876042    
##  Class :character   1st Qu.: 77.0   Class :character   Class :character  
##  Mode  :character   Median :174.0   Mode  :character   Mode  :character  
##                     Mean   :203.3                                        
##                     3rd Qu.:291.0                                        
##                     Max.   :675.0                                        
##       date              month              day                year          
##  Min.   :2019-04-01   Length:3876042     Length:3876042     Length:3876042    
##  1st Qu.:2019-06-22   Class :character   Class :character   Class :character  
##  Median :2019-08-14   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :2019-08-25                                                           
##  3rd Qu.:2019-10-11                                                           
##  Max.   :2020-03-31                                                           
##  day_of_week         ride_length     
##  Length:3876042     Min.   :      1  
##  Class :character   1st Qu.:    412  
##  Mode  :character   Median :    712  
##                     Mean   :   1479  
##                     3rd Qu.:   1289  
##                     Max.   :9387024
```

Descriptive analysis, but it can be done with summary too.

```
mean(all_trips_v2$ride_length) #straight average (total ride length / rides)
```

```
## [1] 1479.139
```

```
median(all_trips_v2$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 712
```

```
max(all_trips_v2$ride_length) #longest ride
```

```
## [1] 9387024
```

```
min(all_trips_v2$ride_length) #shortest ride
```

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

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     412     712    1479    1289 9387024
```

Compare stats for type of members

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

all_trips_v2$member_casual	**all_trips_v2$ride_length**
<chr>	<dbl>
casual	3552.7502
member	850.0662

2 rows

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

all_trips_v2$member_casual	**all_trips_v2$ride_length**
<chr>	<dbl>
casual	1546
member	589

2 rows

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

all_trips_v2$member_casual	**all_trips_v2$ride_length**
<chr>	<dbl>
casual	9387024
member	9056634

2 rows

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

all_trips_v2$member_casual	**all_trips_v2$ride_length**
<chr>	<dbl>
casual	2
member	1

2 rows

###Comparing statistics per day and per type of customer.

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, F
UN = mean)
```

all_trips_v2$member_casual <chr>	all_trips_v2$day_of_week <chr>	all_trips_v2$ride_length <dbl>
casual	Friday	3773.8351
member	Friday	824.5305
casual	Monday	3372.2869
member	Monday	842.5726
casual	Saturday	3331.9138
member	Saturday	968.9337
casual	Sunday	3581.4054
member	Sunday	919.9746
casual	Thursday	3682.9847
member	Thursday	823.9278

1-10 of 14 rows Previous **1** 2 Next

Not sorted

```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "T
uesday", "Wednesday", "Thursday", "Friday", "Saturday"))
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, F
UN = mean)
```

all_trips_v2$member_casual <chr>	all_trips_v2$day_of_week <ord>	all_trips_v2$ride_length <dbl>
casual	Sunday	3581.4054
member	Sunday	919.9746
casual	Monday	3372.2869
member	Monday	842.5726
casual	Tuesday	3596.3599
member	Tuesday	826.1427
casual	Wednesday	3718.6619
member	Wednesday	823.9996

all_trips_v2$member_casual	all_trips_v2$day_of_week	all_trips_v2$ride_length
<chr>	<ord>	<dbl>
casual	Thursday	3682.9847
member	Thursday	823.9278

1-10 of 14 rows Previous **1** 2 Next

analyze by type of member and day

```
df <- all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                           #calculates the number of rides a
nd average duration
  ,average_duration = mean(ride_length)) %>%        # calculates the average duration
  arrange(member_casual, weekday)                               # sorts
```

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

```
df
```

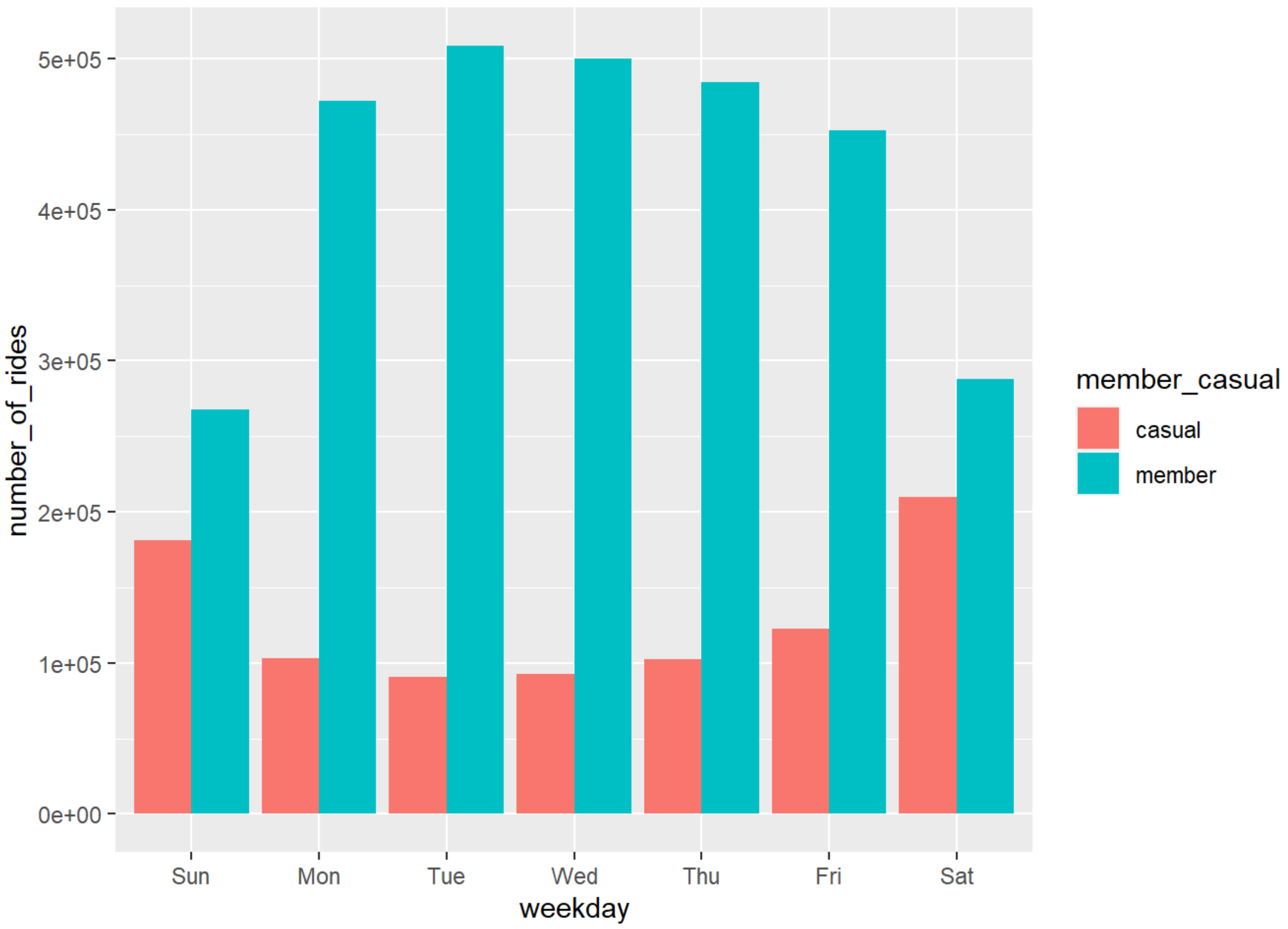
member_casual	weekday	number_of_rides	average_duration
<chr>	<ord>	<int>	<dbl>
casual	Sun	181293	3581.4054
casual	Mon	103296	3372.2869
casual	Tue	90510	3596.3599
casual	Wed	92457	3718.6619
casual	Thu	102679	3682.9847
casual	Fri	122404	3773.8351
casual	Sat	209543	3331.9138
member	Sun	267965	919.9746
member	Mon	472196	842.5726
member	Tue	508445	826.1427

1-10 of 14 rows Previous **1** 2 Next

#Visualize

```
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)  %>%
  #EXtra step to visualize
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

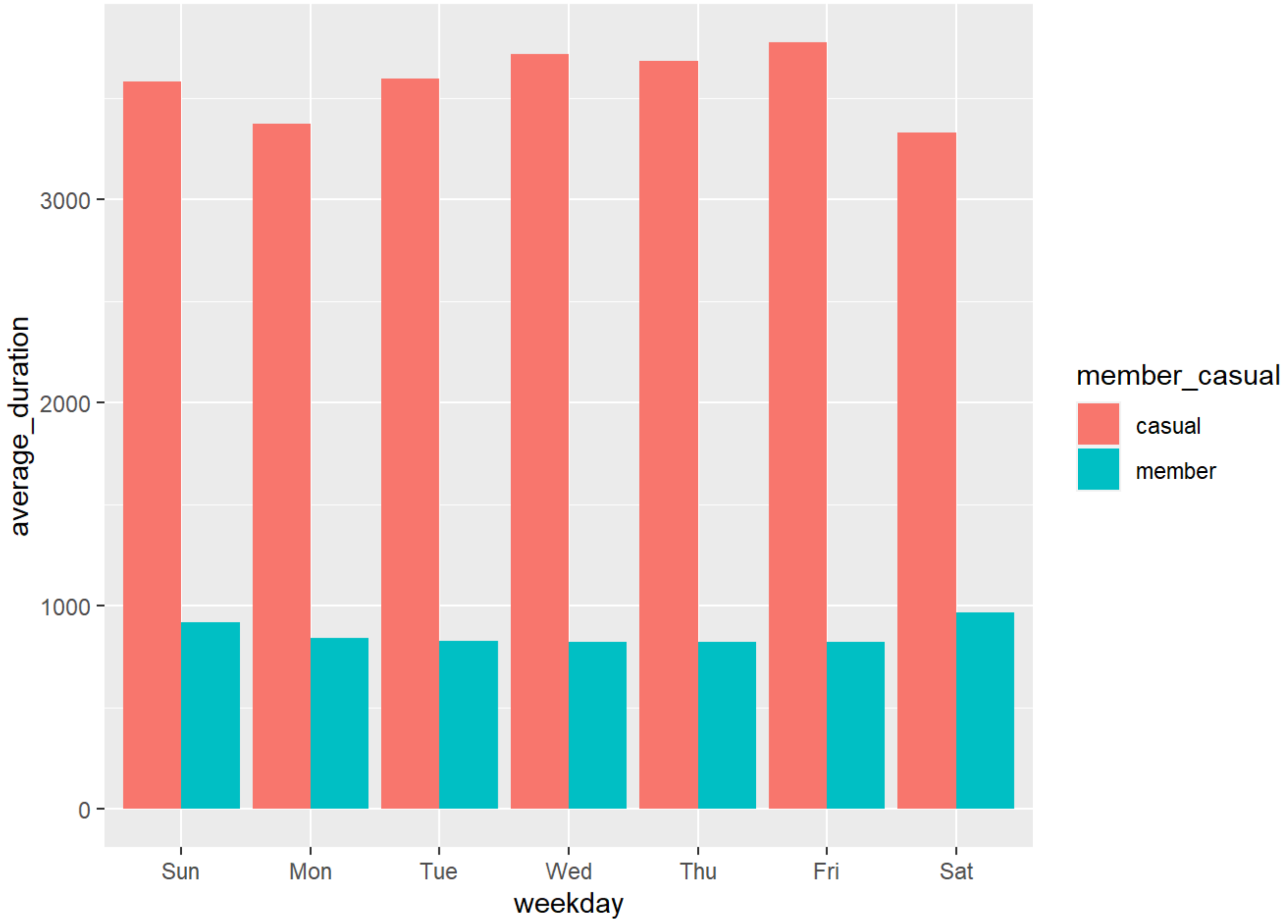
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Visualization for average duration 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)  %>%
  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.
```



### Exporting ### Resources : https://datatofish.com/export-dataframe-to-csv-in-r/ (https://datatofish.com/export-dataframe-to-csv-in-r/)

```
counts <- aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_
of_week, FUN = mean)
write.csv(counts, file = 'report.csv')
write.csv(df, file= "report2.csv")
```