Case Study: How Does a Bike-Share Navigate Speedy Success?

The main goal of this project is to find out differences between regular and casual users of Cyclistic bikes.

Project background:

Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can’t use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

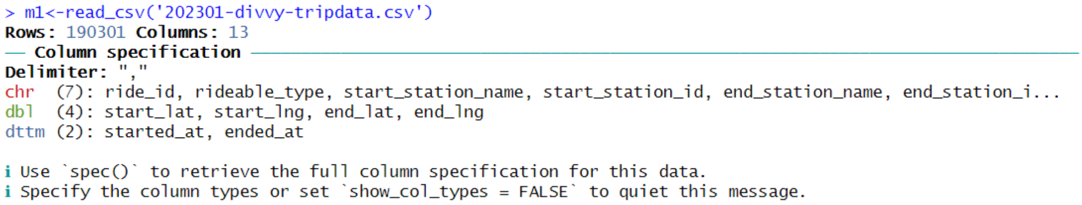
The main source of information is data found under this link. The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable you to answer the business questions. The data has been made available by Motivate International Inc.

The first step was to choose appropriate tools for this task. After some consideration I choose R-studio, because it allows me to relatively quickly process large quantities of data. Any tool that is based on spreadsheets does not have required capacity for this task.

Libraries used in this project are:

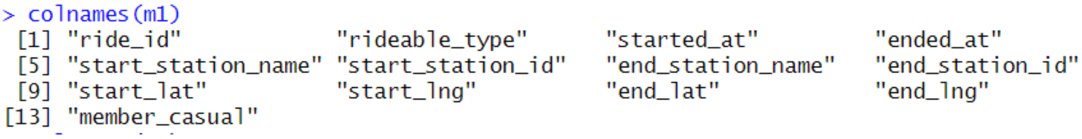
- tidyverse: helps with cleaning and solving conflicts within data,
- lubridate: helps with processing date attributes,
- ggplot2: helps to visualize data.

After setting up the libraries the first step was to import the data:



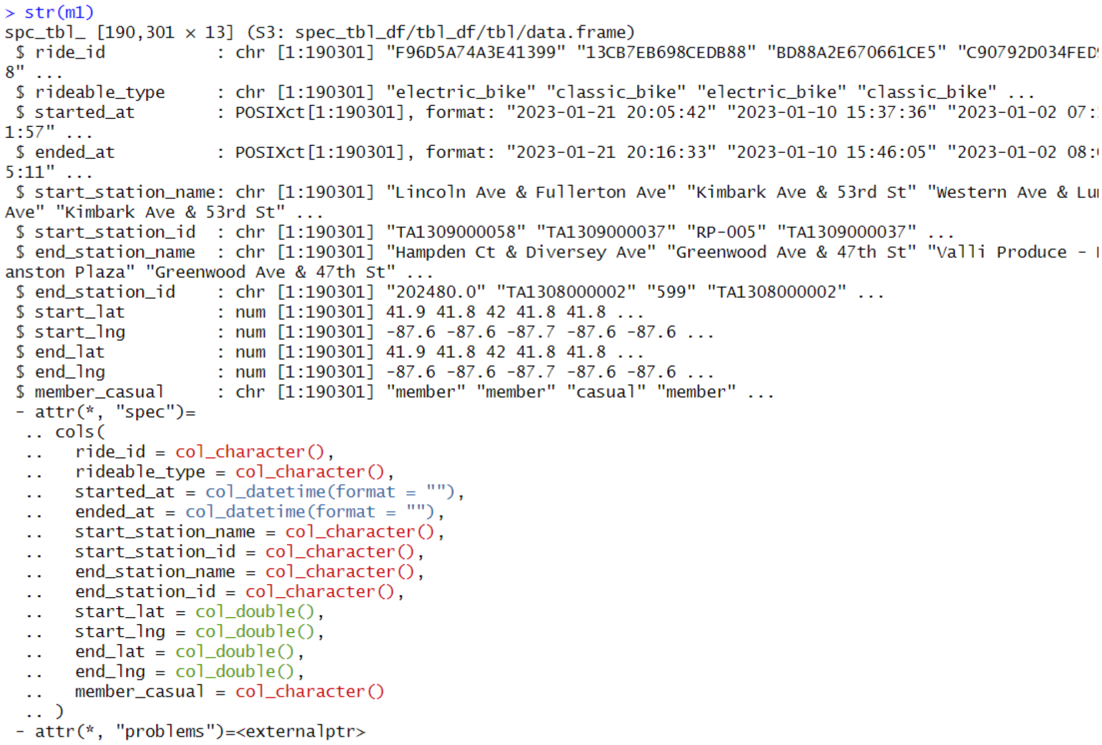
```
> m1<-read_csv('202301-divvy-tripdata.csv')
Rows: 190301 Columns: 13
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_station_name, end_station_i...
dbl  (4): start_lat, start_lng, end_lat, end_lng
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.
```

All of the csv files had the exact same labels:



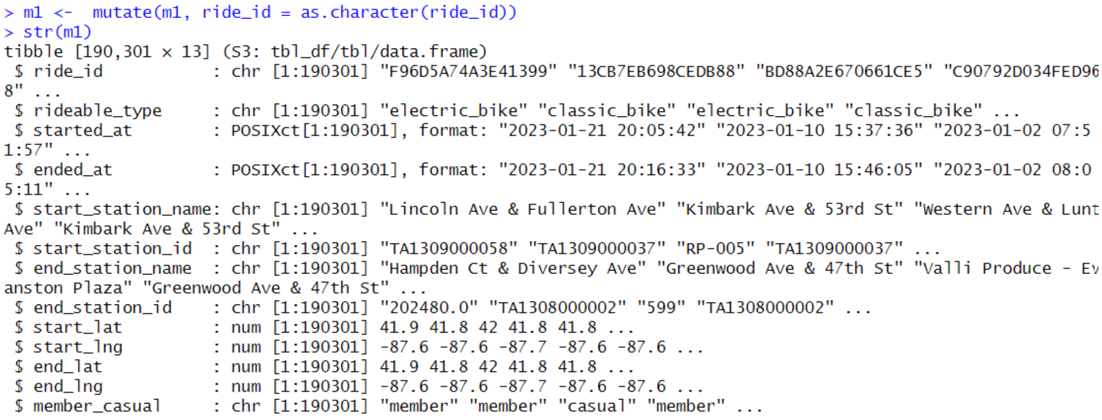
```
> colnames(m1)
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
 [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
 [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
[13] "member_casual"
```

Next part was to check data types in each of the files:



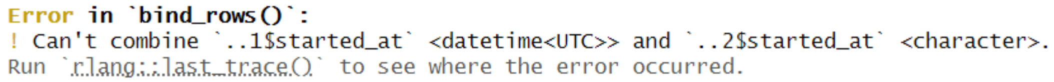
```
> str(m1)
spc_tbl_ [190,301 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ride_id           : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED
8" ...
 $ rideable_type     : chr [1:190301] "electric_bike" "classic_bike" "electric_bike" "classic_bike" ...
 $ started_at        : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" "2023-01-02 07:
1:57" ...
 $ ended_at          : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" "2023-01-02 08:
5:11" ...
 $ start_station_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western Ave & Lu
Ave" "Kimbark Ave & 53rd St" ...
 $ start_station_id  : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
 $ end_station_name  : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli Produce - 
anston Plaza" "Greenwood Ave & 47th St" ...
 $ end_station_id    : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...
 $ start_lat         : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
 $ start_lng         : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ end_lat           : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
 $ end_lng           : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr [1:190301] "member" "member" "casual" "member" ...
 - attr(*, "spec")=
  .. cols(
  ..   ride_id = col_character(),
  ..   rideable_type = col_character(),
  ..   started_at = col_datetime(format = ""),
  ..   ended_at = col_datetime(format = ""),
  ..   start_station_name = col_character(),
  ..   start_station_id = col_character(),
  ..   end_station_name = col_character(),
  ..   end_station_id = col_character(),
  ..   start_lat = col_double(),
  ..   start_lng = col_double(),
  ..   end_lat = col_double(),
  ..   end_lng = col_double(),
  ..   member_casual = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

After discovering that ride_id and rideable_type are not stackable the next step was to convert these columns to char type:



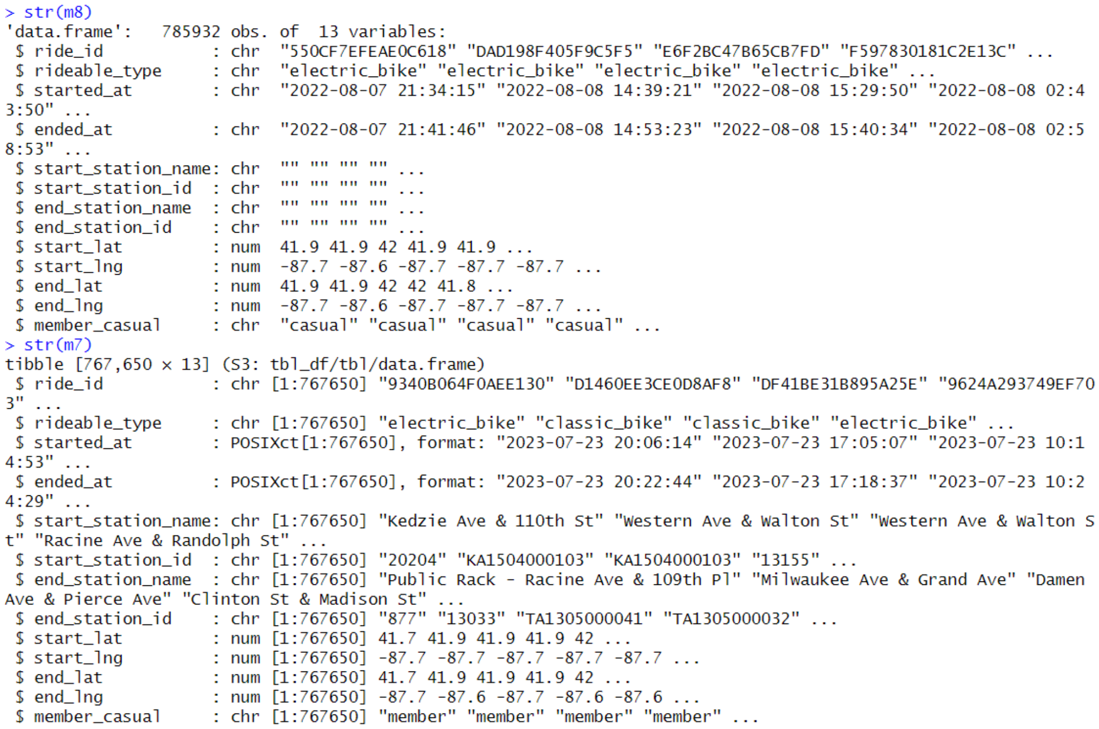
```
> m1 <-  mutate(m1, ride_id = as.character(ride_id))
> str(m1)
tibble [190,301 × 13] (S3: tbl_df/tbl/data.frame)
 $ ride_id           : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED96
8" ...
 $ rideable_type     : chr [1:190301] "electric_bike" "classic_bike" "electric_bike" "classic_bike" ...
 $ started_at        : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" "2023-01-02 07:5
1:57" ...
 $ ended_at          : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" "2023-01-02 08:0
5:11" ...
 $ start_station_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western Ave & Lunt
Ave" "Kimbark Ave & 53rd St" ...
 $ start_station_id  : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
 $ end_station_name  : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli Produce - Ev
anston Plaza" "Greenwood Ave & 47th St" ...
 $ end_station_id    : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...
 $ start_lat         : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
 $ start_lng         : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ end_lat           : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
 $ end_lng           : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr [1:190301] "member" "member" "casual" "member" ...
```

After attempting to merge all of the datasets an error has occurred:

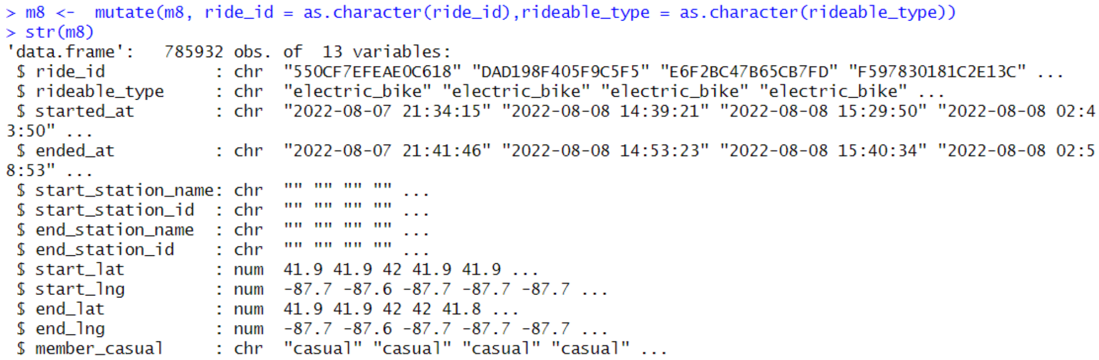


```
Error in `bind_rows()`:
! Can't combine `..1$started_at` <datetime<UTC>> and `..2$started_at` <character>.
Run `rlang::last_trace()` to see where the error occurred.
```

The cause was that file called m8 had columns “started at” and “ ended at” labelled as “char”, rather than POSIXsc (date format), so the next step was to convert it to appropriate format:

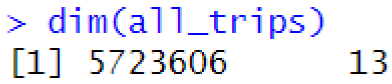


```
> str(m8)
'data.frame':	785932 obs. of  13 variables:
 $ ride_id           : chr  "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB7FD" "F597830181C2E13C" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2022-08-07 21:34:15" "2022-08-08 14:39:21" "2022-08-08 15:29:50" "2022-08-08 02:4
3:50" ...
 $ ended_at          : chr  "2022-08-07 21:41:46" "2022-08-08 14:53:23" "2022-08-08 15:40:34" "2022-08-08 02:5
8:53" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 42 42 41.8 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
> str(m7)
tibble [767,650 × 13] (S3: tbl_df/tbl/data.frame)
 $ ride_id           : chr [1:767650] "9340B064F0AEE130" "D1460EE3CE0D8AF8" "DF41BE31B895A25E" "9624A293749EF70
3" ...
 $ rideable_type     : chr [1:767650] "electric_bike" "classic_bike" "classic_bike" "electric_bike" ...
 $ started_at        : POSIXct[1:767650], format: "2023-07-23 20:06:14" "2023-07-23 17:05:07" "2023-07-23 10:1
4:53" ...
 $ ended_at          : POSIXct[1:767650], format: "2023-07-23 20:22:44" "2023-07-23 17:18:37" "2023-07-23 10:2
4:29" ...
 $ start_station_name: chr [1:767650] "Kedzie Ave & 110th St" "Western Ave & Walton St" "Western Ave & Walton S
t" "Racine Ave & Randolph St" ...
 $ start_station_id  : chr [1:767650] "20204" "KA1504000103" "KA1504000103" "13155" ...
 $ end_station_name  : chr [1:767650] "Public Rack - Racine Ave & 109th Pl" "Milwaukee Ave & Grand Ave" "Damen
Ave & Pierce Ave" "Clinton St & Madison St" ...
 $ end_station_id    : chr [1:767650] "877" "13033" "TA1305000041" "TA1305000032" ...
 $ start_lat         : num [1:767650] 41.7 41.9 41.9 41.9 42 ...
 $ start_lng         : num [1:767650] -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num [1:767650] 41.7 41.9 41.9 41.9 42 ...
 $ end_lng           : num [1:767650] -87.7 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr [1:767650] "member" "member" "member" "member" ...
```

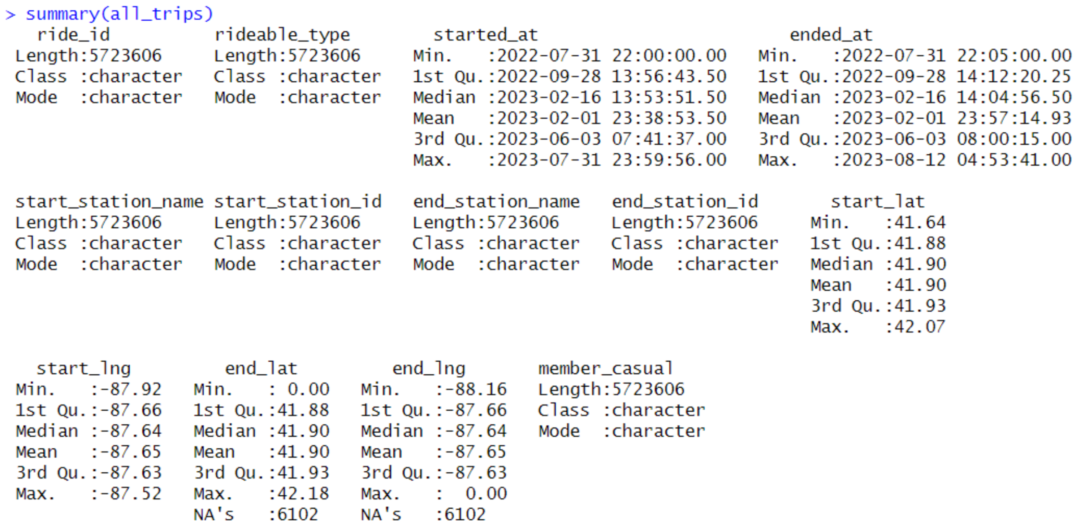


```
> m8 <-  mutate(m8, ride_id = as.character(ride_id),rideable_type = as.character(rideable_type))
> str(m8)
'data.frame':	785932 obs. of  13 variables:
 $ ride_id           : chr  "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB7FD" "F597830181C2E13C" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2022-08-07 21:34:15" "2022-08-08 14:39:21" "2022-08-08 15:29:50" "2022-08-08 02:4
3:50" ...
 $ ended_at          : chr  "2022-08-07 21:41:46" "2022-08-08 14:53:23" "2022-08-08 15:40:34" "2022-08-08 02:5
8:53" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 42 42 41.8 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

After merging I checked the dimensions and summary of the table:

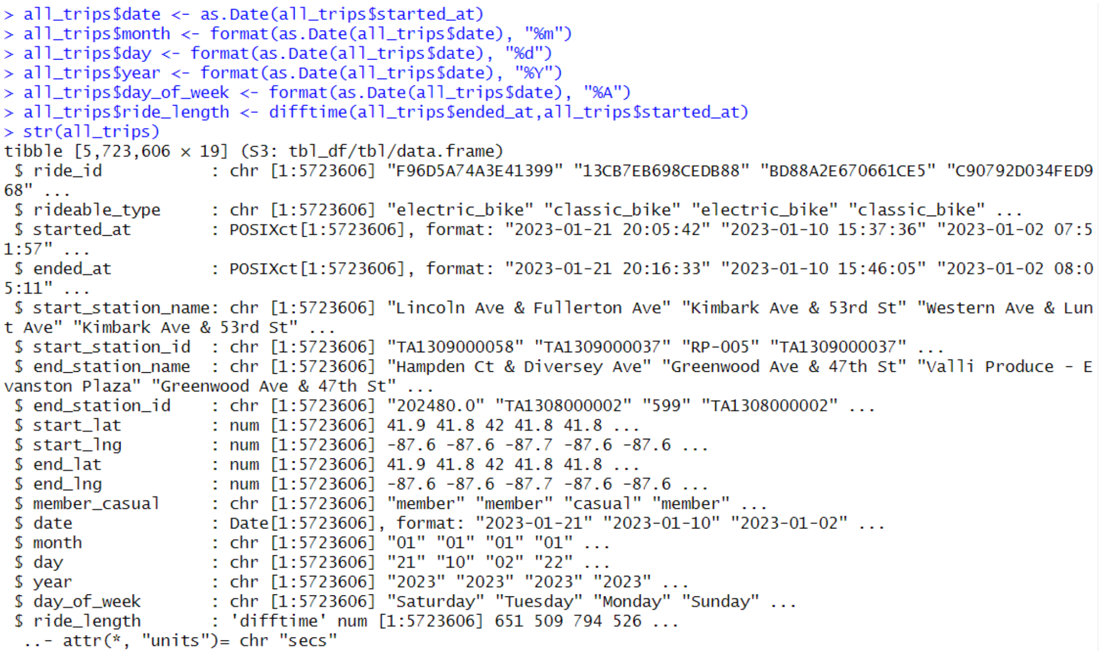


```
> dim(all_trips)
[1] 5723606      13
```



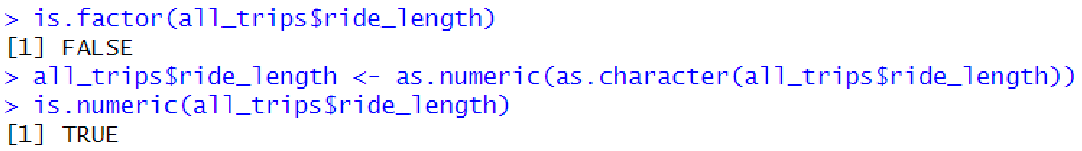
```
> summary(all_trips)
   ride_id          rideable_type        started_at                      ended_at                    
 Length:5723606     Length:5723606     Min.   :2022-07-31 22:00:00.00   Min.   :2022-07-31 22:05:00.00  
 Class :character   Class :character   1st Qu.:2022-09-28 13:56:43.50   1st Qu.:2022-09-28 14:12:20.25  
 Mode  :character   Mode  :character   Median :2023-02-16 13:53:51.50   Median :2023-02-16 14:04:56.50  
                                       Mean   :2023-02-01 23:38:53.50   Mean   :2023-02-01 23:57:14.93  
                                       3rd Qu.:2023-06-03 07:41:37.00   3rd Qu.:2023-06-03 08:00:15.00  
                                       Max.   :2023-07-31 23:59:56.00   Max.   :2023-08-12 04:53:41.00  
                                                                                                        
 start_station_name start_station_id   end_station_name   end_station_id       start_lat    
 Length:5723606     Length:5723606     Length:5723606     Length:5723606     Min.   :41.64  
 Class :character   Class :character   Class :character   Class :character   1st Qu.:41.88  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :41.90  
                                                                             Mean   :41.90  
                                                                             3rd Qu.:41.93  
                                                                             Max.   :42.07  
                                                                                            
   start_lng         end_lat         end_lng        member_casual     
 Min.   :-87.92   Min.   : 0.00   Min.   :-88.16   Length:5723606    
 1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66   Class :character  
 Median :-87.64   Median :41.90   Median :-87.64   Mode  :character  
 Mean   :-87.65   Mean   :41.90   Mean   :-87.65                     
 3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63                     
 Max.   :-87.52   Max.   :42.18   Max.   :  0.00                     
                  NA's   :6102    NA's   :6102                       
```

For easier analysis I decided to add several new columns, namely: year, month, day, day of week and length of ride.

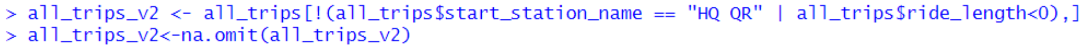


```
> all_trips$date <- as.Date(all_trips$started_at)
> all_trips$month <- format(as.Date(all_trips$date), "%m")
> all_trips$day <- format(as.Date(all_trips$date), "%d")
> all_trips$year <- format(as.Date(all_trips$date), "%Y")
> all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
> all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
> str(all_trips)
tibble [5,723,606 × 19] (S3: tbl_df/tbl/data.frame)
 $ ride_id           : chr [1:5723606] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED9
68" ...
 $ rideable_type     : chr [1:5723606] "electric_bike" "classic_bike" "electric_bike" "classic_bike" ...
 $ started_at        : POSIXct[1:5723606], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" "2023-01-02 07:5
1:57" ...
 $ ended_at          : POSIXct[1:5723606], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" "2023-01-02 08:0
5:11" ...
 $ start_station_name: chr [1:5723606] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western Ave & Lun
t Ave" "Kimbark Ave & 53rd St" ...
 $ start_station_id  : chr [1:5723606] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
 $ end_station_name  : chr [1:5723606] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli Produce - E
vanston Plaza" "Greenwood Ave & 47th St" ...
 $ end_station_id    : chr [1:5723606] "202480.0" "TA1308000002" "599" "TA1308000002" ...
 $ start_lat         : num [1:5723606] 41.9 41.8 42 41.8 41.8 ...
 $ start_lng         : num [1:5723606] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ end_lat           : num [1:5723606] 41.9 41.8 42 41.8 41.8 ...
 $ end_lng           : num [1:5723606] -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr [1:5723606] "member" "member" "casual" "member" ...
 $ date              : Date[1:5723606], format: "2023-01-21" "2023-01-10" "2023-01-02" ...
 $ month             : chr [1:5723606] "01" "01" "01" "01" ...
 $ day               : chr [1:5723606] "21" "10" "02" "22" ...
 $ year              : chr [1:5723606] "2023" "2023" "2023" "2023" ...
 $ day_of_week       : chr [1:5723606] "Saturday" "Tuesday" "Monday" "Sunday" ...
 $ ride_length       : 'difftime' num [1:5723606] 651 509 794 526 ...
  ..- attr(*, "units")= chr "secs"
```

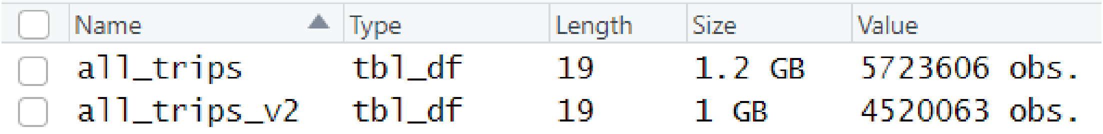
After that the next step was to clean up the data. After converting “ride length ” to numeric values, he first step was to clean all the rows where ride_lenth was negative. It reduced dataset by several hundred rows. The next step was to eliminate all rows with NA values.



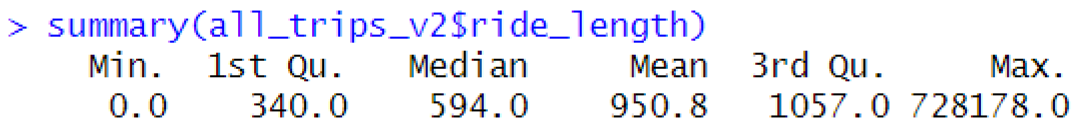
```
> is.factor(all_trips$ride_length)
[1] FALSE
> all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
> is.numeric(all_trips$ride_length)
[1] TRUE
```



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

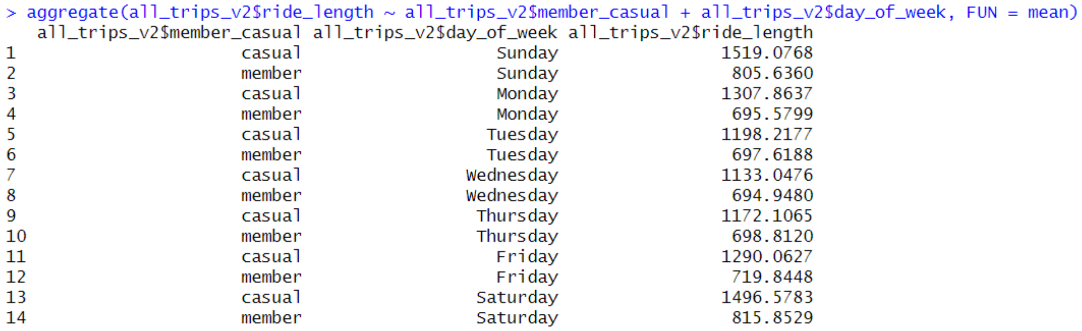


□	Name ▲	Type	Length	Size	Value
□	all_trips	tbl_df	19	1.2 GB	5723606 obs.
□	all_trips_v2	tbl_df	19	1 GB	4520063 obs.



```
> summary(all_trips_v2$ride_length)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
     0.0    340.0    594.0    950.8   1057.0 728178.0 
```

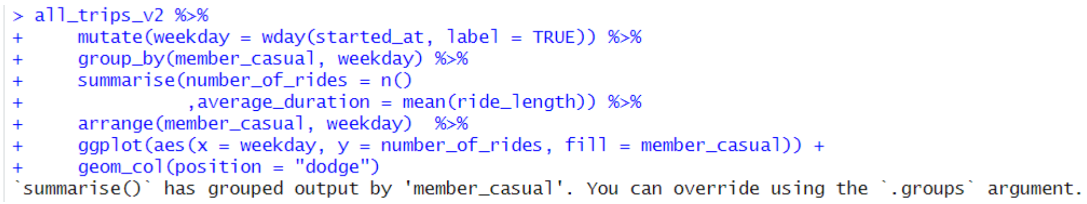
Final step was to calculate number of rides of both casual and member users and group the based on several conditions:



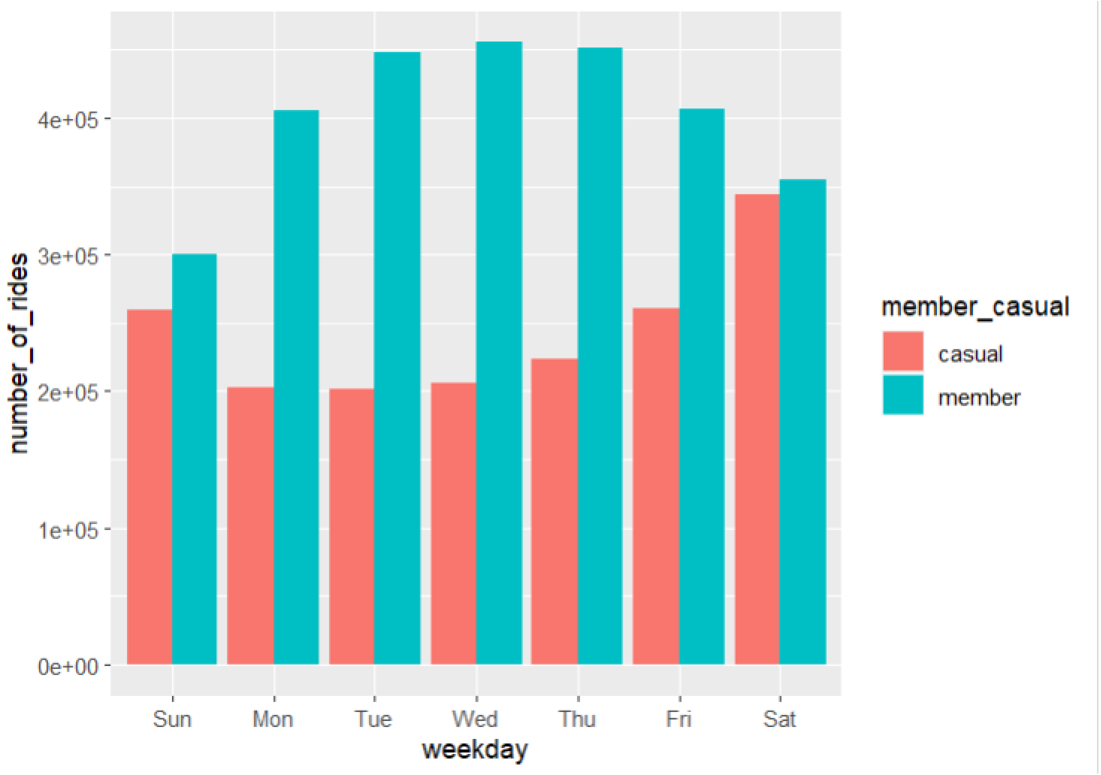
```
> aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
   all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
1                      casual                   Sunday                1519.0768
2                      member                   Sunday                 805.6360
3                      casual                   Monday                1307.8637
4                      member                   Monday                 695.5799
5                      casual                  Tuesday                1198.2177
6                      member                  Tuesday                 697.6188
7                      casual                Wednesday                1133.0476
8                      member                Wednesday                 694.9480
9                      casual                 Thursday                1172.1065
10                     member                 Thursday                 698.8120
11                     casual                   Friday                1290.0627
12                     member                   Friday                 719.8448
13                     casual                 Saturday                1496.5783
14                     member                 Saturday                 815.8529
```

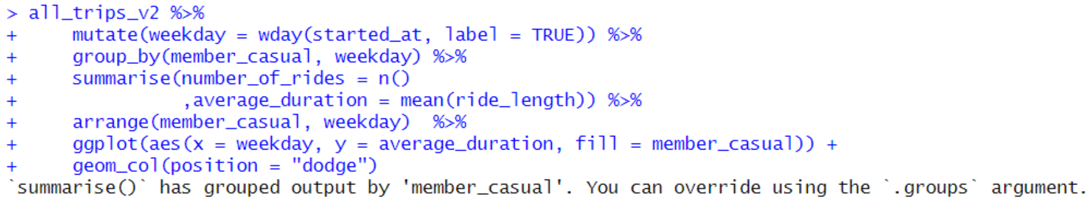
```
> all_trips_v2<-na.omit(all_trips_v2)
```

Plots generated with R-studio:



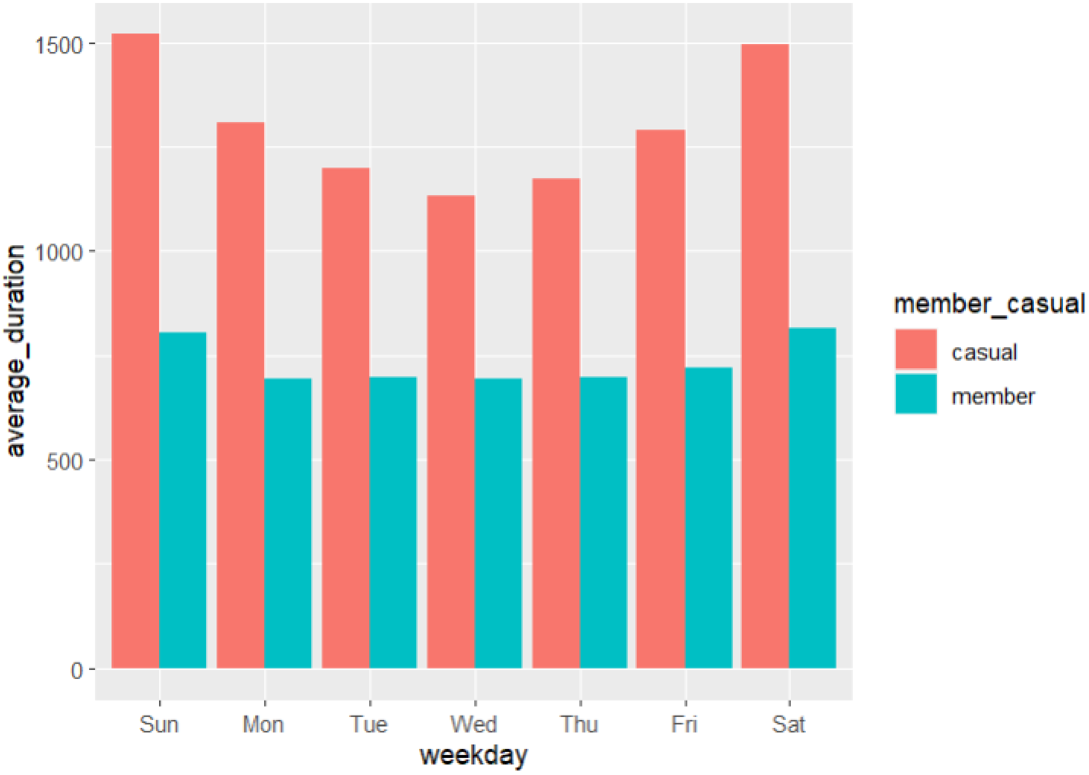
```
> all_trips_v2 %>%
+     mutate(weekday = wday(started_at, label = TRUE)) %>%
+     group_by(member_casual, weekday) %>%
+     summarise(number_of_rides = n()
+               ,average_duration = mean(ride_length)) %>%
+     arrange(member_casual, weekday)  %>%
+     ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
+     geom_col(position = "dodge")
`summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
```





```
> 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.
```

First plot shows the number of rides of casual and member clients depending on day of the week while second shows average length of rides.


1500
1000
500
0
average_duration
Sun
Mon
Tue
Wed
Thu
Fri
Sat
weekday
member_casual
casual
member