Google Data Analytics Capstone Project: Cyclistic Bike - Share Analysis Case Study

Aminuddin Hilmi Ahmad

10/10/2021

About the company

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.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Ask - The Business Task & Objective

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Lily Moreno, the marketing director, believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analysing the Cyclistic historical bike trip data to identify trends.

The main objective, identifying how do annual members and casual riders use Cyclistic bikes differently?

Prepare - The Dataset

The past data trip was obtained from here (https://divvy-tripdata.s3.amazonaws.com/index.html). Its a public data set prepared by the Motivate International Inc ("Motivate"), the bike - sharing company operated in Chicago, Illinois, USA. Since its a first party data sets, the data is considered as fulfilling the ROCCC requirement ie. the data is reliable, original, comprehensive, current, and cited.

The data set chosen were from January 2020 to December 2020. The 12 months data will give a better accuracy analysis in identifying the pattern of usage between the casual and member user.

Process - Cleaning The Data

The datasets were processed in RStudio. R was chosen due to its ability in processing the millions rows of data faster and its powerful 'Tidyverse' package that simplify a lot of data cleaning process. R also have 'ggplot2' package that can easily visualise the data in graphical form that help to comprehend the data faster

and efficiently.

The breakdown of the process are as follow;

```
##################################################

#Running the packages

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
```

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

```
library(here)
```

```
## here() starts at /Users/elina-mac/Documents/Data Analyst - Bootcamp/Google Courser
a/C8/CS1_R/CaseStudy1_v2
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

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

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

```
library(hms)
```

```
##
## Attaching package: 'hms'
```

```
## The following object is masked from 'package:lubridate':
##
##     hms
```

```
################################################

#Importing datasets

Q01_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.
```

```
M04_2020 <- read_csv("202004-divvy-tripdata.csv")
```

```
## Rows: 84776 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.
```

```
M05_2020 <- read_csv("202005-divvy-tripdata.csv")
```

```
## Rows: 200274 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.
```

```
M06_2020 <- read_csv("202006-divvy-tripdata.csv")
```

```
## Rows: 343005 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.
```

```
M07_2020 <- read_csv("202007-divvy-tripdata.csv")
```

```
## Rows: 551480 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.
```

```
M08_2020 <- read_csv("202008-divvy-tripdata.csv")
```

```
## Rows: 622361 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.
```

```
M09_2020 <- read_csv("202009-divvy-tripdata.csv")
```

```
## Rows: 532958 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.
```

```
M10_2020 <- read_csv("202010-divvy-tripdata.csv")
```

```
## Rows: 388653 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.
```

```
M11_2020 <- read_csv("202011-divvy-tripdata.csv")
```

```
## Rows: 259716 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.
```

```
M12_2020 <- read_csv("202012-divvy-tripdata.csv")
```

```
## Rows: 131573 Columns: 13
```

```
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## 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.
```

```
#Inspecting the datasets

str(Q01_2020)
```

```
## spec_tbl_df [426,887 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:426887] "EACB19130B0CDA4A" "8FED874C809DC021" "789F3
C21E472CA96" "C9A388DAC6ABF313" ...
##  $ rideable_type     : chr [1:426887] "docked_bike" "docked_bike" "docked_bike" "d
ocked_bike" ...
##  $ started_at        : POSIXct[1:426887], format: "2020-01-21 20:06:59" "2020-01-3
0 14:22:39" ...
##  $ ended_at          : POSIXct[1:426887], format: "2020-01-21 20:14:30" "2020-01-3
0 14:26:22" ...
##  $ start_station_name: chr [1:426887] "Western Ave & Leland Ave" "Clark St & Montr
ose 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 & Irv
ing 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(M04_2020)
```

```
## spec_tbl_df [84,776 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:84776] "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A
79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:84776] "docked_bike" "docked_bike" "docked_bike" "do
cked_bike" ...
##  $ started_at        : POSIXct[1:84776], format: "2020-04-26 17:45:14" "2020-04-17
17:08:54" ...
##  $ ended_at          : POSIXct[1:84776], format: "2020-04-26 18:12:03" "2020-04-17
17:17:03" ...
##  $ start_station_name: chr [1:84776] "Eckhart Park" "Drake Ave & Fullerton Ave" "M
cClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : num [1:84776] 86 503 142 216 125 173 35 434 627 377 ...
##  $ end_station_name  : chr [1:84776] "Lincoln Ave & Diversey Pkwy" "Kosciuszko Par
k" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : num [1:84776] 152 499 255 657 323 35 635 382 359 508 ...
##  $ start_lat         : num [1:84776] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:84776] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:84776] "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(M05_2020)
```

```
## spec_tbl_df [200,274 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:200274] "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCD
FDB91FE9A52" "58991CF1DB75BA84" ...
##  $ rideable_type     : chr [1:200274] "docked_bike" "docked_bike" "docked_bike" "d
ocked_bike" ...
##  $ started_at        : POSIXct[1:200274], format: "2020-05-27 10:03:52" "2020-05-2
5 10:47:11" ...
##  $ ended_at          : POSIXct[1:200274], format: "2020-05-27 10:16:49" "2020-05-2
5 11:05:40" ...
##  $ start_station_name: chr [1:200274] "Franklin St & Jackson Blvd" "Clark St & Wri
ghtwood Ave" "Kedzie Ave & Milwaukee Ave" "Clarendon Ave & Leland Ave" ...
##  $ start_station_id  : num [1:200274] 36 340 260 251 261 206 261 180 331 219 ...
##  $ end_station_name  : chr [1:200274] "Wabash Ave & Grand Ave" "Clark St & Leland
Ave" "Kedzie Ave & Milwaukee Ave" "Lake Shore Dr & Wellington Ave" ...
##  $ end_station_id    : num [1:200274] 199 326 260 157 206 22 261 180 300 305 ...
##  $ start_lat         : num [1:200274] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:200274] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:200274] 41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:200274] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:200274] "member" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M06_2020)
```

```
## spec_tbl_df [343,005 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:343005] "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D1
4B0B5659BCF" "C41237B506E85FA1" ...
##  $ rideable_type     : chr [1:343005] "docked_bike" "docked_bike" "docked_bike" "d
ocked_bike" ...
##  $ started_at        : POSIXct[1:343005], format: "2020-06-13 23:24:48" "2020-06-2
6 07:26:10" ...
##  $ ended_at          : POSIXct[1:343005], format: "2020-06-13 23:36:55" "2020-06-2
6 07:31:58" ...
##  $ start_station_name: chr [1:343005] "Wilton Ave & Belmont Ave" "Federal St & Pol
k St" "Daley Center Plaza" "Broadway & Cornelia Ave" ...
##  $ start_station_id  : num [1:343005] 117 41 81 303 327 327 41 115 338 84 ...
##  $ end_station_name  : chr [1:343005] "Damen Ave & Clybourn Ave" "Daley Center Pla
za" "State St & Harrison St" "Broadway & Berwyn Ave" ...
##  $ end_station_id    : num [1:343005] 163 81 5 294 117 117 81 303 164 53 ...
##  $ start_lat         : num [1:343005] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:343005] -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:343005] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:343005] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:343005] "casual" "member" "member" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M07_2020)
```

```
## spec_tbl_df [551,480 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:551480] "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8
EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr [1:551480] "docked_bike" "docked_bike" "docked_bike" "d
ocked_bike" ...
##  $ started_at        : POSIXct[1:551480], format: "2020-07-09 15:22:02" "2020-07-2
4 23:56:30" ...
##  $ ended_at          : POSIXct[1:551480], format: "2020-07-09 15:25:52" "2020-07-2
5 00:20:17" ...
##  $ start_station_name: chr [1:551480] "Ritchie Ct & Banks St" "Halsted St & Roscoe
St" "Lake Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id  : num [1:551480] 180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : chr [1:551480] "Wells St & Evergreen Ave" "Broadway & Ridge
Ave" "Clark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id    : num [1:551480] 291 461 156 94 301 289 140 31 191 142 ...
##  $ start_lat         : num [1:551480] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:551480] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:551480] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:551480] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:551480] "member" "member" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M08_2020)
```

```
## spec_tbl_df [622,361 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:622361] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1
D133E8B5816" "C79FBBD412E578A7" ...
##  $ rideable_type     : chr [1:622361] "docked_bike" "electric_bike" "electric_bik
e" "electric_bike" ...
##  $ started_at        : POSIXct[1:622361], format: "2020-08-20 18:08:14" "2020-08-2
7 18:46:04" ...
##  $ ended_at          : POSIXct[1:622361], format: "2020-08-20 18:17:51" "2020-08-2
7 19:54:51" ...
##  $ start_station_name: chr [1:622361] "Lake Shore Dr & Diversey Pkwy" "Michigan Av
e & 14th St" "Columbus Dr & Randolph St" "Daley Center Plaza" ...
##  $ start_station_id  : num [1:622361] 329 168 195 81 658 658 196 67 153 177 ...
##  $ end_station_name  : chr [1:622361] "Clark St & Lincoln Ave" "Michigan Ave & 14t
h St" "State St & Randolph St" "State St & Kinzie St" ...
##  $ end_station_id    : num [1:622361] 141 168 44 47 658 658 49 229 225 305 ...
##  $ start_lat         : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:622361] "member" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M09_2020)
```

```
## spec_tbl_df [532,958 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057
FA01BAC778E" "57F6DC9A153DB98C" ...
##  $ rideable_type     : chr [1:532958] "electric_bike" "electric_bike" "electric_bi
ke" "electric_bike" ...
##  $ started_at        : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-1
7 15:07:31" ...
##  $ ended_at          : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-1
7 15:07:45" ...
##  $ start_station_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N
Broadway" "W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...
##  $ start_station_id  : num [1:532958] 52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : chr [1:532958] "Green St & Randolph St" "W Oakdale Ave & N
Broadway" "W Oakdale Ave & N Broadway" "Montrose Harbor" ...
##  $ end_station_id    : num [1:532958] 112 NA NA 249 24 NA 256 NA NA NA ...
##  $ start_lat         : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:532958] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:532958] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:532958] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M10_2020)
```

```
## spec_tbl_df [388,653 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396
B54A15AC0DF" "44A4AEE261B9E854" ...
##  $ rideable_type     : chr [1:388653] "electric_bike" "electric_bike" "electric_bi
ke" "electric_bike" ...
##  $ started_at        : POSIXct[1:388653], format: "2020-10-31 19:39:43" "2020-10-3
1 23:50:08" ...
##  $ ended_at          : POSIXct[1:388653], format: "2020-10-31 19:57:12" "2020-11-0
1 00:04:16" ...
##  $ start_station_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport A
ve & Waveland Ave" "Stony Island Ave & 67th St" "Clark St & Grace St" ...
##  $ start_station_id  : num [1:388653] 313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : chr [1:388653] "Rush St & Hubbard St" "Kedzie Ave & Milwauk
ee Ave" "University Ave & 57th St" "Broadway & Sheridan Rd" ...
##  $ end_station_id    : num [1:388653] 125 260 423 256 185 53 125 313 199 635 ...
##  $ start_lat         : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:388653] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M11_2020)
```

```
## spec_tbl_df [259,716 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C6152
6D06582BDC5" "E533E89C32080B9E" ...
##  $ rideable_type     : chr [1:259716] "electric_bike" "electric_bike" "electric_bi
ke" "electric_bike" ...
##  $ started_at        : POSIXct[1:259716], format: "2020-11-01 13:36:00" "2020-11-0
1 10:03:26" ...
##  $ ended_at          : POSIXct[1:259716], format: "2020-11-01 13:45:40" "2020-11-0
1 10:14:45" ...
##  $ start_station_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illin
ois St" "Lake Shore Dr & Monroe St" "Leavitt St & Chicago Ave" ...
##  $ start_station_id  : num [1:259716] 110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : chr [1:259716] "St. Clair St & Erie St" "Noble St & Milwauk
ee Ave" "Federal St & Polk St" "Stave St & Armitage Ave" ...
##  $ end_station_id    : num [1:259716] 211 29 41 185 2 76 72 NA 288 273 ...
##  $ start_lat         : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:259716] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:259716] -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:259716] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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(M12_2020)
```

```
## spec_tbl_df [131,573 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "52620
16E0F1F2F9A" "BE119628E44F871E" ...
##  $ rideable_type     : chr [1:131573] "classic_bike" "electric_bike" "electric_bik
e" "electric_bike" ...
##  $ started_at        : POSIXct[1:131573], format: "2020-12-27 12:44:29" "2020-12-1
8 17:37:15" ...
##  $ ended_at          : POSIXct[1:131573], format: "2020-12-27 12:55:06" "2020-12-1
8 17:44:19" ...
##  $ start_station_name: chr [1:131573] "Aberdeen St & Jackson Blvd" NA NA NA ...
##  $ start_station_id  : chr [1:131573] "13157" NA NA NA ...
##  $ end_station_name  : chr [1:131573] "Desplaines St & Kinzie St" NA NA NA ...
##  $ end_station_id    : chr [1:131573] "TA1306000003" NA NA NA ...
##  $ start_lat         : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:131573] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:131573] -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:131573] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#data type conversion

M12_2020$start_station_id <- as.double(M12_2020$start_station_id)
```

```
## Warning: NAs introduced by coercion
```

```
M12_2020$end_station_id <- as.double(M12_2020$end_station_id)
```

```
## Warning: NAs introduced by coercion
```

```
#Combining into single data frame

BikeTrips_2020 <- bind_rows(Q01_2020, M04_2020, M05_2020, M06_2020, M07_2020,
                            M08_2020, M09_2020, M10_2020, M11_2020, M12_2020)

#Preview the new data frame

str(BikeTrips_2020)
```

```
## spec_tbl_df [3,541,683 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:3541683] "EACB19130B0CDA4A" "8FED874C809DC021" "789F
3C21E472CA96" "C9A388DAC6ABF313" ...
##  $ rideable_type     : chr [1:3541683] "docked_bike" "docked_bike" "docked_bike"
"docked_bike" ...
##  $ started_at        : POSIXct[1:3541683], format: "2020-01-21 20:06:59" "2020-01-
30 14:22:39" ...
##  $ ended_at          : POSIXct[1:3541683], format: "2020-01-21 20:14:30" "2020-01-
30 14:26:22" ...
##  $ start_station_name: chr [1:3541683] "Western Ave & Leland Ave" "Clark St & Mont
rose Ave" "Broadway & Belmont Ave" "Clark St & Randolph St" ...
##  $ start_station_id  : num [1:3541683] 239 234 296 51 66 212 96 96 212 38 ...
##  $ end_station_name  : chr [1:3541683] "Clark St & Leland Ave" "Southport Ave & Ir
ving Park Rd" "Wilton Ave & Belmont Ave" "Fairbanks Ct & Grand Ave" ...
##  $ end_station_id    : num [1:3541683] 326 318 117 24 212 96 212 212 96 100 ...
##  $ start_lat         : num [1:3541683] 42 42 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:3541683] -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:3541683] 42 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:3541683] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:3541683] "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>
```

```
#Removing unused columns

BikeTrips_2020 <- BikeTrips_2020%>% select(-c(start_lat, start_lng, end_lat, end_ln
g))

#Renaming columns and converting data types

BikeTrips_2020 <- BikeTrips_2020 %>% rename (
  Ride_ID = ride_id, Bike_Type = rideable_type, DateTime_Start = started_at,
  DateTime_End = ended_at, Start_Station_Name = start_station_name,
  Start_Station_ID = start_station_id, End_Station_Name = end_station_name,
  End_Station_ID = end_station_id, Membership_Type = member_casual)

BikeTrips_2020$Start_Station_ID = as.character(BikeTrips_2020$Start_Station_ID)

BikeTrips_2020$End_Station_ID = as.character(BikeTrips_2020$End_Station_ID)

is.character(BikeTrips_2020$Start_Station_ID)
```

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

```
#Reassign value for naming consistency

BikeTrips_2020 <- BikeTrips_2020 %>% mutate (Membership_Type = recode(Membership_Typ
e,
                                                                      "member" = "Memb
er",
                                                                      "casual" = "Casu
al"))

table (BikeTrips_2020$Membership_Type)
```

```
##
##  Casual  Member
## 1366575 2175108
```

```
BikeTrips_2020 <- BikeTrips_2020 %>% mutate (Bike_Type = recode(Bike_Type,
                                                                "classic_bike" = "Clas
sic_Bike",
                                                                "docked_bike" = "Docke
d_Bike",
                                                                "electric_bike" = "Ele
ctric_Bike"))

table (BikeTrips_2020$Bike_Type)
```

```
##
##  Classic_Bike   Docked_Bike Electric_Bike
##         70616       2966322        504745
```

```
#Separating date-time columns

BikeTrips_2020$Date <- as.Date(BikeTrips_2020$DateTime_Start) #The default format is
 yyyy-mm-dd
BikeTrips_2020$Day <- format(as.Date(BikeTrips_2020$Date), "%d")
BikeTrips_2020$Month <- format(as.Date(BikeTrips_2020$Date), "%m")
BikeTrips_2020$Year <- format(as.Date(BikeTrips_2020$Date), "%Y")
BikeTrips_2020$Day_Of_Week <- format(as.Date(BikeTrips_2020$Date), "%A")

#Calculating ride length in secs

BikeTrips_2020$Ride_Duration <- difftime(BikeTrips_2020$DateTime_End , BikeTrips_2020
$DateTime_Start)

#Convert to h:m:s format

BikeTrips_2020$Ride_Duration <- as.numeric(BikeTrips_2020$Ride_Duration)

BikeTrips_2020$Ride_Duration_HMS <- hms(BikeTrips_2020$Ride_Duration)



#Further cleaning

BikeTrips_2020 <- BikeTrips_2020[!(BikeTrips_2020$Start_Station_Name == "HQ QR" | Bik
eTrips_2020$Ride_Duration <= 0),]

#Preview data frame

skim(BikeTrips_2020)
```

Data summary

Name	BikeTrips_2020
Number of rows	3527000
Number of columns	16
_____	
Column type frequency:	
character	11
Date	1
difftime	1
numeric	1
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Ride_ID	94607	0.97	16	16	0	3432393	0
Bike_Type	94607	0.97	11	13	0	3	0
Start_Station_Name	94607	0.97	10	43	0	693	0
Start_Station_ID	163896	0.95	1	5	0	930	0
End_Station_Name	151249	0.96	5	43	0	692	0
End_Station_ID	216535	0.94	1	5	0	928	0
Membership_Type	94607	0.97	6	6	0	2	0
Day	94607	0.97	2	2	0	31	0
Month	94607	0.97	2	2	0	12	0
Year	94607	0.97	4	4	0	1	0
Day_Of_Week	94607	0.97	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Date	94607	0.97	2020-01-01	2020-12-31	2020-08-07	364

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Ride_Duration_HMS	94607	0.97	1 secs	9387024 secs	00:14:14	25809

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Ride_Duration	94607	0.97	1700.39	19891.83	1	467	854	1570	9387024	▇▁▁▁▁

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
DateTime_Start	94607	0.97	2020-01-01 00:04:44	2020-12-31 23:59:59	2020-08-07 16:04:38	3004497
DateTime_End	94607	0.97	2020-01-01 00:10:54	2021-01-03 08:54:11	2020-08-07 16:31:26	2990244

There original data set contains lots of n/a value. N/a value can lead to data inaccuracy and can skew the result the analysis. Therefore it is decided to remove n/a value for better consistency.

```
#Omit n.a. value
#Original rows = 3,527,002

cleanedBikeTrips_2020_v2 <- na.omit(BikeTrips_2020)


skim(cleanedBikeTrips_2020_v2)
```

Data summary

Name	cleanedBikeTrips_2020_v2
Number of rows	3284635
Number of columns	16
_____	
Column type frequency:	
character	11
Date	1
difftime	1
numeric	1
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Ride_ID	0	1	16	16	0	3284635	0
Bike_Type	0	1	11	13	0	3	0
Start_Station_Name	0	1	10	43	0	689	0
Start_Station_ID	0	1	1	5	0	920	0
End_Station_Name	0	1	5	43	0	689	0
End_Station_ID	0	1	1	5	0	926	0
Membership_Type	0	1	6	6	0	2	0
Day	0	1	2	2	0	31	0
Month	0	1	2	2	0	12	0
Year	0	1	4	4	0	1	0
Day_Of_Week	0	1	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Date	0	1	2020-01-01	2020-12-31	2020-08-03	364

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Ride_Duration_HMS	0	1	1 secs	9387024 secs	00:14:24	25191

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Ride_Duration	0	1	1705.85	20280.78	1	473	864	1583	9387024	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
DateTime_Start	0	1	2020-01-01 00:04:44	2020-12-31 23:59:59	2020-08-03 14:45:30	2873088
DateTime_End	0	1	2020-01-01 00:10:54	2021-01-02 22:03:22	2020-08-03 15:15:07	2858982

The original data, after being combine into single data frame, has 3,541,683 rows. After omitting the n/a value it has 3,284,635 rows which is about 93.13% from original data set, and it said to be good data left for analyse.

Analyse - Where The Data Tell The Stories

After the data has been cleaned, properly stored and backup, its time to analyse the data, as follows;

```
#Total number of user

table(cleanedBikeTrips_2020_v2$Membership_Type)
```

```
##
##  Casual  Member
## 1275052 2009583
```

```
#Descriptive summary

summary(cleanedBikeTrips_2020_v2)
```

```
##    Ride_ID           Bike_Type         DateTime_Start
##  Length:3284635     Length:3284635     Min.   :2020-01-01 00:04:44
##  Class :character   Class :character   1st Qu.:2020-06-14 06:02:28
##  Mode  :character   Mode  :character   Median :2020-08-03 14:45:30
##                                        Mean   :2020-07-21 10:38:33
##                                        3rd Qu.:2020-09-17 06:04:53
##                                        Max.   :2020-12-31 23:59:59
##   DateTime_End                 Start_Station_Name Start_Station_ID
##  Min.   :2020-01-01 00:10:54   Length:3284635     Length:3284635
##  1st Qu.:2020-06-14 07:14:06   Class :character   Class :character
##  Median :2020-08-03 15:15:07   Mode  :character   Mode  :character
##  Mean   :2020-07-21 11:06:59
##  3rd Qu.:2020-09-17 06:26:32
##  Max.   :2021-01-02 22:03:22
##  End_Station_Name   End_Station_ID     Membership_Type         Date
##  Length:3284635     Length:3284635     Length:3284635     Min.   :2020-01-01
##  Class :character   Class :character   Class :character   1st Qu.:2020-06-14
##  Mode  :character   Mode  :character   Mode  :character   Median :2020-08-03
##                                                           Mean   :2020-07-20
##                                                           3rd Qu.:2020-09-17
##                                                           Max.   :2020-12-31
##      Day               Month               Year           Day_Of_Week
##  Length:3284635     Length:3284635     Length:3284635     Length:3284635
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Ride_Duration     Ride_Duration_HMS
##  Min.   :      1   Length:3284635
##  1st Qu.:    473   Class1:hms
##  Median :    864   Class2:difftime
##  Mean   :   1706   Mode  :numeric
##  3rd Qu.:   1583
##  Max.   :9387024
```

```
#In h:m:s format

hms(sum (cleanedBikeTrips_2020_v2$Ride_Duration)) #sum of rides duration
```

```
## 1556414:15:52
```

```
hms(mean(cleanedBikeTrips_2020_v2$Ride_Duration)) # mean average of ride duration per user
```

```
## 00:28:25.849007
```

```
hms(median(cleanedBikeTrips_2020_v2$Ride_Duration)) # median/midpoint
```

```
## 00:14:24
```

```
hms(min(cleanedBikeTrips_2020_v2$Ride_Duration)) # minimum duration
```

```
## 00:00:01
```

```
hms(max(cleanedBikeTrips_2020_v2$Ride_Duration)) # maximum duration
```

```
## 2607:30:24
```

One thing to highlighted here, is the maximum time of user duration. One user was discovered to have illogical riding time ie. 2607 hours 30 minutes and 24 secs.

Now lets see the analysis of ride duration to different membership type.

```
#Analysis of ride duration to membership type

aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type, FUN = sum)
```

```
##   cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
##   cleanedBikeTrips_2020_v2$Ride_Duration
## 1                             3703562627
## 2                             1899528725
```

```
aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type, FUN = mean)
```

```
##   cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
##   cleanedBikeTrips_2020_v2$Ride_Duration
## 1                              2904.6365
## 2                               945.2353
```

```
aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type, FUN = median)
```

```
##   cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
##   cleanedBikeTrips_2020_v2$Ride_Duration
## 1                                   1329
## 2                                    674
```

```
aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type, FUN = min)
```

```
##   cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
##   cleanedBikeTrips_2020_v2$Ride_Duration
## 1                                      1
## 2                                      1
```

```
aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type, FUN = max)
```

```
##   cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
##   cleanedBikeTrips_2020_v2$Ride_Duration
## 1                                9387024
## 2                                5627611
```

Next are the analysis of average ride duration by the membership type to days and months.

```
#Rearranging the Day_Of_Week column

cleanedBikeTrips_2020_v2$Day_Of_Week <- ordered(cleanedBikeTrips_2020_v2$Day_Of_Week,
levels=c("Monday", "Tuesday",

"Wednesday", "Thursday",

"Friday", "Saturday", "Sunday"))

#Mean average of ride duration to membership type and day of ride

aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type +
            cleanedBikeTrips_2020_v2$Day_Of_Week, FUN = mean)
```

```
##    cleanedBikeTrips_2020_v2$Membership_Type
## 1                                   Casual
## 2                                   Member
## 3                                   Casual
## 4                                   Member
## 5                                   Casual
## 6                                   Member
## 7                                   Casual
## 8                                   Member
## 9                                   Casual
## 10                                  Member
## 11                                  Casual
## 12                                  Member
## 13                                  Casual
## 14                                  Member
##    cleanedBikeTrips_2020_v2$Day_Of_Week cleanedBikeTrips_2020_v2$Ride_Duration
## 1                                Monday                                2864.0035
## 2                                Monday                                 898.0897
## 3                               Tuesday                                2631.3315
## 4                               Tuesday                                 871.9058
## 5                             Wednesday                                2630.4427
## 6                             Wednesday                                 884.1340
## 7                              Thursday                                2865.0085
## 8                              Thursday                                 884.0633
## 9                                Friday                                2777.5442
## 10                               Friday                                 928.1886
## 11                             Saturday                                2951.2643
## 12                             Saturday                                1077.1921
## 13                               Sunday                                3295.4414
## 14                               Sunday                                1097.0944
```

```
#Mean average of ride duration to membership type and months

aggregate(cleanedBikeTrips_2020_v2$Ride_Duration ~ cleanedBikeTrips_2020_v2$Membershi
p_Type +
            cleanedBikeTrips_2020_v2$Month, FUN = mean)
```

```
##    cleanedBikeTrips_2020_v2$Membership_Type cleanedBikeTrips_2020_v2$Month
## 1                                   Casual                              01
## 2                                   Member                              01
## 3                                   Casual                              02
## 4                                   Member                              02
## 5                                   Casual                              03
## 6                                   Member                              03
## 7                                   Casual                              04
## 8                                   Member                              04
## 9                                   Casual                              05
## 10                                  Member                              05
## 11                                  Casual                              06
## 12                                  Member                              06
## 13                                  Casual                              07
## 14                                  Member                              07
## 15                                  Casual                              08
## 16                                  Member                              08
## 17                                  Casual                              09
## 18                                  Member                              09
## 19                                  Casual                              10
## 20                                  Member                              10
## 21                                  Casual                              11
## 22                                  Member                              11
## 23                                  Casual                              12
## 24                                  Member                              12
##    cleanedBikeTrips_2020_v2$Ride_Duration
## 1                               9698.9692
## 2                                668.9423
## 3                               7997.1646
## 4                                768.3972
## 5                               4250.2230
## 6                                860.0644
## 7                               4350.1767
## 8                               1282.1822
## 9                               3036.5618
## 10                              1175.3616
## 11                              3073.9902
## 12                              1112.4525
## 13                              3557.3681
## 14                              1054.4568
## 15                              2655.0364
## 16                               994.5785
## 17                              2305.2920
## 18                               916.6625
## 19                              1877.9625
## 20                               836.4760
## 21                              2008.8105
## 22                               809.4879
## 23                              1968.9016
## 24                               772.5130
```

Into the top 10 of stations of start and end.

```
#Top 10 start station

head(count(cleanedBikeTrips_2020_v2, Start_Station_Name, sort = TRUE), n = 10)
```

```
## # A tibble: 10 × 2
##    Start_Station_Name             n
##    <chr>                      <int>
##  1 Streeter Dr & Grand Ave    34769
##  2 Clark St & Elm St          30175
##  3 Theater on the Lake        28681
##  4 Lake Shore Dr & Monroe St  28573
##  5 Lake Shore Dr & North Blvd 25868
##  6 Wells St & Concord Ln      23845
##  7 Millennium Park            23743
##  8 Indiana Ave & Roosevelt Rd 23658
##  9 Dearborn St & Erie St      23324
## 10 Columbus Dr & Randolph St  23018
```

```
head(count(cleanedBikeTrips_2020_v2, Start_Station_Name, Membership_Type = "Member",
 sort = TRUE), n = 10)
```

```
## # A tibble: 10 × 3
##    Start_Station_Name         Membership_Type     n
##    <chr>                      <chr>           <int>
##  1 Streeter Dr & Grand Ave    Member          34769
##  2 Clark St & Elm St          Member          30175
##  3 Theater on the Lake        Member          28681
##  4 Lake Shore Dr & Monroe St  Member          28573
##  5 Lake Shore Dr & North Blvd Member          25868
##  6 Wells St & Concord Ln      Member          23845
##  7 Millennium Park            Member          23743
##  8 Indiana Ave & Roosevelt Rd Member          23658
##  9 Dearborn St & Erie St      Member          23324
## 10 Columbus Dr & Randolph St  Member          23018
```

```
head(count(cleanedBikeTrips_2020_v2, Start_Station_Name, Membership_Type = "Casual",
 sort = TRUE), n = 10)
```

```
## # A tibble: 10 × 3
##    Start_Station_Name         Membership_Type     n
##    <chr>                      <chr>           <int>
##  1 Streeter Dr & Grand Ave    Casual          34769
##  2 Clark St & Elm St          Casual          30175
##  3 Theater on the Lake        Casual          28681
##  4 Lake Shore Dr & Monroe St  Casual          28573
##  5 Lake Shore Dr & North Blvd Casual          25868
##  6 Wells St & Concord Ln      Casual          23845
##  7 Millennium Park            Casual          23743
##  8 Indiana Ave & Roosevelt Rd Casual          23658
##  9 Dearborn St & Erie St      Casual          23324
## 10 Columbus Dr & Randolph St  Casual          23018
```

```
#Top 10 end station

head(count(cleanedBikeTrips_2020_v2, End_Station_Name, sort = TRUE), n = 10)
```

```
## # A tibble: 10 × 2
##    End_Station_Name              n
##    <chr>                     <int>
##  1 Streeter Dr & Grand Ave    37346
##  2 Theater on the Lake        30335
##  3 Clark St & Elm St          30300
##  4 Lake Shore Dr & Monroe St  27926
##  5 Lake Shore Dr & North Blvd 26504
##  6 Millennium Park            25213
##  7 Wells St & Concord Ln      24205
##  8 Dearborn St & Erie St      23815
##  9 Broadway & Barry Ave       23795
## 10 St. Clair St & Erie St     23564
```

```
head(count(cleanedBikeTrips_2020_v2, End_Station_Name, Membership_Type = "Member", so
rt = TRUE), n = 10)
```

```
## # A tibble: 10 × 3
##    End_Station_Name           Membership_Type     n
##    <chr>                      <chr>           <int>
##  1 Streeter Dr & Grand Ave    Member          37346
##  2 Theater on the Lake        Member          30335
##  3 Clark St & Elm St          Member          30300
##  4 Lake Shore Dr & Monroe St  Member          27926
##  5 Lake Shore Dr & North Blvd Member          26504
##  6 Millennium Park            Member          25213
##  7 Wells St & Concord Ln      Member          24205
##  8 Dearborn St & Erie St      Member          23815
##  9 Broadway & Barry Ave       Member          23795
## 10 St. Clair St & Erie St     Member          23564
```

```
head(count(cleanedBikeTrips_2020_v2, End_Station_Name, Membership_Type = "Casual", so
rt = TRUE), n = 10)
```

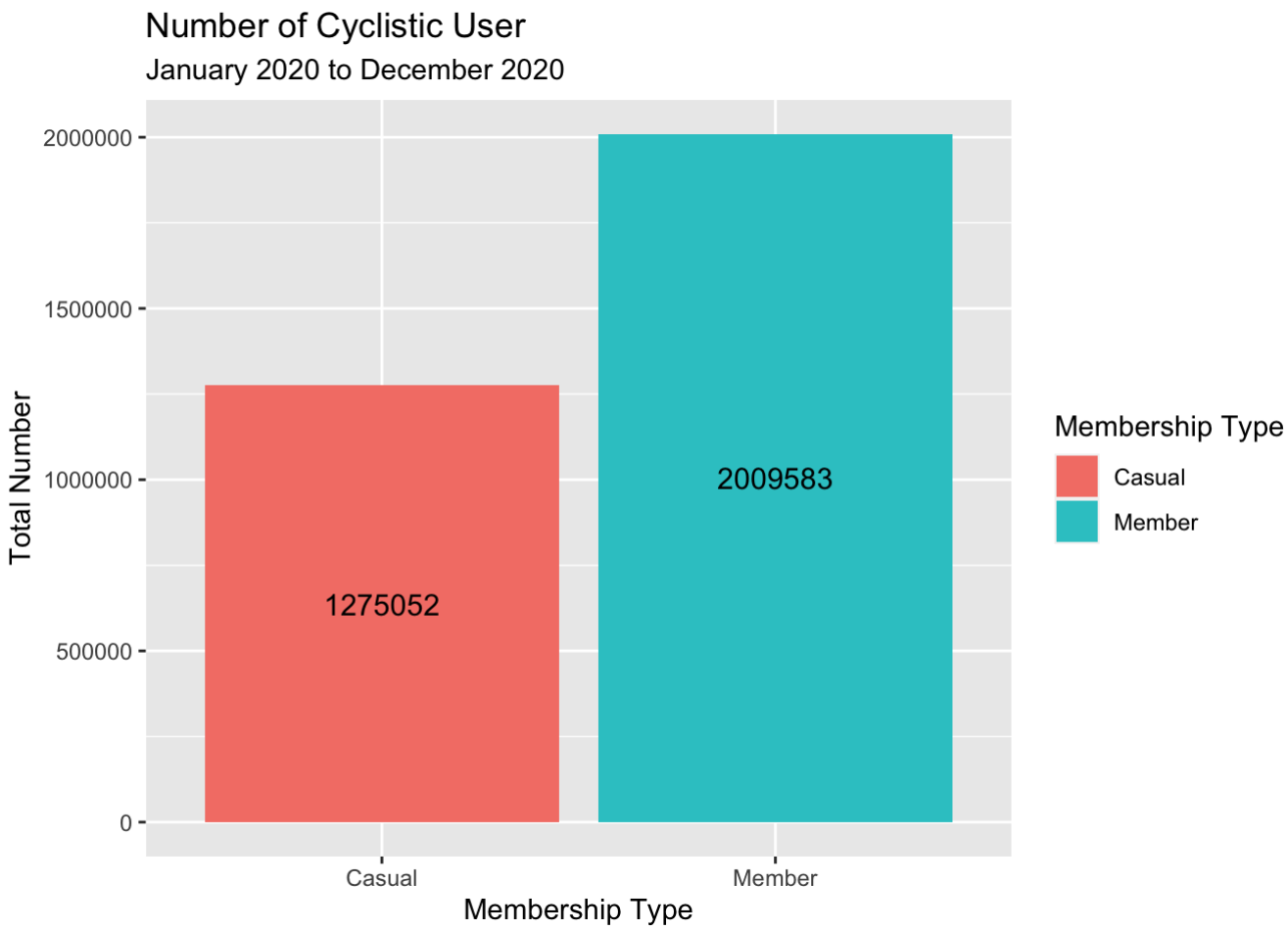
```
## # A tibble: 10 × 3
##    End_Station_Name           Membership_Type     n
##    <chr>                      <chr>           <int>
##  1 Streeter Dr & Grand Ave    Casual          37346
##  2 Theater on the Lake        Casual          30335
##  3 Clark St & Elm St          Casual          30300
##  4 Lake Shore Dr & Monroe St  Casual          27926
##  5 Lake Shore Dr & North Blvd Casual          26504
##  6 Millennium Park            Casual          25213
##  7 Wells St & Concord Ln      Casual          24205
##  8 Dearborn St & Erie St      Casual          23815
##  9 Broadway & Barry Ave       Casual          23795
## 10 St. Clair St & Erie St     Casual          23564
```

This is the top 10 of stations by users. Perhaps marketing department can conduct an aggressive marketing at those stations.

Share - Visualisation

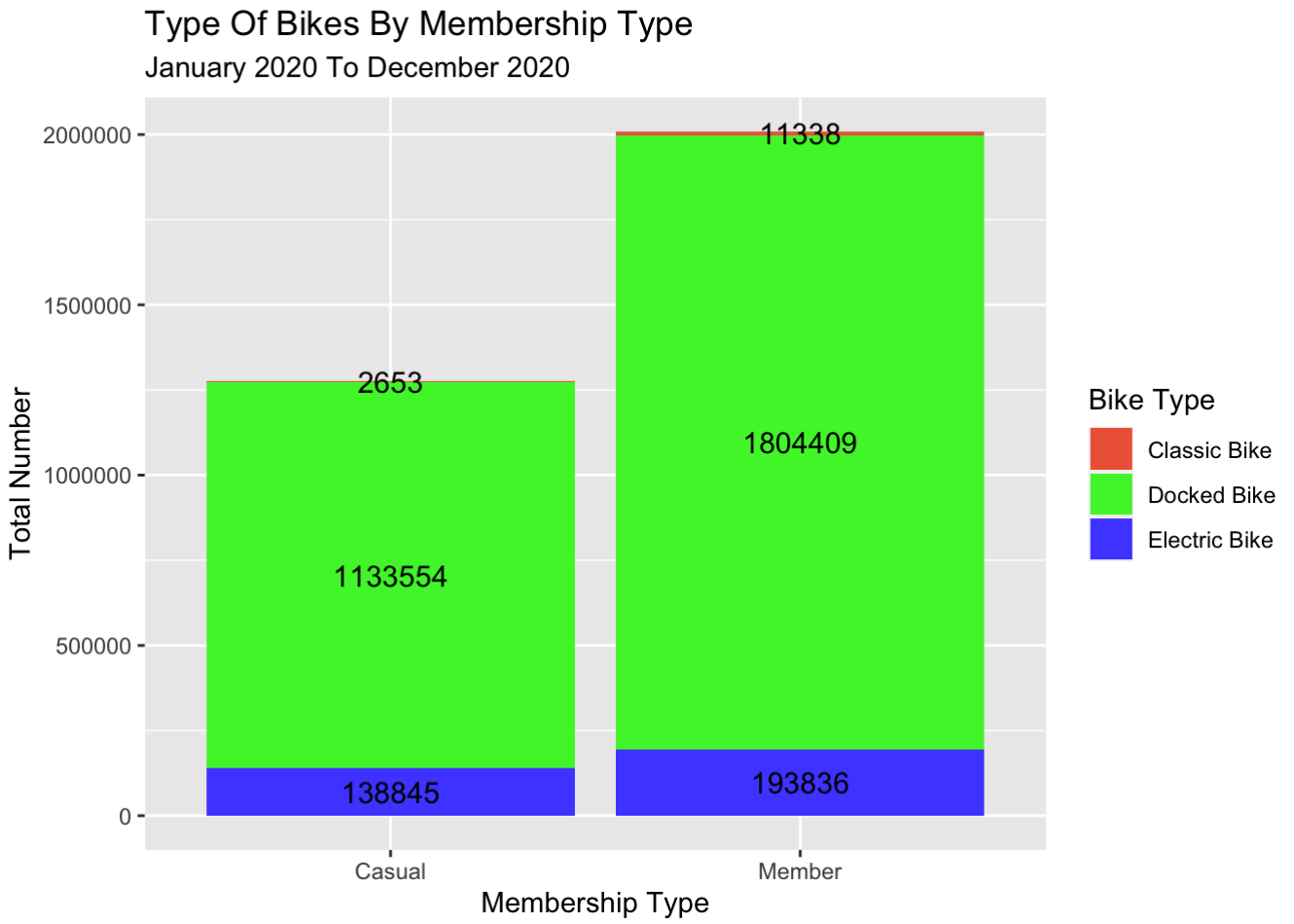
The total number of Cyclistic Bike - Share Service for 2020.

```
#Number of user
cleanedBikeTrips_2020_v2 %>%
  group_by(Membership_Type) %>%
  summarise(Total_Number = n()) %>%
  arrange(Total_Number) %>%

  ggplot(aes(x = Membership_Type, y = Total_Number, fill = Membership_Type)) +
  geom_bar(stat = "identity") +
  stat_identity(geom = "text", colour = "black", size = 4, aes(label = Total_Number),
            position = position_stack(vjust=0.5)) +
  scale_fill_manual(name = "Membership Type", values = c(Casual = '#F48176', Member =
'#30C7CC')) +
  labs(title = "Number of Cyclistic User" , x = "Membership Type", y = "Total Number"
,
       subtitle = "January 2020 to December 2020")
```



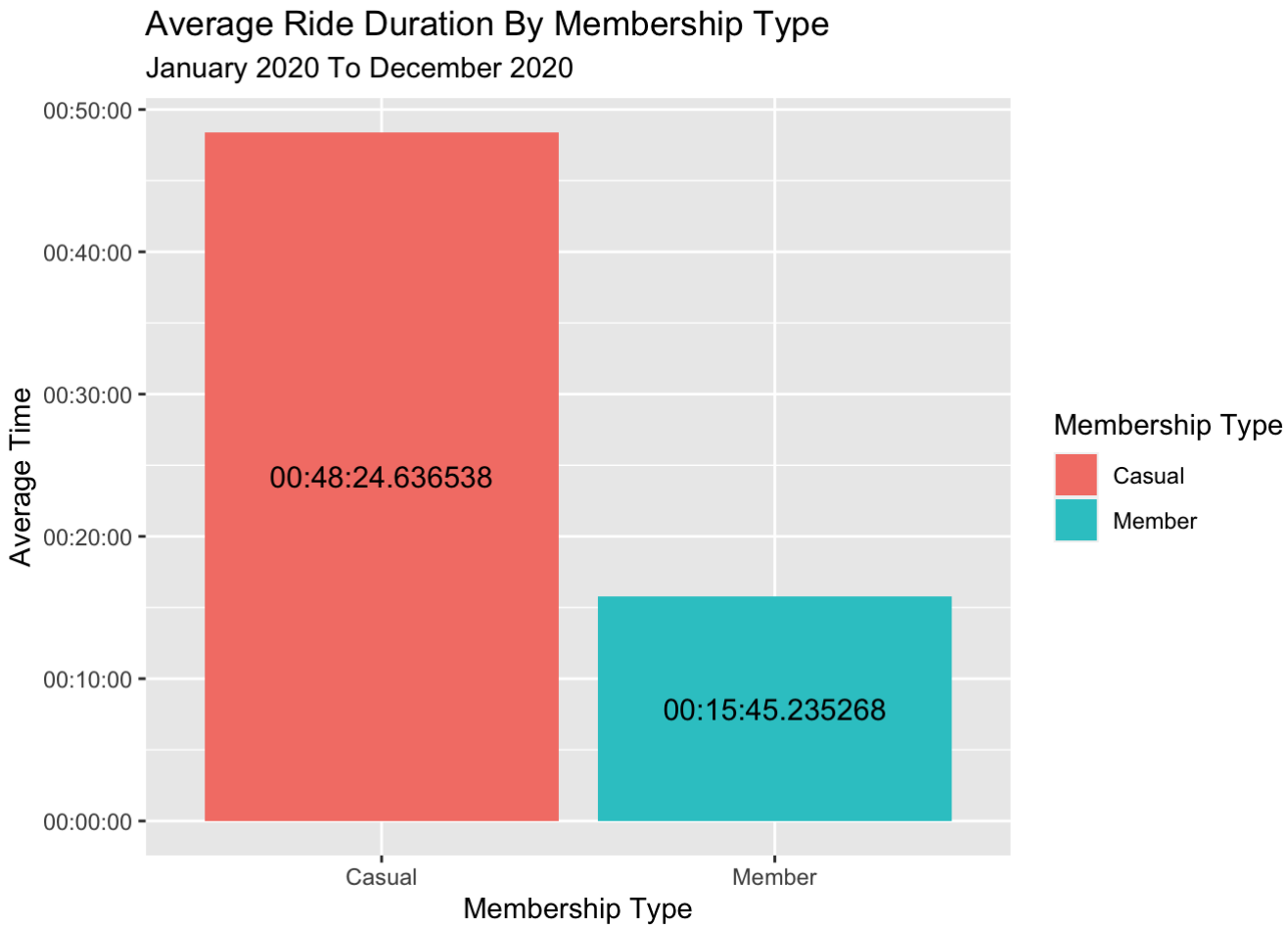
The bike types to membership types.

```
#Bike Types to Membership Type

cleanedBikeTrips_2020_v2 %>%
  group_by(Bike_Type, Membership_Type) %>%
  summarise(Total_Number = n(), .groups = 'drop') %>%
  arrange(Total_Number) %>%

  ggplot(aes(x = Membership_Type , y = Total_Number, fill = Bike_Type)) +
  geom_bar (stat="identity")+
  stat_identity(geom = "text", colour = "black", size = 4, aes(label = Total_Number),
                position = position_stack(vjust=0.5)) +
  scale_fill_manual(name = "Bike Type",
                    labels = c("Classic Bike", "Docked Bike", "Electric Bike"),
                    values = c("#ed6645", "#48f235", "#4f52ff")) +
  labs(title = "Type Of Bikes By Membership Type", x = "Membership Type", y = "Total
 Number",
       subtitle = "January 2020 To December 2020")
```



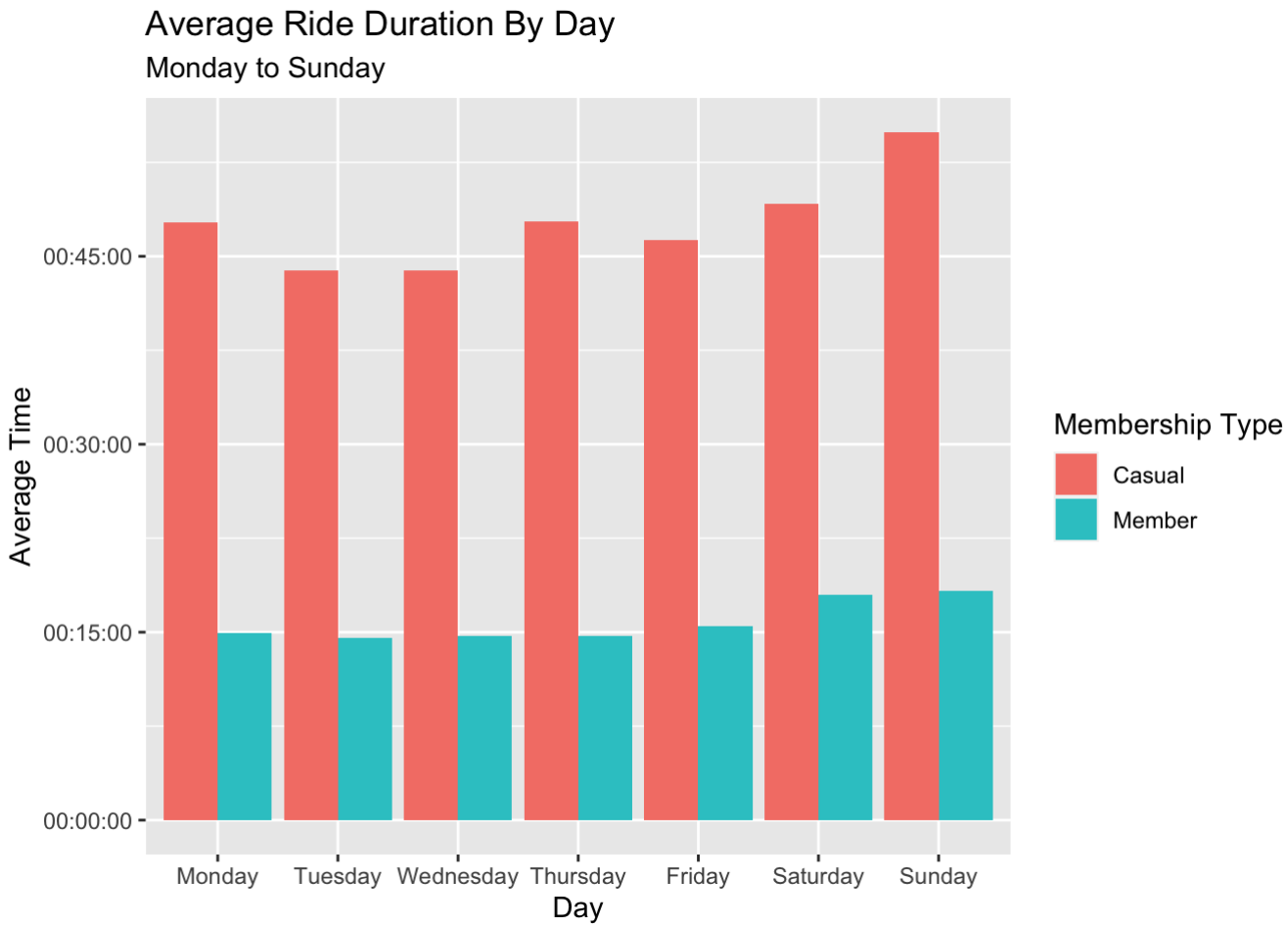
Mean average of membership type to ride duration

```
#Mean average of membership type to ride duration

cleanedBikeTrips_2020_v2 %>%
  mutate(Average_Times = mean(Ride_Duration)) %>%
  group_by(Membership_Type, Average_Times) %>%
  summarise(.groups = 'drop', Average_Times = mean(Ride_Duration)) %>%
  arrange(Membership_Type, Average_Times) %>% mutate (Average_Time= hms(Average_Time
s)) %>%

  ggplot(aes(x = Membership_Type , y = Average_Time, fill = Membership_Type)) +
  geom_bar (stat="identity")+
  stat_identity(geom = "text", colour = "black", size = 4, aes(label = Average_Time),
                position = position_stack(vjust=0.5)) +

  scale_fill_manual(name = "Membership Type",
                    values = c(Casual = '#F48176', Member = '#30C7CC')) +
  labs(title = "Average Ride Duration By Membership Type", x = "Membership Type",
       y = "Average Time",
       subtitle = "January 2020 To December 2020")
```



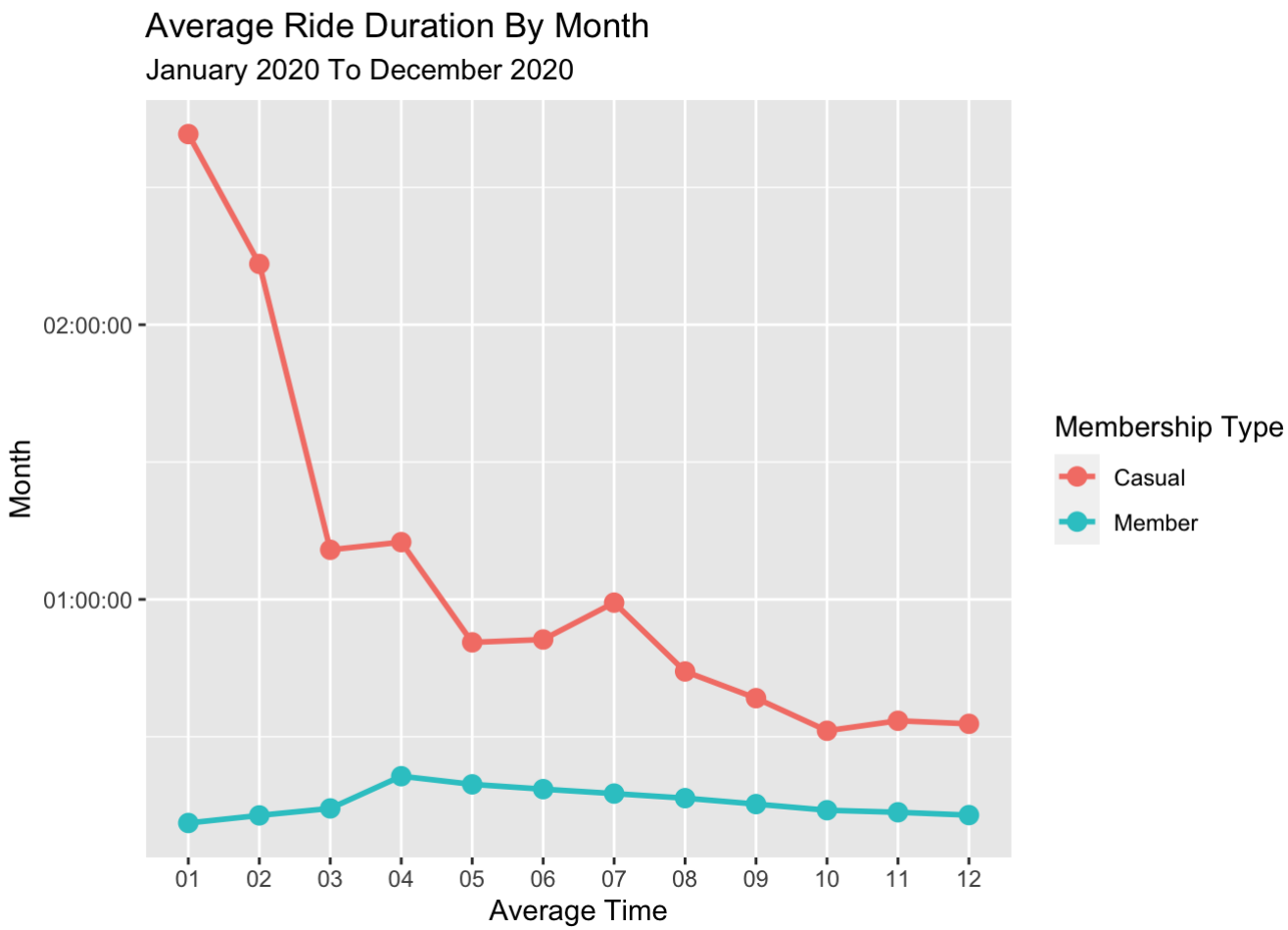
Mean average of ride duration to membership type and day of ride

```
cleanedBikeTrips_2020_v2 %>% mutate(Day = Day_Of_Week) %>%
  group_by(Membership_Type, Day) %>%
  summarise(.groups = 'drop', average_duration = mean(Ride_Duration)) %>%
  arrange(Membership_Type, Day) %>% mutate(Average_Time = hms(average_duration)) %>%

  ggplot(aes(x = Day, y = Average_Time, fill = Membership_Type)) +
  geom_col(position = 'dodge') +
  scale_fill_manual(name = "Membership Type",
                    values = c(Casual = '#F48176', Member = '#30C7CC')) +
  labs(title = "Average Ride Duration By Day", x = "Day", y = "Average Time",
       subtitle = "Monday to Sunday")
```



Mean average of ride duration to membership type and months

```
cleanedBikeTrips_2020_v2 %>% mutate(Month = Month) %>%
  group_by(Membership_Type, Month) %>%
  summarise(.groups = 'drop', average_duration = mean(Ride_Duration)) %>%
  arrange(Membership_Type, Month) %>% mutate(Average_Time = hms(average_duration)) %
>%

  ggplot(aes(x = Month, y = Average_Time, group = Membership_Type, colour = Membershi
p_Type)) +
  geom_line(size = 1) + geom_point(size = 3) +
  scale_colour_manual(name = "Membership Type",
                      values = c(Casual = '#F48176', Member = '#30C7CC')) +
  labs(title = "Average Ride Duration By Month", x = "Average Time", y = "Month",
       subtitle = "January 2020 To December 2020")
```



Act - Conclusion and Recommendation

From the analysis, it is clear that casual membership has a higher average of riding time compare to the member membership. The number is increasing during the weekend, which suggesting many casual member might be a tourists. The member membership usage pattern is almost flat and constant, which suggesting man of member membership are the resident.

To promote more casual member to register for a member membership, an aggressive marketing campaign can be done at top station for casual member. Other than that, casual member should be expose for the benefits of become a member.