Cyclistic Bike case study - Google data analytics project

Nhi Doan Hoai

DATA ANALYSIS PROCESS

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. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments.The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

ASK

The questions we needs to answer:

1. How do annual members and casual rides use Cyclistic bikes differently?

2. Why would casual riders buy Cyclistic annual memberships?

3. How can Cylistic use digital media to influence casual riders to become members?

PREPARE

The dataset The past data trip was obtained from here (https://divvy-tripdata.s3.amazonaws.com/index.html (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.

I chose the data set from April 2020 to March 2021 since it's lighten and fulls of a year which still gives us a better view about their business. However, it takes us a lot of time to download full 12 months and extract them. By default, they are .csv files.

PROCESS

Load library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

```
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
```

Load data

Now, check all the data structure to consider whether their data types are consistent or not

```
str(d1)
```

```
## 'data.frame':    96834 obs. of  13 variables:
##  $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-01-07 13:31:1
3" ...
##  $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-01-07 13:42:5
5" ...
##  $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez S
t" "California Ave & Cortez St" ...
##  $ start_station_id  : chr  "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(d2)
```

```
## 'data.frame':    49622 obs. of  13 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75B" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 19:10:18" "2021-02-02 17:49:4
1" ...
##  $ ended_at          : chr  "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 19:19:10" "2021-02-02 17:54:0
6" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St" "Wood St &
Chicago Ave" ...
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Randolph St" "Hon
ore St & Division St" ...
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num  42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```
str(d3)
```

```
## 'data.frame':    228496 obs. of  13 variables:
##  $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168F2" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-03-16 08:32:30" "2021-03-28 01:26:28" "2021-03-11 21:17:29" "2021-03-11 13:26:4
2" ...
##  $ ended_at          : chr  "2021-03-16 08:36:34" "2021-03-28 01:36:55" "2021-03-11 21:33:53" "2021-03-11 13:55:4
1" ...
##  $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields Ave & 28th Pl"
"Winthrop Ave & Lawrence Ave" ...
##  $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted St & 35th S
t" "Broadway & Sheridan Rd" ...
##  $ end_station_id    : chr  "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num  41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(d4)
```

```
## 'data.frame':    84776 obs. of  13 variables:
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-04-26 17:45:14" "2020-04-17 17:08:54" "2020-04-01 17:54:13" "2020-04-07 12:50:1
9" ...
##  $ ended_at          : chr  "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 18:08:36" "2020-04-07 13:02:3
1" ...
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "California Ave & D
ivision St" ...
##  $ start_station_id  : int  86 503 142 216 125 173 35 434 627 377 ...
##  $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St
& Augusta Blvd" ...
##  $ end_station_id    : int  152 499 255 657 323 35 635 382 359 508 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(d5)
```

```
## 'data.frame':    200274 obs. of  13 variables:
##  $ ride_id           : chr  "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA84" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-05-27 10:03:52" "2020-05-25 10:47:11" "2020-05-02 14:11:03" "2020-05-02 16:25:3
6" ...
##  $ ended_at          : chr  "2020-05-27 10:16:49" "2020-05-25 11:05:40" "2020-05-02 15:48:21" "2020-05-02 16:39:2
8" ...
##  $ start_station_name: chr  "Franklin St & Jackson Blvd" "Clark St & Wrightwood Ave" "Kedzie Ave & Milwaukee Ave"
"Clarendon Ave & Leland Ave" ...
##  $ start_station_id  : int  36 340 260 251 261 206 261 180 331 219 ...
##  $ end_station_name  : chr  "Wabash Ave & Grand Ave" "Clark St & Leland Ave" "Kedzie Ave & Milwaukee Ave" "Lake S
hore Dr & Wellington Ave" ...
##  $ end_station_id    : int  199 326 260 157 206 22 261 180 300 305 ...
##  $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

```
str(d6)
```

```
## 'data.frame':    343005 obs. of  13 variables:
##  $ ride_id           : chr  "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85FA1" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-06-13 23:24:48" "2020-06-26 07:26:10" "2020-06-23 17:12:41" "2020-06-20 01:09:3
5" ...
##  $ ended_at          : chr  "2020-06-13 23:36:55" "2020-06-26 07:31:58" "2020-06-23 17:21:14" "2020-06-20 01:28:2
4" ...
##  $ start_station_name: chr  "Wilton Ave & Belmont Ave" "Federal St & Polk St" "Daley Center Plaza" "Broadway & Co
rnelia Ave" ...
##  $ start_station_id  : int  117 41 81 303 327 327 41 115 338 84 ...
##  $ end_station_name  : chr  "Damen Ave & Clybourn Ave" "Daley Center Plaza" "State St & Harrison St" "Broadway &
Berwyn Ave" ...
##  $ end_station_id    : int  163 81 5 294 117 117 81 303 164 53 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num  -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "member" "member" "casual" ...
```

```
str(d7)
```

```
## 'data.frame':    551480 obs. of  13 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-07-09 15:22:02" "2020-07-24 23:56:30" "2020-07-08 19:49:07" "2020-07-17 19:06:4
2" ...
##  $ ended_at          : chr  "2020-07-09 15:25:52" "2020-07-25 00:20:17" "2020-07-08 19:56:22" "2020-07-17 19:27:3
8" ...
##  $ start_station_name: chr  "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "Lake Shore Dr & Diversey Pkwy" "LaS
alle St & Illinois St" ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : chr  "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "Clark St & Wellington Ave" "Clark
St & Armitage Ave" ...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "casual" "casual" ...
```

```
str(d8)
```

```
## 'data.frame':    622361 obs. of  13 variables:
##  $ ride_id           : chr  "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79FBBD412E578A7" ...
##  $ rideable_type     : chr  "docked_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-08-20 18:08:14" "2020-08-27 18:46:04" "2020-08-26 19:44:14" "2020-08-27 12:05:4
1" ...
##  $ ended_at          : chr  "2020-08-20 18:17:51" "2020-08-27 19:54:51" "2020-08-26 21:53:07" "2020-08-27 12:53:4
5" ...
##  $ start_station_name: chr  "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Columbus Dr & Randolph St"
"Daley Center Plaza" ...
##  $ start_station_id  : int  329 168 195 81 658 658 196 67 153 177 ...
##  $ end_station_name  : chr  "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St & Randolph St" "State St
& Kinzie St" ...
##  $ end_station_id    : int  141 168 44 47 658 658 49 229 225 305 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

```
str(d9)
```

```
## 'data.frame':	532958 obs. of  13 variables:
##  $ ride_id           : chr  "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-09-17 14:27:11" "2020-09-17 15:07:31" "2020-09-17 15:09:04" "2020-09-17 18:10:4
6" ...
##  $ ended_at          : chr  "2020-09-17 14:44:24" "2020-09-17 15:07:45" "2020-09-17 15:09:35" "2020-09-17 18:35:4
9" ...
##  $ start_station_name: chr  "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "A
shland Ave & Belle Plaine Ave" ...
##  $ start_station_id  : int  52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : chr  "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "M
ontrose Harbor" ...
##  $ end_station_id    : int  112 NA NA 249 24 NA 256 NA NA NA ...
##  $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(d10)
```

```
## 'data.frame':	388653 obs. of  13 variables:
##  $ ride_id           : chr  "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE261B9E854" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-10-31 19:39:43" "2020-10-31 23:50:08" "2020-10-31 23:00:01" "2020-10-31 22:16:4
3" ...
##  $ ended_at          : chr  "2020-10-31 19:57:12" "2020-11-01 00:04:16" "2020-10-31 23:08:22" "2020-10-31 22:19:3
5" ...
##  $ start_station_name: chr  "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave" "Stony Island Ave & 67
th St" "Clark St & Grace St" ...
##  $ start_station_id  : int  313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : chr  "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "University Ave & 57th St" "Broad
way & Sheridan Rd" ...
##  $ end_station_id    : int  125 260 423 256 185 53 125 313 199 635 ...
##  $ start_lat         : num  41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(d11)
```

```
## 'data.frame':    259716 obs. of  13 variables:
##  $ ride_id           : chr  "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533E89C32080B9E" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-11-01 13:36:00" "2020-11-01 10:03:26" "2020-11-01 00:34:05" "2020-11-01 00:45:1
6" ...
##  $ ended_at          : chr  "2020-11-01 13:45:40" "2020-11-01 10:14:45" "2020-11-01 01:03:06" "2020-11-01 00:54:3
1" ...
##  $ start_station_name: chr  "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shore Dr & Monroe St" "Leav
itt St & Chicago Ave" ...
##  $ start_station_id  : int  110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : chr  "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal St & Polk St" "Stave St
& Armitage Ave" ...
##  $ end_station_id    : int  211 29 41 185 2 76 72 NA 288 273 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(d12)
```

```
## 'data.frame':    131573 obs. of  13 variables:
##  $ ride_id           : chr  "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE119628E44F871E" ...
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-12-27 12:44:29" "2020-12-18 17:37:15" "2020-12-15 15:04:33" "2020-12-15 15:54:1
8" ...
##  $ ended_at          : chr  "2020-12-27 12:55:06" "2020-12-18 17:44:19" "2020-12-15 15:11:28" "2020-12-15 16:00:1
1" ...
##  $ start_station_name: chr  "Aberdeen St & Jackson Blvd" "" "" "" ...
##  $ start_station_id  : chr  "13157" "" "" "" ...
##  $ end_station_name  : chr  "Desplaines St & Kinzie St" "" "" "" ...
##  $ end_station_id    : chr  "TA1306000003" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

Group d4 -> d11 have start_station_id and end_station_id is int while the rest are chr. Therefore, combine these data sets to change their data type.

```
data_diff <- bind_rows(d4,d5,d6,d7,d8,d9,d10,d11)
data_diff$start_station_id <- as.character(data_diff$start_station_id)
data_diff$end_station_id <- as.character(data_diff$end_station_id)
```

Now, combine all data sets into one

```
Bike <- bind_rows(d1,d2,d3,data_diff, d12)
```

Bike has 3489748 rows Since we do not use all the columns, let’s drop some of them

```
Bike <- Bike %>% select(-c(start_lat, start_lng, end_lat, end_lng))
```

Our main purpose is to compare the member types so we consider the time. Therefore, we need to process the columns relate the time

```
#first, change the data type
Bike$started_at <- as.POSIXct(Bike$started_at, tz ="")
Bike$ended_at <- as.POSIXct(Bike$ended_at, tz ="")
#second, we separate these columns in order to make it easy to analyze
Bike$Date_in <- as.Date(format(Bike$started_at), "%Y-%m-%d")
Bike$Date_month <- format(as.Date(format(Bike$started_at), "%Y-%m-%d"), "%Y-%m")
Bike$Date_wd <- format(as.Date(Bike$started_at), "%A")
```

Calculate the time duration of each trips

```
Bike$Time_duration <- difftime(Bike$ended_at, Bike$started_at)
#diff time in seconds
Bike$Time_duration <- as.numeric(Bike$Time_duration)
Bike$Time_duration_hms <- hms(Bike$Time_duration)
```

Now it’s time for deeper cleaning

```
Bike <- Bike[!(Bike$start_station_name == "HQ QR"| Bike$Time_duration <=0),] #drop trip that has negative time duration
skim(Bike)
```

Data summary

Name	Bike
Number of rows	3478810
Number of columns	14
_______________________	
Column type frequency:	
character	9
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	1.00	16	16	0	3478810	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	0	1.00	0	53	122126	709	0
start_station_id	83576	0.98	0	35	39176	1260	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
end_station_name	0	1.00	0	53	143061	707	0
end_station_id	97995	0.97	0	35	45527	1260	0
member_casual	0	1.00	6	6	0	2	0
Date_month	0	1.00	7	7	0	12	0
Date_wd	0	1.00	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Date_in	0	1	2020-04-01	2021-03-31	2020-08-29	363

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Time_duration_hms	0	1	1 secs	3523202 secs	874 secs	25630

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Time_duration	0	1	1677.24	15171.82	1	476	874	1601	3523202	▇▁▁▁▁

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2020-04-01 00:00:30	2021-03-31 23:59:08	2020-08-29 14:37:40	3035205
ended_at	0	1	2020-04-01 00:10:45	2021-04-06 11:00:11	2020-08-29 15:10:01	3020117

Bike now has only 3478810 rows.The data set has N/A value in *start_station_id* and *end_station_id* which do not affect our analysis so we don’t have to drop these values.

```
summary(Bike$Time_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       1     476     874    1677    1601 3523202
```

ANALYSE -Stories of data

analyze with member casual

```
#COMPARE 2 TYPES OF MEMBER IN TIME_DURATION
aggregate(Bike$Time_duration~Bike$member_casual, FUN = summary)
```

```
##   Bike$member_casual Bike$Time_duration.Min. Bike$Time_duration.1st Qu.
## 1             casual                  1.0000                   694.0000
## 2             member                  1.0000                   391.0000
##   Bike$Time_duration.Median Bike$Time_duration.Mean Bike$Time_duration.3rd Qu.
## 1                 1273.0000               2698.4785                  2413.0000
## 2                  689.0000                967.0123                  1207.0000
##   Bike$Time_duration.Max.
## 1            3341033.0000
## 2            3523202.0000
```

```
aggregate(Bike$Time_duration~Bike$member_casual, FUN = sum)
```

```
##   Bike$member_casual Bike$Time_duration
## 1             casual         3850618205
## 2             member         1984165069
```

```
table(Bike$member_casual)
```

```
## 
##  casual  member 
## 1426959 2051851
```

```
#COMPARE 2 TYPES OF MEMBER IN DATE_WD
#re-arrange the weekday
Bike$Date_wd <- ordered(Bike$Date_wd, levels= c("Monday","Tuesday","Wednesday", "Thursday",

                                                "Friday", "Saturday", "Sunday"))
#note: the ordered function here is to arrange the order of data when you analyze!
aggregate(Bike$Time_duration~Bike$member_casual + Bike$Date_wd,
          FUN = mean)
```

```
##    Bike$member_casual Bike$Date_wd Bike$Time_duration
## 1              casual       Monday          2620.9498
## 2              member       Monday           919.9315
## 3              casual      Tuesday          2492.9497
## 4              member      Tuesday           908.2766
## 5              casual    Wednesday          2377.3631
## 6              member    Wednesday           927.7176
## 7              casual     Thursday          2580.3383
## 8              member     Thursday           904.7152
## 9              casual       Friday          2655.6163
## 10             member       Friday           953.5518
## 11             casual     Saturday          2848.5338
## 12             member     Saturday          1073.7293
## 13             casual       Sunday          2979.8894
## 14             member       Sunday          1077.8262
```

```
# COMPARE 2 TYPES OF MEMBER IN MONTH
aggregate(Bike$Time_duration~Bike$member_casual + Bike$Date_month,
          FUN = mean)
```

```
##    Bike$member_casual Bike$Date_month Bike$Time_duration
## 1              casual         2020-04          4388.5533
## 2              member         2020-04          1288.8205
## 3              casual         2020-05          3073.2645
## 4              member         2020-05          1186.4038
## 5              casual         2020-06          3100.2874
## 6              member         2020-06          1123.9922
## 7              casual         2020-07          3597.2850
## 8              member         2020-07          1066.1054
## 9              casual         2020-08          2696.3853
## 10             member         2020-08          1010.1743
## 11             casual         2020-09          2293.3985
## 12             member         2020-09           932.5190
## 13             casual         2020-10          1815.5378
## 14             member         2020-10           843.0398
## 15             casual         2020-11          1909.3637
## 16             member         2020-11           815.3985
## 17             casual         2020-12          1611.1376
## 18             member         2020-12           764.9993
## 19             casual         2021-01          1541.0754
## 20             member         2021-01           772.3780
## 21             casual         2021-02          2962.6862
## 22             member         2021-02          1081.4072
## 23             casual         2021-03          2289.6601
## 24             member         2021-03           838.2379
```

analyze ridership data by type and weekday

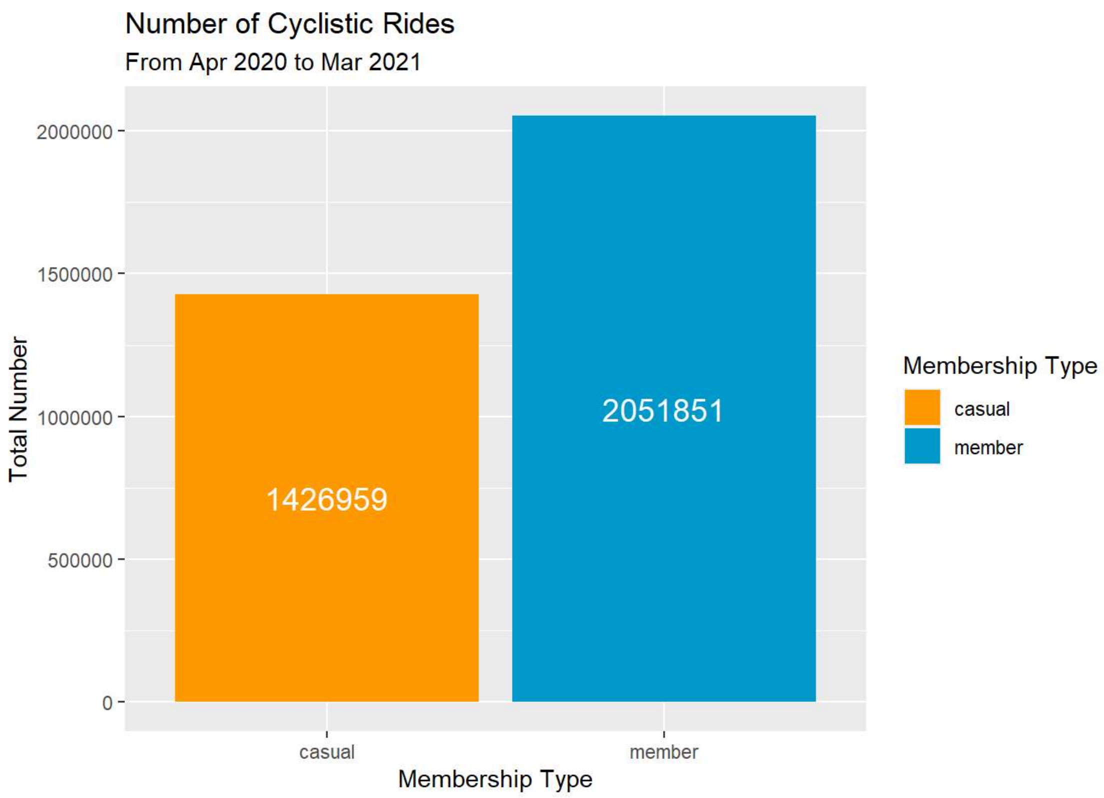
```
Bike %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                            #calculates the number of rides and average duration
            ,average_duration = mean(Time_duration)) %>%        # calculates the average duration
  arrange(member_casual, weekday)
```

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

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              262243            3045.
##  2 casual        Mon              151151            2699.
##  3 casual        Tue              145252            2429.
##  4 casual        Wed              158382            2419.
##  5 casual        Thu              166375            2579.
##  6 casual        Fri              208522            2566.
##  7 casual        Sat              335034            2818.
##  8 member        Sun              265270            1093.
##  9 member        Mon              267311             920.
## 10 member        Tue              284336             908.
## 11 member        Wed              305069             919.
## 12 member        Thu              300426             913.
## 13 member        Fri              306363             948.
## 14 member        Sat              323076            1068.
```

SHARE - Visualization

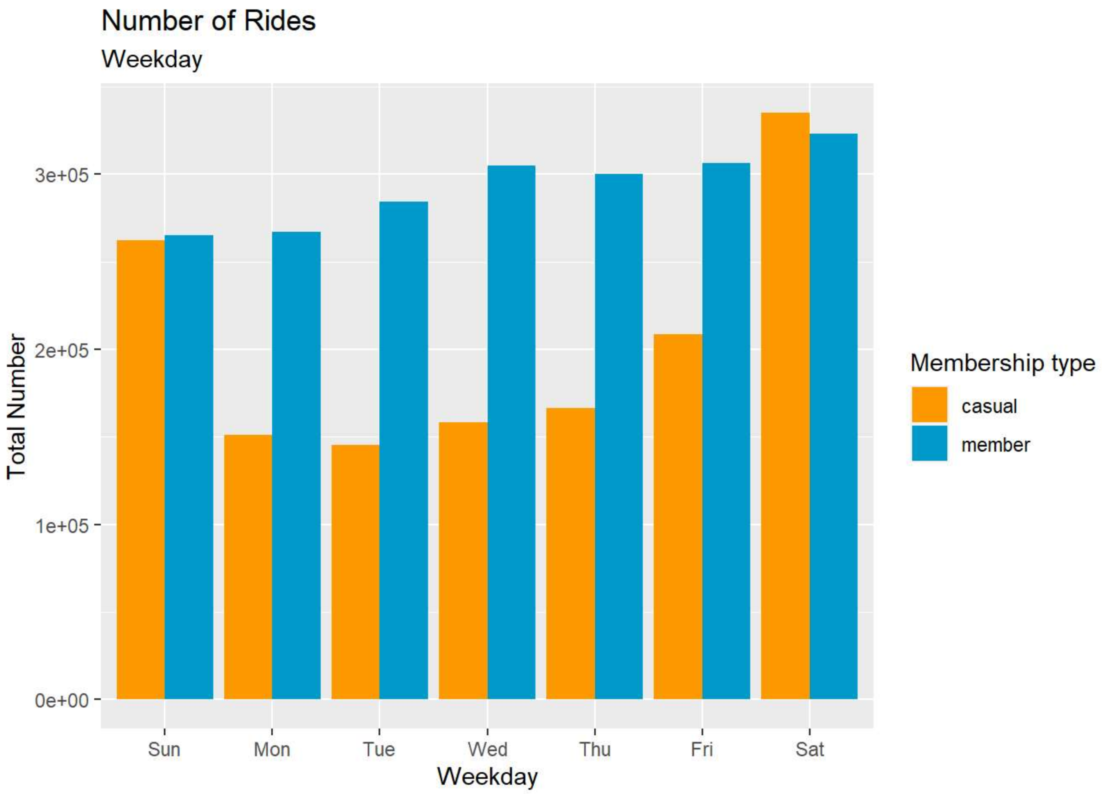
```
#Total users for both types
Bike %>%
  group_by(member_casual) %>%
  summarise(Total_rides = n()) %>%
  arrange(Total_rides) %>%

  ggplot(aes(x = member_casual, y = Total_rides, fill = member_casual)) +
  geom_bar(stat = "identity") +
  stat_identity(geom = "text", color = "white", size = 5, aes(label = Total_rides),
                position = position_stack(vjust = 0.5)) +
  scale_fill_manual(name = "Membership Type", values = c(casual ='#ff9900', member = '#0099cc')) +
  labs(title = "Number of Cyclistic Rides", x = "Membership Type", y = "Total Number", subtitle = "From Apr 2020 to
Mar 2021")
```



Let’s visualize the number of rides by weekday

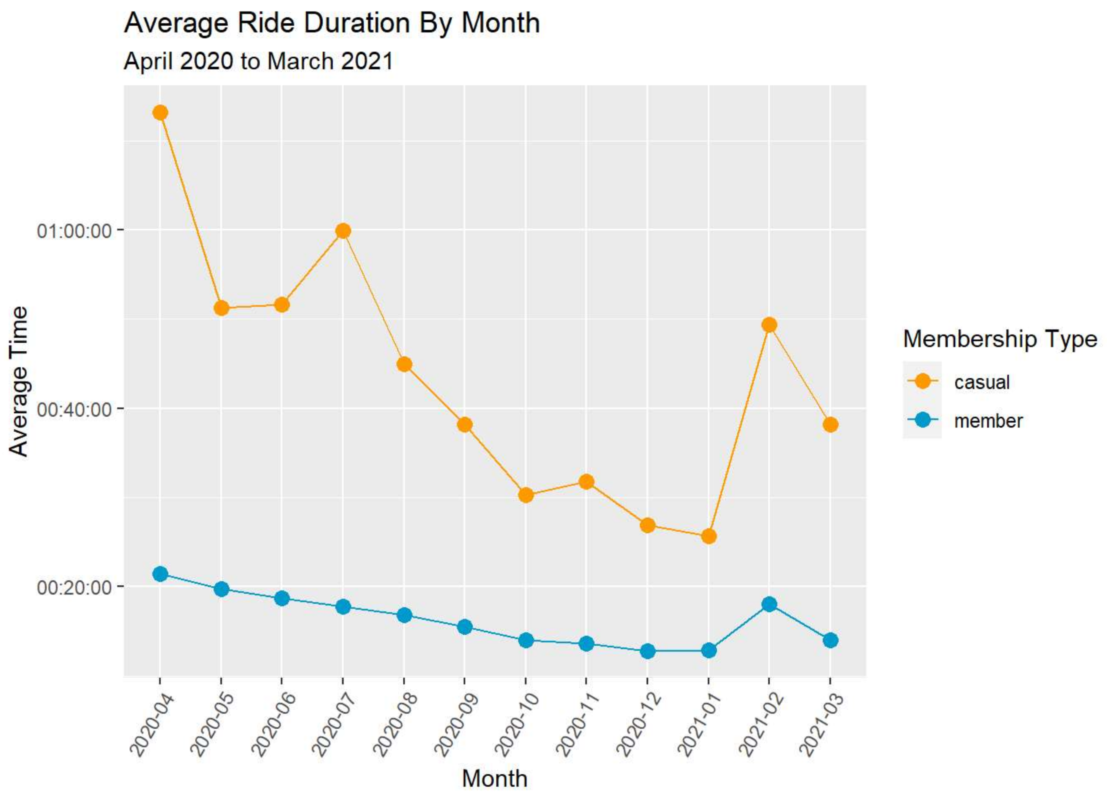
```
#Number of rides per weekday, categorized by membership type
Bike%>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(Time_duration)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill= member_casual)) + geom_col(position = 'dodge') + scale_fill_man
ual(name = "Membership type", values = c(casual = '#ff9900', member = '#0099cc')) +
 labs(title = "Number of Rides" , x = "Weekday", y = "Total Number", subtitle = "Weekday")
```

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



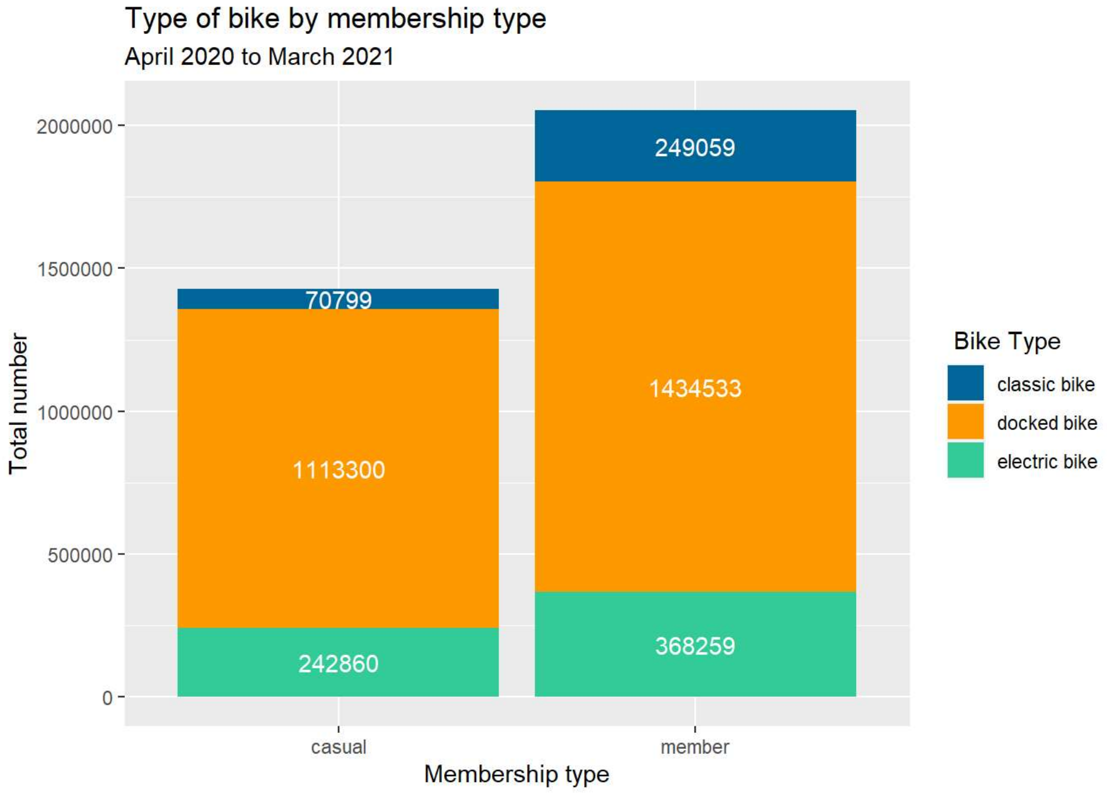
Number of rides to membership type by month

```
Bike%>%
  group_by(member_casual, Date_month) %>%
  summarise(.groups = 'drop', average_duration = mean(Time_duration)) %>%
 arrange(member_casual, Date_month) %>% mutate(Average_Time = hms(average_duration)) %>%

 ggplot(aes(x = Date_month, y = Average_Time, group = member_casual, colour = member_casual)) +
 geom_line() + geom_point(size = 3) +
 scale_colour_manual(name = "Membership Type",
 values = c(casual = '#ff9900', member = '#0099cc')) +
 labs(title = "Average Ride Duration By Month", x = "Month", y = "Average Time",
 subtitle = "April 2020 to March 2021") +
  theme(axis.text.x = element_text(angle = 60, hjust=1))
```



Number of rides to membership type by type of rides

```
Bike %>%
  group_by(rideable_type, member_casual) %>%
  summarise(Total_number = n(), .groups ='drop') %>%
  arrange(Total_number) %>%

  ggplot(aes(x = member_casual, y = Total_number, fill = rideable_type)) +
  geom_bar(stat = "identity") +
  stat_identity(geom = "text", colour = "white", 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("#006699", "#ff9900", "#33cc99")) +
  labs(title = "Type of bike by membership type", x = "Membership type", y = "Total number",
       subtitle = "April 2020 to March 2021")
```



ACT

CONCLUSION

- *Member* membership always have higher total number of rides over the time. However, there is a trend in *Casual* membership that it increases significantly in the weekend(Saturday and Sunday) which suggests that *casual* membership could use their bike to go shopping, travel around, health activities, etc. These activities maybe for entertainment purpose. Also, *member* membership rides decrease slightly in Sunday, which could be implied that they are mostly working people.

- Time duration: *member* has a longer ride duration than *casual*, nearly 40 minutes to approximately 20 minutes

- Bike Type: *casual* prefers to use docked bike and electric bike while these ratios of *member* is lower.

SOLUTION

- Offer more incentives for *member* and increase the renting price(especially for docked and electric bike) to promote them being casual membership
- Besides, we could analyze more about locations(I do not do it in this part) to focus on where have more potential customers(especially for *casual*)