R Notebook

Code

ManikumarKR

1 The Goal

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

The **Dataset**

last 12 months (2020-03 to 2022-04) of bike-sharing data from Motivate International Inc. who operates the City of Chicago's Divvy bicycle sharing service.

Installing packages

Hide

Hide

```
install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")
```

Loading packages

Hide

Hide

```
library(tidyverse)
```

```
Registered S3 methods overwritten by 'dbplyr':
  method         from
  print.tbl_lazy
  print.tbl_sql
── Attaching packages ────────────────────────────────────────────────────────────────────────────
──── tidyverse 1.3.1 ──
✔ ggplot2 3.3.6     ✔ purrr   0.3.4
✔ tibble  3.1.7     ✔ dplyr   1.0.9
✔ tidyr   1.2.0     ✔ stringr 1.4.0
✔ readr   2.1.2     ✔ forcats 0.5.1
── Conflicts ──────────────────────────────────────────────────────────────────────────────── t
idyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

Hide

Hide

```
library(lubridate)
```

```

Attaching package: ‘lubridate’

The following objects are masked from ‘package:base’:

    date, intersect, setdiff, union
```

Hide

Hide

```
library(ggplot2)
library(dbplyr)
```

```

Attaching package: ‘dbplyr’

The following objects are masked from ‘package:dplyr’:

    ident, sql
```

importing data

importing all 12 months dataset.

Hide

Hide

```
one_five <- read.csv("202105-divvy-tripdata.csv")
one_six <- read.csv("202106-divvy-tripdata.csv")
one_seven <- read.csv("202107-divvy-tripdata.csv")
one_eight <- read.csv("202108-divvy-tripdata.csv")
one_nine <- read.csv("202109-divvy-tripdata.csv")
one_ten <- read.csv("202110-divvy-tripdata.csv")
one_eleven <- read.csv("202111-divvy-tripdata.csv")
one_twelve <- read.csv("202112-divvy-tripdata.csv")
two_one <- read.csv("202201-divvy-tripdata.csv")
two_two <- read.csv("202202-divvy-tripdata.csv")
two_three <- read.csv("202203-divvy-tripdata.csv")
two_four <- read.csv("202204-divvy-tripdata.csv")
```

getting familar with dataset

Hide

Hide

```
colnames(one_five)
 colnames(one_six)
 colnames(one_seven)
 colnames(one_eight)
 colnames(one_nine)
 colnames(one_ten)
 colnames(one_ten)
 colnames(one_eleven)
 colnames(one_twelve)
 colnames(two_one)
 colnames(two_two)
 colnames(two_three)
 colnames(two_four)
```

Hide

Hide

```
str(one_five)
```

```
'data.frame':   531633 obs. of  13 variables:
 $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D3911
0C60" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-05-30 11:58:15" "2021-05-30 11:29:14" "2021-05-30 14:24:01" "202
1-05-30 14:25:51" ...
 $ ended_at          : chr  "2021-05-30 12:10:39" "2021-05-30 12:14:09" "2021-05-30 14:25:13" "202
1-05-30 14:41:04" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.8 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

Hide

```
 str(one_six)
```

```
'data.frame':   729595 obs. of  13 variables:
 $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C0FE48C41
2214" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-06-13 14:31:28" "2021-06-04 11:18:02" "2021-06-04 09:49:35" "202
1-06-03 19:56:05" ...
 $ ended_at          : chr  "2021-06-13 14:34:11" "2021-06-04 11:24:19" "2021-06-04 09:55:34" "202
1-06-03 20:21:55" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.8 41.8 41.8 41.8 41.8 ...
 $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
 $ end_lat           : num  41.8 41.8 41.8 41.8 41.8 ...
 $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

Hide

```
 str(one_seven)
```

```
'data.frame':   822410 obs. of  13 variables:
 $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E
8AA5" ...
 $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
 $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 11:30:55" "202
1-07-08 22:08:30" ...
 $ ended_at          : chr  "2021-07-02 15:19:58" "2021-07-07 17:16:09" "2021-07-25 11:48:45" "202
1-07-08 22:23:32" ...
 $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Av
e & 16th St" "California Ave & Cortez St" ...
 $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
 $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubba
rd St" "Carpenter St & Huron St" ...
 $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

Hide

Hide

```
 str(one_eight)
```

```
'data.frame':   804352 obs. of  13 variables:
 $ ride_id           : chr  "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFA
F1DA" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-08-10 17:15:49" "2021-08-10 17:23:14" "2021-08-21 02:34:23" "202
1-08-21 06:52:55" ...
 $ ended_at          : chr  "2021-08-10 17:22:44" "2021-08-10 17:39:24" "2021-08-21 02:50:36" "202
1-08-21 07:08:13" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.8 41.8 42 42 41.8 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ end_lat           : num  41.8 41.8 42 42 41.8 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

Hide

```
 str(one_nine)
```

```
'data.frame':   756147 obs. of  13 variables:
 $ ride_id           : chr  "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1DE133B3D
BF55" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-09-28 16:07:10" "2021-09-28 14:24:51" "2021-09-28 00:20:16" "202
1-09-28 14:51:17" ...
 $ ended_at          : chr  "2021-09-28 16:09:54" "2021-09-28 14:40:05" "2021-09-28 00:23:57" "202
1-09-28 15:00:06" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.8 41.8 41.9 ...
 $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 42 41.8 41.8 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

Hide

```
 str(one_ten)
```

```
'data.frame':   631226 obs. of  13 variables:
 $ ride_id           : chr  "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362947F0437E
1514" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-10-22 12:46:42" "2021-10-21 09:12:37" "2021-10-16 16:28:39" "202
1-10-16 16:17:48" ...
 $ ended_at          : chr  "2021-10-22 12:49:50" "2021-10-21 09:14:14" "2021-10-16 16:36:26" "202
1-10-16 16:19:03" ...
 $ start_station_name: chr  "Kingsbury St & Kinzie St" "" "" "" ...
 $ start_station_id  : chr  "KA1503000043" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

Hide

```
 str(one_eleven)
```

```
'data.frame':   359978 obs. of  13 variables:
 $ ride_id           : chr  "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3BE33085BC
FF02" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "2021-11-27 13:27:38" "2021-11-27 13:38:25" "2021-11-26 22:03:34" "202
1-11-27 09:56:49" ...
 $ ended_at          : chr  "2021-11-27 13:46:38" "2021-11-27 13:56:10" "2021-11-26 22:05:56" "202
1-11-27 10:01:50" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 42 42 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
 $ end_lat           : num  42 41.9 42 41.9 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

Hide

```
 str(one_twelve)
```

```
'data.frame':   247540 obs. of  13 variables:
 $ ride_id           : chr  "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278BA87BF69
8339" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "classic_bike" ...
 $ started_at        : chr  "2021-12-07 15:06:07" "2021-12-11 03:43:29" "2021-12-15 23:10:28" "202
1-12-26 16:16:10" ...
 $ ended_at          : chr  "2021-12-07 15:13:42" "2021-12-11 04:10:23" "2021-12-15 23:23:14" "202
1-12-26 16:30:53" ...
 $ start_station_name: chr  "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted St & North
Branch St" "Halsted St & North Branch St" ...
 $ start_station_id  : chr  "13307" "KP1705001026" "KA1504000117" "KA1504000117" ...
 $ end_station_name  : chr  "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Broadway & Barry A
ve" "LaSalle Dr & Huron St" ...
 $ end_station_id    : chr  "TA1307000130" "TA1307000119" "13137" "KP1705001026" ...
 $ 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 42 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "casual" "member" "member" ...
```

Hide

Hide

```
str(two_one)
```

```
'data.frame':   103770 obs. of  13 variables:
 $ ride_id           : chr  "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB80ED41910
5406" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "classic_bike" "classic_bike" ...
 $ started_at        : chr  "2022-01-13 11:59:47" "2022-01-10 08:41:56" "2022-01-25 04:53:40" "202
2-01-04 00:18:04" ...
 $ ended_at          : chr  "2022-01-13 12:02:44" "2022-01-10 08:46:17" "2022-01-25 04:58:01" "202
2-01-04 00:33:00" ...
 $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffield Ave &
Fullerton Ave" "Clark St & Bryn Mawr Ave" ...
 $ start_station_id  : chr  "525" "525" "TA1306000016" "KA1504000151" ...
 $ end_station_name  : chr  "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave & Fullert
on Ave" "Paulina St & Montrose Ave" ...
 $ end_station_id    : chr  "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
 $ start_lat         : num  42 42 41.9 42 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ end_lat           : num  42 42 41.9 42 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ member_casual     : chr  "casual" "casual" "member" "casual" ...
```

Hide

Hide

```
str(two_two)
```

```
'data.frame':   115609 obs. of  13 variables:
 $ ride_id           : chr  "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A1789BDF844412BE" ...
 $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
 $ started_at        : chr  "2022-02-19 18:08:41" "2022-02-20 17:41:30" "2022-02-25 18:55:56" "2022-02-14 11:57:03" ...
 $ ended_at          : chr  "2022-02-19 18:23:56" "2022-02-20 17:45:56" "2022-02-25 19:09:34" "2022-02-14 12:04:00" ...
 $ start_station_name: chr  "State St & Randolph St" "Halsted St & Wrightwood Ave" "State St & Randolph St" "Southport Ave & Waveland Ave" ...
 $ start_station_id  : chr  "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...
 $ end_station_name  : chr  "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Canal St & Adams St" "Broadway & Sheridan Rd" ...
 $ end_station_id    : chr  "13179" "TA1307000113" "13011" "13323" ...
 $ 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 42 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

Hide

```
str(two_three)
```

```
'data.frame':   284042 obs. of  13 variables:
 $ ride_id           : chr  "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F446FD9DEE3F389" ...
 $ rideable_type     : chr  "classic_bike" "electric_bike" "classic_bike" "classic_bike" ...
 $ started_at        : chr  "2022-03-21 13:45:01" "2022-03-16 09:37:16" "2022-03-23 19:52:02" "2022-03-01 19:12:26" ...
 $ ended_at          : chr  "2022-03-21 13:51:18" "2022-03-16 09:43:34" "2022-03-23 19:54:48" "2022-03-01 19:22:14" ...
 $ start_station_name: chr  "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & Berwyn Ave" "Wabash Ave & Wacker Pl" ...
 $ start_station_id  : chr  "TA1307000131" "13042" "13109" "TA1307000131" ...
 $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apts)" "Broadway & Ridge Ave" "Franklin St & Jackson Blvd" ...
 $ end_station_id    : chr  "KA1503000043" "620" "15578" "TA1305000025" ...
 $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

Hide

```
str(two_four)
```

```
'data.frame':   371249 obs. of  13 variables:
 $ ride_id           : chr  "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F07FF" "84D4751AEB31888D" ...
 $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike" "classic_bike" ...
 $ started_at        : chr  "2022-04-06 17:42:48" "2022-04-24 19:23:07" "2022-04-20 19:29:08" "2022-04-22 21:14:06" ...
 $ ended_at          : chr  "2022-04-06 17:54:36" "2022-04-24 19:43:17" "2022-04-20 19:35:16" "2022-04-22 21:23:29" ...
 $ start_station_name: chr  "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted St & Polk St" "Wentworth Ave & Cermak Rd" ...
 $ start_station_id  : chr  "515" "13075" "TA1307000121" "13075" ...
 $ end_station_name  : chr  "University Library (NU)" "Green St & Madison St" "Green St & Madison St" "Delano Ct & Roosevelt Rd" ...
 $ end_station_id    : chr  "605" "TA1307000120" "TA1307000120" "KA1706005007" ...
 $ start_lat         : num  42 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
 $ end_lat           : num  42.1 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "casual" ...
```

Combining all 12 Data set into one

Hide

Hide

```
all_trips <- rbind(one_five,one_six,one_seven,one_eight,one_nine,one_ten,
      one_eleven,one_twelve,two_one,two_two,two_three,two_four)
```

2 Cleaning Data and formatting

removing unwated columns

Hide

Hide

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

getting familiar with combine dataset

Hide

Hide

```
colnames(new_trips)
```

```
 [1] "ride_id"           "rideable_type"     "started_at"        "ended_at"          "start_station_name"
 [6] "start_station_id"  "end_station_name"  "end_station_id"    "Subcriber_Customer" "ride_length"
[11] "day_of_week"       "date"              "month"             "year"              "day"
```

Hide

Hide

```
head(new_trips)
```

ride_id <chr>	rideable_type <chr>	started_at <chr>	ended_at <chr>	start_station_name <chr>
1 C809ED75D6160B2A	electric_bike	30-05-2021 11:58	30-05-2021 12:10	
2 DD59FDCE0ACACAF3	electric_bike	30-05-2021 11:29	30-05-2021 12:14	
3 0AB83CB88C43EFC2	electric_bike	30-05-2021 14:24	30-05-2021 14:25	
4 7881AC6D39110C60	electric_bike	30-05-2021 14:25	30-05-2021 14:41	
5 853FA701B4582BAF	electric_bike	30-05-2021 18:15	30-05-2021 18:22	
6 F5E63DFD96B2A737	electric_bike	30-05-2021 11:33	30-05-2021 11:57	

6 rows | 1-6 of 15 columns

Hide

Hide

```
dim(new_trips)
```

```
[1] 5757551      15
```

Hide

Hide

```
str(new_trips)
```

```
'data.frame':   5757551 obs. of  15 variables:
 $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D3911
0C60" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : chr  "30-05-2021 11:58" "30-05-2021 11:29" "30-05-2021 14:24" "30-05-2021 1
4:25" ...
 $ ended_at          : chr  "30-05-2021 12:10" "30-05-2021 12:14" "30-05-2021 14:25" "30-05-2021 1
4:41" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ Subcriber_Customer: chr  "Customer" "Customer" "Customer" "Customer" ...
 $ ride_length       : chr  "00:12:24" "00:44:55" "00:01:12" "00:15:13" ...
 $ day_of_week       : int  1 1 1 1 1 1 1 4 4 3 ...
 $ date              : Date, format: "0030-05-20" "0030-05-20" "0030-05-20" "0030-05-20" ...
 $ month             : chr  "05" "05" "05" "05" ...
 $ year              : chr  "0030" "0030" "0030" "0030" ...
 $ day               : chr  "20" "20" "20" "20" ...
```

Hide

Hide

```
summary(new_trips)
```

```
   ride_id          rideable_type       started_at          ended_at         start_station_name st
art_station_id
 Length:5757551     Length:5757551     Length:5757551     Length:5757551     Length:5757551     Le
ngth:5757551
 Class :character   Class :character   Class :character   Class :character   Class :character   Cl
ass :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mo
de  :character



 end_station_name   end_station_id     Subcriber_Customer ride_length         day_of_week       da
te
 Length:5757551     Length:5757551     Length:5757551     Length:5757551     Min.   :1.0   Min.
:0001-01-20
 Class :character   Class :character   Class :character   Class :character   1st Qu.:2.0   1st Q
u.:0008-08-20
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.0   Median
:0016-01-20
                                                                             Mean   :4.1   Mean
:0016-01-17
                                                                             3rd Qu.:6.0   3rd Q
u.:0023-04-20
                                                                             Max.   :7.0   Max.
:0031-12-20
    month              year               day
 Length:5757551     Length:5757551     Length:5757551
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Hide

Hide

```
View(new_trips)
nrow(new_trips)
```

```
[1] 5757551
```

Naming variables properly

Hide

Hide

```
 new_trips <- new_trips %>% rename("Subcriber_Customer" =  "member_casual")
```

Hide

Hide

```
new_trips <-  new_trips %>%
  mutate(Subcriber_Customer = recode(Subcriber_Customer,
                                "member" = "Subcriber"
                                ,"casual" = "Customer"))
```

Hide

Hide

```
table(new_trips$Subcriber_Customer)
```

```

 Customer Subcriber
  2536358   3221193
```

Formatting Dates

Hide

Hide

```
new_trips$started_at <-as.POSIXct( as.character(new_trips$started_at), format = "%d-%m-%Y %H:%M")
new_trips$ended_at <-as.POSIXct( as.character(new_trips$ended_at), format = "%d-%m-%Y %H:%M")
```

Add columns that list the date, month, day, and year of each ride

Hide

Hide

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

Hide

Hide

```
new_trips$month<-format(as.Date(new_trips$date),"%m")
```

Hide

Hide

```
new_trips$year<-format(as.Date(new_trips$date),"%Y")
```

Hide

Hide

```
new_trips$day<-format(as.Date(new_trips$date),"%d")
```

Hide

Hide

```
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
```

Hide

Hide

```
new_trips$ride_length_sescs <- difftime(new_trips$ended_at,new_trips$started_at)
```

removing the word sec

Hide

Hide

```
new_trips$ride_length_secs<-gsub("sec","",as.character(new_trips$ride_length_secs)) #revoming the
 word secs
```

formate data type for ride_length_sec

Hide

Hide

```
new_trips$ride_length_secs <- as.numeric(as.character(new_trips$ride_length_secs))
```

Inspect the structure of the columns

Hide

Hide

```
str(new_trips)
```

```
'data.frame':   5757551 obs. of  16 variables:
 $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D3911
0C60" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : POSIXct, format: "2021-05-30 11:58:00" "2021-05-30 11:29:00" "2021-05-30 1
4:24:00" "2021-05-30 14:25:00" ...
 $ ended_at          : POSIXct, format: "2021-05-30 12:10:00" "2021-05-30 12:14:00" "2021-05-30 1
4:25:00" "2021-05-30 14:41:00" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ Subcriber_Customer: chr  "Customer" "Customer" "Customer" "Customer" ...
 $ ride_length       : chr  "00:12:24" "00:44:55" "00:01:12" "00:15:13" ...
 $ day_of_week       : int  1 1 1 1 1 1 1 4 4 3 ...
 $ date              : Date, format: "2021-05-30" "2021-05-30" "2021-05-30" "2021-05-30" ...
 $ month             : chr  "05" "05" "05" "05" ...
 $ year              : chr  "2021" "2021" "2021" "2021" ...
 $ day               : chr  "30" "30" "30" "30" ...
 $ ride_length_secs  : num  720 2700 60 960 420 1440 900 1020 180 1560 ...
```

Hide

Hide

```
head(new_trips)
```

	ride_id <chr>	rideable_type <chr>	started_at <S3: POSIXct>	ended_at <S3: POSIXct>	start_station_name <chr>
1	C809ED75D6160B2A	electric_bike	2021-05-30 11:58:00	2021-05-30 12:10:00	
2	DD59FDCE0ACACAF3	electric_bike	2021-05-30 11:29:00	2021-05-30 12:14:00	
3	0AB83CB88C43EFC2	electric_bike	2021-05-30 14:24:00	2021-05-30 14:25:00	
4	7881AC6D39110C60	electric_bike	2021-05-30 14:25:00	2021-05-30 14:41:00	
5	853FA701B4582BAF	electric_bike	2021-05-30 18:15:00	2021-05-30 18:22:00	
6	F5E63DFD96B2A737	electric_bike	2021-05-30 11:33:00	2021-05-30 11:57:00	

6 rows | 1-6 of 16 columns

Remove “bad” data

The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or ride_length_secs was negative and removed NA values and We will create a new version of the dataframe

Hide

Hide

```
new_trips_v2 <- new_trips[!(new_trips$start_station_name == "HQ QR" |
        new_trips$ride_length_secs<0),] %>% na.omit(new_trips$ride_length_secs)
```

Removed duplicates

Hide

Hide

```
new_trips_v2[!duplicated(new_trips$ride_id),]
```

Inspect the structure of the columns

Hide

Hide

```
nrow(new_trips_v2)
```

```
[1] 5757465
```

Hide

Hide

```
head(new_trips_v2)
```

	ride_id <chr>	rideable_type <chr>	started_at <S3: POSIXct>	ended_at <S3: POSIXct>	start_station_name <chr>
1	C809ED75D6160B2A	electric_bike	2021-05-30 11:58:00	2021-05-30 12:10:00	
2	DD59FDCE0ACACAF3	electric_bike	2021-05-30 11:29:00	2021-05-30 12:14:00	
3	0AB83CB88C43EFC2	electric_bike	2021-05-30 14:24:00	2021-05-30 14:25:00	
4	7881AC6D39110C60	electric_bike	2021-05-30 14:25:00	2021-05-30 14:41:00	
5	853FA701B4582BAF	electric_bike	2021-05-30 18:15:00	2021-05-30 18:22:00	
6	F5E63DFD96B2A737	electric_bike	2021-05-30 11:33:00	2021-05-30 11:57:00	

6 rows | 1-6 of 16 columns

3 ** CONDUCT DESCRIPTIVE ANALYSIS **

Summary of ride_length_secs

Hide

Hide

```
summary(new_trips_v2$ride_length_secs)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     360     720    1268    1260 3356640 
```

Compare members and casual users

find mean ,median,min,max

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs ~ new_trips_v2$Subcriber_Customer, FUN = mean)
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$ride_length_secs <dbl>
Customer	1877.5268
Subcriber	788.7073

2 rows

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs~new_trips_v2$Subcriber_Customer,FUN  = median)
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$ride_length_secs <dbl>
Customer	960
Subcriber	540

2 rows

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs~new_trips_v2$Subcriber_Customer,FUN  = max)
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$ride_length_secs <dbl>
Customer	3356640
Subcriber	93600

2 rows

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs~new_trips_v2$Subcriber_Customer, FUN = min )
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$ride_length_secs <dbl>
Customer	0
Subcriber	0

2 rows

Notice that the days of the week are out of order. Let’s fix that.

Hide

Hide

```
new_trips_v2 <- new_trips_v2 %>% arrange(day_of_week)
```

Compare members and casual users by weekdays

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs~new_trips_v2$Subcriber_Customer+new_trips_v2$day_of_week,
FUN = mean)
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$day_of_week <int>	new_trips_v2$ride_length_secs <dbl>

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$day_of_week <int>	new_trips_v2$ride_length_secs <dbl>
Customer	1	2218.1078
Subcriber	1	903.7507
Customer	2	1863.8911
Subcriber	2	762.7730
Customer	3	1587.8929
Subcriber	3	735.6076
Customer	4	1625.6574
Subcriber	4	744.9751
Customer	5	1673.2172
Subcriber	5	746.2709

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

Compare members and casual users by bike_type

Hide

Hide

```
aggregate(new_trips_v2$ride_length_secs~new_trips_v2$Subcriber_Customer+new_trips_v2$rideable_type, FUN = mean)
```

new_trips_v2$Subcriber_Customer <chr>	new_trips_v2$rideable_type <chr>	new_trips_v2$ride_length_secs <dbl>
Customer	classic_bike	1730.2096
Subcriber	classic_bike	827.8173
Customer	docked_bike	4992.6058
Customer	electric_bike	1159.2506
Subcriber	electric_bike	727.2593

5 rows

analyze Subcriber_Customer data by type and weekday

aggregate

Hide

Hide

```
new_trips_v2 %>%
   mutate(weekday=wday(started_at,label=TRUE))%>%

     group_by(Subcriber_Customer,weekday)%>%

     summarise(number_of_rides = n(),average_duration = mean(ride_length_secs))%>%

     arrange(Subcriber_Customer,weekday)
```

```
`summarise()` has grouped output by 'Subcriber_Customer'. You can override using the `.groups` arg
ument.
```

Subcriber_Customer <chr>	**weekday** <ord>	**number_of_rides** <int>	**average_duration** <dbl>
Customer	Sun	477000	2218.1078
Customer	Mon	289028	1863.8911
Customer	Tue	270546	1587.8929
Customer	Wed	284868	1625.6574
Customer	Thu	298061	1673.2172
Customer	Fri	358200	1752.5647
Customer	Sat	558614	2051.6201
Subcriber	Sun	388020	903.7507
Subcriber	Mon	445635	762.7730
Subcriber	Tue	498680	735.6076

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

4 Visualization

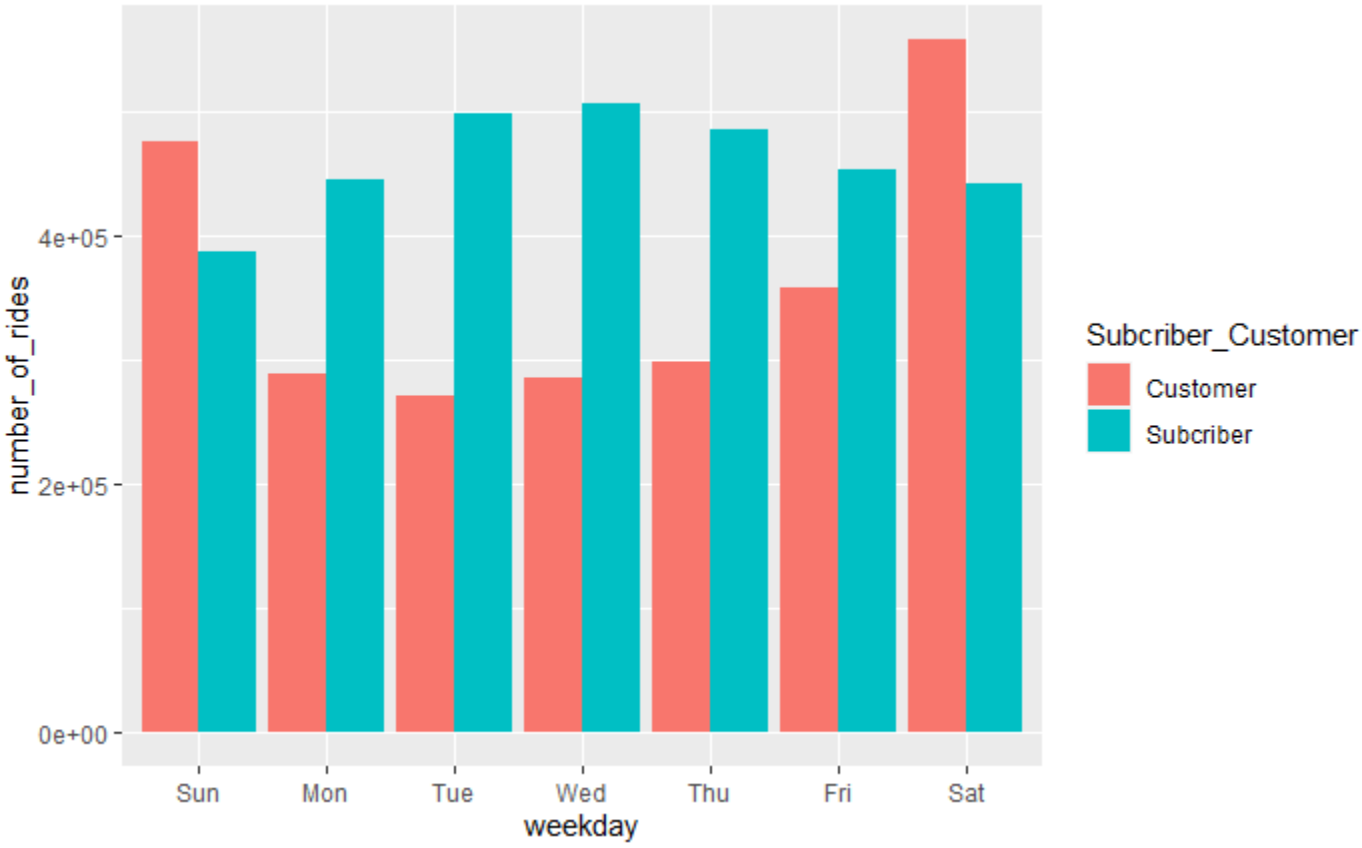
Let’s visualize the number of rides by rider type

Hide

Hide

```
  new_trips_v2 %>%
   mutate(weekday=wday(started_at,label=TRUE))%>%

     group_by(Subcriber_Customer,weekday)%>%

     summarise(number_of_rides = n(),average_duration = mean(ride_length_secs))%>%

     arrange(Subcriber_Customer,weekday)%>%

 ggplot(aes(x = weekday, y = number_of_rides, fill = Subcriber_Customer)) +
    geom_col(position = "dodge")
```

```
`summarise()` has grouped output by 'Subcriber_Customer'. You can override using the `.groups` arg
ument.
```



Hide

Hide

```
NA
```

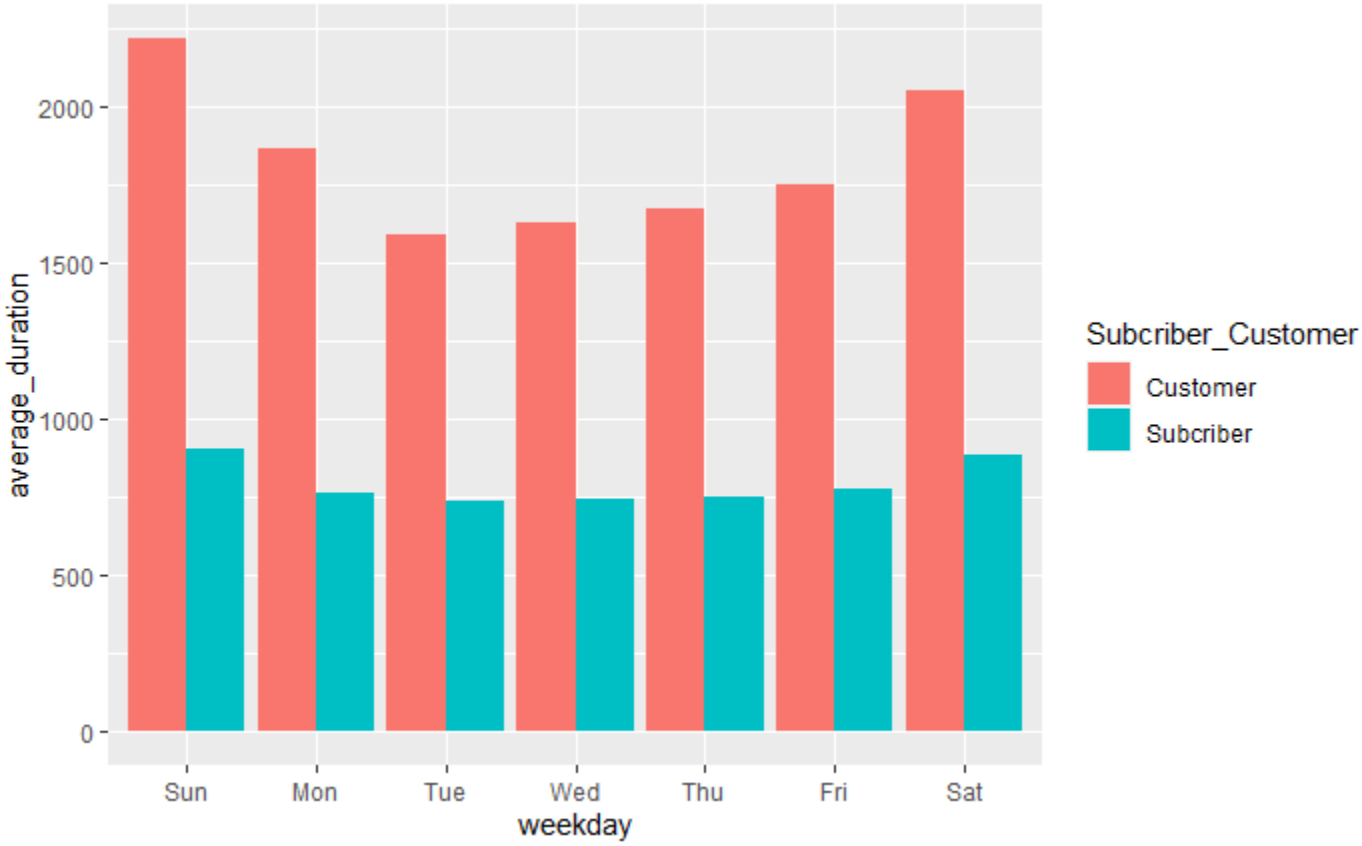
Let’s create a visualization for average duration

Hide

Hide

```
new_trips_v2 %>%
   mutate(weekday=wday(started_at,label=TRUE))%>%

     group_by(Subcriber_Customer,weekday)%>%

     summarise(number_of_rides = n(),average_duration = mean(ride_length_secs))%>%

     arrange(Subcriber_Customer,weekday)%>%

 ggplot(aes(x = weekday, y = average_duration, fill = Subcriber_Customer)) +
    geom_col(position = "dodge")
```

```
`summarise()` has grouped output by 'Subcriber_Customer'. You can override using the `.groups` arg
ument.
```



More analysis of the data using charts created with Tableau can be found in my presentation.