CYCLISTIC BIKE SHARE: CASE STUDY WITH R

Quang Cao Phan

2023-02-20

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. 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

- Question
    1. What is the difference between Subscriber and Custumer using Cyclistic bikes?
    2. Why did Custumer buy an annual Cyclistic membership?
    3. How can Cyclistic use digital media to influence casual riders to become members?

PREPARE

Dataset: https://divvy-tripdata.s3.amazonaws.com/index.html - For this analysis, I will use Q1 2019 to Q4 2019 data

```r
#install packages
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Pearls/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.2:
##   download from 'http://cran.us.r-project.org/bin/windows/contrib/4.2/PACKAGES' failed
```

```
## installing the source package 'tidyverse'
```

```
## Warning in download.file(url, destfile, method, mode = "wb", ...): downloaded
## length 413495 != reported length 702514
```

```
## Warning in download.file(url, destfile, method, mode = "wb", ...): URL
## 'http://lib.stat.cmu.edu/R/CRAN/src/contrib/tidyverse_1.3.2.tar.gz': Timeout of
## 60 seconds was reached
## Error in download.file(url, destfile, method, mode = "wb", ...) :
##   download from 'http://cran.us.r-project.org/src/contrib/tidyverse_1.3.2.tar.gz' failed

## Warning in download.packages(pkgs, destdir = tmpd, available = available, :
## download of package 'tidyverse' failed
```

```
install.packages("lubridate", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Pearls/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.2:
##   download from 'http://cran.us.r-project.org/bin/windows/contrib/4.2/PACKAGES' failed

## Package which is only available in source form, and may need
##   compilation of C/C++/Fortran: 'lubridate'

##   These will not be installed
```

```
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Pearls/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.2:
##   download from 'http://cran.us.r-project.org/bin/windows/contrib/4.2/PACKAGES' failed

## installing the source package 'ggplot2'

## Warning in download.file(url, destfile, method, mode = "wb", ...): downloaded
## length 1348581 != reported length 3150856

## Warning in download.file(url, destfile, method, mode = "wb", ...): URL
## 'http://lib.stat.cmu.edu/R/CRAN/src/contrib/ggplot2_3.4.1.tar.gz': Timeout of
## 60 seconds was reached

## Error in download.file(url, destfile, method, mode = "wb", ...) :
##   download from 'http://cran.us.r-project.org/src/contrib/ggplot2_3.4.1.tar.gz' failed

## Warning in download.packages(pkgs, destdir = tmpd, available = available, :
## download of package 'ggplot2' failed
```

```
install.packages("dplyr", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Pearls/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.2:
##   download from 'http://cran.us.r-project.org/bin/windows/contrib/4.2/PACKAGES' failed
## Package which is only available in source form, and may need
##   compilation of C/C++/Fortran: 'dplyr'

##   These will not be installed
```

```r
install.packages('plyr', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Pearls/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.2:
##   download from 'http://cran.us.r-project.org/bin/windows/contrib/4.2/PACKAGES' failed

## Package which is only available in source form, and may need
##   compilation of C/C++/Fortran: 'plyr'

##   These will not be installed
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2
## --

## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

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

```r
library(ggplot2)
library(dplyr)
library(knitr)
```

```r
#import datasets
q1_2019 <- read_csv("C:/Users/Pearls/Documents/Google R Project/Divvy_Trips_2019_Q1.csv")
```

```
## Rows: 365069 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
q2_2019 <- read_csv("C:/Users/Pearls/Documents/Google R Project/Divvy_Trips_2019_Q2.csv")
```

```
## Rows: 1108163 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): 03 - Rental Start Station Name, 02 - Rental End Station Name, User...
## dbl  (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID, 03 - R...
## num  (1): 01 - Rental Details Duration In Seconds Uncapped
## dttm (2): 01 - Rental Details Local Start Time, 01 - Rental Details Local En...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
q3_2019 <- read_csv("C:/Users/Pearls/Documents/Google R Project/Divvy_Trips_2019_Q3.csv")
```

```
## Rows: 1640718 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
q4_2019 <- read_csv("C:/Users/Pearls/Documents/Google R Project/Divvy_Trips_2019_Q4.csv")
```

```
## Rows: 704054 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Process

First, we need to check the column names first before merging the four datasets. This is important because all column names must be the same.

```
colnames(q1_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"         
##  [4] "bikeid"            "tripduration"      "from_station_id"  
##  [7] "from_station_name" "to_station_id"     "to_station_name"  
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q2_2019)
```

```
##  [1] "01 - Rental Details Rental ID"                   
##  [2] "01 - Rental Details Local Start Time"            
##  [3] "01 - Rental Details Local End Time"              
##  [4] "01 - Rental Details Bike ID"                     
##  [5] "01 - Rental Details Duration In Seconds Uncapped"
##  [6] "03 - Rental Start Station ID"                    
##  [7] "03 - Rental Start Station Name"                  
##  [8] "02 - Rental End Station ID"                      
##  [9] "02 - Rental End Station Name"                    
## [10] "User Type"                                       
## [11] "Member Gender"                                   
## [12] "05 - Member Details Member Birthday Year"
```

```
colnames(q3_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"         
##  [4] "bikeid"            "tripduration"      "from_station_id"  
##  [7] "from_station_name" "to_station_id"     "to_station_name"  
## [10] "usertype"          "gender"            "birthyear"
```

```
colnames(q4_2019)
```

```
##  [1] "trip_id"           "start_time"        "end_time"         
##  [4] "bikeid"            "tripduration"      "from_station_id"  
##  [7] "from_station_name" "to_station_id"     "to_station_name"  
## [10] "usertype"          "gender"            "birthyear"
```

Because the columns are not identical with other datasets, we will proceed to rename the columns.

```
q2_2019fixed <- rename(q2_2019
                   ,ride_id = "01 - Rental Details Rental ID"
                   ,rideable_type = "01 - Rental Details Bike ID"
                   ,started_at = "01 - Rental Details Local Start Time"
                   ,ended_at = "01 - Rental Details Local End Time"
                   ,start_station_name = "03 - Rental Start Station Name"
                   ,start_station_id = "03 - Rental Start Station ID"
                   ,end_station_name = "02 - Rental End Station Name"
                   ,end_station_id = "02 - Rental End Station ID"
                   ,tripduration = "01 - Rental Details Duration In Seconds Uncapped"
                   ,birthyear = "05 - Member Details Member Birthday Year"
                   ,gender = "Member Gender"
                   ,member_casual = "User Type")

q4_2019fixed <- rename(q4_2019
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype)

q3_2019fixed <- rename(q3_2019
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype)

q1_2019fixed <- rename(q1_2019
                   ,ride_id = trip_id
                   ,rideable_type = bikeid
                   ,started_at = start_time
                   ,ended_at = end_time
                   ,start_station_name = from_station_name
                   ,start_station_id = from_station_id
                   ,end_station_name = to_station_name
                   ,end_station_id = to_station_id
                   ,member_casual = usertype)
```

Next, we need to check the data types of each column to make sure all the data is formatted correctly.

```
str(q1_2019fixed)
```

```
## spc_tbl_ [365,069 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id          : num [1:365069] 21742443 21742444 21742445 21742446 21742447 ...
##  $ started_at       : POSIXct[1:365069], format: "2019-01-01 00:04:37" "2019-01-01 00:08:13" ...
##  $ ended_at         : POSIXct[1:365069], format: "2019-01-01 00:11:07" "2019-01-01 00:15:34" ...
##  $ rideable_type    : num [1:365069] 2167 4386 1524 252 1170 ...
##  $ tripduration     : num [1:365069] 390 441 829 1783 364 ...
##  $ start_station_id : num [1:365069] 199 44 15 123 173 98 98 211 150 268 ...
##  $ start_station_name: chr [1:365069] "Wabash Ave & Grand Ave" "State St & Randolph St" "Racine Ave &
##  $ end_station_id   : num [1:365069] 84 624 644 176 35 49 49 142 148 141 ...
##  $ end_station_name  : chr [1:365069] "Milwaukee Ave & Grand Ave" "Dearborn St & Van Buren St (*)" "W
##  $ member_casual     : chr [1:365069] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:365069] "Male" "Female" "Female" "Male" ...
##  $ birthyear         : num [1:365069] 1989 1990 1994 1993 1994 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2019fixed)
```

```
## spc_tbl_ [1,108,163 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:1108163] 22178529 22178530 22178531 22178532 22178533 ...
##  $ started_at        : POSIXct[1:1108163], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" ...
##  $ ended_at          : POSIXct[1:1108163], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" ...
##  $ rideable_type     : num [1:1108163] 6251 6226 5649 4151 3270 ...
##  $ tripduration      : num [1:1108163] 446 1048 252 357 1007 ...
##  $ start_station_id  : num [1:1108163] 81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr [1:1108163] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jacks
##  $ end_station_id    : num [1:1108163] 56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr [1:1108163] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal
##  $ member_casual     : chr [1:1108163] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:1108163] "Male" "Female" "Male" "Male" ...
##  $ birthyear         : num [1:1108163] 1975 1984 1990 1993 1992 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ‘01 - Rental Details Rental ID‘ = col_double(),
##   ..   ‘01 - Rental Details Local Start Time‘ = col_datetime(format = ""),
##   ..   ‘01 - Rental Details Local End Time‘ = col_datetime(format = ""),
##   ..   ‘01 - Rental Details Bike ID‘ = col_double(),
##   ..   ‘01 - Rental Details Duration In Seconds Uncapped‘ = col_number(),
##   ..   ‘03 - Rental Start Station ID‘ = col_double(),
##   ..   ‘03 - Rental Start Station Name‘ = col_character(),
##   ..   ‘02 - Rental End Station ID‘ = col_double(),
##   ..   ‘02 - Rental End Station Name‘ = col_character(),
##   ..   ‘User Type‘ = col_character(),
##   ..   ‘Member Gender‘ = col_character(),
##   ..   ‘05 - Member Details Member Birthday Year‘ = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(q3_2019fixed)
```

```
## spc_tbl_ [1,640,718 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:1640718] 23479388 23479389 23479390 23479391 23479392 ...
##  $ started_at        : POSIXct[1:1640718], format: "2019-07-01 00:00:27" "2019-07-01 00:01:16" ...
##  $ ended_at          : POSIXct[1:1640718], format: "2019-07-01 00:20:41" "2019-07-01 00:18:44" ...
##  $ rideable_type     : num [1:1640718] 3591 5353 6180 5540 6014 ...
##  $ tripduration      : num [1:1640718] 1214 1048 1554 1503 1213 ...
##  $ start_station_id  : num [1:1640718] 117 381 313 313 168 300 168 313 43 43 ...
##  $ start_station_name: chr [1:1640718] "Wilton Ave & Belmont Ave" "Western Ave & Monroe St" "Lakeview
##  $ end_station_id    : num [1:1640718] 497 203 144 144 62 232 62 144 195 195 ...
##  $ end_station_name  : chr [1:1640718] "Kimball Ave & Belmont Ave" "Western Ave & 21st St" "Larrabee
##  $ member_casual     : chr [1:1640718] "Subscriber" "Customer" "Customer" "Customer" ...
##  $ gender            : chr [1:1640718] "Male" NA NA NA ...
##  $ birthyear         : num [1:1640718] 1992 NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(q4_2019fixed)
```

```
## spc_tbl_ [704,054 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : num [1:704054] 25223640 25223641 25223642 25223643 25223644 ...
##  $ started_at        : POSIXct[1:704054], format: "2019-10-01 00:01:39" "2019-10-01 00:02:16" ...
##  $ ended_at          : POSIXct[1:704054], format: "2019-10-01 00:17:20" "2019-10-01 00:06:34" ...
##  $ rideable_type     : num [1:704054] 2215 6328 3003 3275 5294 ...
##  $ tripduration      : num [1:704054] 940 258 850 2350 1867 ...
##  $ start_station_id  : num [1:704054] 20 19 84 313 210 156 84 156 156 336 ...
##  $ start_station_name: chr [1:704054] "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St
##  $ end_station_id    : num [1:704054] 309 241 199 290 382 226 142 463 463 336 ...
##  $ end_station_name  : chr [1:704054] "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave &
##  $ member_casual     : chr [1:704054] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr [1:704054] "Male" "Male" "Female" "Male" ...
##  $ birthyear         : num [1:704054] 1987 1998 1991 1990 1987 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   trip_id = col_double(),
##   ..   start_time = col_datetime(format = ""),
##   ..   end_time = col_datetime(format = ""),
##   ..   bikeid = col_double(),
##   ..   tripduration = col_number(),
##   ..   from_station_id = col_double(),
##   ..   from_station_name = col_character(),
##   ..   to_station_id = col_double(),
##   ..   to_station_name = col_character(),
##   ..   usertype = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

We need to convert gender to characters so we can merge the tables.

```
q4_2019fixed <-  mutate(q4_2019fixed, gender = as.character(gender))
q3_2019fixed <-  mutate(q3_2019fixed, gender = as.character(gender))
q2_2019fixed <-  mutate(q2_2019fixed, gender = as.character(gender))
q1_2019fixed <-  mutate(q1_2019fixed, gender = as.character(gender))
```

Now, I will merge the dataframes into one dataframe.

```
trips = bind_rows(q1_2019fixed, q2_2019fixed, q3_2019fixed, q4_2019fixed)
```

The next step is to clean the data.

We will add columns that list the day, month, day, and year of each trip. This will allow us to aggregate trip data for each month, day or year.

```
trips$date <- as.Date(trips$started_at)
trips$month <- format(as.Date(trips$date), "%m")
trips$day <- format(as.Date(trips$date), "%d")
trips$year <- format(as.Date(trips$date), "%Y")
trips$day_of_week <- format(as.Date(trips$date), "%A")
colnames(trips)
```

```
##  [1] "ride_id"           "started_at"         "ended_at"
##  [4] "rideable_type"     "tripduration"       "start_station_id"
##  [7] "start_station_name" "end_station_id"     "end_station_name"
## [10] "member_casual"     "gender"             "birthyear"
## [13] "date"              "month"              "day"
## [16] "year"              "day_of_week"
```

Now, we will add a column for the trip length for each trip by finding the time difference between the start time and end time of the trip.

```
trips$ride_length = difftime(trips$ended_at,trips$started_at)
```

There is some “bad” data to remove when ride_length is a negative number due to the maintenance of removing the bike for quality check. We will create a new dataframe to remove these negative trip length trips.

```r
trip_data_clean <- trips[!(trips$ride_length <= 0),]
glimpse(trip_data_clean)
```

```
## Rows: 3,817,991
## Columns: 18
## $ ride_id            <dbl> 21742443, 21742444, 21742445, 21742446, 21742447, 2~
## $ started_at         <dttm> 2019-01-01 00:04:37, 2019-01-01 00:08:13, 2019-01--~
## $ ended_at           <dttm> 2019-01-01 00:11:07, 2019-01-01 00:15:34, 2019-01--~
## $ rideable_type      <dbl> 2167, 4386, 1524, 252, 1170, 2437, 2708, 2796, 6205~
## $ tripduration       <dbl> 390, 441, 829, 1783, 364, 216, 177, 100, 1727, 336,~
## $ start_station_id   <dbl> 199, 44, 15, 123, 173, 98, 98, 211, 150, 268, 299, ~
## $ start_station_name <chr> "Wabash Ave & Grand Ave", "State St & Randolph St",~
## $ end_station_id     <dbl> 84, 624, 644, 176, 35, 49, 49, 142, 148, 141, 295, ~
## $ end_station_name   <chr> "Milwaukee Ave & Grand Ave", "Dearborn St & Van Bur~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Female", "Female", "Male", "Male", "Female~
## $ birthyear          <dbl> 1989, 1990, 1994, 1993, 1994, 1983, 1984, 1990, 199~
## $ date               <date> 2019-01-01, 2019-01-01, 2019-01-01, 2019-01-01, 20~
## $ month              <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01~
## $ day                <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01~
## $ year               <chr> "2019", "2019", "2019", "2019", "2019", "2019", "20~
## $ day_of_week        <chr> "Tuesday", "Tuesday", "Tuesday", "Tuesday", "Tuesda~
## $ ride_length        <drtn> 6.500000 mins, 7.350000 mins, 13.816667 mins, 29.7~
```

Analyze

We will now perform a descriptive analysis of the data to find patterns between Customer and Subscriber. Before we begin the analysis, it is a good idea to review basic descriptive statistics about the data.

```r
mean(trip_data_clean$ride_length)
```

```
## Time difference of 24.17443 mins
```

```r
median(trip_data_clean$ride_length)
```

```
## Time difference of 11.81667 mins
```

```r
max(trip_data_clean$ride_length)
```

```
## Time difference of 177200.4 mins
```

```r
min(trip_data_clean$ride_length)
```

```
## Time difference of 1.016667 mins
```

First, we'll compare Customer and Subscriber trip stats.

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

```
##   trip_data_clean$member_casual trip_data_clean$ride_length
## 1                      Customer              57.01802 mins
## 2                    Subscriber              14.32780 mins
```

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

```
##   trip_data_clean$member_casual trip_data_clean$ride_length
## 1                      Customer              25.83333 mins
## 2                    Subscriber               9.80000 mins
```

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

```
##   trip_data_clean$member_casual trip_data_clean$ride_length
## 1                      Customer              177200.4 mins
## 2                    Subscriber              150943.9 mins
```

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

```
##   trip_data_clean$member_casual trip_data_clean$ride_length
## 1                      Customer              1.016667 mins
## 2                    Subscriber              1.016667 mins
```

Before continuing, arrange the day_of_week column in the correct order.

```
trip_data_clean$day_of_week <- ordered(trip_data_clean$day_of_week, levels=c("Sunday", "Monday", "Tuesda
```

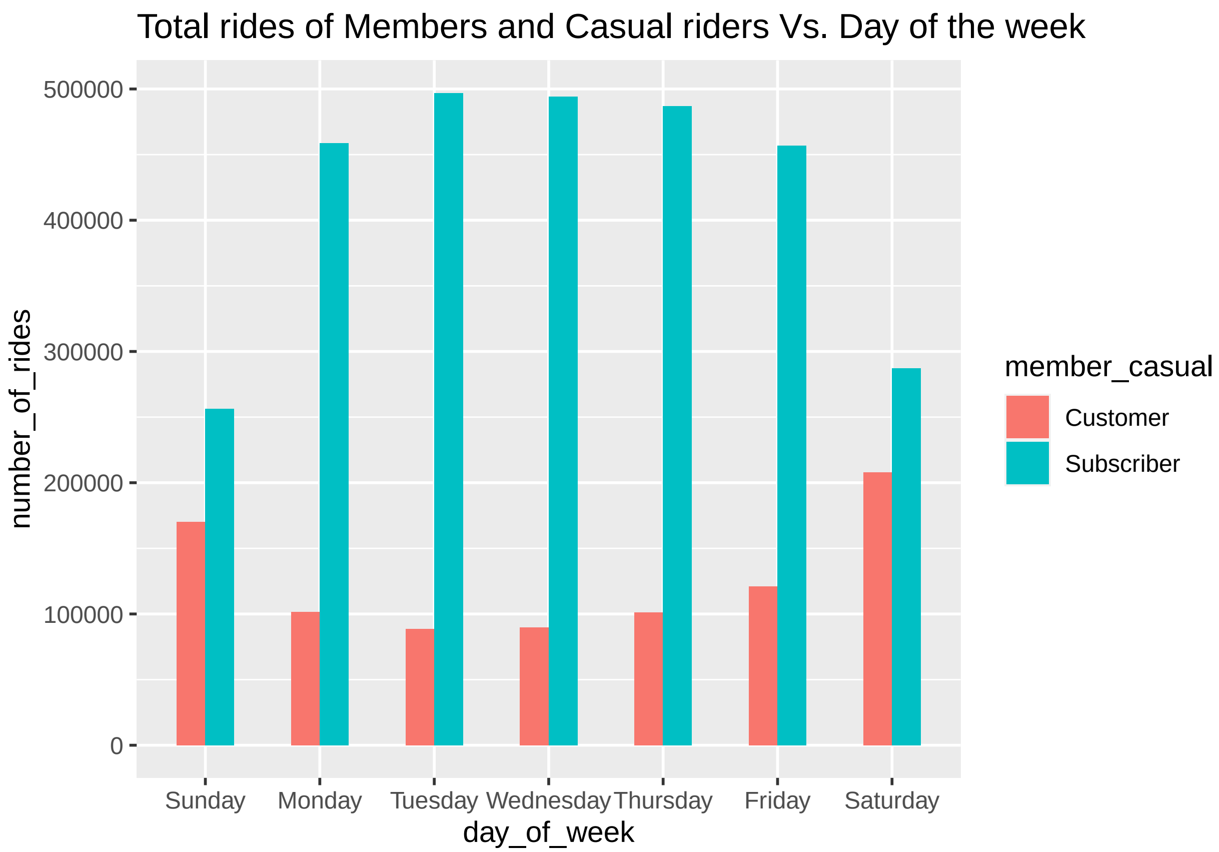
Next, we will check the average ride time per day and the total number of trips for Customer and Subscriber

```
plot <- trip_data_clean %>%
  group_by(member_casual, day_of_week) %>%  #groups by member_casual
  summarise(number_of_rides = n() #calculates the number of rides and average duration
  ,average_ride_length = mean(ride_length),.groups="drop") %>% # calculates the average duration
  arrange(member_casual, day_of_week) #sort
```

Share

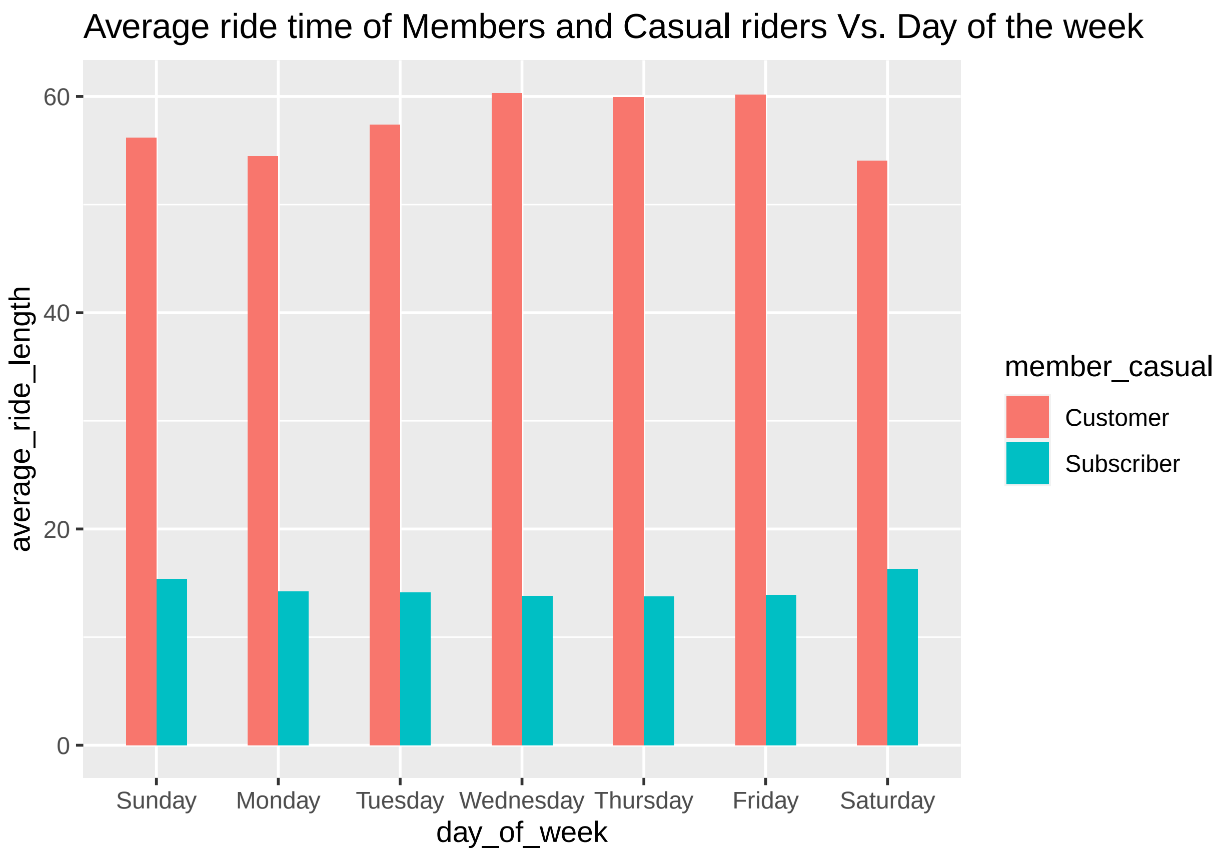
Before making recommendations to the marketing department, we will create some visualizations to share with stakeholders as well as give us a better idea of what insights to share.

```
 ggplot(plot,aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
 labs(title ="Total rides of Members and Casual riders Vs. Day of the week") +
 geom_col(width=0.5, position = position_dodge(width=0.5))+
 scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



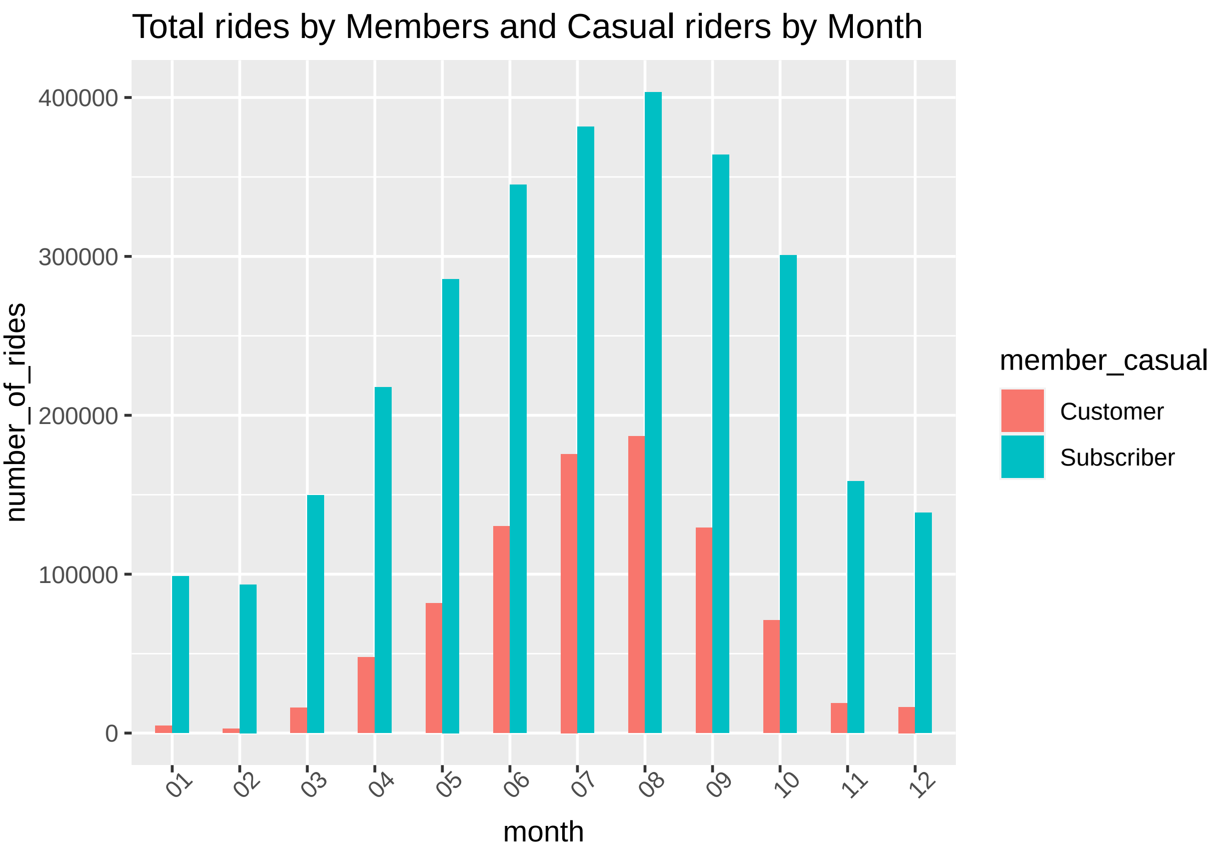
From the chart above, it can be seen that Subscribe is the group with the most number of rides on weekdays.

```r
ggplot(plot,aes(x = day_of_week, y = average_ride_length, fill = member_casual)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title ="Average ride time of Members and Casual riders Vs. Day of the week")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



From the chart above, we can observe that the Custumer group cycled for longer periods of the week with the highest number of rides on weekends while the Subscribers drove at a steady pace during the week with the highest number of rides. highest on weekends.

```
trip_data_clean %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(),.groups="drop") %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  labs(title ="Total rides by Members and Casual riders by Month") +
  theme(axis.text.x = element_text(angle = 45)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



From the chart above, we can see that the Subscriber group has a higher number of trips throughout the year

Act

For the last step in the data analysis process, we will make three recommendations to increase the number of Subscribers every year. But first, we'll lay out three key insights.

Key Findings:

1. Custumer rides the most on weekends. In contrast, Subscriber makes the most trips during the week.
2. On average, Subscriber rides shorter than Custumer.
3. There is no difference between Custumer and Subscriber in terms of the number of trips they make per month. Both Custumer and Subscriber have the highest number of trips in the summer months and the least number of trips in late winter and early spring.

Recommendations

1. Target Custumer bike rentals for weekend fun.
2. Create a big summer campaign when more people can afford to rent bikes.