Google Data Analytics Capstone Project.

Shruti Ghiya

2022-12-21

Introduction

I have recently completed the Google Data Analytics Certification Program on Coursera. The Final module of the program is a capstone project which is a showcase of my learning so far. The tools I chose to use in this project is R programming.

Scenario You are a junior data analyst working in the marketing

Analyst team at Cyclistic, a bike-share company in Chicago. 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. 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, The director of marketing 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, she 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.

Objective:

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

Prepare

About Dataset

This is public data that was use to explore how different customer types are using Cyclistic bikes. This data has been made available by Motivate International Inc. Dataset downloaded from the below link: https://divvy-tripdata.s3.amazonaws.com/index.htm For the analysis, I used just Divvy_trips data for quarter Q2 2019 – Q2 2020. The datasets have a different name because Cyclistic is a fictional company. The data is reliable because it was directly downloaded from AWS server and it is comprehensive current and cited. Data has some limitations and privacy issues that prohibit from using riders' personally identifiable information. This means that we won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Files Used :

- Divvy_Trips_2019_Q2
- Divvy_Trips_2019_Q3
- Divvy_Trips_2019_Q4
- Divvy_Trips_2020_Q1

File format .csv and has the following column names:

- ride_id • started_at • ended_at • rideable_type • duration • start_station_id • start_station_name • end_station_id • end_station_name • member_causal • gender • birthyear

Process

Tools used : RStudio is used for data cleaning and analyzing. Markdown report is created to verify data is clean and ready to analyze.

Collect Data

** Installing Package **

```
install.packages("tidyverse", repo="http://cran.rstudio.com/")
```

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

## package ’tidyverse’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Dell\AppData\Local\Temp\Rtmpeq1VC8\downloaded_packages
```

```
install.packages("lubridate", repo="http://cran.rstudio.com/")
```

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

## package ’lubridate’ successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package ’lubridate’

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Dell\AppData\Local\R\win-library\4.2\00LOCK\lubridate\libs\x64\lubridate.dll
## to
## C:\Users\Dell\AppData\Local\R\win-library\4.2\lubridate\libs\x64\lubridate.dll:
## Permission denied

## Warning: restored ’lubridate’

##
## The downloaded binary packages are in
##  C:\Users\Dell\AppData\Local\Temp\Rtmpeq1VC8\downloaded_packages
```

```
library(tidyverse)
```

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

## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

** Loading Data **

```
q2_2019 <- read.csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read.csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read.csv("Divvy_Trips_2019_Q4.csv")
q1_2020 <- read.csv("Divvy_Trips_2020_Q1.csv")
```

Warangling Data and Combining it into single file

Checking the column names for each data set.

```
colnames(q2_2019)
```

```
##  [1] "X01...Rental.Details.Rental.ID"
##  [2] "X01...Rental.Details.Local.Start.Time"
##  [3] "X01...Rental.Details.Local.End.Time"
##  [4] "X01...Rental.Details.Bike.ID"
##  [5] "X01...Rental.Details.Duration.In.Seconds.Uncapped"
##  [6] "X03...Rental.Start.Station.ID"
##  [7] "X03...Rental.Start.Station.Name"
##  [8] "X02...Rental.End.Station.ID"
##  [9] "X02...Rental.End.Station.Name"
## [10] "User.Type"
## [11] "Member.Gender"
## [12] "X05...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"
```

```
colnames(q1_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

While checking the structure and column names have noticed that the column names are not consistent in the data. And to join the data, columns should match. So renaming the columns matching q1_2020 below:

```
q4_2019 <- 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_2019 <- 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)
```

```
q2_2019 <- rename(q2_2019
                  ,ride_id = "X01...Rental.Details.Rental.ID"
                  ,rideable_type = "X01...Rental.Details.Bike.ID"
```

```
         ,started_at = "X01...Rental.Details.Local.Start.Time"
         ,ended_at = "X01...Rental.Details.Local.End.Time"
         ,start_station_name = "X03...Rental.Start.Station.Name"
         ,start_station_id = "X03...Rental.Start.Station.ID"
         ,end_station_name = "X02...Rental.End.Station.Name"
         ,end_station_id = "X02...Rental.End.Station.ID"
         ,member_casual = "User.Type")
```

Inspecting the data frame and looking for incongruities

```
str(q2_2019)
```

```
## 'data.frame':    1108163 obs. of  12 variables:
##  $ ride_id                                       : int  22178529 22178530 22178531 22178532 221785
##  $ started_at                                    : chr  "2019-04-01 00:02:22" "2019-04-01 00:03:0
##  $ ended_at                                      : chr  "2019-04-01 00:09:48" "2019-04-01 00:20:3
##  $ rideable_type                                 : int  6251 6226 5649 4151 3270 3123 6418 4513 32
##  $ X01...Rental.Details.Duration.In.Seconds.Uncapped: chr  "446.0" "1,048.0" "252.0" "357.0" ...
##  $ start_station_id                              : int  81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name                            : chr  "Daley Center Plaza" "Wood St & Taylor St
##  $ end_station_id                                : int  56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name                              : chr  "Desplaines St & Kinzie St" "Wabash Ave &
##  $ member_casual                                 : chr  "Subscriber" "Subscriber" "Subscriber" "Su
##  $ Member.Gender                                 : chr  "Male" "Female" "Male" "Male" ...
##  $ X05...Member.Details.Member.Birthday.Year     : int  1975 1984 1990 1993 1992 1999 1969 1991 N
```

```
str(q3_2019)
```

```
## 'data.frame':    1640718 obs. of  12 variables:
##  $ ride_id           : int  23479388 23479389 23479390 23479391 23479392 23479393 23479394 23479395 2
##  $ started_at        : chr  "2019-07-01 00:00:27" "2019-07-01 00:01:16" "2019-07-01 00:01:48" "2019-0
##  $ ended_at          : chr  "2019-07-01 00:20:41" "2019-07-01 00:18:44" "2019-07-01 00:27:42" "2019-0
##  $ rideable_type     : int  3591 5353 6180 5540 6014 4941 3770 5442 2957 6091 ...
##  $ tripduration      : chr  "1,214.0" "1,048.0" "1,554.0" "1,503.0" ...
##  $ start_station_id  : int  117 381 313 313 168 300 168 313 43 43 ...
##  $ start_station_name: chr  "Wilton Ave & Belmont Ave" "Western Ave & Monroe St" "Lakeview Ave & Full
##  $ end_station_id    : int  497 203 144 144 62 232 62 144 195 195 ...
##  $ end_station_name  : chr  "Kimball Ave & Belmont Ave" "Western Ave & 21st St" "Larrabee St & Webste
##  $ member_casual     : chr  "Subscriber" "Customer" "Customer" "Customer" ...
##  $ gender            : chr  "Male" "" "" "" ...
##  $ birthyear         : int  1992 NA NA NA NA 1990 NA NA NA NA ...
```

```
str(q4_2019)
```

```
## 'data.frame':    704054 obs. of  12 variables:
##  $ ride_id           : int  25223640 25223641 25223642 25223643 25223644 25223645 25223646 25223647 2
##  $ started_at        : chr  "2019-10-01 00:01:39" "2019-10-01 00:02:16" "2019-10-01 00:04:32" "2019-1
##  $ ended_at          : chr  "2019-10-01 00:17:20" "2019-10-01 00:06:34" "2019-10-01 00:18:43" "2019-1
##  $ rideable_type     : int  2215 6328 3003 3275 5294 1891 1061 1274 6011 2957 ...
##  $ tripduration      : chr  "940.0" "258.0" "850.0" "2,350.0" ...
##  $ start_station_id  : int  20 19 84 313 210 156 84 156 156 336 ...
##  $ start_station_name: chr  "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St" "Milwauke
##  $ end_station_id    : int  309 241 199 290 382 226 142 463 463 336 ...
##  $ end_station_name  : chr  "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave & Grand Av
##  $ member_casual     : chr  "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender            : chr  "Male" "Male" "Female" "Male" ...
##  $ birthyear         : int  1987 1998 1991 1990 1987 1994 1991 1995 1993 NA ...
```

```
str(q1_2020)
```

```
## 'data.frame':    426887 obs. of  13 variables:
##  $ ride_id           : chr  "EACB19130B0CDA4A" "8FED874C809DC021" "789F3C21E472CA96" "C9A388DAC6ABF3
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-01-21 20:06:59" "2020-01-30 14:22:39" "2020-01-09 19:29:26" "2020-0
##  $ ended_at          : chr  "2020-01-21 20:14:30" "2020-01-30 14:26:22" "2020-01-09 19:32:17" "2020-0
##  $ start_station_name: chr  "Western Ave & Leland Ave" "Clark St & Montrose Ave" "Broadway & Belmont
##  $ start_station_id  : int  239 234 296 51 66 212 96 96 212 38 ...
##  $ end_station_name  : chr  "Clark St & Leland Ave" "Southport Ave & Irving Park Rd" "Wilton Ave & B
##  $ end_station_id    : int  326 318 117 24 212 96 212 212 96 100 ...
##  $ start_lat         : num  42 42 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  42 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

ride_id and ride-able_type have inconsistent data type within the quarterly data. Converting them to character so that can be stacked properly.

```
q4_2019 <- mutate(q4_2019, ride_id = as.character(ride_id),
                  rideable_type = as.character(rideable_type))

q3_2019 <- mutate(q3_2019, ride_id = as.character(ride_id),
                  rideable_type = as.character(rideable_type))

q2_2019 <- mutate(q2_2019, ride_id = as.character(ride_id),
                  rideable_type = as.character(rideable_type))
```

Joining the individual quarter data into 1 single big data frame.

```
all_trips <- bind_rows(q2_2019, q3_2019, q4_2019, q1_2020)
```

There were few columns which were dropped beginning in 2020: * lat * long * birthyear * gender

Removing those columns for consistency.

```
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng,birthyear, gender,
            "X01...Rental.Details.Duration.In.Seconds.Uncapped",
            "X05...Member.Details.Member.Birthday.Year","Member.Gender",
            tripduration))
```

Cleaning and adding data to prepare for analysis

Inspecting the new table that has been created

```
nrow(all_trips)
```

```
## [1] 3879822
```

```
ncol(all_trips)
```

```
## [1] 9
```

```
dim(all_trips)
```

```
## [1] 3879822       9
```

```
head(all_trips)
```

```
##    ride_id          started_at            ended_at rideable_type
## 1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48          6251
## 2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30          6226
## 3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19          5649
## 4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58          4151
## 5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13          3270
## 6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56          3123
##   start_station_id        start_station_name end_station_id
## 1               81        Daley Center Plaza             56
## 2              317       Wood St & Taylor St             59
## 3              283 LaSalle St & Jackson Blvd            174
## 4               26  McClurg Ct & Illinois St            133
## 5              202      Halsted St & 18th St            129
## 6              420       Ellis Ave & 55th St            426
##            end_station_name member_casual
## 1 Desplaines St & Kinzie St    Subscriber
## 2 Wabash Ave & Roosevelt Rd    Subscriber
## 3     Canal St & Madison St    Subscriber
## 4  Kingsbury St & Kinzie St    Subscriber
## 5 Blue Island Ave & 18th St    Subscriber
## 6       Ellis Ave & 60th St    Subscriber
```

```
tail(all_trips)
```

```
##                  ride_id          started_at            ended_at rideable_type
## 3879817 6F4D221BDDFD943F 2020-03-10 10:40:27 2020-03-10 10:40:29   docked_bike
## 3879818 ADDAA33CEBCAE733 2020-03-10 10:40:06 2020-03-10 10:40:07   docked_bike
## 3879819 82B10FA3994BC66A 2020-03-07 15:25:55 2020-03-07 16:14:03   docked_bike
## 3879820 AA0D5AAA0B59C8AA 2020-03-01 13:12:38 2020-03-01 13:38:29   docked_bike
## 3879821 3296360A7BC20FB8 2020-03-07 18:02:45 2020-03-07 18:13:18   docked_bike
## 3879822 064EC7698E4FF9B3 2020-03-08 13:03:57 2020-03-08 13:32:27   docked_bike
##         start_station_id        start_station_name end_station_id
## 3879817              675                        HQ QR            675
## 3879818              675                        HQ QR            675
## 3879819              161     Rush St & Superior St            240
## 3879820              141    Clark St & Lincoln Ave            210
## 3879821              672 Franklin St & Illinois St            264
## 3879822              110     Dearborn St & Erie St             85
##                      end_station_name member_casual
## 3879817                        HQ QR        casual
## 3879818                        HQ QR        casual
## 3879819 Sheridan Rd & Irving Park Rd        member
## 3879820    Ashland Ave & Division St        casual
## 3879821 Stetson Ave & South Water St        member
## 3879822        Michigan Ave & Oak St        casual
```

```
str(all_trips)
```

```
## ’data.frame':    3879822 obs. of  9 variables:
##  $ ride_id           : chr  "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : chr  "2019-04-01 00:02:22" "2019-04-01 00:03:02" "2019-04-01 00:11:07" "2019-0
##  $ ended_at          : chr  "2019-04-01 00:09:48" "2019-04-01 00:20:30" "2019-04-01 00:15:19" "2019-0
##  $ rideable_type     : chr  "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : int  81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr  "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jackson Blvd" "
##  $ end_station_id    : int  56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr  "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madis
##  $ member_casual     : chr  "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
```

```
summary(all_trips)
```

```
##    ride_id           started_at          ended_at         rideable_type
##  Length:3879822     Length:3879822     Length:3879822     Length:3879822
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_id start_station_name end_station_id  end_station_name
##  Min.   :  1.0    Length:3879822     Min.   :  1.0   Length:3879822
##  1st Qu.: 77.0    Class :character   1st Qu.: 77.0   Class :character
##  Median :174.0    Mode  :character   Median :174.0   Mode  :character
##  Mean   :202.9                       Mean   :203.8
##  3rd Qu.:291.0                       3rd Qu.:291.0
##  Max.   :675.0                       Max.   :675.0
##                                      NA’s   :1
##  member_casual
##  Length:3879822
##  Class :character
##  Mode  :character
##
##
##
##
```

While checking data, have noticed few problems which we need to fix:

- The member_casual column: There are different names for members (" Subscriber", "member") and for causal riders ("Customer", "casual")

```
unique(all_trips$member_casual)
```

```
## [1] "Subscriber" "Customer"   "member"     "casual"
```

Consolidating four labels into two labels.

```
all_trips <- all_trips %>%
  mutate(member_casual = recode(member_casual, "Subscriber" = "member",
                               "Customer" = "casual"))
unique(all_trips$member_casual)
```

```
## [1] "member" "casual"
```

- We can aggregate the data at the ride-level. And for that, we need to add some additional columns such as day, month, year using started_at column.

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

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

- Now adding a calculated field to calculate the length of the ride as this column was removed from the data starting 2020.

```
all_trips$ride_length <- difftime(all_trips$ended_at, all_trips$started_at)
```

```
str(all_trips)
```

```
## 'data.frame':    3879822 obs. of  15 variables:
##  $ ride_id           : chr  "22178529" "22178530" "22178531" "22178532" ...
##  $ started_at        : chr  "2019-04-01 00:02:22" "2019-04-01 00:03:02" "2019-04-01 00:11:07" "2019-0
##  $ ended_at          : chr  "2019-04-01 00:09:48" "2019-04-01 00:20:30" "2019-04-01 00:15:19" "2019-0
##  $ rideable_type     : chr  "6251" "6226" "5649" "4151" ...
##  $ start_station_id  : int  81 317 283 26 202 420 503 260 211 211 ...
##  $ start_station_name: chr  "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jackson Blvd" "
##  $ end_station_id    : int  56 59 174 133 129 426 500 499 211 211 ...
##  $ end_station_name  : chr  "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madi
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  $ date              : Date, format: "2019-04-01" "2019-04-01" ...
##  $ month             : chr  "04" "04" "04" "04" ...
##  $ day               : chr  "01" "01" "01" "01" ...
##  $ year              : chr  "19" "19" "19" "19" ...
##  $ day_of_week       : chr  "Monday" "Monday" "Monday" "Monday" ...
##  $ ride_length       : 'difftime' num  446 1048 252 357 ...
##   ..- attr(*, "units")= chr "secs"
```

Ride length is different datatype so converting it to numeric so that can perform calculations.

```
is.factor(all_trips$ride_length)
```

```
## [1] FALSE
```

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

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

- There are few rides where trip duration was negative, which includes rides where the Divvy took bikes out of circulation for maintenance. We want to remove those many rides. So creating new version of data frame(v2)

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

Descriptive Analysis

Descriptive analysis on ride_length column(in seconds).

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     412     712    1479    1289 9383424
```

Descriptive analysis on member_casual Column (Char type)

```
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual, FUN=mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                3552.7741
## 2                     member                 850.0783
```

```
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual, FUN=median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                     1546
## 2                     member                      589
```

```
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual, FUN=max)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                  9383424
## 2                     member                  9056634
```

```
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual, FUN=min)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                        2
## 2                     member                        1
```

Average ride time for members vs casual riders by each day

```
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual+ all_trips_v2$day_of_week, FUN=mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Friday                3773.8351
## 2                      member                   Friday                 824.5385
## 3                      casual                   Monday                3372.2869
## 4                      member                   Monday                 842.5649
## 5                      casual                 Saturday                3331.8795
## 6                      member                 Saturday                 968.9962
## 7                      casual                   Sunday                3581.4054
## 8                      member                   Sunday                 920.0284
## 9                      casual                 Thursday                3683.0548
## 10                     member                 Thursday                 823.9278
## 11                     casual                  Tuesday                3596.3599
## 12                     member                  Tuesday                 826.1498
## 13                     casual                Wednesday                3718.8955
## 14                     member                Wednesday                 823.9996
```

Days of the week seems out of order so fixing it

```
all_trips_v2$day_of_week<- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday","Tuesday","Wed
# rechecking the order
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual+ all_trips_v2$day_of_week, FUN=mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Sunday                3581.4054
## 2                      member                   Sunday                 920.0284
## 3                      casual                   Monday                3372.2869
## 4                      member                   Monday                 842.5649
## 5                      casual                  Tuesday                3596.3599
## 6                      member                  Tuesday                 826.1498
## 7                      casual                Wednesday                3718.8955
## 8                      member                Wednesday                 823.9996
## 9                      casual                 Thursday                3683.0548
## 10                     member                 Thursday                 823.9278
## 11                     casual                   Friday                3773.8351
## 12                     member                   Friday                 824.5385
## 13                     casual                 Saturday                3331.8795
## 14                     member                 Saturday                 968.9962
```

Now analyzing riders data based on Type and weekday

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,day_of_week)
```

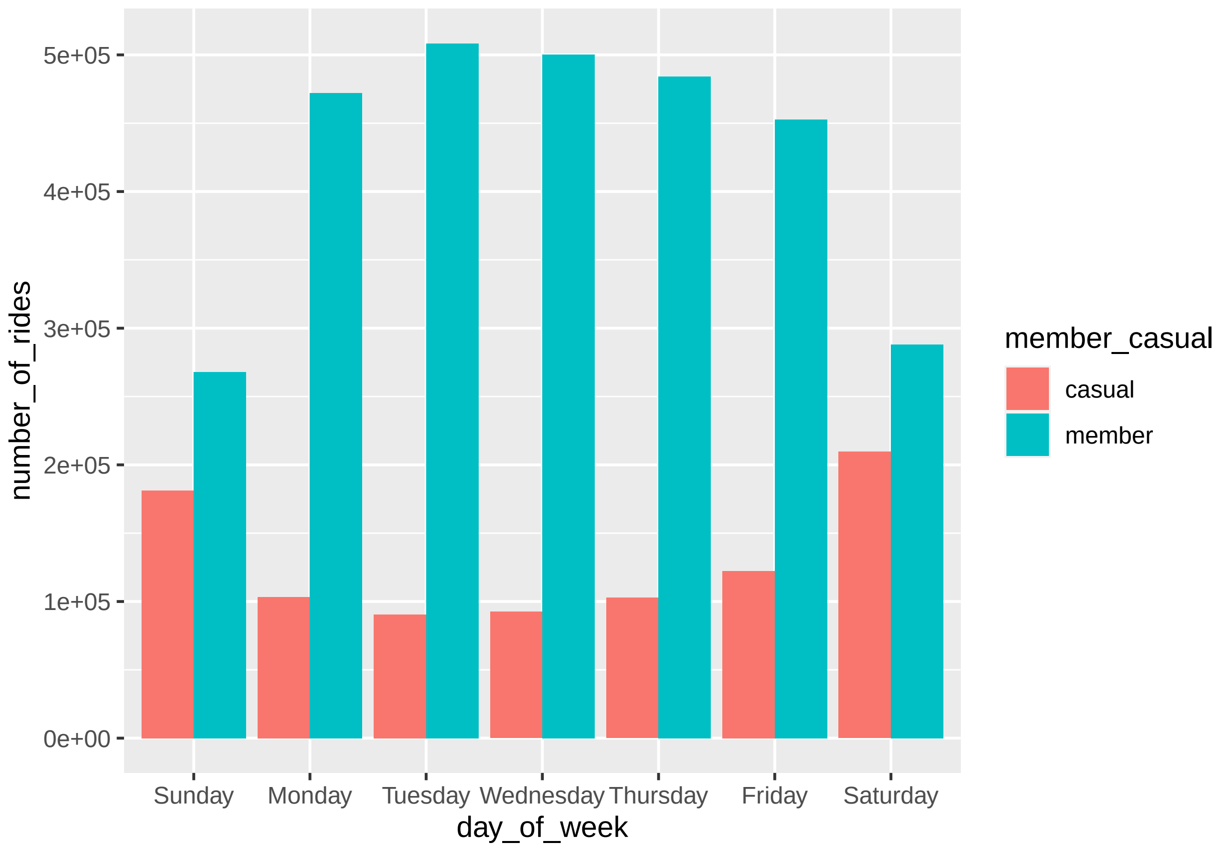
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                 <int>            <dbl>
##  1 casual        Sunday               181293            3581.
##  2 casual        Monday               103296            3372.
##  3 casual        Tuesday               90510            3596.
##  4 casual        Wednesday             92457            3719.
##  5 casual        Thursday             102679            3683.
##  6 casual        Friday               122404            3774.
##  7 casual        Saturday             209543            3332.
##  8 member        Sunday               267965             920.
##  9 member        Monday               472196             843.
## 10 member        Tuesday              508445             826.
## 11 member        Wednesday            500329             824.
## 12 member        Thursday             484177             824.
## 13 member        Friday               452790             825.
## 14 member        Saturday             287958             969.
```

Let's visualize these numbers

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

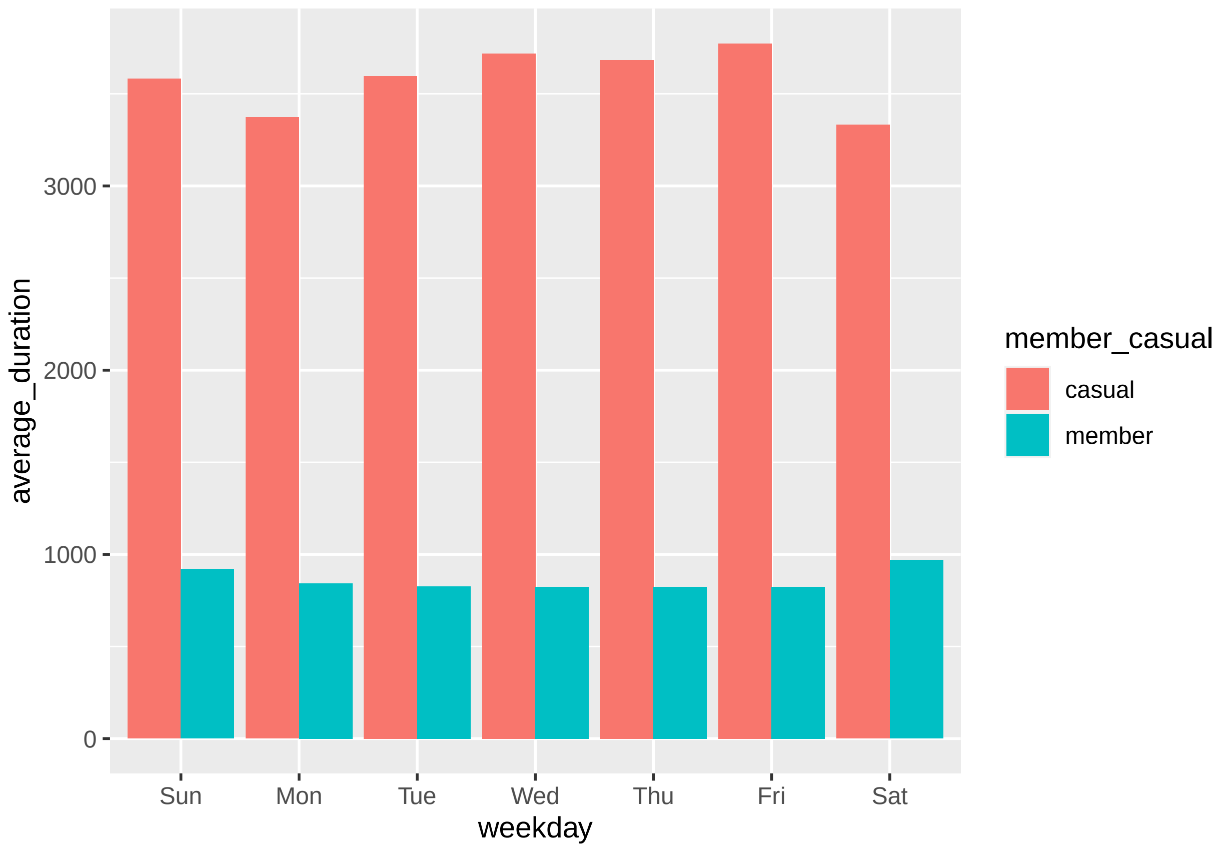
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Let’s visualize the number of rides by average duration

```
all_trips_v2 %>%
  mutate(weekday=wday(started_at,label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,weekday) %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

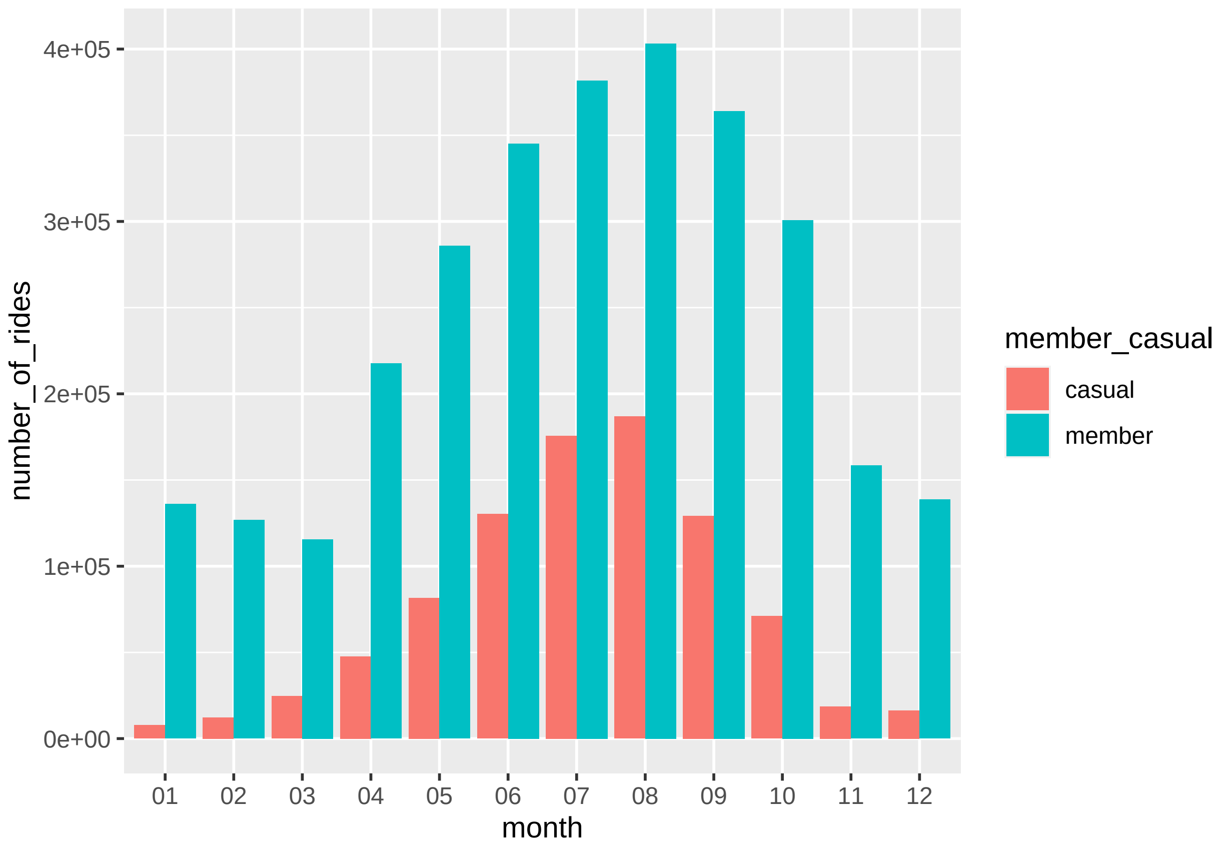
```
## ‘summarise()‘ has grouped output by ’member_casual’. You can override using the
## ‘.groups‘ argument.
```



Let's visualize by rider type and month

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,month) %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

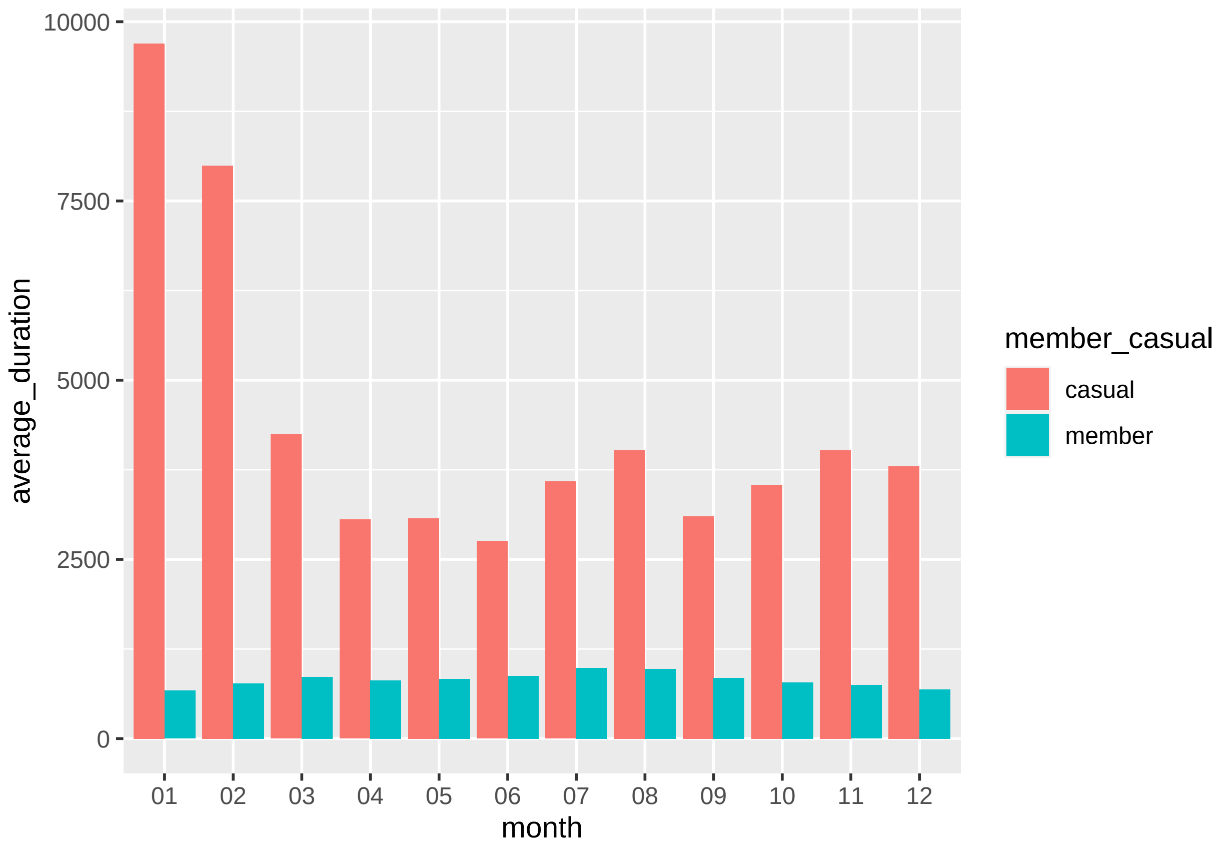
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Let's visualize the by average duration throught out the year.

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,month) %>%
  ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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



Exporting summary file for further analysis

```
counts <- aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FU
write.csv(counts, file = 'avg_ride_length.csv')
```

Conclusion

- Number of rides through out the week is more for members vs causal
- Less number of rides but more length duration for causal vs members
- Number of rides are more from Apr-Oct for both member and causal.

Recommendations

- *Can start the rewards program for the membership sign- up. And start some campaigns to attract more users to sign up.*
- As casual riders take longer trips, we can offer cheaper ride when member vs causal.