Cyclistic_case

Abdallah Zubedi

2023-03-18

Contents

Introduction

This analysis report is based on case study 1 provided by Google. It ultimately showcases my analysis and visualization skills using R and Tableau.

Cyclistic is a company providing bicycles in a rented capacity. The 3 bikes include **electrical**, **docked** and the **classic bikes**.

There are two categories of customers:-

- **Members**: customers who pay a premium annually for perks
- **Casual**: no membership payments

The case describes a scenario whereby Cyclistic wants to convert casual riders to members. The main question asked is "How do both segments differ in usage of Cyclistic?"

The following will cover the **preparation**, **processing**, **analyzing**, **visualizing** and **recommendation** phases.

The source of the data is located: here

The license is located here: here

Preparation

The files received are in a csv format. 12 files for 12 months, starting from February 2022 to January 2023.

The files were massive in size, which is why I used R desktop to process the data. BigQuery required an upgrade form the free tier (I am cheap) while excel become extremely slow once the data was imported.

```
Y2022_02 <- read.csv("Y2022_02.csv")
Y2022_03 <- read.csv("Y2022_03.csv")
Y2022_04 <- read.csv("Y2022_04.csv")
Y2022_05 <- read.csv("Y2022_05.csv")
Y2022_06 <- read.csv("Y2022_06.csv")
Y2022_07 <- read.csv("Y2022_07.csv")
Y2022_08 <- read.csv("Y2022_08.csv")
Y2022_09 <- read.csv("Y2022_09.csv")
Y2022_10 <- read.csv("Y2022_10.csv")
Y2022_11 <- read.csv("Y2022_11.csv")
Y2022_12 <- read.csv("Y2022_12.csv")
Y2023_01 <- read.csv("Y2023_01.csv")
```

I loaded the necessary packages and looked into the structures of the files to get an understanding of the layout and classifications of each column.

```
##Initializing setup post installing packages

library("tidyverse")

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.1.8
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

library("readr")
library("ggplot2")
library("dplyr")
library("tibble")
library("lubridate")
library("tidyr")
library("pryr")

##
## Attaching package: 'pryr'
##
## The following object is masked from 'package:dplyr':
##
##     where
##
## The following objects are masked from 'package:purrr':
##
##     compose, partial

library("janitor")

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

##Checking column headers for inconsistencies
colnames(Y2022_02)
```

```
##  [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"
```

```
colnames(Y2022_03)
```

```
##  [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"
```

```
colnames(Y2022_04)
```

```
##  [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"
```

```
colnames(Y2022_05)
```

```
##  [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"
```

```
colnames(Y2022_06)
```

```
##  [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"
```

```
colnames(Y2022_07)
```

```
##  [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"
```

```
colnames(Y2022_08)
```

```
##  [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"

colnames(Y2022_09)

##  [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"

colnames(Y2022_10)

##  [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"

colnames(Y2022_11)

##  [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"

colnames(Y2022_12)

##  [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"

colnames(Y2023_01)

##  [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"

##Checking structure
str(Y2022_02)

## '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" ...

str(Y2022_03)

## '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" ...

str(Y2022_04)

## '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" ...

str(Y2022_05)

## 'data.frame':    634858 obs. of  13 variables:
##  $ ride_id           : chr  "EC2DE40644C6B0F4" "1C31AD03897EE385"
"1542FBEC830415CF" "6FF59852924528F8" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike"
"classic_bike" ...
##  $ started_at        : chr  "2022-05-23 23:06:58" "2022-05-11 08:53:28"
"2022-05-26 18:36:28" "2022-05-10 07:30:07" ...
##  $ ended_at          : chr  "2022-05-23 23:40:19" "2022-05-11 09:31:22"
"2022-05-26 18:58:18" "2022-05-10 07:38:49" ...
##  $ start_station_name: chr  "Wabash Ave & Grand Ave" "DuSable Lake Shore
Dr & Monroe St" "Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr  "TA1307000117" "13300" "TA1305000032"
"TA1305000032" ...
##  $ end_station_name  : chr  "Halsted St & Roscoe St" "Field Blvd & South
Water St" "Wood St & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr  "TA1309000025" "15534" "13221" "TA1305000030"
...
##  $ 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 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...

str(Y2022_06)

## 'data.frame':    769204 obs. of  13 variables:
##  $ ride_id           : chr  "600CFD130D0FD2A4" "F5E6B5C1682C6464"
"B6EB6D27BAD771D2" "C9C320375DE1D5C6" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2022-06-30 17:27:53" "2022-06-30 18:39:52"
"2022-06-30 11:49:25" "2022-06-30 11:15:25" ...
##  $ ended_at          : chr  "2022-06-30 17:35:15" "2022-06-30 18:47:28"
"2022-06-30 12:02:54" "2022-06-30 11:19:43" ...
```

```
##  $ 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 41.8 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...

str(Y2022_07)

## 'data.frame':    823488 obs. of  13 variables:
##  $ ride_id           : chr  "954144C2F67B1932" "292E027607D218B6"
"57765852588AD6E0" "B5B6BE44314590E6" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike"
"classic_bike" ...
##  $ started_at        : chr  "2022-07-05 08:12:47" "2022-07-26 12:53:38"
"2022-07-03 13:58:49" "2022-07-31 17:44:21" ...
##  $ ended_at          : chr  "2022-07-05 08:24:32" "2022-07-26 12:55:31"
"2022-07-03 14:06:32" "2022-07-31 18:42:50" ...
##  $ start_station_name: chr  "Ashland Ave & Blackhawk St" "Buckingham
Fountain (Temp)" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)"
...
##  $ start_station_id  : chr  "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Michigan Ave & 8th
St" "Michigan Ave & 8th St" "Woodlawn Ave & 55th St" ...
##  $ end_station_id    : chr  "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...

str(Y2022_08)

## 'data.frame':    785932 obs. of  13 variables:
##  $ ride_id           : chr  "550CF7EFEAE0C618" "DAD198F405F9C5F5"
"E6F2BC47B65CB7FD" "F597830181C2E13C" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2022-08-07 21:34:15" "2022-08-08 14:39:21"
"2022-08-08 15:29:50" "2022-08-08 02:43:50" ...
##  $ ended_at          : chr  "2022-08-07 21:41:46" "2022-08-08 14:53:23"
"2022-08-08 15:40:34" "2022-08-08 02:58:53" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
```

```
##  $ end_lat           : num  41.9 41.9 42 42 41.8 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(Y2022_09)
```

```
## 'data.frame':    701339 obs. of  13 variables:
##  $ ride_id           : chr  "5156990AC19CA285" "E12D4A16BF51C274"
"A02B53CD7DB72DD7" "C82E05FEE872DF11" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2022-09-01 08:36:22" "2022-09-01 17:11:29"
"2022-09-01 17:15:50" "2022-09-01 09:00:28" ...
##  $ ended_at          : chr  "2022-09-01 08:39:05" "2022-09-01 17:14:45"
"2022-09-01 17:16:12" "2022-09-01 09:10:32" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "California Ave & Milwaukee Ave" "" "" "" ...
##  $ end_station_id    : chr  "13084" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(Y2022_10)
```

```
## 'data.frame':    558685 obs. of  13 variables:
##  $ ride_id           : chr  "A50255C1E17942AB" "DB692A70BD2DD4E3"
"3C02727AAF60F873" "47E653FDC2D99236" ...
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike"
"electric_bike" ...
##  $ started_at        : chr  "2022-10-14 17:13:30" "2022-10-01 16:29:26"
"2022-10-19 18:55:40" "2022-10-31 07:52:36" ...
##  $ ended_at          : chr  "2022-10-14 17:19:39" "2022-10-01 16:49:06"
"2022-10-19 19:03:30" "2022-10-31 07:58:49" ...
##  $ start_station_name: chr  "Noble St & Milwaukee Ave" "Damen Ave &
Charleston St" "Hoyne Ave & Balmoral Ave" "Rush St & Cedar St" ...
##  $ start_station_id  : chr  "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr  "Larrabee St & Division St" "Damen Ave &
Cullerton St" "Western Ave & Leland Ave" "Orleans St & Chestnut St (NEXT
Apts)" ...
##  $ end_station_id    : chr  "KA1504000079" "13089" "TA1307000140" "620"
...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```
str(Y2022_11)
```

```
## 'data.frame':    337735 obs. of  13 variables:
##  $ ride_id           : chr  "BCC66FC6FAB27CC7" "772AB67E902C180F"
"585EAD07FDEC0152" "91C4E7ED3C262FF9" ...
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike"
"classic_bike" ...
##  $ started_at        : chr  "2022-11-10 06:21:55" "2022-11-04 07:31:55"
"2022-11-21 17:20:29" "2022-11-25 17:29:34" ...
##  $ ended_at          : chr  "2022-11-10 06:31:27" "2022-11-04 07:46:25"
"2022-11-21 17:34:36" "2022-11-25 17:45:15" ...
##  $ start_station_name: chr  "Canal St & Adams St" "Canal St & Adams St"
"Indiana Ave & Roosevelt Rd" "Indiana Ave & Roosevelt Rd" ...
##  $ start_station_id  : chr  "13011" "13011" "SL-005" "SL-005" ...
##  $ end_station_name  : chr  "St. Clair St & Erie St" "St. Clair St & Erie
St" "St. Clair St & Erie St" "St. Clair St & Erie St" ...
##  $ end_station_id    : chr  "13016" "13016" "13016" "13016" ...
##  $ 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 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...

str(Y2022_12)

## 'data.frame':    181806 obs. of  13 variables:
##  $ ride_id           : chr  "65DBD2F447EC51C2" "0C201AA7EA0EA1AD"
"E0B148CCB358A49D" "54C5775D2B7C9188" ...
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "electric_bike"
"classic_bike" ...
##  $ started_at        : chr  "2022-12-05 10:47:18" "2022-12-18 06:42:33"
"2022-12-13 08:47:45" "2022-12-13 18:50:47" ...
##  $ ended_at          : chr  "2022-12-05 10:56:34" "2022-12-18 07:08:44"
"2022-12-13 08:59:51" "2022-12-13 19:19:48" ...
##  $ start_station_name: chr  "Clifton Ave & Armitage Ave" "Broadway &
Belmont Ave" "Sangamon St & Lake St" "Shields Ave & 31st St" ...
##  $ start_station_id  : chr  "TA1307000163" "13277" "TA1306000015"
"KA1503000038" ...
##  $ end_station_name  : chr  "Sedgwick St & Webster Ave" "Sedgwick St &
Webster Ave" "St. Clair St & Erie St" "Damen Ave & Madison St" ...
##  $ end_station_id    : chr  "13191" "13191" "13016" "13134" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -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 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...

str(Y2023_01)

## 'data.frame':    190301 obs. of  13 variables:
##  $ ride_id           : chr  "F96D5A74A3E41399" "13CB7EB698CEDB88"
"BD88A2E670661CE5" "C90792D034FED968" ...
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "electric_bike"
```

```
"classic_bike" ...
##  $ started_at        : chr  "2023-01-21 20:05:42" "2023-01-10 15:37:36"
"2023-01-02 07:51:57" "2023-01-22 10:52:58" ...
##  $ ended_at          : chr  "2023-01-21 20:16:33" "2023-01-10 15:46:05"
"2023-01-02 08:05:11" "2023-01-22 11:01:44" ...
##  $ start_station_name: chr  "Lincoln Ave & Fullerton Ave" "Kimbark Ave &
53rd St" "Western Ave & Lunt Ave" "Kimbark Ave & 53rd St" ...
##  $ start_station_id  : chr  "TA1309000058" "TA1309000037" "RP-005"
"TA1309000037" ...
##  $ end_station_name  : chr  "Hampden Ct & Diversey Ave" "Greenwood Ave &
47th St" "Valli Produce - Evanston Plaza" "Greenwood Ave & 47th St" ...
##  $ end_station_id    : chr  "202480.0" "TA1308000002" "599" "TA1308000002"
...
##  $ start_lat         : num  41.9 41.8 42 41.8 41.8 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.8 42 41.8 41.8 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "casual" "member" ...
```

After a quick glimpse into the data, I decided to merge the files.

```
#Merged data files to one
Full_year <-
bind_rows(Y2022_02,Y2022_03,Y2022_04,Y2022_05,Y2022_06,Y2022_07,Y2022_08,Y202
2_09,Y2022_10,Y2022_11,Y2022_12,Y2023_01)
```

I then calculated the ride time and converted it to numeric for analysis.

```
#Ride_length calculation (sec)
Full_year$ride_length <- difftime(Full_year$ended_at, Full_year$started_at,
units = "secs")

#Convert "ride_length" from factor to numeric so we can run calculations on
the data
is.factor(Full_year$ride_length)

## [1] FALSE

Full_year$ride_length <- as.numeric(as.character(Full_year$ride_length))
is.numeric(Full_year$ride_length)

## [1] TRUE
```

From the datetime column, I created more columns on date, month, day, year and day of the week. This would help in the analysis section.

```
##Creating date,month,day,year and day of the week column
Full_year$date <- as.Date(Full_year$started_at)
Full_year$month <- format(as.Date(Full_year$date), "%m")
Full_year$day <- format(as.Date(Full_year$date), "%d")
Full_year$year <- format(as.Date(Full_year$date), "%y")
```

```
Full_year$day_of_week <- format(as.Date(Full_year$date), "%a")
View(Full_year)
```

A quick note: I also had a backup named Full_year_v1. you might be asking why? well I worked with the two files just so that if one goes wrong (loss of data,error in calculation) I would revert to the other file as a back up.

To organize the data set, I arranged the day of the week and ride length in ascending order.

```
##Arrange by ascending order
Full_year <- arrange(Full_year,day_of_week)
Full_year <- arrange(Full_year,ride_length)
View(Full_year)
```

I noticed blank/NA and negative values which were missing in data collection. I decided to remove the rows with missing data as we had enough sample to run an analysis.

```
##Removing n.a values
Full_year <- na.omit(Full_year)

##Remove blanks from rows
##Cannot use complete.cases for whole dataframe due to differing data formats
Full_year <- Full_year[Full_year$start_station_id != "", ]
Full_year <- Full_year[Full_year$start_station_name != "", ]
Full_year <- Full_year[Full_year$end_station_id != "", ]
Full_year <- Full_year[Full_year$end_station_name !="", ]
```

I then looked into the min and max values of ride length where I found some discrepancies such as ride length having 0sec and 12sec ride lengths.

To clean the data I removed anything that was 30seconds and below. Assuming riders were taking short rides. I used 30seconds as the cutoff as I communicated with some friends who use services such as Cyclistic to find out why would anyone use a rented bicycle for 30seconds. The answer was “30 seconds on a bicycle can take you a long way”.

```
#Removing values that are below 20s in ride_length_sec
Full_year <- Full_year[!(Full_year$ride_length <= 30), ]
```

Analysis

Started of with descriptive analysis

```
##Descriptive analysis
mean(Full_year$ride_length)

## [1] 1029.53

median(Full_year$ride_length)

## [1] 639
```

```
max(Full_year$ride_length)

## [1] 2057644

min(Full_year$ride_length)

## [1] 31

summary(Full_year$ride_length)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      31     368     639    1030    1142 2057644
```

The descriptive analysis showed that there might be outliers such as the max time. Upon running further analysis using order function, I saw there were some other entries similar to the time the max was in which I then assumed was not an outlier.

I compared how differently both segments use bicycles.

```
##Comparing members and casual riders
aggregate(Full_year$ride_length ~ Full_year$member_casual, FUN = mean)

##   Full_year$member_casual Full_year$ride_length
## 1                  casual             1442.6330
## 2                  member              753.2122

aggregate(Full_year$ride_length ~ Full_year$member_casual, FUN = median)

##   Full_year$member_casual Full_year$ride_length
## 1                  casual                   834
## 2                  member                   542

aggregate(Full_year$ride_length ~ Full_year$member_casual, FUN = max)

##   Full_year$member_casual Full_year$ride_length
## 1                  casual               2057644
## 2                  member                 89872

aggregate(Full_year$ride_length ~ Full_year$member_casual, FUN = min)

##   Full_year$member_casual Full_year$ride_length
## 1                  casual                    31
## 2                  member                    31
```

It was clear to see the casual riders were riding for longer lengths in time compared to the members. This made me ask if it might be due to having more casual riders compared to members?

To confirm I used the function table to count how many casuals and members have taken rides.

```
##How many members and casuals are there using table
table(Full_year$member_casual)
```

```
##
##  casual  member
## 1758311 2628726
```

There are more members than casuals, which can mean casual riders use bikes longer while members are more frequent but have a shorter ride time.

I decided to complete the ride length calculations first

```
##Average ride length for the day of the week vs type of user
aggregate(Full_year$ride_length ~ Full_year$member_casual +
Full_year$day_of_week, FUN = mean)
```

```
##    Full_year$member_casual Full_year$day_of_week Full_year$ride_length
## 1                   casual                   Fri             1345.9908
## 2                   member                   Fri              740.3594
## 3                   casual                   Mon             1492.3934
## 4                   member                   Mon              726.8219
## 5                   casual                   Sat             1613.4828
## 6                   member                   Sat              848.0498
## 7                   casual                   Sun             1643.5631
## 8                   member                   Sun              839.4142
## 9                   casual                   Thu             1278.7436
## 10                  member                   Thu              727.8500
## 11                  casual                   Tue             1289.4242
## 12                  member                   Tue              711.7332
## 13                  casual                   Wed             1240.0087
## 14                  member                   Wed              717.0111
```

This validated that the casuals do definitely have more ride time compared to members.

I also looked into the use of the different bikes offered

```
Full_year %>%
  group_by(rideable_type) %>%
  summarise(mean_ride_length = mean(ride_length))
```

```
## # A tibble: 3 × 2
##   rideable_type mean_ride_length
##   <chr>                   <dbl>
## 1 classic_bike            1030.
## 2 docked_bike             3018.
## 3 electric_bike            813.
```

The classic bikes were used the most then the docked bikes and electric bikes.

I also wanted to see which segment used different bikes on average

```
Full_year %>%
  group_by(rideable_type, member_casual) %>%
  summarise(mean_ride_length = mean(ride_length))
```

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

## # A tibble: 5 × 3
## # Groups:   rideable_type [3]
##   rideable_type member_casual mean_ride_length
##   <chr>         <chr>                   <dbl>
## 1 classic_bike  casual                  1475.
## 2 classic_bike  member                   800.
## 3 docked_bike   casual                  3018.
## 4 electric_bike casual                  1006.
## 5 electric_bike member                   665.
```

What stood out was that the docked bikes were only used by casual riders. This will be described more on the limitations sections.

```
##Analyze ride behavior by type and day of week
Full_year  %>%
  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)

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

## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              301525            1644.
##  2 casual        Mon              211401            1492.
##  3 casual        Tue              198038            1289.
##  4 casual        Wed              204495            1240.
##  5 casual        Thu              229752            1279.
##  6 casual        Fri              248314            1346.
##  7 casual        Sat              364786            1613.
##  8 member        Sun              298896             839.
##  9 member        Mon              377916             727.
## 10 member        Tue              418992             712.
## 11 member        Wed              417052             717.
## 12 member        Thu              417251             728.
## 13 member        Fri              362039             740.
## 14 member        Sat              336580             848.
```

The top 5 in ride length were the casual riders. I noticed members had shorter ride times but had more frequently taken bikes from bike stations.

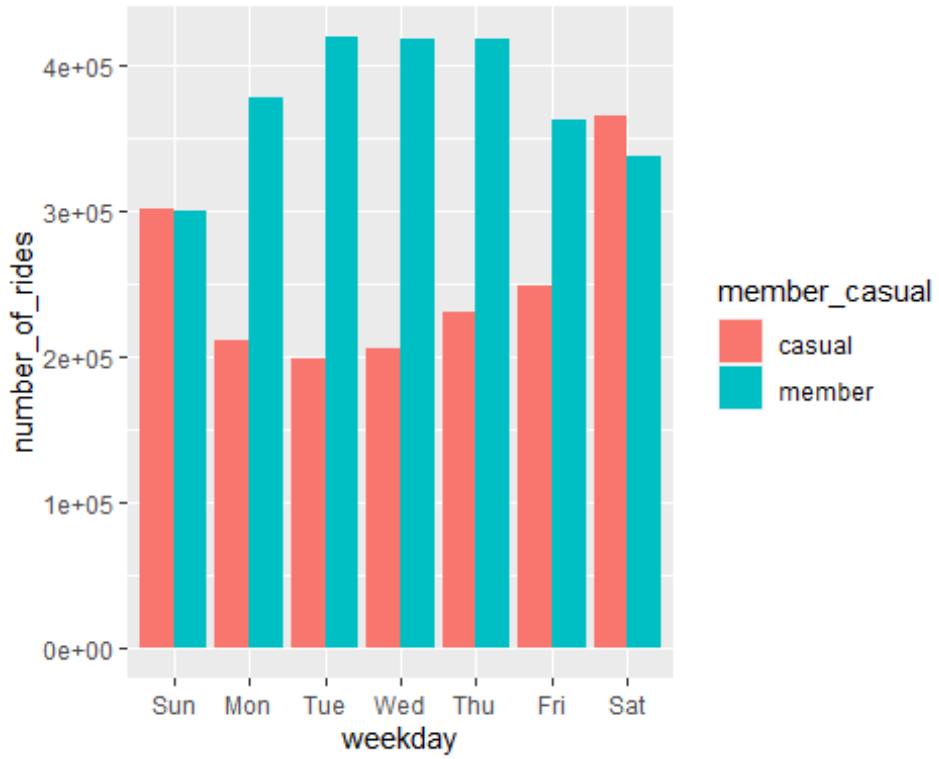
From the analysis, I understood that casual riders had more time spent riding bicycles but would less frequently take bikes from the stations, while members rented bicycles more frequently. This can be due to certain characteristics that members have, which led to visualizing the data for patterns.

Visualization

Started looking into the number of rides taken by rider type to establish visual differences.

```
##Visualize: number of rides by rider type
Full_year %>%
  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 = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

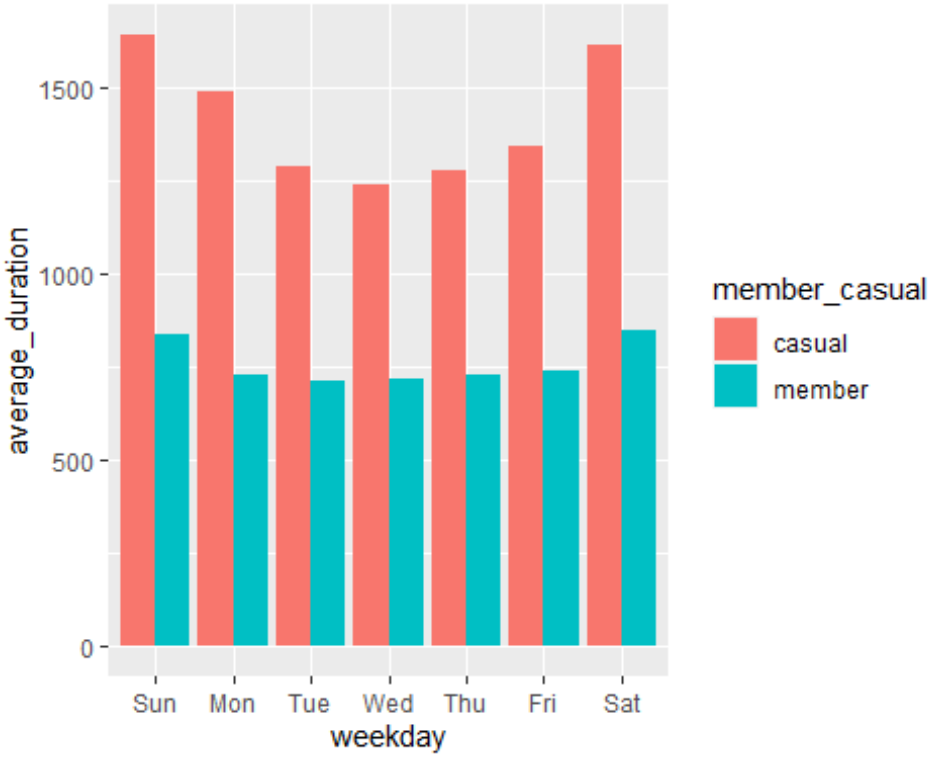
```
## `summarise()` has grouped output by 'member_casual'. You can override
using the
## `.groups` argument.
```



It is clear that members take more rides than casuals, with the peaks being Tuesday - Thursday. While casuals had would take rides during the weekends.

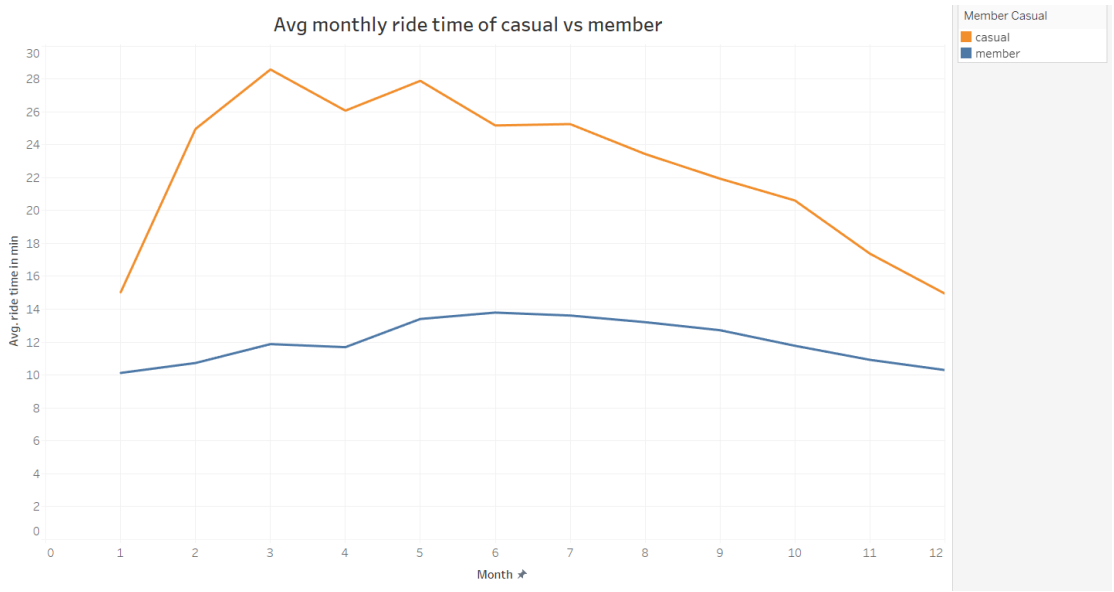
```r
##Visualize: average ride length for rider type
Full_year %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = 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.
```



In terms of ride time, casuals would ride the most on weekends, but still had more ride time than members every day of the week.

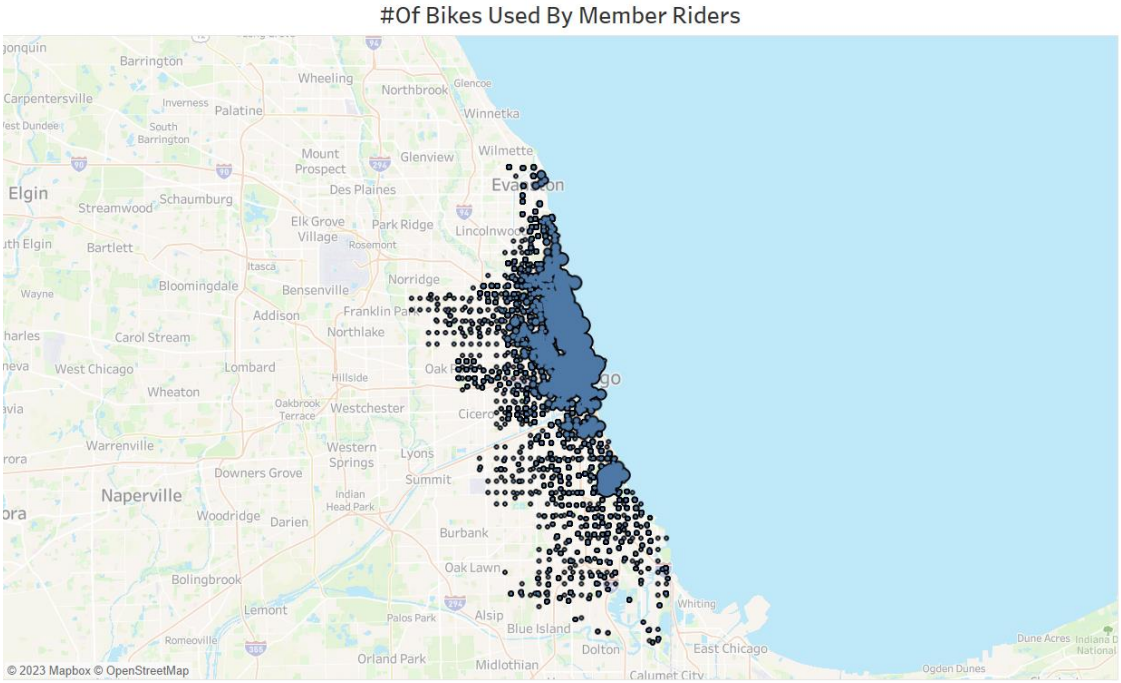
I also used Tableau to gain more insights.



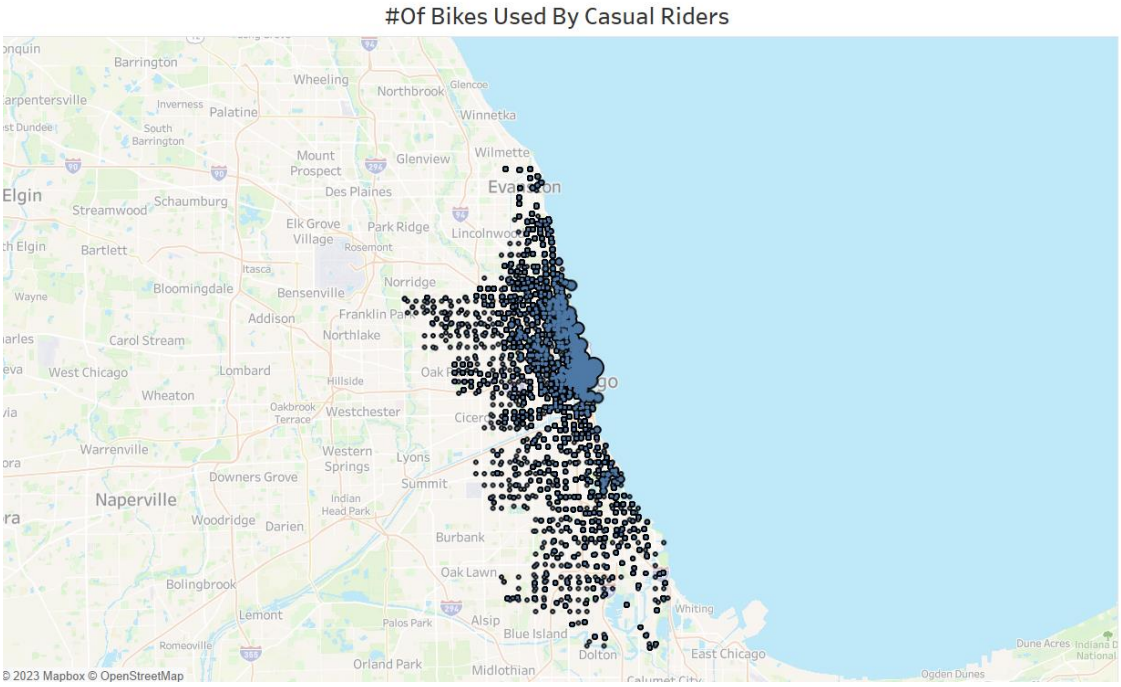
Avg.Ride Time monthly

While members were riding consistently throughout the month, casuals had 3 peaks which were March, May and July. Overall there was more riding time with casuals compared to members throughout the year.

I decided to map out where the riders would take bikes the most in Chicago.



Number of bikes used by Member riders



Number of bikes used by Casual riders

Member riders would take bikes out of stations that were more in the business district compared to casual riders taking bikes out of stations at parks and the pier.

Thoughts and Recommendations

Based on the analysis, we look at the objective which is to find the difference and convert casual riders to members via marketing strategies. We have discovered that casual riders have more ride time while members take more bicycles frequently. We also know that casual riders take rides on the weekend the most, with Sunday being the peak. While members ride the most on the weekdays.

Some assumptions we can make with the information provided and the analysis are the following:

- Member riders are using bikes the most for work since it's a weekday and also consistently through out the months.
- Member riders are more situated in the central area of Chicago (Central Business District).
- Member riders take bikes more frequently.
- Member riders ride have short ride times.
- Casual riders are using bikes the most for leisure time during the weekend and also have high peaks in March to the summer.

- Casual riders start rides locations are mostly situated at the pier and parks.
- Casual riders take bikes less frequently.
- Casual riders have longer ride times.

The following are the **top recommendations** on what marketing strategies to take to convert casual riders to members:

- Marketing campaigns towards casual members should go active mainly on the weekends (Friday to Sunday) as this is when the casual riders use the bikes most of the time.
- Marketing campaigns can be implemented toward leisure.
  Here are some examples of what can be done:
  - Getting a membership would provide a friend discount (which is capped), so if one of the casual riders finds value in the annual membership, word of mouth takes place
  - Partnering with the piers and parks during the weekend to provide more value for the parties involved
- Marketing campaigns aimed to target the months of March to July as casual riders used bikes most during the 4 months. This can be complemented with products/services that are themed for the month.
- Showcase membership as being a discounted rate in the long run

Limitations:

- Docked bikes has not been described in detail, as there is around 8000+ entries of docked bikes being used by casuals. The decision to not remove this is due to not having sufficient data to disregard the docked bikes category.
- Asking questions is a big part of the data analysis process, with a case such as this we can only make assumptions via researching information.
- Having limited knowledge of Chicago and the culture as I am not based there. Having general understanding in regards to the location with the data analysis process, I find is vital to be more accurate.
- Rows with missing data had to be removed. In this case Cyclsitic should figure out what went wrong and attempt to minimize the missing data. I mention this as there was around 20% of missing data.

Overall the case was interesting to process and analyze. There would be some focus points mentioned in the limitations sections that Cyclisitic needs to look into, such as the recording of data. This would greatly streamline the processing stages. Personally, it has

been a great experience and I look forward to learning more about how to leverage R and Tableau more effectively.