Google Data Analytics Capstone

Tony Xie

25/04/2022

Google Data Analytics Capstone

Cyclistic Bike-share Analysis Case Study

Introduction

The Cyclistic Bike-share Analysis Case Study has been selected as the capstone project for the Google Data Analytics Professional Certificate Program. This case study involves analyzing the dataset provided by the fictional company named Cyclistic following the steps of the data analysis process (ask, prepare, process, analyze, share and act) to develop a new marketing strategy for the company.

Ask

The business task of Cyclistic is to maximize the number of annual memberships by converting casual riders into annual member. The data analysis goal is to gain understanding of the differences between annual members and casual riders, the reason why casual riders would decide to purchase annual membership and how digital media could influence their marketing strategies.

Prepare

The dataset consist of bike ride data from April 2021 to March 2022. The raw data is stored in csv format which includes details of each ride such as start time, end time, start station name, end station name, start location coordinates, end location coordinates and membership type.

Load R packages required for analysis.

```
# Install packages
library(tidyverse) # collection of R packages for data science
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(lubridate) # package to work with date-time
```

```
##
## Attaching package: 'lubridate'

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

```
library(janitor) # package for cleaning data
```

```
##
## Attaching package: 'janitor'

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

```
library(ggplot2) # package to create data visualizations
```

Import bike ride raw datasets from April 2021 to March 2022.

```
m04_2021 <- read.csv('202104-divvy-tripdata.csv')
m05_2021 <- read.csv('202105-divvy-tripdata.csv')
m06_2021 <- read.csv('202106-divvy-tripdata.csv')
m07_2021 <- read.csv('202107-divvy-tripdata.csv')
m08_2021 <- read.csv('202108-divvy-tripdata.csv')
m09_2021 <- read.csv('202109-divvy-tripdata.csv')
m10_2021 <- read.csv('202110-divvy-tripdata.csv')
m11_2021 <- read.csv('202111-divvy-tripdata.csv')
m12_2021 <- read.csv('202112-divvy-tripdata.csv')
m01_2022 <- read.csv('202201-divvy-tripdata.csv')
m02_2022 <- read.csv('202202-divvy-tripdata.csv')
m03_2022 <- read.csv('202203-divvy-tripdata.csv')
```

Examine the column name of the April 2021 bike ride dataset.

```
colnames(m04_2021)
```

```
##  [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"
```

Check if column names are consistent across all files.

```
colnames(m04_2021) == colnames(m05_2021)
```

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

```
colnames(m05_2021) == colnames(m06_2021)
```

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

```
colnames(m06_2021) == colnames(m07_2021)
```

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

```
colnames(m07_2021) == colnames(m08_2021)
```

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

```
colnames(m08_2021) == colnames(m09_2021)
```

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

```
colnames(m09_2021) == colnames(m10_2021)
```

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

```
colnames(m10_2021) == colnames(m11_2021)
```

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

```
colnames(m11_2021) == colnames(m12_2021)
```

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

```
colnames(m12_2021) == colnames(m01_2022)
```

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

```
colnames(m01_2022) == colnames(m02_2022)
```

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

```
colnames(m02_2022) == colnames(m03_2022)
```

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

Combine data from each month into one data frame.

```
bike_trips <- rbind(m04_2021, m05_2021, m06_2021, m07_2021, m08_2021, m09_2021, m10_2021, m11_2021, m12_
```

Preview combined data frame.

```
head(bike_trips)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual
```

Check internal structure of the combined data frame.

```
str(bike_trips)
```

```
## 'data.frame':    5723532 obs. of  13 variables:
##  $ ride_id           : chr  "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887262AD101C6
##  $ rideable_type     : chr  "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-04-12 18:25:36" "2021-04-27 17:27:11" "2021-04-03 12:42:45" "2021-
##  $ ended_at          : chr  "2021-04-12 18:56:55" "2021-04-27 18:31:29" "2021-04-07 11:40:24" "2021-
##  $ start_station_name: chr  "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blvd & 84th S
##  $ start_station_id  : chr  "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr  "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loomis Blvd &
##  $ end_station_id    : chr  "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "member" ...
```

Process

Although the datasets have been combined into one data frame, it needs to be cleaned for analysis. The columns "started_at" and "ended_at" are not in the right format and therefore need to be converted from character to date-time.

```
bike_trips$started_at <- as.POSIXct(bike_trips$started_at, format = "%Y-%m-%d %H:%M:%S", tz = "America/
bike_trips$ended_at <- as.POSIXct(bike_trips$ended_at, format = "%Y-%m-%d %H:%M:%S",  tz = "America/Chi
```

Preview the combined data frame after the character to date-time conversion.

```
head(bike_trips)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual
```

Split "started_at" column into different columns for ease of analysis.

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

Preview data frame again to ensure that the new date and time columns have been added.

```
head(bike_trips)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
```

```
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual year
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member 2021
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual 2021
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member 2021
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual 2021
##   month day day_of_week hour minute       date
## 1    04  12      Monday   18     25 2021-04-12
## 2    04  27     Tuesday   17     27 2021-04-27
## 3    04  03    Saturday   12     42 2021-04-03
## 4    04  17    Saturday    9     17 2021-04-17
## 5    04  03    Saturday   12     42 2021-04-03
## 6    04  25      Sunday   18     43 2021-04-25
```

Add a new column to for the bike ride duration by calculating the difference between the "started_at" and "ended_at".

```
bike_trips$duration_hour <- as.numeric(difftime(bike_trips$ended_at, bike_trips$started_at, units = "hou
bike_trips$duration_minute <- as.numeric(difftime(bike_trips$ended_at, bike_trips$started_at, units = "m
```

Preview data frame again to ensure that the duration column has been added.

```
head(bike_trips)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual year
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member 2021
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual 2021
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member 2021
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual 2021
##   month day day_of_week hour minute       date duration_hour duration_minute
## 1    04  12      Monday   18     25 2021-04-12    0.52194444      31.3166667
## 2    04  27     Tuesday   17     27 2021-04-27    1.07166667      64.3000000
## 3    04  03    Saturday   12     42 2021-04-03   94.96083333    5697.6500000
## 4    04  17    Saturday    9     17 2021-04-17    0.41833333      25.1000000
```

```
## 5    04  03    Saturday   12     42 2021-04-03    1.52138889      91.2833333
## 6    04  25      Sunday   18     43 2021-04-25    0.01138889       0.6833333
```

Check the dimension of the data frame.

```
dim(bike_trips)
```

```
## [1] 5723532      22
```

Check the summary of the data frame.

```
summary(bike_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5723532     Length:5723532     Min.   :2021-04-01 00:03:18.00
##  Class :character   Class :character   1st Qu.:2021-06-22 15:20:26.50
##  Mode  :character   Mode  :character   Median :2021-08-17 18:25:49.00
##                                        Mean   :2021-08-26 22:34:00.07
##                                        3rd Qu.:2021-10-14 19:48:10.50
##                                        Max.   :2022-03-31 23:59:47.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2021-04-01 00:14:29.00   Length:5723532     Length:5723532
##  1st Qu.:2021-06-22 15:47:37.00   Class :character   Class :character
##  Median :2021-08-17 18:44:32.50   Mode  :character   Mode  :character
##  Mean   :2021-08-26 22:55:32.66
##  3rd Qu.:2021-10-14 20:03:28.50
##  Max.   :2022-04-01 22:10:12.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5723532     Length:5723532     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :45.64   Max.   :-73.80
##
##     end_lat         end_lng       member_casual          year
##  Min.   :41.39   Min.   :-88.97   Length:5723532     Length:5723532
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character   Class :character
##  Median :41.90   Median :-87.64   Mode  :character   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.49
##  NA's   :4716    NA's   :4716
##     month               day            day_of_week             hour
##  Length:5723532     Length:5723532     Length:5723532     Min.   : 0.00
##  Class :character   Class :character   Class :character   1st Qu.:11.00
##  Mode  :character   Mode  :character   Mode  :character   Median :15.00
##                                                           Mean   :14.23
##                                                           3rd Qu.:18.00
##                                                           Max.   :23.00
```

```
##
##      minute           date            duration_hour      duration_minute
##  Min.   : 0.00   Min.   :2021-04-01   Min.   : -1.9672   Min.   : -118.03
##  1st Qu.:14.00   1st Qu.:2021-06-22   1st Qu.:  0.1094   1st Qu.:    6.57
##  Median :29.00   Median :2021-08-17   Median :  0.1953   Median :   11.72
##  Mean   :29.48   Mean   :2021-08-26   Mean   :  0.3591   Mean   :   21.54
##  3rd Qu.:44.00   3rd Qu.:2021-10-15   3rd Qu.:  0.3556   3rd Qu.:   21.33
##  Max.   :59.00   Max.   :2022-04-01   Max.   :932.4025   Max.   :55944.15
##
```

Check if there are any negative duration values in the data frame.

```
nrow(bike_trips[bike_trips$duration_hour < 0,]) # negative duration counts
```

```
## [1] 145
```

Preview rows with negative duration.

```
head(bike_trips[bike_trips$duration_hour < 0,])
```

```
##                 ride_id rideable_type          started_at            ended_at
## 22362  BC53ECCBC76278FD  classic_bike 2021-04-07 16:11:33 2021-04-07 16:11:26
## 31845  209C097828F9CD43 electric_bike 2021-04-27 17:13:44 2021-04-27 17:11:32
## 292567 6E81034B446FC2FD electric_bike 2021-04-23 09:43:39 2021-04-23 09:43:29
## 292679 318DD838369AEA61  classic_bike 2021-04-30 10:56:32 2021-04-30 10:56:30
## 293034 8ADD13BD8F6A7567  classic_bike 2021-04-17 12:43:36 2021-04-17 12:43:27
## 650362 3EC1B5A4D4B9AB99  classic_bike 2021-05-05 16:10:04 2021-05-05 16:09:51
##             start_station_name start_station_id        end_station_name
## 22362  Ashland Ave & Grand Ave            13434 Ashland Ave & Grand Ave
## 31845
## 292567   Dayton St & North Ave            13058   Dayton St & North Ave
## 292679   Dayton St & North Ave            13058   Dayton St & North Ave
## 293034   Dayton St & North Ave            13058   Dayton St & North Ave
## 650362   Dayton St & North Ave            13058   Dayton St & North Ave
##        end_station_id start_lat start_lng  end_lat   end_lng member_casual year
## 22362           13434  41.89107 -87.66661 41.89107 -87.66661        member 2021
## 31845                  41.91000 -87.64000 41.91000 -87.64000        member 2021
## 292567          13058  41.91064 -87.64937 41.91065 -87.64939        member 2021
## 292679          13058  41.91058 -87.64942 41.91058 -87.64942        member 2021
## 293034          13058  41.91058 -87.64942 41.91058 -87.64942        member 2021
## 650362          13058  41.91058 -87.64942 41.91058 -87.64942        member 2021
##        month day day_of_week hour minute       date duration_hour
## 22362     04  07   Wednesday   16     11 2021-04-07 -0.0019444444
## 31845     04  27     Tuesday   17     13 2021-04-27 -0.0366666667
## 292567    04  23      Friday    9     43 2021-04-23 -0.0027777778
## 292679    04  30      Friday   10     56 2021-04-30 -0.0005555556
## 293034    04  17    Saturday   12     43 2021-04-17 -0.0025000000
## 650362    05  05   Wednesday   16     10 2021-05-05 -0.0036111111
##        duration_minute
## 22362      -0.11666667
## 31845      -2.20000000
## 292567     -0.16666667
```

```
## 292679     -0.03333333
## 293034     -0.15000000
## 650362     -0.21666667
```

Remove rows with negative duration values.

```
bike_trips <- bike_trips[bike_trips$duration_hour >= 0,]
```

Check dimension of the data frame again after removing rows with negative duration values.

```
dim(bike_trips)
```

```
## [1] 5723387      22
```

Check the summary of data frame again after removing negative duration values.

```
summary(bike_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5723387     Length:5723387     Min.   :2021-04-01 00:03:18.00
##  Class :character   Class :character   1st Qu.:2021-06-22 15:19:19.50
##  Mode  :character   Mode  :character   Median :2021-08-17 18:25:00.00
##                                        Mean   :2021-08-26 22:33:10.97
##                                        3rd Qu.:2021-10-14 19:47:23.00
##                                        Max.   :2022-03-31 23:59:47.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2021-04-01 00:14:29.00   Length:5723387     Length:5723387
##  1st Qu.:2021-06-22 15:46:31.50   Class :character   Class :character
##  Median :2021-08-17 18:43:54.00   Mode  :character   Mode  :character
##  Mean   :2021-08-26 22:54:43.65
##  3rd Qu.:2021-10-14 20:02:47.00
##  Max.   :2022-04-01 22:10:12.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5723387     Length:5723387     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :45.64   Max.   :-73.80
##
##     end_lat         end_lng       member_casual          year
##  Min.   :41.39   Min.   :-88.97   Length:5723387     Length:5723387
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character   Class :character
##  Median :41.90   Median :-87.64   Mode  :character   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.49
##  NA's   :4716    NA's   :4716
##     month              day            day_of_week             hour
##  Length:5723387     Length:5723387     Length:5723387     Min.   : 0.00
```

```
##  Class :character   Class :character   Class :character   1st Qu.:11.00
##  Mode  :character   Mode  :character   Mode  :character   Median :15.00
##                                                          Mean   :14.23
##                                                          3rd Qu.:18.00
##                                                          Max.   :23.00
##
##      minute           date            duration_hour      duration_minute
##  Min.   : 0.00   Min.   :2021-04-01   Min.   :  0.0000   Min.   :    0.00
##  1st Qu.:14.00   1st Qu.:2021-06-22   1st Qu.:  0.1094   1st Qu.:    6.57
##  Median :29.00   Median :2021-08-17   Median :  0.1953   Median :   11.72
##  Mean   :29.48   Mean   :2021-08-26   Mean   :  0.3591   Mean   :   21.54
##  3rd Qu.:44.00   3rd Qu.:2021-10-15   3rd Qu.:  0.3556   3rd Qu.:   21.33
##  Max.   :59.00   Max.   :2022-04-01   Max.   :932.4025   Max.   :55944.15
##
```

Determine the number of NA values in the data frame.

```
sum(is.na(bike_trips))
```

```
## [1] 9432
```

Calculate the percentage of NA values in the data frame.

```
sum(is.na(bike_trips))/(nrow(bike_trips)*ncol(bike_trips))*100
```

```
## [1] 0.007490796
```

Since only a small percentage of the data frame contains NA values, rows with NA values are removed from the data frame.

```
bike_trips <- bike_trips[complete.cases(bike_trips),]
```

Check to ensure all NA values have been removed.

```
sum(is.na(bike_trips))
```

```
## [1] 0
```

Check the dimension of data frame after removing rows with NA values.

```
dim(bike_trips)
```

```
## [1] 5718671      22
```

Check summary of data frame after removing rows with NA values.

```
summary(bike_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5718671     Length:5718671     Min.   :2021-04-01 00:03:18.00
##  Class :character   Class :character   1st Qu.:2021-06-22 15:24:01.50
##  Mode  :character   Mode  :character   Median :2021-08-17 18:27:51.00
##                                        Mean   :2021-08-26 22:39:39.67
##                                        3rd Qu.:2021-10-14 19:57:02.00
##                                        Max.   :2022-03-31 23:59:47.00
##     ended_at                      start_station_name start_station_id
##  Min.   :2021-04-01 00:14:29.00   Length:5718671     Length:5718671
##  1st Qu.:2021-06-22 15:49:58.50   Class :character   Class :character
##  Median :2021-08-17 18:46:29.00   Mode  :character   Mode  :character
##  Mean   :2021-08-26 23:00:09.95
##  3rd Qu.:2021-10-14 20:11:49.00
##  Max.   :2022-04-01 22:10:12.00
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5718671     Length:5718671     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :45.64   Max.   :-73.80
##     end_lat         end_lng       member_casual          year
##  Min.   :41.39   Min.   :-88.97   Length:5718671     Length:5718671
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character   Class :character
##  Median :41.90   Median :-87.64   Mode  :character   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.49
##     month               day            day_of_week            hour
##  Length:5718671     Length:5718671     Length:5718671     Min.   : 0.00
##  Class :character   Class :character   Class :character   1st Qu.:11.00
##  Mode  :character   Mode  :character   Mode  :character   Median :15.00
##                                                           Mean   :14.23
##                                                           3rd Qu.:18.00
##                                                           Max.   :23.00
##      minute           date            duration_hour      duration_minute
##  Min.   : 0.00   Min.   :2021-04-01   Min.   :  0.0000   Min.   :    0.00
##  1st Qu.:14.00   1st Qu.:2021-06-22   1st Qu.:  0.1094   1st Qu.:    6.57
##  Median :29.00   Median :2021-08-17   Median :  0.1953   Median :   11.72
##  Mean   :29.48   Mean   :2021-08-26   Mean   :  0.3417   Mean   :   20.50
##  3rd Qu.:44.00   3rd Qu.:2021-10-15   3rd Qu.:  0.3550   3rd Qu.:   21.30
##  Max.   :59.00   Max.   :2022-04-01   Max.   :932.4025   Max.   :55944.15
```

Label month with proper names and put month and day of week in the right order.

```
bike_trips$month_name <- factor(bike_trips$month, levels = c("04", "05", "06", "07", "08", "09", "10",

bike_trips$day_of_week <- factor(bike_trips$day_of_week, levels = c("Monday", "Tuesday", "Wednesday", "
```

Preview data frame after data cleaning.

```
head(bike_trips)
```

```
##            ride_id rideable_type          started_at            ended_at
```

```
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual year
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member 2021
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual 2021
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member 2021
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual 2021
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual 2021
##   month day day_of_week hour minute       date duration_hour duration_minute
## 1    04  12      Monday   18     25 2021-04-12    0.52194444      31.3166667
## 2    04  27     Tuesday   17     27 2021-04-27    1.07166667      64.3000000
## 3    04  03    Saturday   12     42 2021-04-03   94.96083333    5697.6500000
## 4    04  17    Saturday    9     17 2021-04-17    0.41833333      25.1000000
## 5    04  03    Saturday   12     42 2021-04-03    1.52138889      91.2833333
## 6    04  25      Sunday   18     43 2021-04-25    0.01138889       0.6833333
##   month_name
## 1 April 2021
## 2 April 2021
## 3 April 2021
## 4 April 2021
## 5 April 2021
## 6 April 2021
```

Analyze

Since the data cleaning process has been completed, the cleaned data frame will be analyzed to determine the differences between annual members and casual riders.

Here are the basic statistics of the ride duration.

```
cat("The average amount of time spent in a bike trip is", mean(bike_trips$duration_minute), " minutes.\n
```

```
## The average amount of time spent in a bike trip is 20.50459  minutes.
```

```
cat("The median amount of time spent in a bike trip is ", median(bike_trips$duration_minute), " minutes
```

```
## The median amount of time spent in a bike trip is  11.71667  minutes.
```

```
cat("The minimum amount of time spent in a bike trip is ", min(bike_trips$duration_minute), " minutes.\n
```

```
## The minimum amount of time spent in a bike trip is  0  minutes.
```

```
cat("The maximum amount of time spent in a bike trip is ", max(bike_trips$duration_minute), " minutes.\n
```

```
## The maximum amount of time spent in a bike trip is  55944.15  minutes.
```

```
cat("The standard deviation of the amount of time spent in a bike trip is ", sd(bike_trips$duration_minu
```

```
## The standard deviation of the amount of time spent in a bike trip is  163.8598  minutes.
```

The results show that the bike rider spends around 20.5 minutes in average per trip. However, there is a large difference between the maximum and minimum value. In addition, the amount of variation is substantial as shown by the standard deviation.

Here is the comparison between the average amount of time spent by casual riders and annual members.

```
setNames(aggregate(bike_trips$duration_minute ~ bike_trips$member_casual, FUN = mean), c("Casual Rider o
```

```
##   Casual Rider or Annual Member Trip Duration (Minute)
## 1                        casual               29.73109
## 2                        member               13.11712
```

The casual riders spends more than twice amount of time than the annual members.

Here is the average amount of time spent by casual riders and annual members grouped by month.

```
aggregate(bike_trips$duration_minute ~ bike_trips$member_casual + bike_trips$month_name, FUN = mean)
```

```
##    bike_trips$member_casual bike_trips$month_name bike_trips$duration_minute
## 1                    casual            April 2021                   36.63454
## 2                    member            April 2021                   14.43342
## 3                    casual              May 2021                   36.90602
## 4                    member              May 2021                   14.41202
## 5                    casual             June 2021                   35.73165
## 6                    member             June 2021                   14.34121
## 7                    casual             July 2021                   31.34836
## 8                    member             July 2021                   13.99954
## 9                    casual           August 2021                   27.57607
## 10                   member           August 2021                   13.85466
## 11                   casual        September 2021                   26.66806
## 12                   member        September 2021                   13.48662
## 13                   casual          October 2021                   24.13636
## 14                   member          October 2021                   12.26170
## 15                   casual         November 2021                   20.20385
## 16                   member         November 2021                   11.10347
## 17                   casual         December 2021                   21.11610
## 18                   member         December 2021                   10.81803
## 19                   casual          January 2022                   24.01745
## 20                   member          January 2022                   11.64887
## 21                   casual         February 2022                   22.10774
## 22                   member         February 2022                   11.04789
## 23                   casual            March 2022                   25.74071
## 24                   member            March 2022                   11.74894
```

Note that the bike ride duration for annual members are more consistent across each month; whereas, more variability is observed with casual riders. Both the casual riders and annual members spent more time on bike rides from April to July 2021.

Here is the average amount of time spent by casual riders and annual members grouped by day of week.

```
aggregate(bike_trips$duration_minute ~ bike_trips$member_casual + bike_trips$day_of_week, FUN = mean)
```

```
##    bike_trips$member_casual bike_trips$day_of_week bike_trips$duration_minute
## 1                    casual                 Monday                    30.73983
## 2                    member                 Monday                    12.87064
## 3                    casual                Tuesday                    26.43995
## 4                    member                Tuesday                    12.33598
## 5                    casual              Wednesday                    26.17930
## 6                    member              Wednesday                    12.38300
## 7                    casual               Thursday                    25.54278
## 8                    member               Thursday                    12.34915
## 9                    casual                 Friday                    27.73022
## 10                   member                 Friday                    12.72428
## 11                   casual               Saturday                    31.44049
## 12                   member               Saturday                    14.49131
## 13                   casual                 Sunday                    34.66303
## 14                   member                 Sunday                    15.03502
```

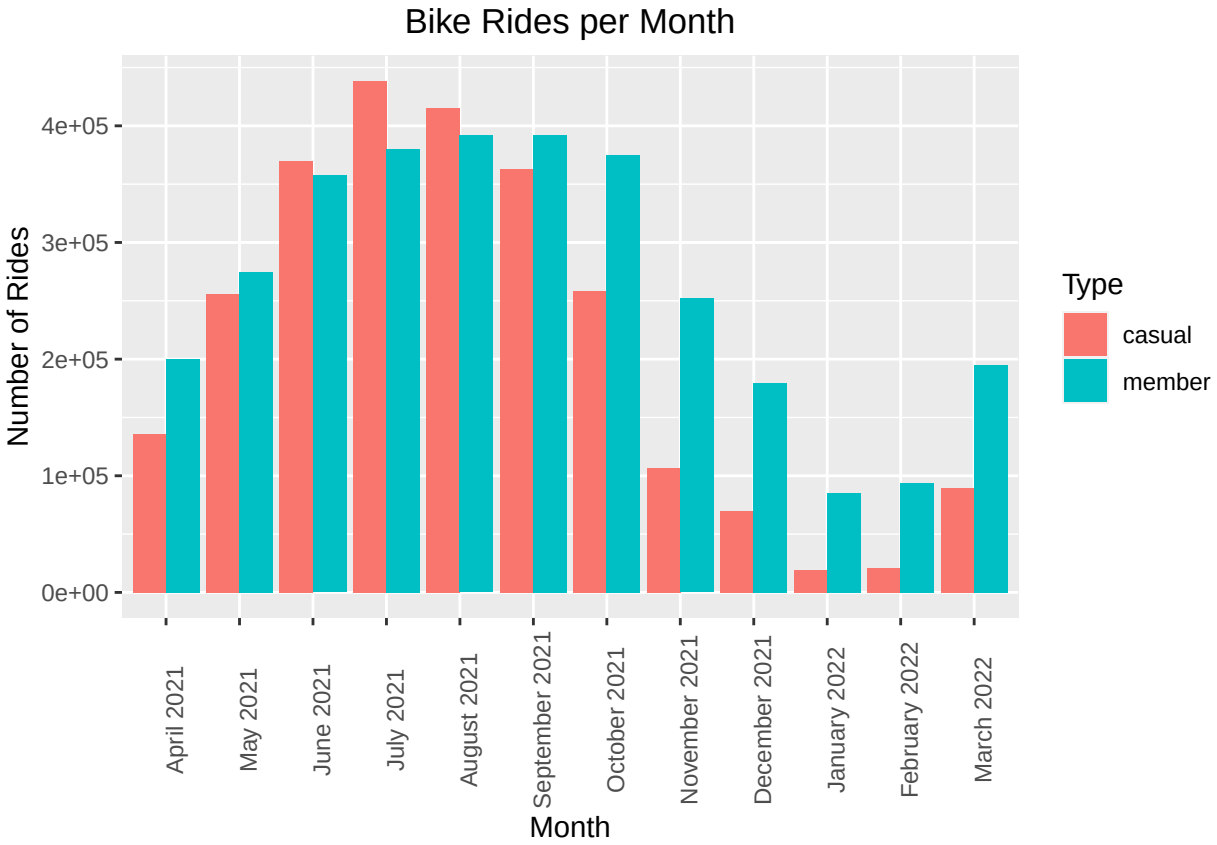
The bike ride duration for annual members are more consistent, especially in the weekdays. Casual riders spend more time using bike ride service on Saturday, Sunday and Monday.

Share

Visualizations are developed to further identify patterns and trends between casual bike riders and annual members.

Here is the number of bike rides by month for casual riders and annual members.

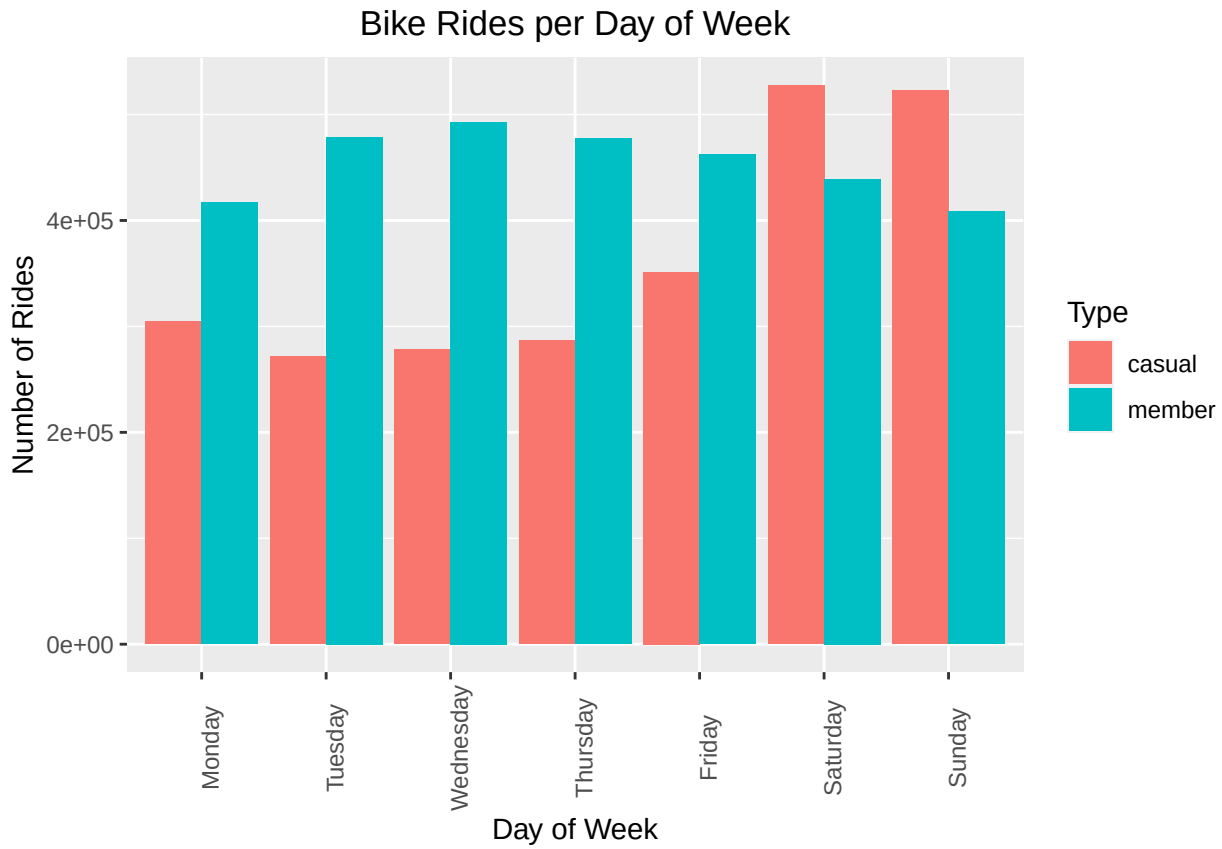
```
ggplot(bike_trips, aes(x = month_name, group = member_casual, fill = member_casual)) + geom_bar(positio
```



This chart shows that there are considerably more bike rides in the summer months than the winter month. Although there are less number of rides in the winter months for both members and non-members (casual riders), the number of bike rides for non-members diminishes much more significantly, indicating that annual members are more likely to utilize Cyclistic's bike riding service than non-members in the winter months.

Here is the number of bike rides by day of week for casual riders and annual members.

```
ggplot(bike_trips, aes(x = day_of_week, group = member_casual, fill = member_casual)) + geom_bar(positi
```

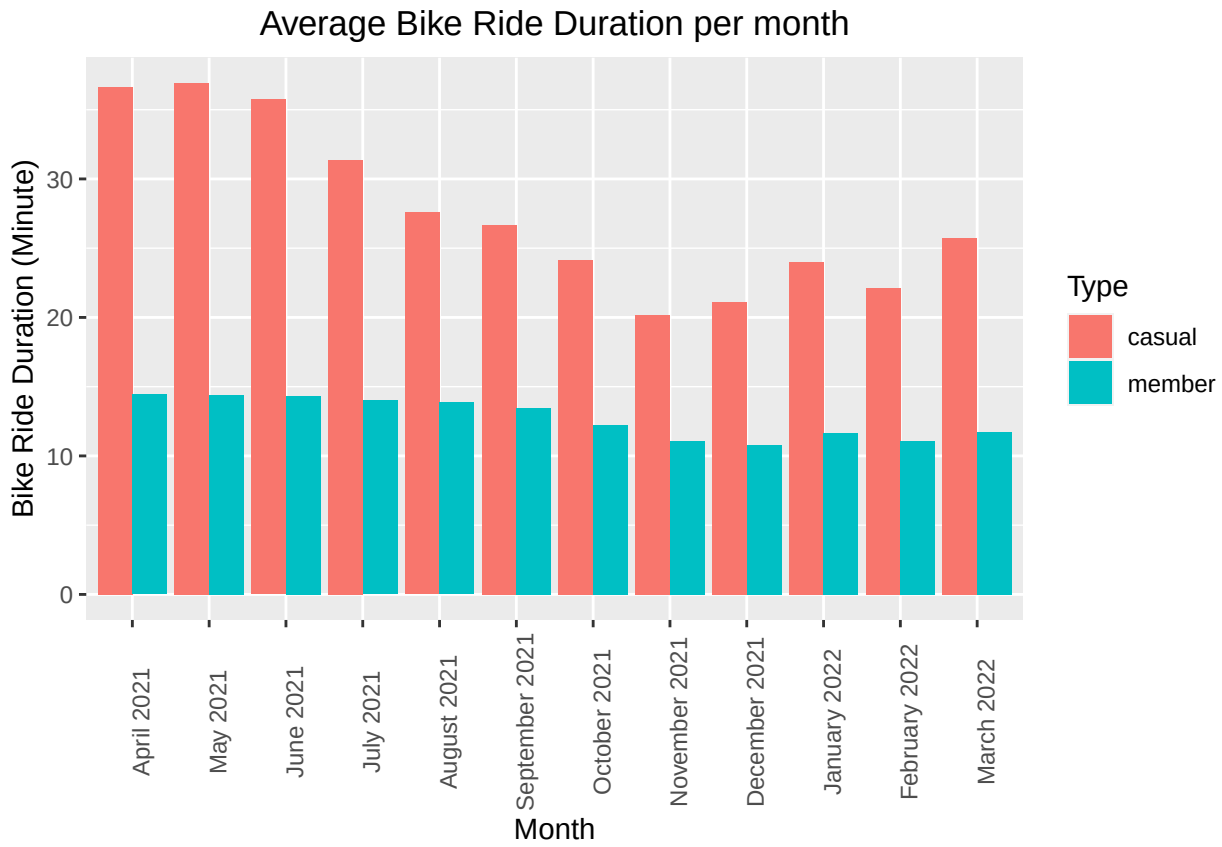


This chart shows that annual members uses the bike rides services more consistently than the non-members on a day-to-day basis. The number of bike rides for non-members is substantially higher on the weekends, whereas the number of bike rides for annual members are similar across each day of the week. This results could imply that the annual members mainly uses Cyclistic's bike riding service for daily commute and the non-members mainly uses it for leisure activities.

Here is the duration of bike rides by month for casual riders and annual members.

```
ggplot(bike_trips, aes(x = month_name, y = duration_minute, fill = member_casual)) + stat_summary(geom =
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

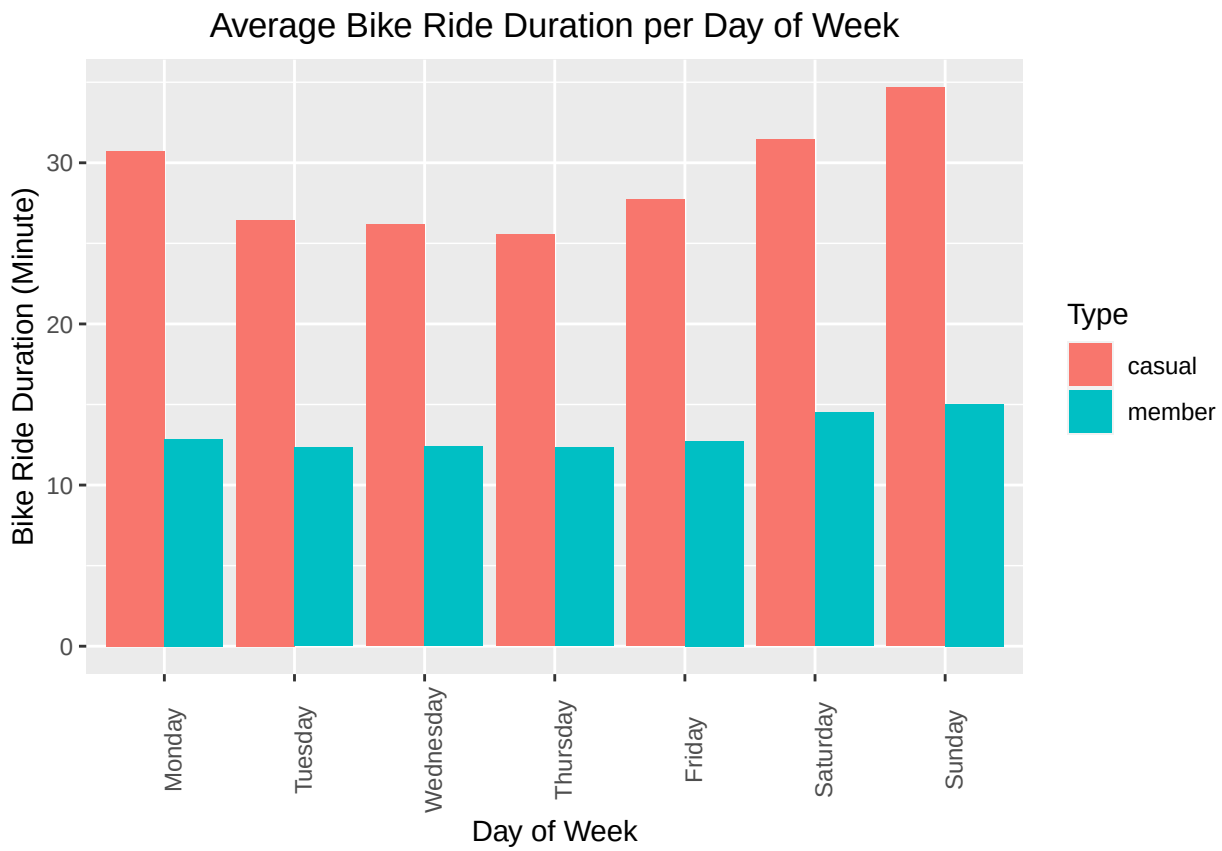


This chart shows that the average time spent on bike riding for annual members is fairly consistent over each month; however, non-members spent much more time using Cyclistic's bike riding service in the summer months compare to the winter months.

Here is the duration of bike rides by day of week for casual riders and annual members.

```
ggplot(bike_trips, aes(x = day_of_week, y = duration_minute, fill = member_casual)) + stat_summary(geom
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



This chart shows that the average bike ride duration for annual members is fairly consistent over each day of week with slight increase on the weekends; whereas, the average bike ride duration for non-members are relatively higher on Saturday, Sunday and Monday.

Here are the maps that mark the start and end locations of the annual members and non-members using a sampled dataset.

```
library(mapview)
library(leafsync)
bike_trips_sample <- {set.seed(1); bike_trips[sample(nrow(bike_trips), 100), ]} # randomly sample 100 d
bike_trips_sample_casual <- subset(bike_trips_sample, member_casual == "casual")
bike_trips_sample_member <- subset(bike_trips_sample, member_casual == "member")
map_start <- mapview(bike_trips_sample_casual, xcol = "start_lng", ycol = "start_lat", crs = 4326, colo
map_end <- mapview(bike_trips_sample_casual, xcol = "end_lng", ycol = "end_lat", crs = 4326, color = "c

sync(map_start, map_end)
```

These maps indicate that most of the rides started and ended in the city center area; however, the start and end locations for annual members are more scattered than the start and end locations the non-members.

Act

Summary

1. Annual members use the bike ride service more consistently than non-members throughout each week. Non-members use the bike ride service more frequently on the weekends.

2. Although both members' and non-members' bike usage follows a seasonal trend, non-members use the bike ride service more frequently than annual members in the summer months and less frequently in the winter months.
3. On average, non-members spend around twice amount time than the annual members per bike ride.
4. The majority of the bike rides started and ended in the city center area; however, annual members' start and end locations are more scattered than the start and end locations of the non-members.

Recommendations

1. Provide discount or coupons to members who use the service over a set amount of hours per trip to attract more membership purchases.
2. Run more advertisements of Cyclistic's annual membership during the weekends of summer months to gain more recognition from casual riders.
3. Focus promoting Cyclistic's annual membership in the city center area.