Cyclistic Bike Share Case Study

Aarti Singh

2023-02-21

Introduction

This analysis is one of the capstone projects for Google Data Analytics Certification. This project consists of a scenario about a bike-share company in Chicago named Cyclistic. Cyclistic have 5,824 bicycles are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system at anytime. 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.

The director of marketing Lily Moreno, who is responsible for the development of campaigns and initiatives to promote the bike-share program, believes that the company's future success depends on maximizing the number of annual memberships. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into Cylistic members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their transportation needs. Therefore, she has set a clear goal of designing marketing strategies aimed at converting casual riders into annual members. A team of data analysts has been assigned a task that will guide the future marketing program and achieve Moreno's goal. The approach of this case study will have six phases of data analysis which are ask, prepare, process, analyze, share, and act.

Ask

Business task:

- How do annual members and casual riders use Cyclistic bikes differently?

Key Stakeholders:

- Lily Moreno: the director of marketing and my manager
- Cyclistic marketing analytics team
- Cyclistic executive team

Prepare

The public dataset used for this case study is obtained from Divvy (https://divvy-tripdata.s3.amazonaws.com/index.html). The dataset has been made available by Motivate International Inc. under Divvy license (https://ride.divvybikes.com/data-license-agreement). The credibility of the data is neutral but reliable as each bicycle is geotracked and locked into the network of multiple stations across the city. The data in the dataset is original but not 100% comprehensive as some information is missing, which will not be included in this analysis. Moreover, the 12 csv files downloaded are cited and current, as it identifies the trends, is from February 2022 to January 2023 (12 months).

Process

I will be using R programming for data cleaning and analysis for this case study. First, I will be installing all the required packages and import all 12 csv files.

```
#Install and load the necessary packages
options(repos = c(CRAN = "http://cran.rstudio.com"))
install.packages("tidyverse")
```

```
##
##   There is a binary version available but the source version is later:
##           binary source needs_compilation
## tidyverse  1.3.2  2.0.0             FALSE
```

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

```
##
## The downloaded binary packages are in
##  /var/folders/pr/md14sq0926j_0yc3hyjhbfrh0000gn/T//Rtmp1fSgi0/downloaded_packages
```

```
install.packages("knitr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/pr/md14sq0926j_0yc3hyjhbfrh0000gn/T//Rtmp1fSgi0/downloaded_packages
```

```
install.packages("skimr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/pr/md14sq0926j_0yc3hyjhbfrh0000gn/T//Rtmp1fSgi0/downloaded_packages
```

```
install.packages("janitor")
```

```
##
## The downloaded binary packages are in
##  /var/folders/pr/md14sq0926j_0yc3hyjhbfrh0000gn/T//Rtmp1fSgi0/downloaded_packages
```

```
library(tidyverse)
library(lubridate)
library(knitr)
library(skimr)
library(janitor)

#Import data
setwd("~/Desktop/cyclistic")
month_01 <- read_csv("~/Desktop/cyclistic/202202-divvy-tripdata.csv")
month_02 <- read_csv("~/Desktop/cyclistic/202203-divvy-tripdata.csv")
month_03 <- read_csv("~/Desktop/cyclistic/202204-divvy-tripdata.csv")
month_04 <- read_csv("~/Desktop/cyclistic/202205-divvy-tripdata.csv")
```

```r
month_05 <- read_csv("~/Desktop/cyclistic/202206-divvy-tripdata.csv")
month_06 <- read_csv("~/Desktop/cyclistic/202207-divvy-tripdata.csv")
month_07 <- read_csv("~/Desktop/cyclistic/202208-divvy-tripdata.csv")
month_08 <- read_csv("~/Desktop/cyclistic/202209-divvy-tripdata.csv")
month_09 <- read_csv("~/Desktop/cyclistic/202210-divvy-tripdata.csv")
month_10 <- read_csv("~/Desktop/cyclistic/202211-divvy-tripdata.csv")
month_11 <- read_csv("~/Desktop/cyclistic/202212-divvy-tripdata.csv")
month_12 <- read_csv("~/Desktop/cyclistic/202301-divvy-tripdata.csv")
```

Now I will examine the structure of each data set before combining into one large data set.

```r
for(i in list(month_01, month_02, month_03, month_04, month_05, month_06, month_07,
              month_08,month_09,month_10,month_11,month_12))
print(str(i))
```

```
## spc_tbl_ [115,609 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:115609] "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A178
##  $ rideable_type     : chr [1:115609] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:115609], format: "2022-02-19 18:08:41" "2022-02-20 17:41:30" ...
##  $ ended_at          : POSIXct[1:115609], format: "2022-02-19 18:23:56" "2022-02-20 17:45:56" ...
##  $ start_station_name: chr [1:115609] "State St & Randolph St" "Halsted St & Wrightwood Ave" "State S
##  $ start_station_id  : chr [1:115609] "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...
##  $ end_station_name  : chr [1:115609] "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Cana
##  $ end_station_id    : chr [1:115609] "13179" "TA1307000113" "13011" "13323" ...
##  $ start_lat         : num [1:115609] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:115609] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:115609] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:115609] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:115609] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [284,042 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:284042] "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F44
##  $ rideable_type     : chr [1:284042] "classic_bike" "electric_bike" "classic_bike" "classic_bike" .
##  $ started_at        : POSIXct[1:284042], format: "2022-03-21 13:45:01" "2022-03-16 09:37:16" ...
##  $ ended_at          : POSIXct[1:284042], format: "2022-03-21 13:51:18" "2022-03-16 09:43:34" ...
##  $ start_station_name: chr [1:284042] "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & Be
```

```
##  $ start_station_id  : chr [1:284042] "TA1307000131" "13042" "13109" "TA1307000131" ...
##  $ end_station_name  : chr [1:284042] "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apts
##  $ end_station_id    : chr [1:284042] "KA1503000043" "620" "15578" "TA1305000025" ...
##  $ start_lat         : num [1:284042] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:284042] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:284042] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [371,249 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:371249] "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F07FF" "84D4
##  $ rideable_type     : chr [1:371249] "electric_bike" "classic_bike" "classic_bike" "classic_bike" .
##  $ started_at        : POSIXct[1:371249], format: "2022-04-06 17:42:48" "2022-04-24 19:23:07" ...
##  $ ended_at          : POSIXct[1:371249], format: "2022-04-06 17:54:36" "2022-04-24 19:43:17" ...
##  $ start_station_name: chr [1:371249] "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted S
##  $ start_station_id  : chr [1:371249] "515" "13075" "TA1307000121" "13075" ...
##  $ end_station_name  : chr [1:371249] "University Library (NU)" "Green St & Madison St" "Green St & M
##  $ end_station_id    : chr [1:371249] "605" "TA1307000120" "TA1307000120" "KA1706005007" ...
##  $ start_lat         : num [1:371249] 42 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:371249] 42.1 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:371249] "member" "member" "member" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
```

```
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [634,858 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF5
##  $ rideable_type     : chr [1:634858] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
##  $ ended_at          : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
##  $ start_station_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St" "
##  $ start_station_id  : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
##  $ end_station_name  : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood S
##  $ end_station_id    : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:634858] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [769,204 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:769204] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD771D2" "C9C3
##  $ rideable_type     : chr [1:769204] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:769204], format: "2022-06-30 17:27:53" "2022-06-30 18:39:52" ...
##  $ ended_at          : POSIXct[1:769204], format: "2022-06-30 17:35:15" "2022-06-30 18:47:28" ...
##  $ start_station_name: chr [1:769204] NA NA NA NA ...
##  $ start_station_id  : chr [1:769204] NA NA NA NA ...
##  $ end_station_name  : chr [1:769204] NA NA NA NA ...
##  $ end_station_id    : chr [1:769204] NA NA NA NA ...
##  $ start_lat         : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
##  $ start_lng         : num [1:769204] -87.6 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num [1:769204] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:769204] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
```

```
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [823,488 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:823488] "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6
##  $ rideable_type     : chr [1:823488] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:823488], format: "2022-07-05 08:12:47" "2022-07-26 12:53:38" ...
##  $ ended_at          : POSIXct[1:823488], format: "2022-07-05 08:24:32" "2022-07-26 12:55:31" ...
##  $ start_station_name: chr [1:823488] "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buc
##  $ start_station_id  : chr [1:823488] "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr [1:823488] "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan A
##  $ end_station_id    : chr [1:823488] "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num [1:823488] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:823488] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:823488] 41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num [1:823488] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:823488] "member" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [785,932 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:785932] "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB7FD" "F597
##  $ rideable_type     : chr [1:785932] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:785932], format: "2022-08-07 21:34:15" "2022-08-08 14:39:21" ...
##  $ ended_at          : POSIXct[1:785932], format: "2022-08-07 21:41:46" "2022-08-08 14:53:23" ...
##  $ start_station_name: chr [1:785932] NA NA NA NA ...
##  $ start_station_id  : chr [1:785932] NA NA NA NA ...
##  $ end_station_name  : chr [1:785932] NA NA NA NA ...
```

```
##  $ end_station_id    : chr [1:785932] NA NA NA NA ...
##  $ start_lat         : num [1:785932] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:785932] 41.9 41.9 42 42 41.8 ...
##  $ end_lng           : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:785932] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [701,339 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:701339] "5156990AC19CA285" "E12D4A16BF51C274" "A02B53CD7DB72DD7" "C82E
##  $ rideable_type     : chr [1:701339] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:701339], format: "2022-09-01 08:36:22" "2022-09-01 17:11:29" ...
##  $ ended_at          : POSIXct[1:701339], format: "2022-09-01 08:39:05" "2022-09-01 17:14:45" ...
##  $ start_station_name: chr [1:701339] NA NA NA NA ...
##  $ start_station_id  : chr [1:701339] NA NA NA NA ...
##  $ end_station_name  : chr [1:701339] "California Ave & Milwaukee Ave" NA NA NA ...
##  $ end_station_id    : chr [1:701339] "13084" NA NA NA ...
##  $ start_lat         : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:701339] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
```

```
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [558,685 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:558685] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E65
##  $ rideable_type     : chr [1:558685] "classic_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:558685], format: "2022-10-14 17:13:30" "2022-10-01 16:29:26" ...
##  $ ended_at          : POSIXct[1:558685], format: "2022-10-14 17:19:39" "2022-10-01 16:49:06" ...
##  $ start_station_name: chr [1:558685] "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne A
##  $ start_station_id  : chr [1:558685] "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr [1:558685] "Larrabee St & Division St" "Damen Ave & Cullerton St" "Western
##  $ end_station_id    : chr [1:558685] "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat         : num [1:558685] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:558685] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:558685] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:558685] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:558685] "member" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [337,735 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:337735] "BCC66FC6FAB27CC7" "772AB67E902C180F" "585EAD07FDEC0152" "91C4
##  $ rideable_type     : chr [1:337735] "electric_bike" "classic_bike" "classic_bike" "classic_bike" .
##  $ started_at        : POSIXct[1:337735], format: "2022-11-10 06:21:55" "2022-11-04 07:31:55" ...
##  $ ended_at          : POSIXct[1:337735], format: "2022-11-10 06:31:27" "2022-11-04 07:46:25" ...
##  $ start_station_name: chr [1:337735] "Canal St & Adams St" "Canal St & Adams St" "Indiana Ave & Roos
##  $ start_station_id  : chr [1:337735] "13011" "13011" "SL-005" "SL-005" ...
##  $ end_station_name  : chr [1:337735] "St. Clair St & Erie St" "St. Clair St & Erie St" "St. Clair St
##  $ end_station_id    : chr [1:337735] "13016" "13016" "13016" "13016" ...
##  $ start_lat         : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:337735] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
```

```
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [181,806 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:181806] "65DBD2F447EC51C2" "0C201AA7EA0EA1AD" "E0B148CCB358A49D" "54C5
##  $ rideable_type     : chr [1:181806] "electric_bike" "classic_bike" "electric_bike" "classic_bike"
##  $ started_at        : POSIXct[1:181806], format: "2022-12-05 10:47:18" "2022-12-18 06:42:33" ...
##  $ ended_at          : POSIXct[1:181806], format: "2022-12-05 10:56:34" "2022-12-18 07:08:44" ...
##  $ start_station_name: chr [1:181806] "Clifton Ave & Armitage Ave" "Broadway & Belmont Ave" "Sangamo
##  $ start_station_id  : chr [1:181806] "TA1307000163" "13277" "TA1306000015" "KA1503000038" ...
##  $ end_station_name  : chr [1:181806] "Sedgwick St & Webster Ave" "Sedgwick St & Webster Ave" "St. C
##  $ end_station_id    : chr [1:181806] "13191" "13191" "13016" "13134" ...
##  $ start_lat         : num [1:181806] 41.9 41.9 41.9 41.8 41.9 ...
##  $ start_lng         : num [1:181806] -87.7 -87.6 -87.7 -87.6 -87.7 ...
##  $ end_lat           : num [1:181806] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:181806] -87.6 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:181806] "member" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
## spc_tbl_ [190,301 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C907
##  $ rideable_type     : chr [1:190301] "electric_bike" "classic_bike" "electric_bike" "classic_bike"
##  $ started_at        : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" ...
##  $ ended_at          : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" ...
##  $ start_station_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western
##  $ start_station_id  : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
##  $ end_station_name  : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli P
##  $ end_station_id    : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...
##  $ start_lat         : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
```

```
##  $ start_lng         : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
##  $ end_lng           : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:190301] "member" "member" "casual" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
## NULL
```

After examining all 12 datasets, all of them have 13 common attributes and types of data such as chr, num, and POSIXct. Now, we are going to combine all 12 csv files into one large dataset called "final_data".

```
final_data <- bind_rows(month_01, month_02, month_03, month_04, month_05, month_06,
                        month_07, month_08, month_09, month_10, month_11, month_12)
```

Next, I will use the "as.POSIXct()" function to convert the column to a date-time format for both columns "started_at" and "ended_at". To make sure all the combined data's data-time column is recognized by R for smoother analysis process.

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

The data set is now ready to be clean and format in a correct order.

```
summary(final_data) #shows the data structure and how many NA values is in the combined data set.
```

```
##    ride_id          rideable_type        started_at
##  Length:5754248     Length:5754248     Min.   :2022-02-01 00:03:18.00
##  Class :character   Class :character   1st Qu.:2022-06-02 15:18:09.50
##  Mode  :character   Mode  :character   Median :2022-07-27 22:50:40.50
##                                        Mean   :2022-07-29 13:28:03.16
##                                        3rd Qu.:2022-09-22 20:34:47.25
##                                        Max.   :2023-01-31 23:56:09.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-02-01 00:09:37.00   Length:5754248     Length:5754248
##  1st Qu.:2022-06-02 15:37:50.50   Class :character   Class :character
##  Median :2022-07-27 23:09:33.00   Mode  :character   Mode  :character
```

```
##  Mean   :2022-07-29 13:47:21.50
##  3rd Qu.:2022-09-22 20:53:25.25
##  Max.   :2023-02-04 04:27:03.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5754248     Length:5754248     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.   :42.07   Max.   :-87.52
##
##     end_lat          end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5754248
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.37   Max.   :  0.00
##  NA's   :5899    NA's   :5899
```

There are 5899 NA values in the combined dataset as shown on the above summary results. Now, all those NA values will be removed from the combined dataset.

```
final_data_clean <- na.omit(final_data)
summary(final_data_clean) # shows that all NA values in the combines data set has been removed.
```

```
##    ride_id          rideable_type        started_at
##  Length:4437516     Length:4437516     Min.   :2022-02-01 00:03:18.00
##  Class :character   Class :character   1st Qu.:2022-06-02 16:56:53.75
##  Mode  :character   Mode  :character   Median :2022-07-26 14:48:41.50
##                                        Mean   :2022-07-28 22:01:52.88
##                                        3rd Qu.:2022-09-21 15:23:24.25
##                                        Max.   :2023-01-31 23:53:18.00
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-02-01 00:09:37.00   Length:4437516     Length:4437516
##  1st Qu.:2022-06-02 17:13:37.00   Class :character   Class :character
##  Median :2022-07-26 15:08:15.00   Mode  :character   Mode  :character
##  Mean   :2022-07-28 22:18:50.80
##  3rd Qu.:2022-09-21 15:39:36.25
##  Max.   :2023-02-01 00:28:12.00
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:4437516     Length:4437516     Min.   :41.65   Min.   :-87.83
##  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   :-87.64
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.06   Max.   :-87.53
##     end_lat          end_lng       member_casual
##  Min.   : 0.00   Min.   :-87.83   Length:4437516
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.64
```

```
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.06   Max.   :  0.00
```

Temporary removing variables that are not used for this analysis before adding new ones. Moreover, Filtering out the "docked_bike" from column "rideable_type" as those are times where bike were parked at a certain station across the city and not riden by a member or non-member (casual) of Cyclistic

```
#Temporary removing variables
final_data_clean <- final_data_clean %>%
  select(-c(start_station_name,start_station_id,end_station_name,
            end_station_id,start_lat,start_lng,end_lat,end_lng))

#Filtering out the "docked_bike" from column "rideable_type"
final_data_clean <- final_data_clean %>%
  filter(rideable_type != "docked_bike")
```

Extract day,day of week, month, and year from "started_at" column and add those as new variables to the data frame. Also put the days of the week and month in the correct order before starting any descriptive analysis. Since the data analysis period is from February 2022 to January 2033. Month of February will come first and January will be last.

```
#Extract day,day of week, month, and year from "started_at" column and
#add those as new variables to the data frame.
final_data_clean <- final_data_clean %>%
  mutate(date = as.Date(started_at),
         day = day(started_at),
         day_of_week = weekdays(started_at),
         month = month.name[month(started_at)],
         year = year(started_at))

# putting the days of the week in the correct order
final_data_clean$day_of_week <- factor(final_data_clean$day_of_week,
                                       levels = c("Sunday", "Monday", "Tuesday", "Wednesday",
                                                  "Thursday", "Friday", "Saturday"))

# putting the months in the correct order.
final_data_clean$month <- factor(final_data_clean$month,
                                 levels = c("February", "March", "April", "May", "June",
                                            "July", "August", "September", "October", "November",
                                            "December", "January"))
```

Add another variable called "ride_length" to the dataset and filter out any negative ride length and ride length of zero secs as either they could a wrong data entry or glitch in the geo-tracked.

```
# Adding "ride_length" variable to the data set
final_data_clean <- final_data_clean %>%
  mutate(ride_length = difftime(ended_at,started_at,units = "secs"))

#filtering out any negative ride length and ride length of zero secs
final_data_clean <- subset(final_data_clean, ride_length >= 1)
```

Analyze

Now we will be performing some descriptive average or mean, max, minimum, and median of ride length riden by the type of Cyclistic's customers.

```
#Calculating the descriptive analysis such as average or mean of the ride length and
#number of ride while grouping members or non-members(casual) of Cyclistic and rideable_type
final_data_clean_analysis1 <- final_data_clean %>%
  group_by(member_casual,rideable_type) %>%
  summarize(average_ride_length = mean(ride_length), number_of_rides = n()) %>%
  arrange(member_casual,rideable_type)
print(final_data_clean_analysis1)
```

```
## # A tibble: 4 x 4
## # Groups:   member_casual [2]
##   member_casual rideable_type average_ride_length number_of_rides
##   <chr>         <chr>         <drtn>                        <int>
## 1 casual        classic_bike  1462.0031 secs               895693
## 2 casual        electric_bike  995.0446 secs               703776
## 3 member        classic_bike   791.6420 secs              1737130
## 4 member        electric_bike  654.2940 secs               925014
```

```
#Calculating the descriptive analysis such as average or mean of the ride length and number of rides
#for users by day_of_week while grouping members or non-members(casual) of Cyclistic and day of week
final_data_clean_analysis2 <- final_data_clean %>%
  group_by(member_casual, day_of_week) %>%
  summarize (average_ride_length = mean(ride_length), number_of_rides = n()) %>%
  arrange(member_casual, day_of_week)
print(final_data_clean_analysis2)
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual day_of_week average_ride_length number_of_rides
##    <chr>         <fct>       <drtn>                        <int>
##  1 casual        Sunday      1430.5024 secs               269002
##  2 casual        Monday      1267.3021 secs               190966
##  3 casual        Tuesday     1125.1633 secs               182245
##  4 casual        Wednesday   1087.0923 secs               189262
##  5 casual        Thursday    1122.5673 secs               212376
##  6 casual        Friday      1187.9134 secs               227586
##  7 casual        Saturday    1412.7138 secs               328032
##  8 member        Sunday       828.0908 secs               302975
##  9 member        Monday       718.0805 secs               382610
## 10 member        Tuesday      703.4451 secs               424032
## 11 member        Wednesday    708.4556 secs               422192
## 12 member        Thursday     719.0872 secs               422441
## 13 member        Friday       731.1152 secs               366707
## 14 member        Saturday     836.8321 secs               341187
```

```
#Calculating the descriptive analysis such as average or mean of the ride length and number of rides
#for users by month while grouping members or non-members(casual) of Cyclistic and month
final_data_clean_analysis3 <- final_data_clean %>%
```

```
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n(), average_ride_length = mean(ride_length)) %>%
  arrange(member_casual, month)
print(final_data_clean_analysis3)
```

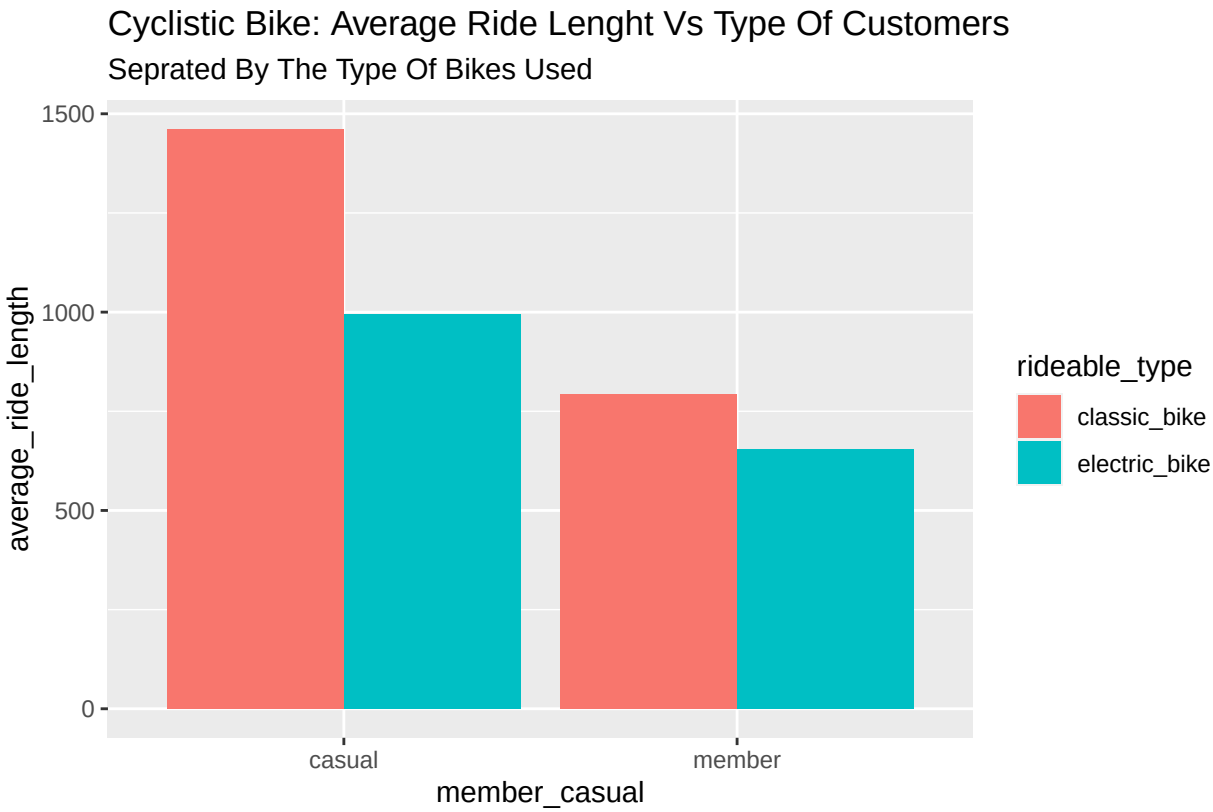
```
## # A tibble: 24 x 4
## # Groups:   member_casual [2]
##    member_casual month     number_of_rides average_ride_length
##    <chr>         <fct>               <int> <drtn>
##  1 casual        February            13799 1145.2507 secs
##  2 casual        March               58929 1416.5704 secs
##  3 casual        April               79909 1354.6937 secs
##  4 casual        May                194112 1466.2668 secs
##  5 casual        June               261842 1335.4696 secs
##  6 casual        July               281054 1315.1628 secs
##  7 casual        August             244186 1235.1688 secs
##  8 casual        September          201417 1172.4876 secs
##  9 casual        October            138916 1093.7494 secs
## 10 casual        November            67739  919.5622 secs
## # ... with 14 more rows
```

```
#Calculating the descriptive analysis such as max,minimum, and median of
#ride length while grouping members or non-members(casual) of Cyclistic.
final_data_clean_analysis4 <- final_data_clean %>%
  group_by(member_casual) %>%
  summarize(max_ride_length = max(ride_length), min_ride_length = min(ride_length),
            median_ride_length = median(ride_length))%>%
  arrange(member_casual)
print(final_data_clean_analysis4)
```

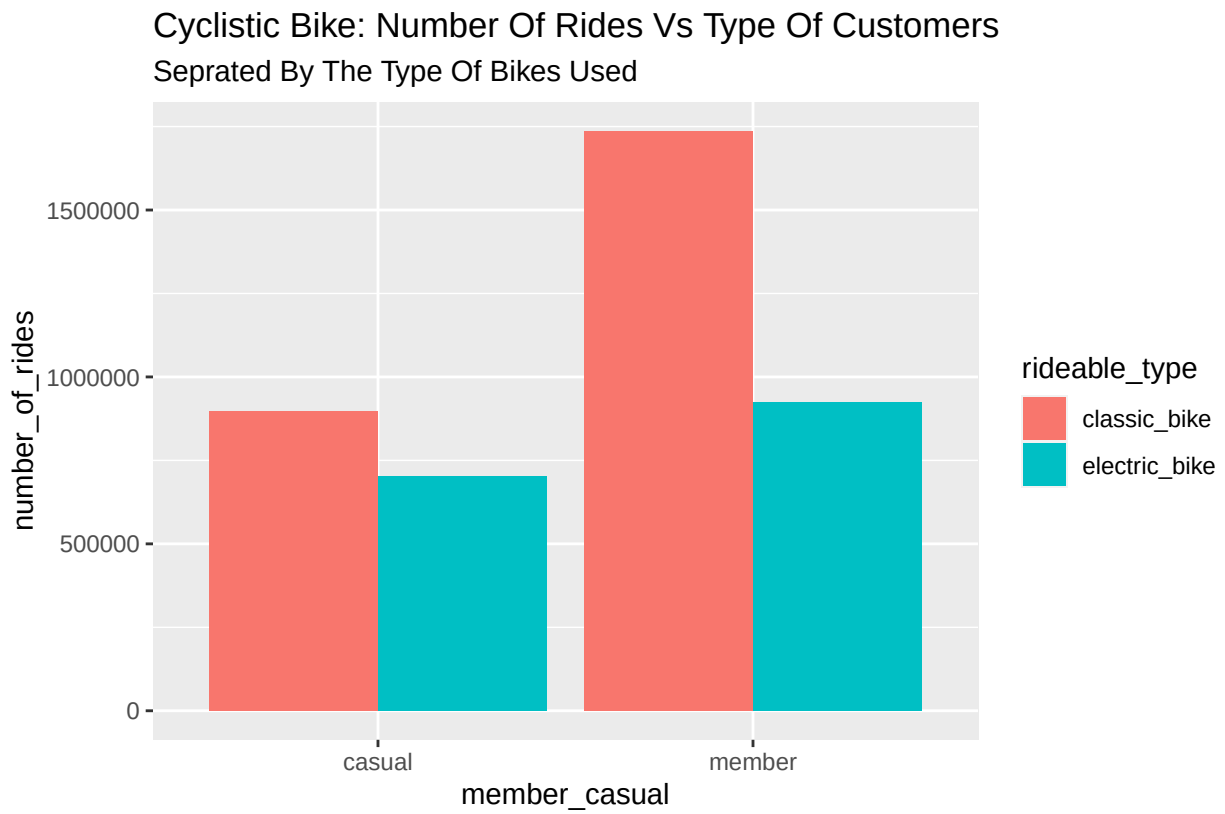
```
## # A tibble: 2 x 4
##   member_casual max_ride_length min_ride_length median_ride_length
##   <chr>         <drtn>          <drtn>          <drtn>
## 1 casual        89965 secs      1 secs          771 secs
## 2 member        89872 secs      1 secs          535 secs
```

Share (Data Visualization)

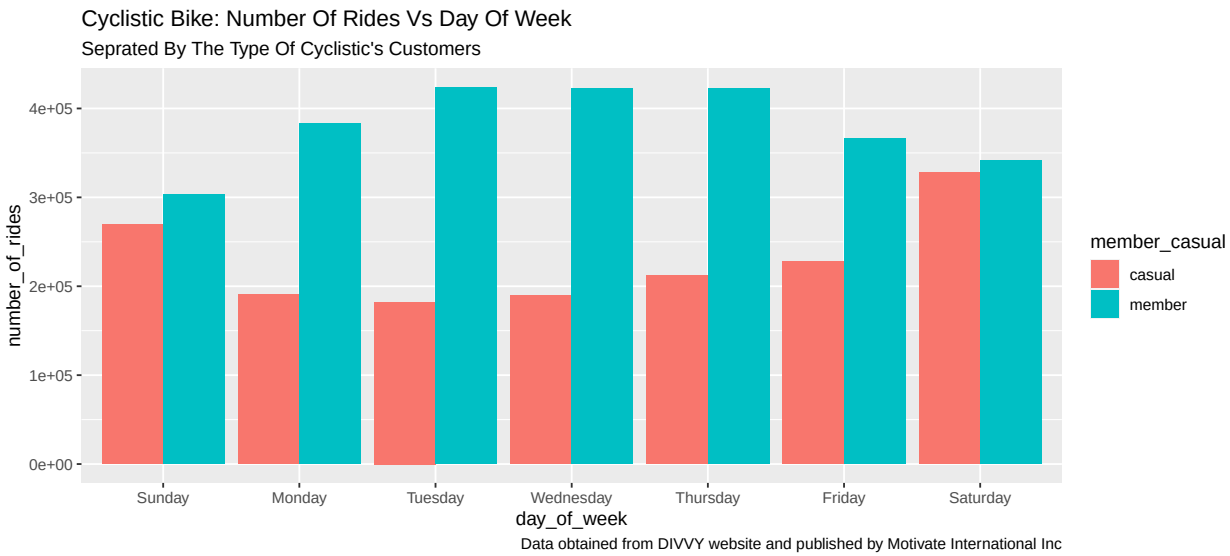
```
ggplot(final_data_clean_analysis1) +
  geom_col(mapping = aes(x = member_casual, y = average_ride_length, fill = rideable_type),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Average Ride Lenght Vs Type Of Customers",
       subtitle = "Seprated By The Type Of Bikes Used",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```



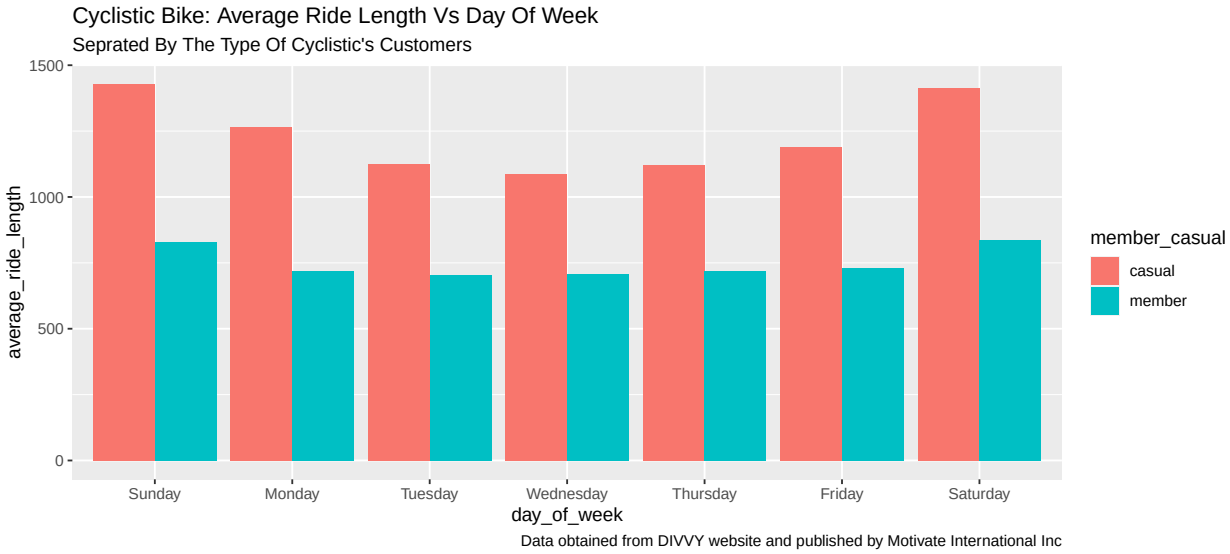

```
ggplot(final_data_clean_analysis1) +
  geom_col(mapping = aes(x = member_casual, y = number_of_rides, fill = rideable_type),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Number Of Rides Vs Type Of Customers",
       subtitle = "Seprated By The Type Of Bikes Used",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```



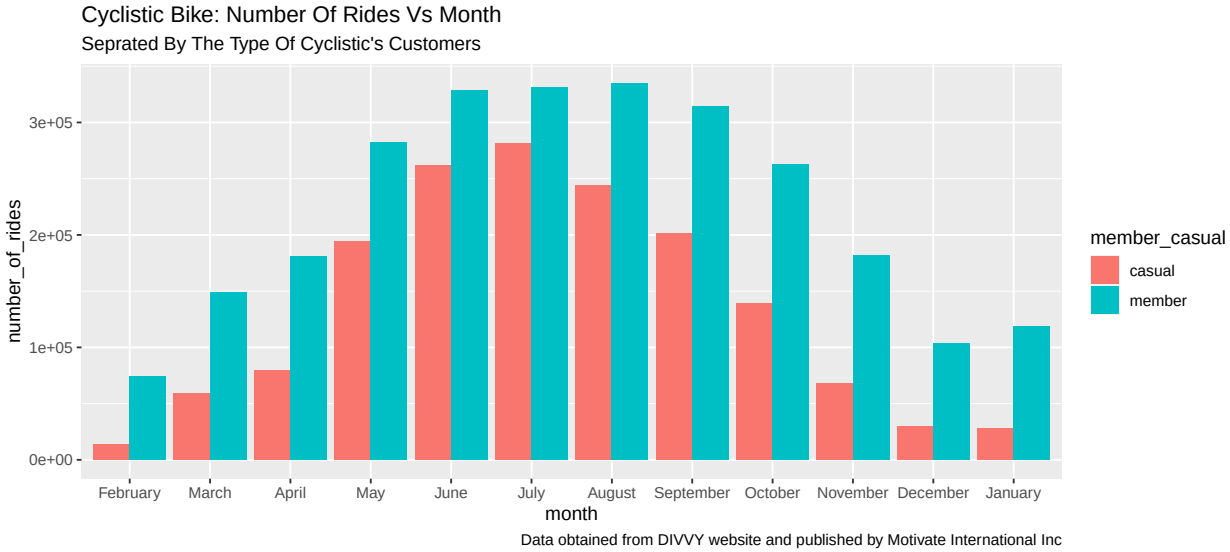

```
ggplot(final_data_clean_analysis2) +
  geom_col(mapping = aes(x = day_of_week, y = number_of_rides, fill = member_casual),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Number Of Rides Vs Day Of Week",
       subtitle = "Seprated By The Type Of Cyclistic's Customers",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```



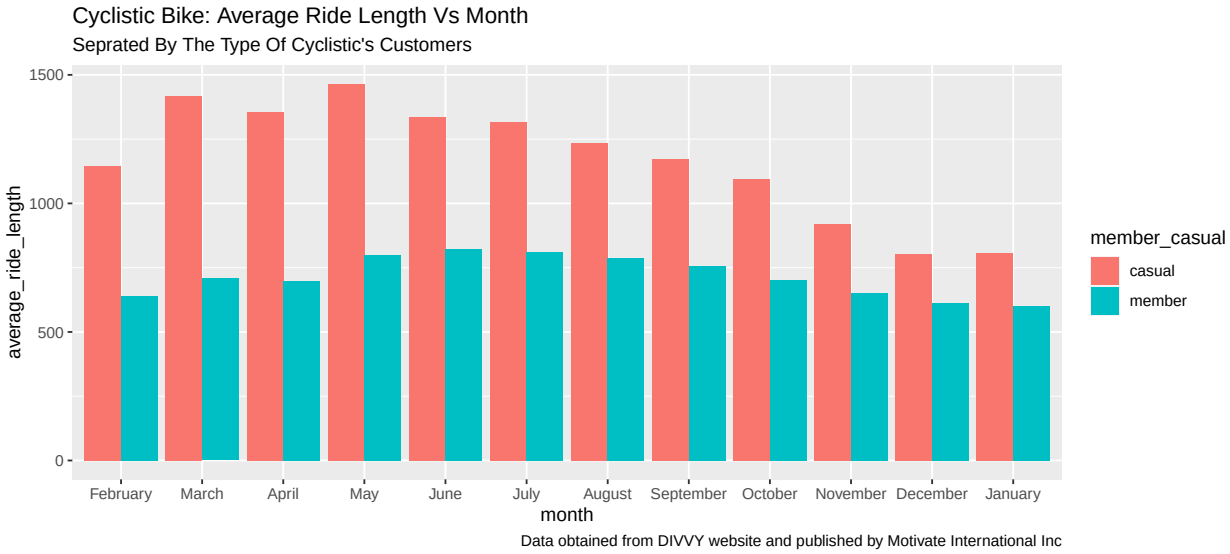
```
ggplot(final_data_clean_analysis2) +
  geom_col(mapping = aes(x = day_of_week, y = average_ride_length, fill = member_casual),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Average Ride Length Vs Day Of Week",
       subtitle = "Seprated By The Type Of Cyclistic's Customers",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```




```
ggplot(final_data_clean_analysis3) +
  geom_col(mapping = aes(x = month, y = number_of_rides, fill = member_casual),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Number Of Rides Vs Month",
       subtitle = "Seprated By The Type Of Cyclistic's Customers",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```



```r
ggplot(final_data_clean_analysis3) +
  geom_col(mapping = aes(x = month, y = average_ride_length, fill = member_casual),
           position = "dodge") +
  labs(title = "Cyclistic Bike: Average Ride Length Vs Month",
       subtitle = "Seprated By The Type Of Cyclistic's Customers",
       caption = "Data obtained from DIVVY website and published by Motivate International Inc")
```



Act

For the last step of data analysis process, we will go over some key findings based on the results of the analysis that we got. Moreover, three recommendations will be made to support a clear goal of designing marketing strategies aimed at converting casual riders into annual members.

Key Findings:

- Based on the average ride length, casual or nonmembers of Cyclistic have ridden the bikes longer than members regardless of the month or day of the week.
- Based on the number of rides used among members is higher than casual regardless of the month or day of the week
- Casuals have ridden bikes longer during the months of spring and summer (March to July). As for the day of the week, Saturdays and Sundays are when nonmembers have ridden the bike the longest. Moreover, they also used the bikes more often on Saturday and Sunday compared to the rest of the days.
- Weekdays especially between Tuesdays to Thursdays are when members have used the bikes more often.
- Between May to September when both members and casuals used the bikes more often compared to the rest of the months.

Recommendations:

- Advertise the annual membership more between May to September. Also, increase the availability of bikes during these months.

- The marketing team should focus more on promoting membership at the most popular stations where bikes have been used.
- The team should also consider a referral program where members refer other customers to Cyclistic to get a discount on their upcoming renewal price.
- The company should also consider designing a campaign that highlights the health and environmental benefits of using the bikes more.