**THE CYCLISTIC CASE STUDY REPORT**

## The case study consists of 6 phases: Ask, Prepare, Process, Analyse, Share and Act.

## **Ask**

### **1.Introduction**

Cyclistic is a bike-sharing company based in Chicago, Illinois. They possess more than 5800 bikes (and also cargo-bikes, hand-tricycles and reclining bikes) and 600 docked stations throughout the city.

### **2.Business task**

Analysing historical trip data from Cyclistic to identify trends and connections between casual riders and annual members.

### **3.Key Stakeholders**

**Lily Moreno**: She is the manager and the director of marketing. She is handling the development of campaigns and the ways to promote the bike sharing program.

**Cyclistic Executive team**: A team of data analysts who are responsible for collecting, analysing, and reporting data that helps guide Cyclistic marketing strategy.

**Prepare**

The data has been shared publicly by Motivate International inc under strict licence. The dataset includes previous 24 months of Cyclistic trip data from May, 2022 through May, 2024. Each row of data is a unique ride with defined start and end data including date, time, station name and ID, lat/long coordinates, the type of bike and whether or not the user was a casual rider or an annual member. It’s also available publicly on Kaggle, thus showing strong integrity.

**Does the data ROCCC?**

A good data source is **ROCCC** which stands for Reliable, Original, Comprehensive, Current, and Cited.

* Reliable - High - It has millions of rides.
* Original - High - It is first-party data.
* Comprehensive - Med - It matches most of Cyclistic's product parameters.
* Current - High - The data is from 2022 to 2024.
* Cited - High- It is first-party data.

# Process

I used RStudio to conduct my analysis due to the efficiency and accessibility of the program, huge amount of data and being able to create the visualisations.

### **1.Installing packages and library**

install.packages("tidyverse")

install.packages(“dplyr”)

install.packages("ggplot2")

install.packages("lubridate")

install.packages("janitor")

install.packages("ggpubr")

install.packages("skimr")

install.packages("here")

install.packages("ggrepel")

install.packages(“readr”)

install.packages(“scales”)

library(tidyverse)

library(dplyr)

library(ggplot2)

library(lubridate)

library(janitor)

library(ggpubr)

library(skimr)

library(here)

library(ggrepel)

library(readr)

library(scales)

I chose these packages to help me with my analysis.

### **2.Importing and preparing the data**

202205-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202205-divvy-tripdata.csv"),

202206-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202206-divvy-tripdata.csv"),

202207-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202207-divvy-tripdata.csv"),

202208-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202208-divvy-tripdata.csv"),

202209-divvy-publictripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202209-divvy-publictripdata.csv"),

202210-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202210-divvy-tripdata.csv"),

202211-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202211-divvy-tripdata.csv"),

202212-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202212-divvy-tripdata.csv"),

202301-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202301-divvy-tripdata.csv"),

202302-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202302-divvy-tripdata.csv"),

202303-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202303-divvy-tripdata.csv"),

202304-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202304-divvy-tripdata.csv")

202305-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202305-divvy-tripdata.csv"),

202306-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202306-divvy-tripdata.csv"),

202307-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202307-divvy-tripdata.csv"),

202308-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202308-divvy-tripdata.csv"),

202309-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202309-divvy-tripdata.csv"),

202310-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202310-divvy-tripdata.csv"),

202311-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202311-divvy-tripdata.csv"),

202312-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202312-divvy-tripdata.csv"),

202401-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202401-divvy-tripdata.csv"),

202402-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202402-divvy-tripdata.csv"),

202403-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202403-divvy-tripdata.csv"),

202404-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202404-divvy-tripdata.csv"),

202405-divvy-tripdata` <- read.csv("C:/Users/Computer/Desktop/cyclistic.csv/202405-divvy-tripdata.csv")

To maintain the consistency of data formats of the columns of each dataset, I have deleted

The last column ‘X’ consisting of null values of the dataset ‘202405-divvy-tripdata’ using the code:

`202405-divvy-tripdata`<-select(`202405-divvy-tripdata`,-X)

### **3.Data cleaning**

-I checked the type of each bike offered using unique() function for all tables.

Eg.-unique(`202405-divvy-tripdata`$rideable\_type)

It was observed that up to Aug 2023, 3 types of bikes were in service ie. classic\_bike, docked\_bike and electric\_bike. Whereas from Sep 2023 docked\_bikes were stopped as the users increasingly preferred dock less, e-assist bikes over traditional docked bikes according to various sources.

-Then I checked the member status of every table using unique()

unique(`202405-divvy-tripdata`$member\_casual)

for all tables, there are 2 values ie. casual and member. Hence, it’s consistent.

-I checked for duplicated rides in every table.

sum(duplicated(`202405-divvy-tripdata`))

it’s 0 in every case, therefore no duplicates found.

-I removed **negative time** of every table, using codes,

`202205-divvy-tripdata` %>% filter(ended\_at < started\_at) %>% count()

`202205-divvy-tripdata` <- `202205-divvy-tripdata` %>% filter(ended\_at > started\_at)

-I removed rows with NA values in every table.

`202205-divvy-tripdata`<-`202205-divvy-tripdata` %>% drop\_na()

-I cleaned and formatted the column names of every table to make sure of its accuracy and consistency.

clean\_names(`202405-divvy-tripdata`)

`202405-divvy-tripdata` <- rename\_with(`202405-divvy-tripdata`, tolower)

-I formatted the datetime column from “chr” to “datetime” format in every table.

`202205-divvy-tripdata` <- `202205-divvy-tripdata` %>% mutate(started\_at = ymd\_hms(started\_at), ended\_at = ymd\_hms(ended\_at))

-Then I merged and renamed the file using bind\_rows()

bike\_tripdata <- bind\_rows(`202205-divvy-tripdata`,`202206-divvy-tripdata`,`202207-divvy-tripdata`,`202208-divvy-tripdata`,`202209-divvy-publictripdata`,`202210-divvy-tripdata`,`202211-divvy-tripdata`,`202212-divvy-tripdata`,`202301-divvy-tripdata`,`202302-divvy-tripdata`,`202303-divvy-tripdata`,`202304-divvy-tripdata`,`202305-divvy-tripdata`,`202306-divvy-tripdata`,`202307-divvy-tripdata`,`202308-divvy-tripdata`,`202309-divvy-tripdata`,`202310-divvy-tripdata`,`202311-divvy-tripdata`,`202312-divvy-tripdata`,`202401-divvy-tripdata`,`202402-divvy-tripdata`,`202403-divvy-tripdata`,`202404-divvy-tripdata`,`202405-divvy-tripdata`)

-I extracted the day from the started\_at column.

bike\_tripdata <- bike\_tripdata %>% mutate(weekday = wday(started\_at, label = TRUE, abbr = TRUE))

-I extracted the month from started\_at column.

bike\_tripdata <- bike\_tripdata %>% mutate(month = month(started\_at,label=TRUE,abbr=TRUE))

- I extracted the time from the started\_at and ended\_at columns.

bike\_tripdata <- bike\_tripdata %>% mutate(start\_time = format(started\_at, "%H:%M:%S"))%>% mutate(end\_time = format(ended\_at, "%H:%M:%S")) %>% mutate(start\_time = hms(start\_time))%>% mutate(end\_time = hms(end\_time)

-I extracted the hour from the started\_at column.

bike\_tripdata <- bike\_tripdata %>% mutate(hour = hour(start\_time))

-Then, I created a duration column.

bike\_tripdata$duration <- difftime(bike\_tripdata$ended\_at, bike\_tripdata$started\_at, units = "mins")

# 

# Analyse and Share(viz)

### **1.Determine the number of members vs casual riders.**

table(bike\_tripdata$member\_casual)

**casual member**

4,624,654 7,556,352

table(bike\_tripdata$rideable\_type)

**classic\_bike docked\_bike electric\_bike**

5,728,999 2,28,488 6,223,519

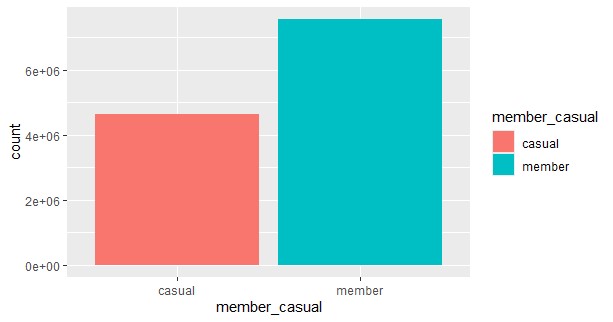
- I plotted the results.

From May 2022 to May 2024, the company had more members (customers that pay a membership fee) than casual riders.

ggplot(data=bike\_tripdata)+

geom\_bar(mapping=aes(x=member\_casual, fill=member\_casual))

Number of Member and Casual riders



### **2. Number of rides per user and type of bikes.**

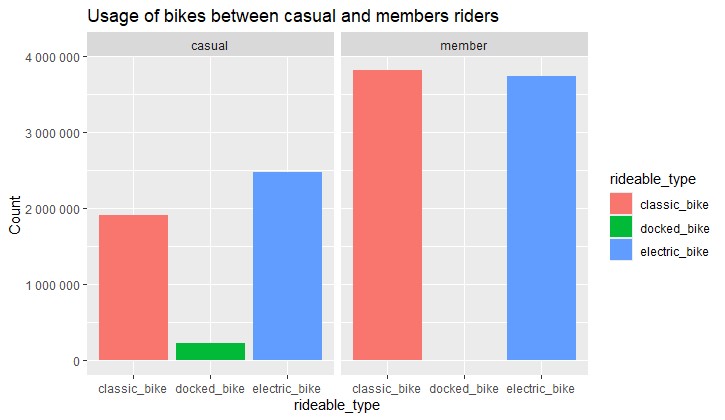
Then, I wanted to see the number of rides per user (member and casual) by rides offered (electric, classic and docked bikes).

bike\_tripdata\_number <- bike\_tripdata %>% select(ride\_id, member\_casual, rideable\_type) %>% group\_by(rideable\_type, member\_casual) %>% count()

|  |  |  |  |
| --- | --- | --- | --- |
|  | rideable\_type | member\_casual | n |
| 1 | classic\_bike | member | 3815113 |
| 2 | electric\_bike | member | 3741239 |
| 3 | electric\_bike | casual | 2482280 |
| 4 | classic\_bike | casual | 1913886 |
| 5 | docked\_bike | casual | 228488 |

-I plotted the results.

ggplot(data=bike\_tripdata)+ geom\_bar(mapping=aes(x=rideable\_type,fill=rideable\_type))+facet\_wrap(~member\_casual)+ scale\_y\_continuous(labels = scales::number\_format(accuracy = 1)) + labs(y = "Count", title = "Usage of bikes between casual and members riders")



-Electric bikes are the go-to bikes from **casual riders**, and Classic bikes from **member riders.**

-Classic bikes comes at 2nd and docked bikes at 3rd(least popular out of the three) from **Casual riders.**

-Electric bikes a close 2nd (next popular choice) and no/very few docked bikes from **member riders.**

-It shows that docked bikes are no more in the picture, as the users increasingly preferred dock less, e-assist bikes over traditional docked bikes according to various sources.

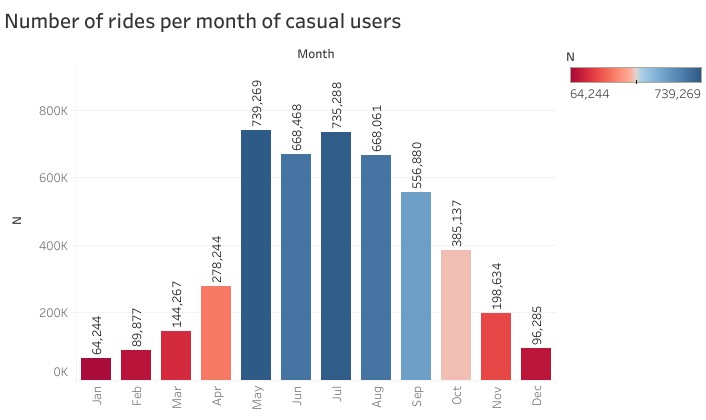
### **3. The number of rides per month.**

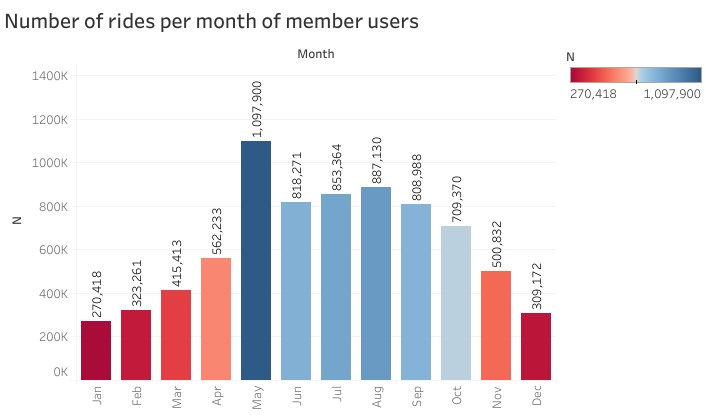
I wondered which months had the most users.

bike\_trip\_month <- bike\_tripdata %>% select(member\_casual, month) %>% group\_by(member\_casual, month)%>% count()%>% arrange(member\_casual)

member\_month <- filter(bike\_trip\_month, member\_casual =="member")

casual\_month <- filter(bike\_trip\_month, member\_casual =="casual")





-**May** to **September** resp. are busiest months for **casual** riders. It is no surprise to witness spring and summer months being the busiest month for the bike sharing company. The members are again more consistent in their use of the bike. And, for **members**, **May** is the busiest month followed by comparatively busy months from **June** to **September.**

-In both cases, **May** is the busiest month out of all the months. As it’s the cusp of Spring and Summer.

-In members, all **12 months** have at least **200,000** rides which is surprising.

### **4.Determine the number of rides for each day.**

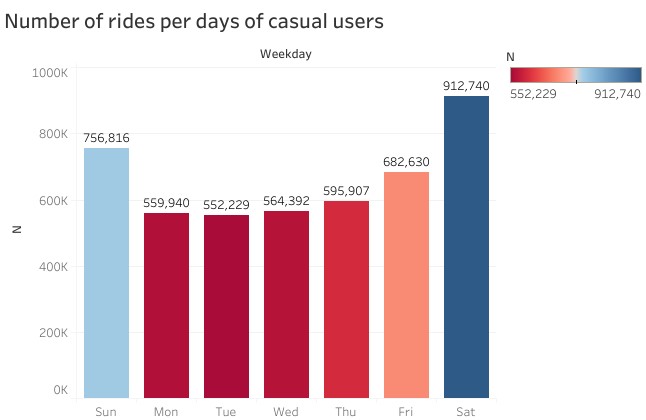
I wanted to determine which day of the week was the busiest for each type of membership.

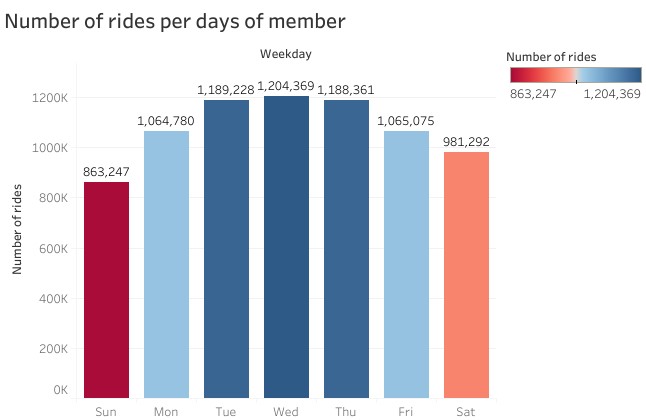
bike\_trip\_weekdays <- bike\_tripdata %>% select(member\_casual, weekday) %>% group\_by(member\_casual, weekday)%>% count()%>% arrange(member\_casual)

member\_weekday <- filter(bike\_trip\_weekdays, member\_casual =="member")

casual\_weekday <- filter(bike\_trip\_weekdays, member\_casual =="casual")

And I plotted the results.





-According to the plots, **member riders** are the most consistent in using the bikes. 5 days of the week exceeding 1,000,000 rides in the member category with **Tuesday, Wednesday, Thursday** in the top 3.

-The casual riders mostly use the bikes on **Saturday** followed by **Sunday** (the weekends).

-The difference between member and casual riders can be explained through the idea of commute + fitness routine+ leisure for the member, but only leisure for the casual riders.

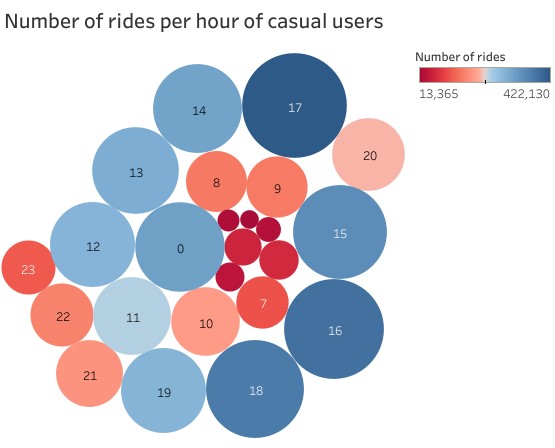
### **5.Getting the number of rides per hour.**

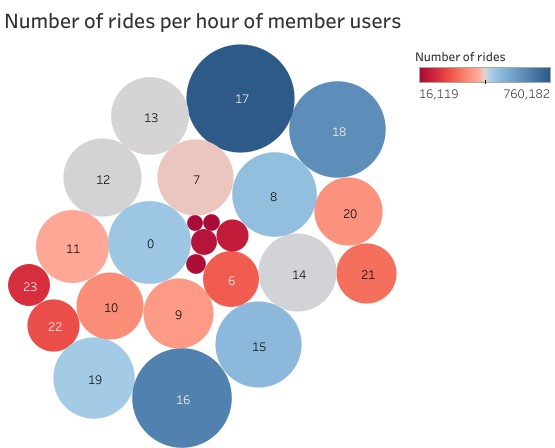
I wanted to be more precise by having the number of rides per hour for each membership.

bike\_trip\_hour <- bike\_tripdata %>% select(member\_casual, hour) %>% group\_by(member\_casual, hour)%>% count()%>% arrange(hour)

member\_hour <- filter(bike\_trip\_hour, member\_casual =="member")

casual\_hour <- filter(bike\_trip\_hour, member\_casual =="casual")





# -Member riders mostly use the bike to commute back home with a massive peak of use between 4 pm and 6pm and 12 am (15h to 19h and 00h). They also ride at 8 am in the morning most likely due to commute and fitness.

# -For casual riders, the peak hours are from 2 pm to 6 pm and 12 am, slight variation in timings from that of members. It can be said that members use bikes throughout the day with some consistency to the difference of the casual users that have a growing pattern from 11 am to 5 pm.

# 6.Finding the top 10 start stations for member and casual riders.

# I have calculated the top 10 most common starting points for both the users to gain some insightful insights.

# I have used the following code to do the calculation.

station\_counts <- bike\_tripdata %>% select(start\_station\_name,start\_station\_id,member\_casual) %>% group\_by(start\_station\_name,start\_station\_id,member\_casual) %>%summarise(user\_count = n()) %>% arrange(desc(user\_count))

member\_counts<- filter(station\_counts, member\_casual =="member")

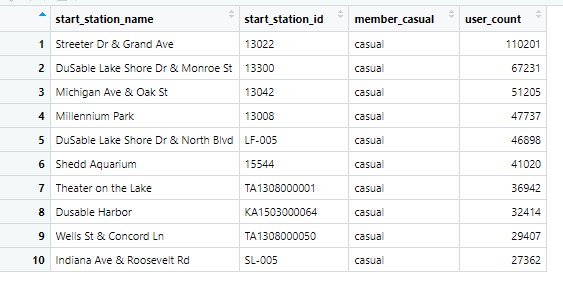
casual\_counts <- filter(station\_counts, member\_casual =="casual")

top\_10\_stations\_member <- member\_counts[2:11, ]

top\_10\_stations\_casual <- casual\_counts[2:11, ]

These are the results for both members and casuals.





### **7.Mean duration for member and casual riders.**

# bike\_trip\_mean\_duration <- bike\_tripdata %>% select(member\_casual, duration)%>% summarise(mean\_duration = mean(duration))

# =114.9405 mins

# casual\_mean\_duration <- bike\_tripdata %>% filter(member\_casual == 'casual')

# casual\_mean\_duration %>% select(member\_casual, duration) %>% summarise(mean\_duration = mean(duration))

= **mean\_duration**

180.5822 mins

member\_mean\_duration <- bike\_tripdata %>% filter(member\_casual == 'member')

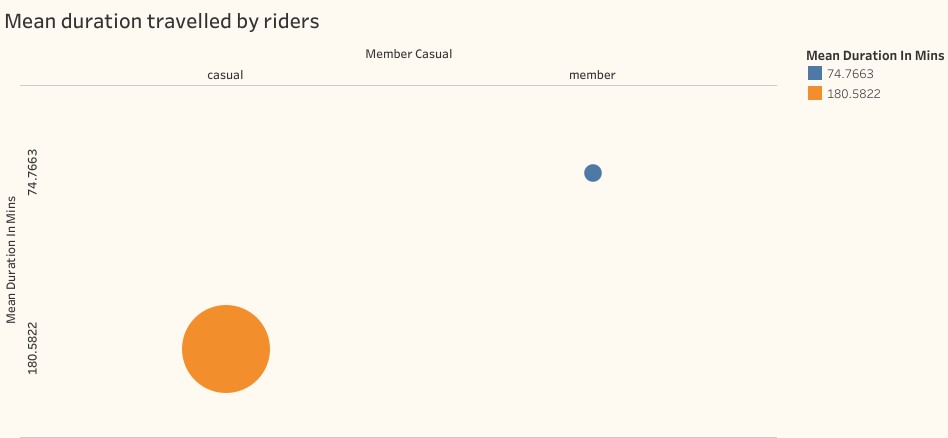
member\_mean\_duration %>% select(member\_casual, duration)%>% summarise(mean\_duration = mean(duration))

= **mean\_duration**

74.7663 mins

-I put the values in tabular form.

|  |  |
| --- | --- |
| member\_casual | mean\_duration |
| member | |  | | --- | | 74.7663 mins | |
| casual | 180.5822 mins |



# I can confirm that the mean duration of a bike trip by casual riders is more than 180 minutes (3 hrs) long compared to the 75 minutes (1.25 hrs) of the member riders. Along with the info that casual riders only ride during the weekends on average confirms that the casual riders use the bike for leisure.

# 

# Act phase

# As I said, over the two-year period that I covered (2022-2024), there were more members than casual riders. The members preferred the classic bikes and the electric bikes being close second. Whereas, casual riders preferred electric bikes.

# Members ride more during the week than casual members that ride mostly during the weekend, but casual members have a longer ride duration than members. It can be explained by the fact that members ride to commute to work or to go home while casuals ride for pure leisure.

# Members and casual riders ride more in the evening period, but members also ride a lot during the morning of the work week.

# Recommendations:

# Plan campaigns to promote the health benefits of using bikes to commute to work and the environmental benefits .

* Create an app that could track the activity of riders (ie. calories burned or CO2 rejections prevented).

# Create partnerships with local companies to promote the usage of bikes through discount memberships or stations nearby.

# Increase bike availability in the top 10 start stations.