Cyclistic Bike Share Capstone Project Final Report

==========

Damtew, Wasihun Eshetu

Google Data Analytics Professional Certificate Capstone Project

December 22, 2021. 06:19 PM

==========

Case Study: How does a Bike-Share Navigate Speedy Success?

Introduction and Goal of the Capstone Project

In 2016, a Cyclistic (a fictional bike-share company) launched a bike-share offering program that benefited Chicago residents and visitors by availing around a fleet of 5,824 bicycles (include the standard two wheeled bikes, reclining bikes, hand tricycles, and cargo bikes) that are geo-tracked and locked into a network of 692 stations across the city. The program runs with serving two types of customers namely the casual riders who use the service temporarily and the cyclistic members who utilize the service on regular basis. The bikes can be unlocked from one station and returned to any other station in the system anytime. The price of the service is flexible ranging from a single-ride passes, full-day passes as well as annual memberships.

Moreno, the finance analyst of the company, concluded that annual membership is much more profitable than casual ride. Instead of creating a marketing campaign that targets new customers, Moreno believes that maximizing the number of annual members will be key to future growth. Moreno's goal is to design marketing strategies aimed at converting casual riders into annual members. For the company to achieve its goal, there is a need to conduct a study to understand how annual members and casual riders differ, why casual riders would buy a cyclistic membership, and how digital media could affect their marketing tactics to come up with an evidence-based or informed strategies. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends and come up with results and recommendations.

As a junior data analyst, who is currently working in the marketing analyst team at Cyclistic, a bike-share company in Chicago, I got access to a one year data sets from the company, through Coursera, to conduct simple and rapid analysis using the six phases of data analysis (Ask, Prepare, Process, Analyze, Share, and Act) to address the three basic questions stipulated above and to produce results of standard quality backed up with a compelling data insights and professional data visualizations to convince the Cyclistic marketing analytics team, Lily Moreno (the director of marketing and my manager) and Cyclistic executive team.

PHASE 1: Ask

The statement of the business task, in other words, the key questions I would like to address that will help guide the future project include:

1. How do annual members and casual riders differ in utilization of the cyclist biking service?
2. Why would casual riders buy Cyclist annual memberships?
3. What comparative advantages does it exist for the company and the clients to become annual members of the cycling service
4. How can Cyclist use digital media to influence casual riders to become members

Key Stakeholders

The three key actors of the project are the Cyclistic marketing analytics team, Lily Moreno (the director of marketing) and Cyclistic executive team.

PHASE 2: Data Preparation

Data Source:

Motivate international Inc., a licensed company (https://www.divvybikes.com/data-license-agreement) which operates Divvy bicycle sharing services in the City of Chicago presented the public data set to conduct the study. A representative sample of the data taken, compiled quarterly, covered a period of one full year - the first, second and third quarters of 1999 and the first quarter of 2000 which runs from April 2019 through March 2020 to conduct the analysis.

Description of Data Sets:

The datasets, compiled on quarterly basis and saved as Divvy_Trip_2019_Q2, Divvy_Trip_2019_Q3, Divvy_Trip_2019_Q4 and Divvy_Trip_2020_Q1, were downloaded from Coursera online training website where Google is offering Data Analytics Professional Training to my personal computer and saved under the folder entitled “Bike_Share Company”. A replica of the same data set was also created to preserve the original data from being tampered and to serve as a backup file.

PHASE 3: Process

Since the data is collected and shared by the same company and the project is intended for internal consumption, it is assumed that the data set satisfies the standard criteria of ROCCC (Reliable, Original, Comprehensive, Current and Cited). Hence, the issue of bias or credibility, licensing, privacy, security, accessibility, and integrity is not an issue that needs to be discussed or justified.

Data Wrangling, Cleaning and Transformation in EXCEL SHEET.

The four data sets were open in my PC using Microsoft excel one after the other and checked for errors, cleaning, completeness, and data consistencies. After checking the contents of the datasets, column heads, number of columns and rows, and for any missing or ‘na’ values and noted what needs to be done. The following actions were taken on the excel sheet formats before even importing the data sets to R-programming software.

1. The ‘started_at’ and ‘ended_at’ columns had to be further extrapolated to obtain a new numeric column called a ‘ride length’ to calculate descriptive statistics. Using time difference function, a ride length was calculated by subtracting time value registered at “ended_at” from column “started_at” and saved into newly created column entitled “ride_length”. The measuring unit for ride length was calculated and/or converted into seconds.

2. Few columns which are not important for the analysis such as start_latitude, start_longitude, end_latitude, end_longitude, birth year and gender were deleted from all the data frames using ‘delete’ command.

3. The column head entitled “member_casual” had two types of names for each category, ‘subscriber’ or ‘member’ and ‘customer’ or ‘casual’; hence using the “find and replace” command all ‘subscribers’ were reassigned into ‘members’ and all ‘customer’ were reassigned into ‘casual’ as well.

4. The four quarterly data sets were saved as CBS_2019_Q2, CBS_2019_Q3, CBS_2019_Q4 and CBS_2020_Q1, where CBS represents Cyclistic Bike Share. The data sets were ready for importing into R-programming. . The relevant R-programming packages such as tidyverse, lubridate, and Rmarkdown were installed using the 'installed.packages' command indicated below and incorporated into the R programming to conduct the task of data cleaning, manipulation, visualization, and analysis.

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

All the three packages were also loaded into the R-Studio using 'library' command to conduct data wrangling, cleaning, and transformation.

```
library(tidyverse)
## Warning: package 'tidyverse' was built under R version 4.1.2
## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## Warning: package 'tibble' was built under R version 4.1.2
## Warning: package 'readr' was built under R version 4.1.2
## -- Conflicts ------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
library(lubridate)
## Warning: package 'lubridate' was built under R version 4.1.2
## ## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
## date, intersect, setdiff, union
library(rmarkdown)
## Warning: package 'rmarkdown' was built under R version 4.1.2
```

By opening the R-script pane in R-studio and using 'readr' command, all (four) data sets were imported into the software or 'environment pane'.

```
library(readr) CBS_2019_Q2 <- read_csv("CBS_2019_Q2.csv") View(CBS_2019_Q2)
library(readr) CBS_2019_Q3 <- read_csv("CBS_2019_Q3.csv") View(CBS_2019_Q3)
```

```
library(readr) CBS_2019_Q4 <- read_csv("CBS_2019_Q4.csv") View(CBS_2019_Q4)
library(readr) CBS_2020_Q1 <- read_csv("CBS_2020_Q1.csv") View(CBS_2020_Q1)
```

Using 'str' command of the R studio, columns names, data across columns and rows were viewed for any discrepancies and to clean it further and improve the data quality. The data sets for all the three quarters of 2019 had of 12 variables while the first quarter of 2020 had 13 variables. Since the data set for first quarter of 2020 were different and had a better format than the other three data sets of 2019, it is used as a template to ensure all data sets do match and merge perfectly as one and single clean and complete document.

```
str(CBS_2019_Q2)
str(CBS_2019_Q3)
str(CBS_2019_Q4)
str(CBS_2020_Q1)
```

To ensure data sets perfectly match and merge as one, single complete document, 'mutate' function were used to manipulate the data. Data types such as the 'ride_id', 'rideable_type' columns of all the 2019 datasets were converted into "character". Besides, in all datasets, the data type in column 'ride_length' were also converted from 'character' to "numeric" to ensure matching and merging the data sets into one perfect data frame.

```
CBS_2019_Q2 <-mutate(CBS_2019_Q2,ride_id=as.character(ride_id))
CBS_2019_Q2 <-mutate(CBS_2019_Q2,rideable_type=as.character(rideable_type))
CBS_2019_Q3 <-mutate(CBS_2019_Q3,ride_id=as.character(ride_id))
CBS_2019_Q3 <-mutate(CBS_2019_Q3,rideable_type=as.character(rideable_type))
CBS_2019_Q4 <-mutate(CBS_2019_Q4,ride_id=as.character(ride_id))
CBS_2019_Q4 <-mutate(CBS_2019_Q4,rideable_type=as.character(rideable_type))
CBS_2019_Q2 <-mutate(CBS_2019_Q2,ride_length=as.numeric(ride_length))
CBS_2019_Q3 <-mutate(CBS_2019_Q3,ride_length=as.numeric(ride_length))
CBS_2019_Q4 <-mutate(CBS_2019_Q4,ride_length=as.numeric(ride_length))
CBS_2020_Q1 <-mutate(CBS_2020_Q1,ride_length=as.numeric(ride_length))
```

Finally, using the 'bind_rows' command the four data sets were merged into one big data frame and saved as mdf.csv (merged data frame) file. Up on merging the data frames, one extra column under the "start_station_name" that has "NA" values under it were created and deleted from the fine using the second command found below.

```
mdf <-bind_rows(CBS_2019_Q2, CBS_2019_Q3, CBS_2019_Q4, CBS_2020_Q1)
mdf <- mdf[ , -c(11)]
```

Results of the merged data frame yield 10 variable or field names that include:
1. ride_id: a unique ID assigned for each ride.

2. rideable_type: bicycle type used for the ride.

3. started_at : the year, date and time of the service commenced.

4. ended_at: the year, date and time of the service terminated.

5. ride_length: the time taken to ride a bike from beginning to the end of the station

6. start_station_name: name of the station where the bicycle is picked.

7. start_station_id: a unique ID given for the starting point of the docking station.

8. end_station_name: name of the station the where the bicycle is returned.

9. end_station_id: a unique ID given for the destination of the docking station.

10. member_casual: indicates if whether a person is a member or a casual of the service.

Using the 'str' (data structure), 'head', 'colnames', 'nrows', 'dim', 'summary' function of the r-studio, all data sets were inspected for any discrepancies.

```
str(mdf)

head(mdf)

colnames(mdf)

nrow(mdf)

dim(mdf)

summary(mdf)
```

Immediately after finishing the data inspection, cleaning, manipulation, the data frame entitled 'mdf' was converted into 'mdf_final' using the following command.

```
mdf_final <-mdf
```

Rows and dimension of data frame were inspected again. Data cells that are found to be blank or cells designated as 'NA' were deleted using 'filter' command. Up on inspecting the data frame for the second time, the following additional action were taken to address data discrepancies or incompleteness.

The date command stipulated below were used to split the column entitled "started_at" into five new columns namely the 'date' with yyyy-mm-dd format, 'month', 'day', 'year' and 'day_of_week data formats' to obtain additional in-depth insights to address the four fundamental questions developed in the "Ask Phase" of the data analysis.

```
mdf_final <- mdf%>%

  mutate(started_at=as_datetime(mdf$started_at, format = "%d/%m/%Y %H:%M"))%>
% mutate(ended_at=as_datetime(mdf$ended_at, format = "%d/%m/%Y %H:%M"))

mdf_final$date <-as.Date(mdf_final$started_at)

mdf_final$month <-format(as.Date(mdf_final$date),"%m")

mdf_final$day <-format(as.Date(mdf_final$date), "%d")

mdf_final$year <-format(as.Date(mdf_final$date), "%y")
```

```
mdf_final$day_of_week <-format(as.Date(mdf_final$date), "%A")
```

Ride_length that depicted zero or negative values due to removing the bicycles out of docks for maintenance and any other reasons were deleted from the data frame using 'filter' command.

```
mdf_final <- mdf_final %>%
  filter(ride_length > 0)
```

Finally, the latest and cleaned version of the data frame named as "mdf_final" that contains 15 columns and 3,228,091 rows was ready for analysis phase. The five new or additional column include 'date' with yyyy-mm-dd format, 'month', 'day', 'year' and 'day_of_week'. All newly created columns except 'day_of_week' are numeric data type while the other contains character type.

PHASE 4: Analyzing Data and Summarizing the Result

The final version of data frame entitled "mdf_fianal" was subjected to inspection and cleaning one more time using the following command to ensure data quality.

```
str(mdf_final)
head(mdf_final)
colnames(mdf_final)
nrow(mdf_final)
dim(mdf_final)
summary(mdf_final)
```

Descriptive statistics were used to extrapolate some variables to obtain additional information. The first action in data analysis were to work on the ride_length. Using the following command, the mean, median, minimum, and maximum value of the aggregated data of the ride_length was analyzed.

```
summary(mdf_final$ride_length)


aggregate(mdf_final$ride_length ~mdf_final$member_casual, FUN = mean)
aggregate(mdf_final$ride_length ~mdf_final$member_casual, FUN = median)
aggregate(mdf_final$ride_length ~mdf_final$member_casual, FUN = max)
aggregate(mdf_final$ride_length ~mdf_final$member_casual, FUN = min)
```

Then value of ride_length was cross tabulated with membership type (member_casual) and weekdays (day_of_week) using the following command. Result of the analysis is presented in Table # 1.

```
mdf_final$member_casual~mdf_final$day_of_week~mdf_final$ride_length
aggregate(mdf_final$ride_length ~mdf_final$member_casual+mdf_final$day_of_wee
k, FUN = mean)
```

```
counts<-aggregate(mdf_final$ride_length ~mdf_final$member_casual+mdf_final$da
y_of_week, FUN = mean)
```

Table # 1: Mean ride length in seconds		
	Casual	Member
Moday	3087	883
Tuesday	3496	794
Wednesday	3628	818
Thursday	3256	846
Friday	3726	837
Saturday	4312	795
Sunday	4133	793
Mean ride length for the week	3663	824

The number of bike users throughout the weekdays and the number of rides and average duration ofthe ride were also cross tabulated with membership type (member_casual) and weekdays(day_of_week) using the following command. The data is presented in Table # 2.

```
mdf_final %>%
  mutate(weekday = wday(started_at, label=TRUE))%>%
  group_by(member_casual, weekday)%>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length))%>%
  arrange(member_casual, weekday)
```

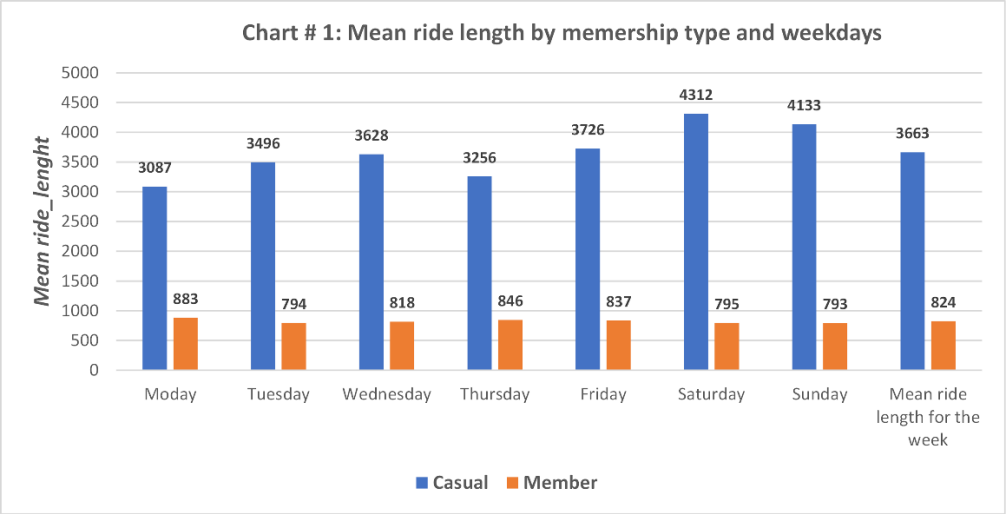
Table # 2: Number of Bike users per day		
Number of Bike Users per Day	Casual	Member
Moday	44506	147494
Tuesday	35619	166887
Wednesday	47414	158604
Thursday	31152	164977
Friday	44058	163176
Saturday	50472	139251
Sunday	56623	135376
Mean total for the week	44263	153681

The minimum, first quartile, mean, median, third quartile and maximum value of the ride length for membership type (casual and member) as well as for both clients were also computed and summarized in Table # 3

Table # 3:Ride length aggregated by membership type			
	Casual	Member	Both
Mean	3668	844	1474
Medain	1560	600	720
Minimum	60	60	-3360
Maximum	9387000	9056640	9387000
First Quartile			420
Third Quartile			1260

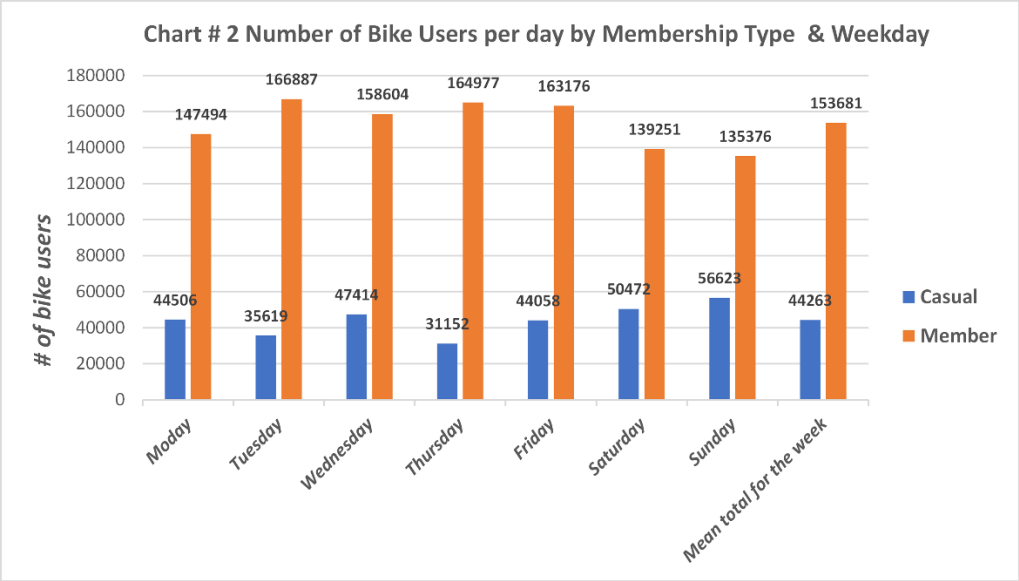
Visual representation of the two tables mentioned above were also obtained using ggplot (see command below) and excel sheet. For clarity of visualization purposes, the two chart obtained from excel sheet were presented below

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

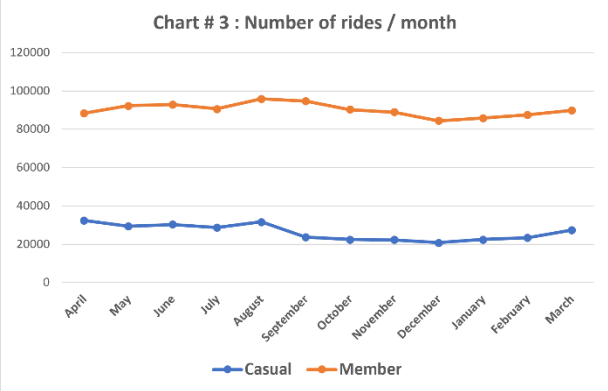


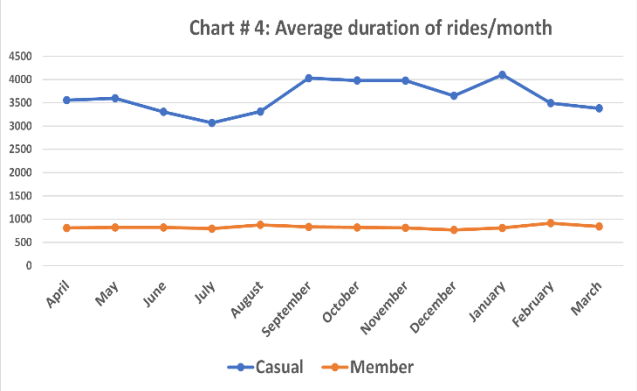

```
mdf_final %>%
  mutate(weekday = wday(started_at, label=TRUE))%>%
  group_by(member_casual, weekday)%>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length))%>%
  arrange(member_casual, weekday)%>%
```

```
  ggplot(aes(x=weekday, y=average_duration, fill = member_casual))+ geom_col(
position="dodge")
```



Summary of the aggregate data analysis for number of rides and average duration rides per month across the entire year (April 2019 through March 2020) was also computed for weekdays and months using the following two commands (cross tabulation of membership type by weekday, number of rides and average duration of ride).






```
summary_week_day <- mdf_final%>%
  mutate(weekday = wday(started_at, label=TRUE))%>%
  group_by(member_casual, weekday)%>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length))%>%
  arrange(member_casual, weekday)
```

Aggregated data analysis (a cross tabulation of membership type by month, number of rides and average duration of ride)

```
summary_month <- mdf_final%>%
  mutate(month = month(started_at, label=TRUE))%>%
  group_by(month, member_casual)%>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length))%>%
  arrange(month, member_casual)
```

Percentage value of membership type by weekday (computation of percentage value for members and casual across weekday)

```
xtabs(~member_casual+day_of_week, data = mdf_final)
crosstabs <-xtabs(~member_casual+day_of_week, data = mdf_final)
prop.table(crosstabs, 1)
```

Percentage value of membership type by day (computation of percentage value for members and casual by day)

```
xtabs(~member_casual+day_of_week, data = mdf_final)
crosstabs <-xtabs(~member_casual+day_of_week, data = mdf_final)
prop.table(crosstabs, 2)
```

RESULT of the Data Analysis

The final version of the cyclist bike share data frame is consisted of 15 columns and 3,228,091 rows.

As presented in Table 1 and Chart 1, in all the weekdays, the mean ride length for casual bike users exceeded by far that of regular customers. The weekly mean ride length of casual is 3663 seconds while for members is 824 seconds, depicting the mean ride length for casuals exceeded by four folds than members. However, Casuals, on average, traveled long distance than members.

The ride length for members across all weekdays (Table 1 & 3 and Chart 1) has very little or no significance difference; however, it showed declining trend over the weekends. The highest value was on Monday (883 sec) followed by Thursday (846 sec). The maximum service utilization by members during weekdays indicates that they are using the service to commute for work. The ride length for casuals showed an increasing trend over the weekend. Ride length for Saturday scored the highest (4312 sec) followed by Sunday (4133 sec), this figure is indicating that casuals are weekend service users most likely for picnic and recreational activities. The number of bike users per day for members is by far higher than the casuals in all the seven days of the week. The weekly average or mean bike use value of members is 3.5 times higher than that of casuals.

As depicted in Table # 2 and Chart # 2, there was a very significance difference as far as the number of daily bike users between members and casuals. Throughout the whole week, the average number bike users for members were 3.5 times higher than that of casual customer. The daily average of bike users for members scored 153,681 while that of casuals was 44,263. The data showed that there was a decline in the number of bike users for members during the weekends but the number of bike users for casuals over the weekend increased significantly.

The number and average duration of rides were aggregated in months for the entire year (April 2019 through March 2020) and result of the analysis (as shown in Chart # 3) showed that casuals used bike

rides more between the month of March through April (scored around 30,000 users a month) then sharp declined in September with the lowest value from October through February. The number of bike users for regular members showed an increasing trend from the month of April through September (averagely around 90,000) then declined starting from October through March. The line graph for members as shown in Chart # 4 yielded almost a straight line indicating there was similar duration of rides in all months, while the line graph for casuals showed up and down trend. The average duration of rides for casuals showed declining trend from April to end of July and then an increasing trend from beginning of July through November.

PHASE 5: Share

Findings of the data analysis is compiled in R markdown. Result of the analysis is presented in the form of narratives, summarized figures in table format and visual data in charts using ggplot and excel sheet. The final version of the cyclist bike share data frame was saved to R programming and to my desktop for future references using the commands shown below.

```
write.csv(mdf, file="mdf_final.csv")


View(mdf_final)


write.csv(mdf_final, "C:/Users/wasih/OneDrive/Desktop/capstone_project.cSV",
sep=",", row.names = FALSE)
```

Preliminary findings of the data analysis that addressed the original key questions with its recommendation will be presented in Power Point format to the relevant key actors seeking for input to improve output. Any constructive feedback from team members (key actors) will be incorporated into the report. The final version of the report will finally be shared among relevant key actors for decision making and further actions. Both print out and soft copies of the report will be shared among team members and the executives in person and via e-mail address. Key findings covering all steps and processes of data analysis with its recommendations, final version of data frame, tables and charts has been saved in my computer in a new folder entitled “Cyclist Bike Share Capstone Project”. The final output and the row data is also accessible for the audience on git-Hub website through the following links: file:///C:/Users/wasih/OneDrive/Desktop/CSV/Rmarkdown-for-Bike-Share_Final.html

PHASE 6: Act

Based on the results from the analysis, the following conclusion and recommendations can be drawn to assist the marketing department to design a marketing strategy that will convert casual riders into annual members.

Casual use bike share rides much less frequent than regular members but take longer rides, in each trip, than members. Casuals use the bike service more over the weekend than weekdays mainly for recreational purposes. The use of bike service for Casuals during the month of March through April higher than the rest of months. Members use bike service throughout the week in similar manner indicating that they use the bike service to commute to work.

Based on the conclusion the following six recommendation can be made to address the original question.

1. Design, promote and market different types of bike riding sales packages such as seasonal and annual membership bike riding packages to attract casual as well as new customers that can address their immediate and long-term needs.
2. Conduct promotion at the right time where most casual clients go out and about, especially over the weekends, spring and summertime (between March and April) and appropriate sites targeting the busiest bike stations to reach out to casual members and encourage them sign in for annual membership.
3. Liaise with event organizers (sport, parks, theatres, and music concert & festivals etc,) and take advantage to reach large number of people to introduce the packages. Use different types of marketing strategy such advertisements on TV and radio, websites, leaflets and billboards. Use different type of promotion as offering discount rates, prizes, and coupons to encourage and attract more customers.
4. Up on conducting needs assessment, increasing number of bikes and parking stations would most likely attract both casual and new clients due to creating awareness, an increase of service accessibility and availability issues.
5. There also needs to be analyzing latest cyclistic bike share data sets to get more insights and to see trends and changes of the bike users over time and to design need-based strategy to meet both the client's and the company's objectives.
6. While campaigning for casual members, it would also be wise to target new members with very minimal or no additional costs incurred to benefit the company to meet its objectives.