Case Study: Use the Bike-Sharing company's historical data to find customer's insight

Thanawat Riencharoen

2022-04-04

This is my analysis for Google Data Analytics Professional Certificate Course's study case, please read the study case question here

Business Task

Use Cyclistic's historical bike trip data to analyze the different behavior between annual members and causal riders and use the information to recommend the new marketing strategies to convert causal riders to annual members.

Key Questions

1. How do annual members and casual riders use Cyclist bikes differently?
2. Why would casual riders buy Cyclistic annual membership?
3. How can Cyclisitc use digital media to influence casual riders to become members?

Data preparation

The public data that is used in the analysis, is provided by Motivate International Inc. here

For this study case, I will use the R language for analysis because the data is exceed Excel row limitation (more than 5 million rows). First, I load all the data inside a new folder, installed the necessary packages, and Imported all the CSV files for every month in 2021.

```
library(tidyverse)

## -- 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_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(lubridate)

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

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

library(ggplot2)
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor

library(dplyr)

trip_data_2022_01 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202101-divvy-tripdata.csv")
trip_data_2022_02 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202102-divvy-tripdata.csv")
trip_data_2022_03 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202103-divvy-tripdata.csv")
trip_data_2022_04 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202104-divvy-tripdata.csv")
trip_data_2022_05 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202105-divvy-tripdata.csv")
trip_data_2022_06 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202106-divvy-tripdata.csv")
trip_data_2022_07 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202107-divvy-tripdata.csv")
trip_data_2022_08 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202108-divvy-tripdata.csv")
trip_data_2022_09 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202109-divvy-tripdata.csv")
trip_data_2022_10 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202110-divvy-tripdata.csv")
trip_data_2022_11 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
```

```
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202111-divvy-tripdata.csv")
trip_data_2022_12 <- read.csv("C:/Users/ADMIN/Desktop/Prin/Google Data Analys
is/Lesson 8 Capstone Complete a Case Study/Study Case How Does a Bike-Share N
avigate Speedy data/202112-divvy-tripdata.csv")
```

Then I combine all of the loaded data into a single data frame.

```
trip_data_2022 <- rbind(trip_data_2022_01, trip_data_2022_02, trip_data_2022_
03, trip_data_2022_04, trip_data_2022_05, trip_data_2022_06, trip_data_2022_0
7, trip_data_2022_08, trip_data_2022_09, trip_data_2022_10, trip_data_2022_11
, trip_data_2022_12)
```

Clean data

After having a data frame to work with, I clean the data by removed any row that have missing value and rename member_causal column to member_type for clarification.

```
trip_data_2022_clean <- na.omit(trip_data_2022)
trip_data_2022_clean <- rename(trip_data_2022_clean, member_type = member_cas
ual)
```

Processing data

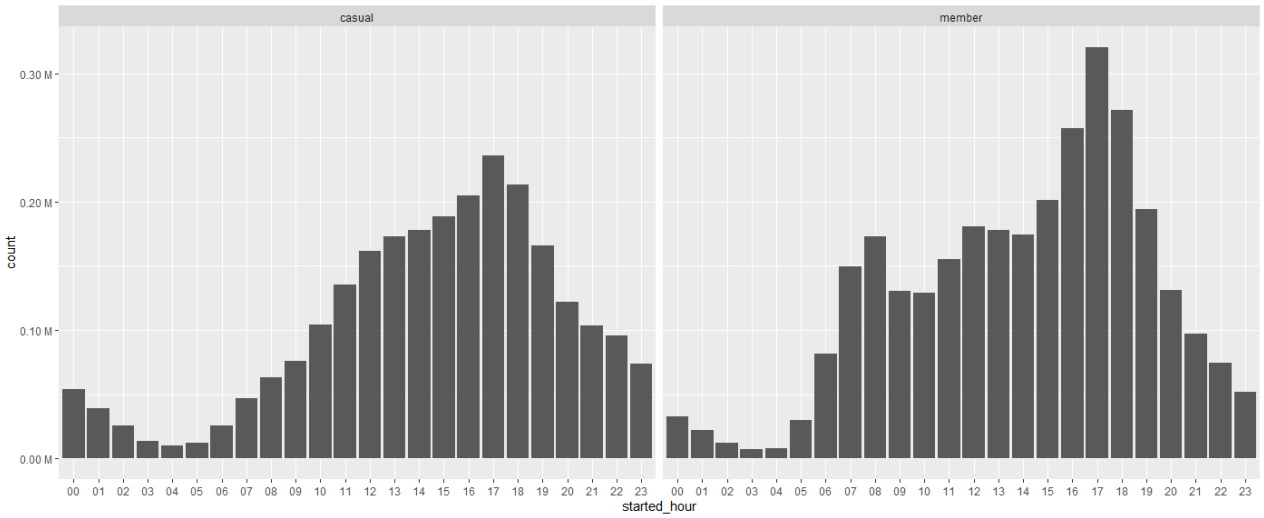
After reviewing the cleaned data I notice that there is still lack of the information for behavior analysis, so I create new columns for biking duration, day of the week, and another column for extracted hour from the start_hours column.

```
trip_data_process <- mutate(trip_data_2022_clean, duration = round((as.numeri
c(ymd_hms(ended_at) - ymd_hms(started_at)))/60))
trip_data_process <- mutate(trip_data_process, day=weekdays(ymd_hms(started_a
t)))
trip_data_process <- mutate(trip_data_process, started_hour=format((ymd_hms(s
tarted_at)), format = "%H"))
```

Analyze

I start analyzing the hours first to see the difference between the time causal riders and annual members using the bike.

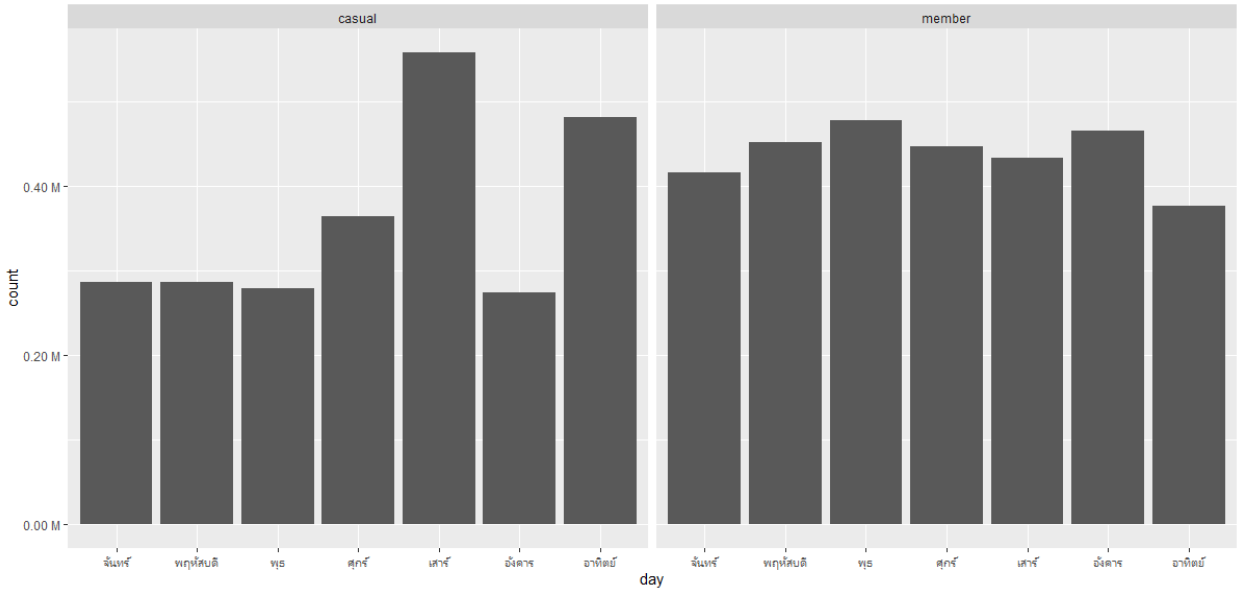
```
ggplot(data = trip_data_process) + geom_bar(mapping = aes(x = started_hour))
+ facet_wrap(~member_type) + scale_y_continuous(labels = unit_format(unit = "
M", scale = 1e-6))
```



As you see, there is a significant difference in the time both types of users use the bike. Causal riders use the bike lowest at 4 pm and peak at 5 pm, however, annual members have a small peak at 8 am and then gradually increase to the peak usage after work at 5 pm.

I investigate further by create the bar graphs comparing the difference both groups using during the day of work.

```
ggplot(data = trip_data_process) + geom_bar(mapping = aes(x = day)) + facet_wrap(~member_type) + scale_y_continuous(labels = unit_format(unit = "M", scale = 1e-6))
```



While annual members use bike usage almost the same amount everyday, casual rider's usage is highest on Saturday and Sunday

both of this information suggests that many annual members riders use the bike for commuting to work that why the usage hour peak at 7 am and then higher peak at 5 pm, while causal riders may use bike for the bike free-time activities.

Now I want to find the average duration that the annual members ride the bike to work, so I use the code to group the rider type together and find the average riding duration during 8am and 5pm.

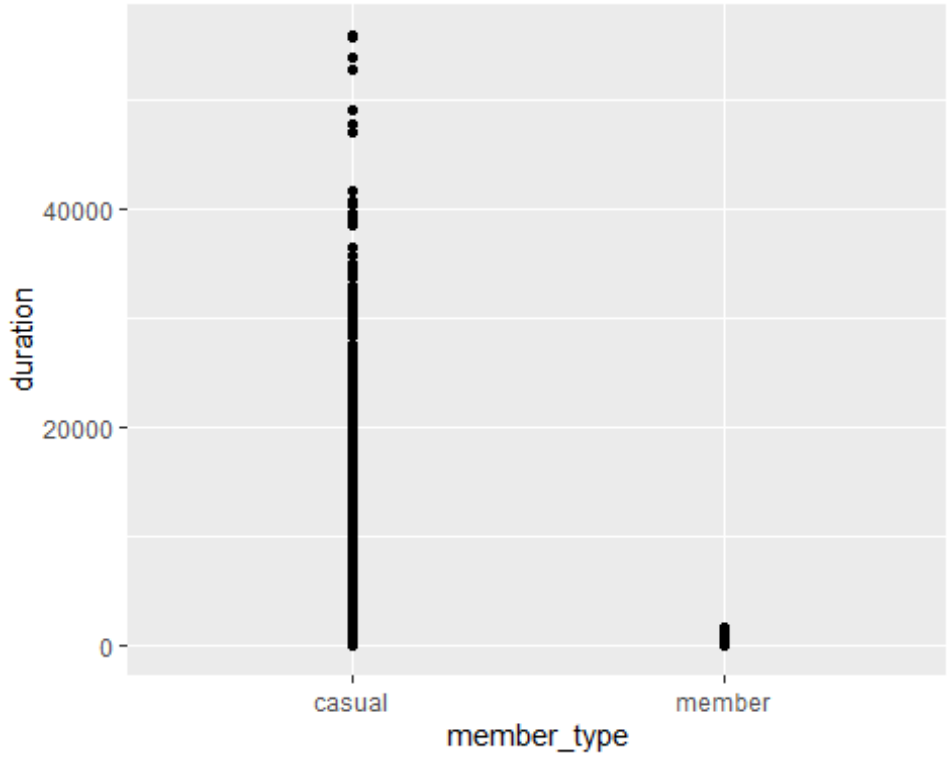
```
work_duration_summary <- trip_data_process %>%
  group_by(member_type) %>%
  filter(started_hour == 8 | 17) %>%
  summarise(mean(duration))
```

	member_type	mean(duration)
1	casual	30.23548
2	member	13.35251

The avenge duration that annual members ride to work is 13 min or calculate to roughly 3.5 km. in cycling distance.

Now, let's compare both groups overall riding duration and frequency to see if there is any difference in cycling behavior.

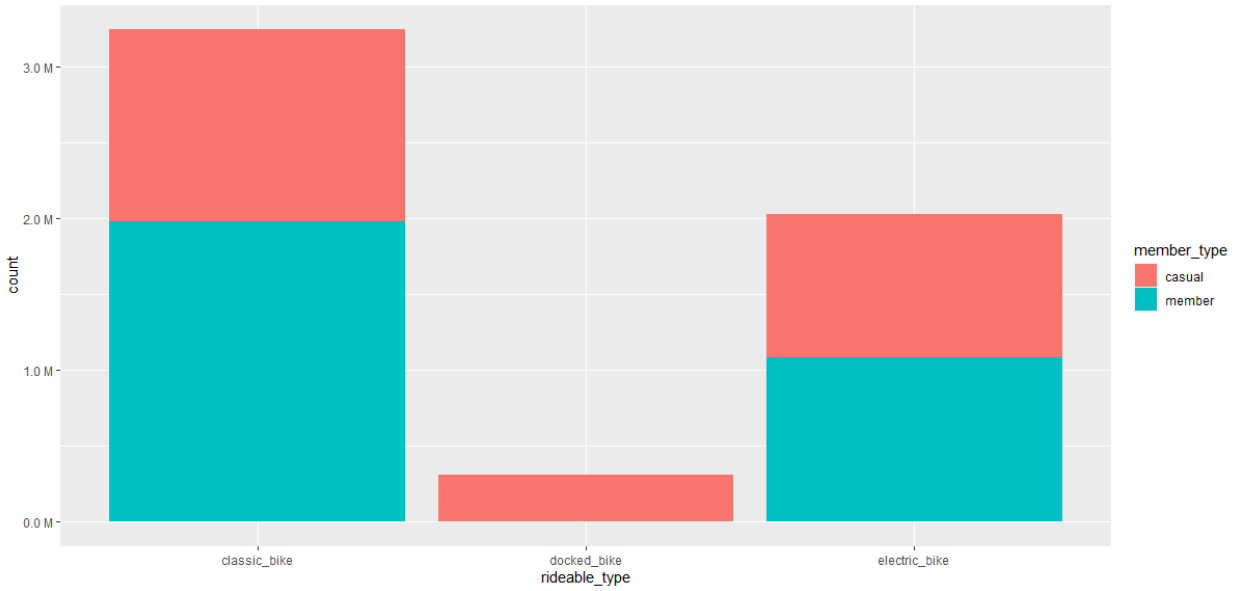
```
work_duration_summary <- trip_data_process %>%
  group_by(member_type) %>%
  summarise(count = n() )

ggplot(data = trip_data_process) + geom_point(mapping = aes(x = member_type,
y = duration))
```



The annual members cycle more but shorter duration, while causal riders cycle less often but usually cycle for longer.

Then I Create a bar graph to see which bike type is the most use for each member type.

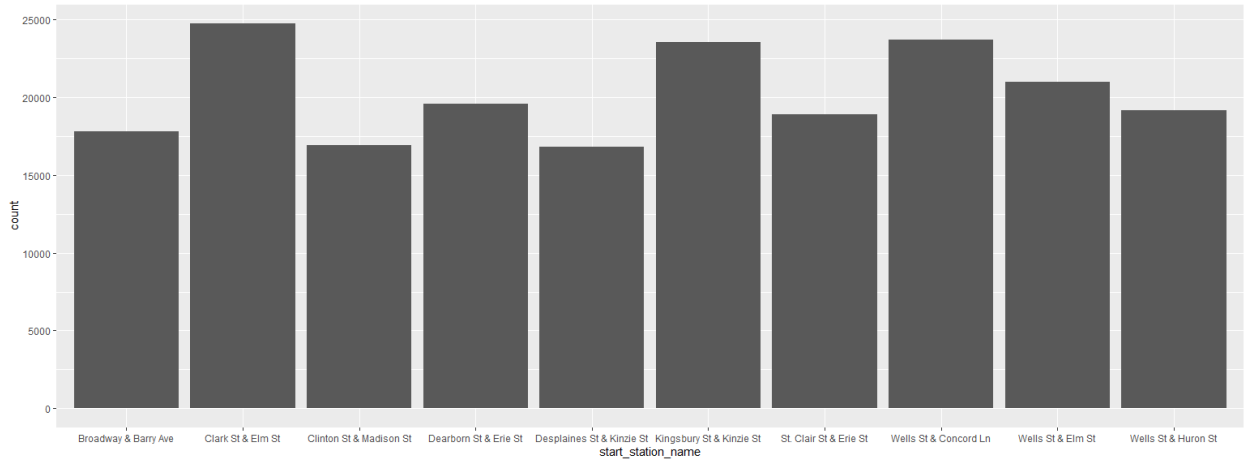
```
ggplot(data = trip_data_process) + geom_bar(mapping = aes(x = rideable_type, fill = member_type)) + scale_y_continuous(labels = unit_format(unit = "M", scale = 1e-6))
```



The most popular type of bike for both groups is the classic bike, and the annual members rarely use the docked bike.

finally, I create a bar graph to see which location is the most popular cycling location for the annual members.

```
top_location_summary <- trip_data_process %>%
  group_by(start_station_name) %>%
  filter(start_station_name != "")  %>%
  filter(member_type == "member")  %>%
  summarise(count = n()) %>%
  top_n(n = 10, wt = count) %>%
  arrange(desc(count))

  ggplot(data = top_location_summary) + geom_bar(stat='identity', mapping = a
es(x =    start_station_name, y = count))
```



Shere

Key insight to share

- Annual members usually use the bike to commute to work, while causal riders use the bike for much longer activities.

additional useful information

1. Both Annual members and causal riders ride the bike most at 5 pm.
2. The duration of most annual members' rides to work is 13 minutes or about 3.5km.
3. Annual members ride more often but shorter duration, causal riders ride less often but for longer.
4. Classical bike is the most popular type of bike for both user types.
5. Clark St & Elm St is the most popular start cycling location for the annual members.

See my PowerPoint here

Act

my recommended strategies to convert causal riders to Annual members are.

1. Promote more people around Clark St & Elm St and in the cities area to cycle to work
2. Decrease renting price on workdays to convert more causal members to cycling during the workday.
3. Promote by targeting office workers who have office syndrome to exercise more frequently.
4. Investigate further for any worries or obstructions that prevent causal riders from cycling to work.