Capstone_1

Amadi Chinevu

2/27/2022

Case Problem

This a solution to the capstone project on the Google Analytics Course on Coursera.

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations………..

The data can be accessed using the link below:

https://divvy-tripdata.s3.amazonaws.com/index.html

This analysis covers a period of 12 months, from February 2021 to January 2022. I downloaded the files to my pc and extracted them. I loaded the required R libraries, imported the files to R, and did a little cleaning by replacing empty cells with NA

```
library(scales)
library(kableExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:kableExtra':
##
##     group_rows

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(ggplot2)
library(lubridate)
```

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

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

```
library(tidyverse)
```

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

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x readr::col_factor()      masks scales::col_factor()
## x lubridate::date()        masks base::date()
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::group_rows()      masks kableExtra::group_rows()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(tidyr)
library(ggthemes)
```

Then importing the csv files

```
feb_21_data = read.csv("202102-divvy-tripdata.csv") #import csv
feb_21_data[feb_21_data == "" | feb_21_data == " "] <- NA #replace empty cells with NA
march_21_data = read.csv("202103-divvy-tripdata.csv")#import csv
march_21_data[march_21_data == "" | march_21_data == " "] <- NA  #replace empty cells with NA
april_21_data = read.csv("202104-divvy-tripdata.csv")#import csv
april_21_data[april_21_data == "" | april_21_data == " "] <- NA  #replace empty cells with NA
may_21_data = read.csv("202105-divvy-tripdata.csv") #import csv
may_21_data[may_21_data == "" | may_21_data == " "] <- NA  #replace empty cells with NA
june_21_data = read.csv("202106-divvy-tripdata.csv")#import csv
june_21_data[june_21_data == "" | june_21_data == " "] <- NA   #replace empty cells with NA
july_21_data = read.csv("202107-divvy-tripdata.csv")#import csv
july_21_data[july_21_data == "" | july_21_data == " "] <- NA   #replace empty cells with NA
august_21_data = read.csv("202108-divvy-tripdata.csv")#import csv
august_21_data[august_21_data == "" | august_21_data == " "] <- NA   #replace empty cells with NA
sept_21_data = read.csv("202109-divvy-tripdata.csv")#import csv
sept_21_data[sept_21_data == "" | sept_21_data == " "] <- NA   #replace empty cells with NA
oct_21_data = read.csv("202110-divvy-tripdata.csv")#import csv
oct_21_data[oct_21_data == "" | oct_21_data == " "] <- NA   #replace empty cells with NA
nov_21_data = read.csv("202111-divvy-tripdata.csv")#import csv
nov_21_data[nov_21_data == "" | nov_21_data == " "] <- NA   #replace empty cells with NA
dec_21_data = read.csv("202112-divvy-tripdata.csv")#import csv
dec_21_data[dec_21_data == "" | dec_21_data == " "] <- NA   #replace empty cells with NA
jan_22_data = read.csv("202201-divvy-tripdata.csv")#import csv
jan_22_data[jan_22_data == "" | jan_22_data == " "] <- NA   #replace empty cells with NA
```

Then combining the data sets into a single data frame object

```
all_datasets = rbind(feb_21_data,march_21_data,april_21_data,may_21_data,june_21_data
                     ,july_21_data,august_21_data,sept_21_data,oct_21_data,nov_21_data
                     ,dec_21_data,jan_22_data)
use_data = all_datasets
```

I took a glimpse to have a visual inspection of the data

```
head(use_data)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 89E7AA6C29227EFF  classic_bike 2021-02-12 16:14:56 2021-02-12 16:21:43
## 2 0FEFDE2603568365  classic_bike 2021-02-14 17:52:38 2021-02-14 18:12:09
## 3 E6159D746B2DBB91 electric_bike 2021-02-09 19:10:18 2021-02-09 19:19:10
## 4 B32D3199F1C2E75B  classic_bike 2021-02-02 17:49:41 2021-02-02 17:54:06
## 5 83E463F23575F4BF electric_bike 2021-02-23 15:07:23 2021-02-23 15:22:37
## 6 BDAA7E3494E8D545 electric_bike 2021-02-24 15:43:33 2021-02-24 15:49:05
##           start_station_name start_station_id           end_station_name
## 1   Glenwood Ave & Touhy Ave              525 Sheridan Rd & Columbia Ave
## 2   Glenwood Ave & Touhy Ave              525   Bosworth Ave & Howard St
## 3         Clark St & Lake St     KA1503000012     State St & Randolph St
## 4      Wood St & Chicago Ave              637    Honore St & Division St
## 5         State St & 33rd St            13216      Emerald Ave & 31st St
## 6 Fairbanks St & Superior St            18003      LaSalle Dr & Huron St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1            660  42.01270 -87.66606 42.00458 -87.66141        member
## 2          16806  42.01270 -87.66606 42.01954 -87.66956        casual
## 3   TA1305000029  41.88579 -87.63110 41.88487 -87.62750        member
## 4   TA1305000034  41.89563 -87.67207 41.90312 -87.67394        member
## 5   TA1309000055  41.83473 -87.62583 41.83816 -87.64512        member
## 6   KP1705001026  41.89581 -87.62025 41.89489 -87.63198        casual
```

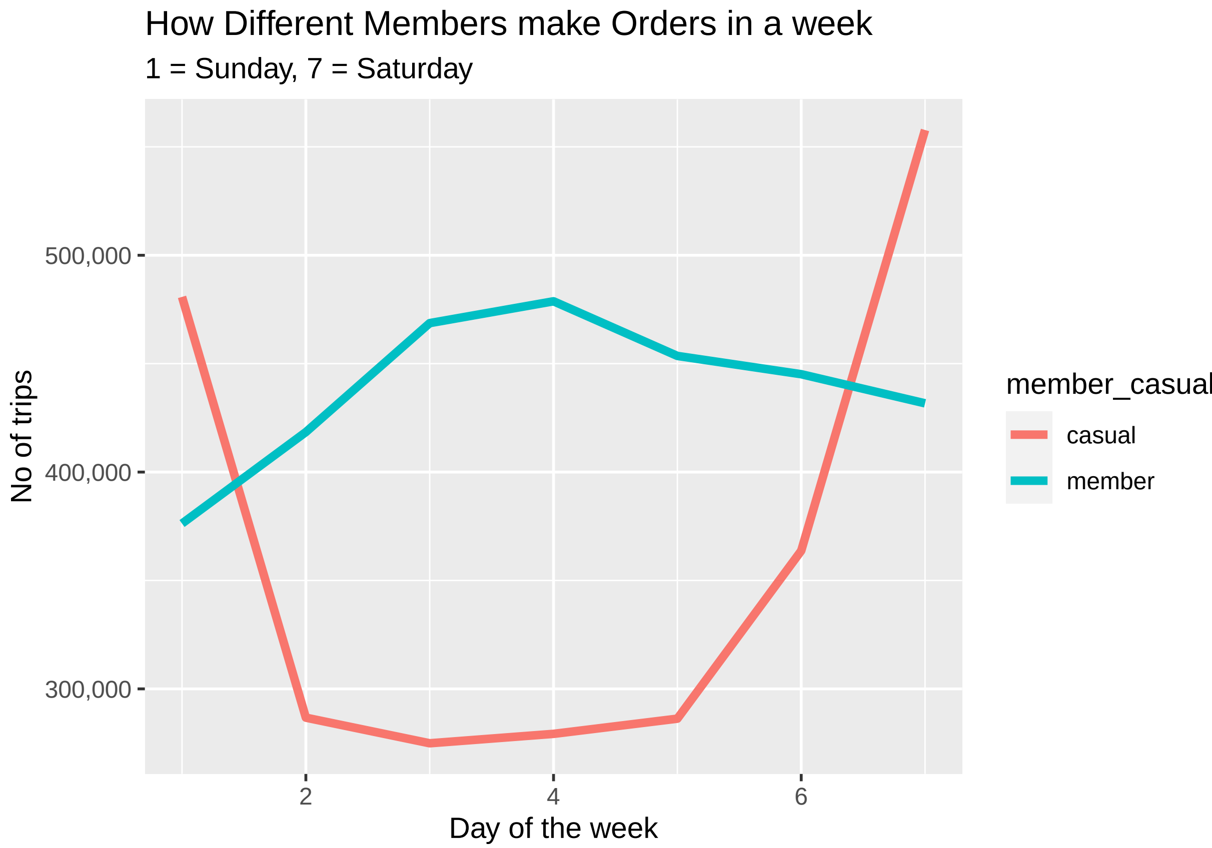
I took note that the columns that contain the dates appear as char. I tried to see if there are difference between the two membership classes in terms of seasonality

```
data_set_ = mutate(all_datasets, start_time = ymd_hms(started_at)) %>%
  mutate(end_time = ymd_hms(ended_at)) %>%
  mutate(start_month = round_date(start_time, "month"))
## I checked if there is seasonality among both categories
xy = data_set_ %>%
  group_by(start_month)%>%
  count(member_casual)
```

From the above, there are more of member rides except for the months around summer (July to August). This is showing that possible casual member order rides for leisure rather than for work or daily activities.

In order to establish this hypothesis, I had to do a further investigation to see the differences in the order behavior for the different days of the week

```
use_data = all_datasets
data_set_ = data_set_ %>%
  mutate(day_of_week = wday(start_time))
xy = data_set_ %>%
  group_by(day_of_week)%>%
  count(member_casual)
```

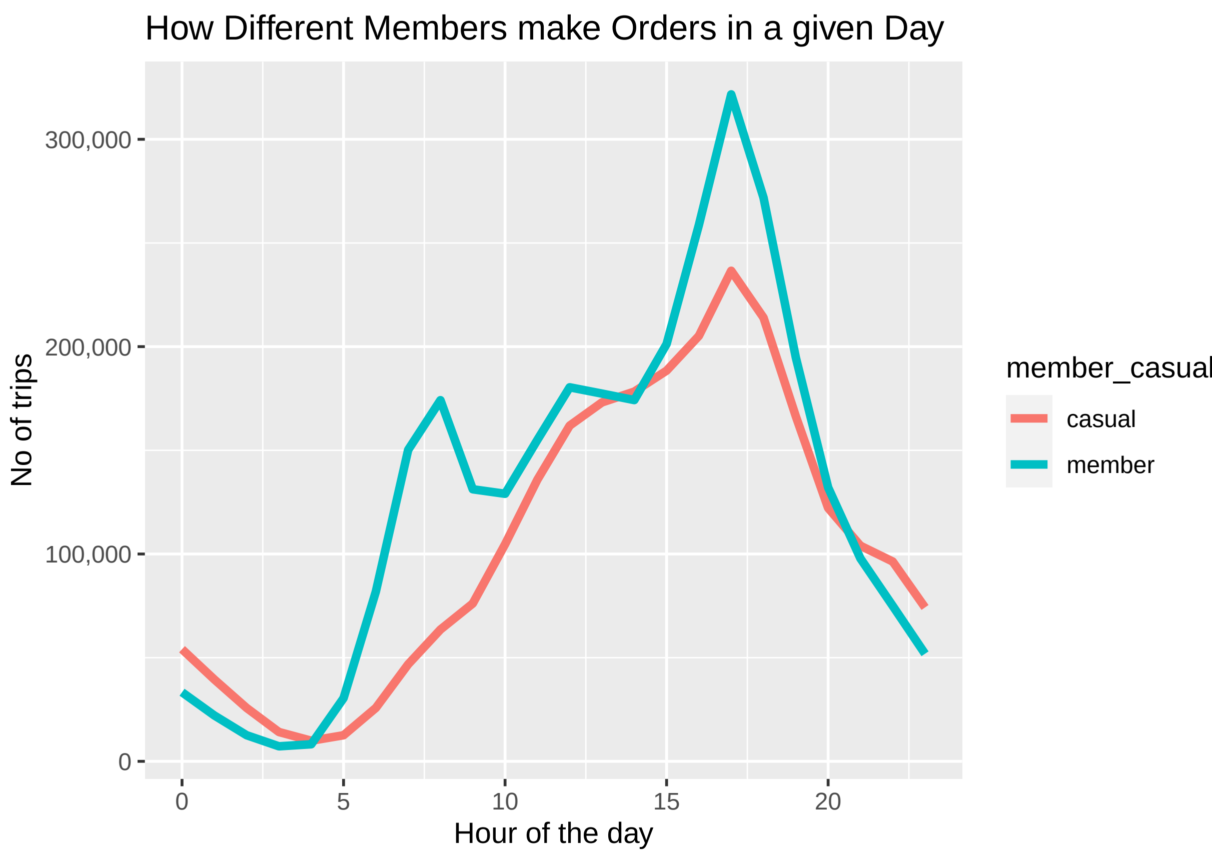


The number of rides by casual members tend to spike during the weekends but decline substantially during the week days. This is unlike the member where the number of trips remain relatively constant. This further validates that casual members make use of the bikes for leisure.

I tried to check if there are any differences in the time of the day when the two different membership make order for trips

```
use_data = all_datasets
data_set_ = data_set_ %>%
  mutate(hour_of = hour(start_time))
xy = data_set_ %>%
  group_by(hour_of)%>%
  count(member_casual)
```

member_casual	Average_min
casual	32.03430
member	13.60652



There is no noticeable difference between the two different membership.

I tried to see the average duration of rides between the two different types of membership

```
df = all_datasets %>%
  mutate(start_time = ymd_hms(started_at)) %>%
  mutate(end_time = ymd_hms(ended_at)) %>%
  mutate(ride_length = as.duration((end_time - start_time)))%>%
  mutate(ride_length = (ride_length/60)) %>%
  group_by(member_casual)%>%
  summarise(Average_min = mean(ride_length))
df %>%
  kbl() %>%
  kable_styling()
```

The casual member has average ride duration that is more than double that of the members.

Finally, I checked the busiest stations for both types of membership to see if there are any differences.

The top 10 busiest station for both members are as follow

start_station_name	member_casual	n
Streeter Dr & Grand Ave	casual	66395
Millennium Park	casual	33539
Michigan Ave & Oak St	casual	29758
Clark St & Elm St	member	24615
Kingsbury St & Kinzie St	member	23847
Wells St & Concord Ln	member	23691
Shedd Aquarium	casual	23285
Theater on the Lake	casual	21309
Wells St & Elm St	member	20982
Wells St & Concord Ln	casual	19886
Lake Shore Dr & Monroe St	casual	19421
Dearborn St & Erie St	member	19289
Wells St & Huron St	member	18970
St. Clair St & Erie St	member	18817
Broadway & Barry Ave	member	17757
Clinton St & Madison St	member	17120
Clark St & Lincoln Ave	casual	16982
Desplaines St & Kinzie St	member	16752
Wabash Ave & Grand Ave	member	16678
Wells St & Elm St	casual	16644

```
data_set_ = all_datasets
## I plotted the result
xy = data_set_ %>%
  filter(!is.na(start_station_name)) %>%
  #filter( member_casual == "casual")%>%
  group_by(start_station_name)%>%
  count(member_casual)%>%
  arrange(desc(n))%>%
  head(20)

xy %>%
  kbl() %>%
  kable_styling()
```

Conclusion & Recommendations

Casual members use the bikes more for leisure while members use it for work and business. The following strategies can help increase membership subscription from casual members.

1. A targeted advertisement on travelling and vacation sites
2. A targeted advertisement on the busiest stations for the casual members
3. Possibly a different membership class that would suit them could be created. This could be membership status that spans for half of the year rather than a whole year