Google Data Analytics Capstone Project : Case Study

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-sharing company in Chicago. Moreno (Director of marketing) believes the company's future success depends on maximizing the number of annual memberships. Our goal is to design marketing strategies aimed at converting casual riders into annual members. In order to do that, we need to understand how casual riders and annual members use Cyclistic bikes differently.

About the company

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Moreno believes that **maximizing the number of annual members will be key to future growth**. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.Moreno has set a clear goal:Design marketing strategies aimed at converting casual riders into annual members.

In this case study, we will follow the six steps Data Analysis Process which we learned in this course i.e. Ask, Prepare, Process, Analyze, Share, and Act.

1)Ask

"How can we convert casuals to members?" Here, the director of marketing believes the company's future success depends on maximizing the number of annual memberships. As a Junior Data Analyst my business task is to understand the behavior of Casual bike riders and annual Members, also provide insights that going to help the marketing team to launch a campaign to convert casual bike rider to annual members.

2)Prepare

Then the second phase Prepare which means to collect or use data relevant to the problem we are trying to solve. In this case, we will be using Cyclistic historical trip data click here. We have to download twelve csv files, one file represents one month of trip data.

Firstly, we would need to install & load the packages required for this process, which in this case is: "Tidyverse", "hydroTSM" & "Lubridate".

```
library(tidyverse)
library(lubridate)
library("hydroTSM")
```

I used (August 2020 -July2021). I imported the 12 csv files into 12 data frames then I merge the 12 data frames into 1-year data frame. After I removed the 12-month data frame to clear up space in the environment.(RAM management).

```
jan <- read.csv("january.csv")
feb <- read.csv("febrnuary.csv")
mar <- read.csv("march.csv")
april <- read.csv("202104-divvy-tripdata.csv")
may <- read.csv("202105-divvy-tripdata.csv")
june <- read.csv("202106-divvy-tripdata.csv")
july <- read.csv("202107-divvy-tripdata.csv")
aug <- read.csv("august.csv")
sep <- read.csv("september.csv")
oct <- read.csv("october.csv")
nov <- read.csv("november.csv")
dec <- read.csv("december.csv")
bike_ride <- rbind(jan,feb,mar,april,may,june,july,aug,sep,oct,nov,dec)
remove(jan,feb,mar,april,may,june,july,aug,sep,oct,nov,dec)
```

3) Process

The third phase is to process the data. Data processing is to find various inaccuracies, errors, inconsistencies in the data and get rid of them so that our business problem is not affected. In order to process the **46,49,054** observations, spreadsheets wouldn't be able to handle the sheer amount of data. In this case, we would be using RStudio instead.

During this phase,

1. I removed both rows & columns with Not Available Values (NA Values)
2. I removed duplicated rows
3. I removed unwanted columns for our analysis
4. I created (start hour of trip, month of trip, season of trip, weekday of trip and finally trip duration) columns for our analysis
5. I removed negative values from trip duration

```
row_before_cleaning <- nrow(bike_ride)
na.omit(bike_ride)
unique(bike_ride)
bike_ride <- bike_ride %>% select(-c(ride_id,start_station_id,end_station_id,start_lat,start_lng,end_lat,end_lng))
bike_ride <- bike_ride %>% rename(Type_of_Membership = member_casual)
bike_ride$start_hour <- hour(bike_ride$started_at)
bike_ride$end_hour <- hour(bike_ride$ended_at)
bike_ride$hour <- (bike_ride$end_hour - bike_ride$start_hour)
bike_ride$month <- month(bike_ride$started_at)
bike_ride$weekday <- weekdays(as.Date(bike_ride$started_at))
bike_ride$season <- time2season(as.Date(bike_ride$started_at), out.fmt = "seasons")
bike_ride$trip_duration <- difftime(bike_ride$ended_at, bike_ride$started_at,units = "mins")
bike_ride$trip_duration <- as.integer(bike_ride$trip_duration)
bike_ride <- bike_ride %>% filter(trip_duration>0) %>% drop_na()
row_after_cleaning <- nrow(bike_ride)
cleaned_rows <- row_before_cleaning - row_after_cleaning
print(cleaned_rows)
```

Before cleaning 4731081 rows | After cleaning 4649054 rows

4) Analyze

The fourth phase is to analyze data by organizing, sorting and filtering and transforming data. Here, I used *ggplot2* library in R studio to create various charts to understand the behavior of our bike riders.

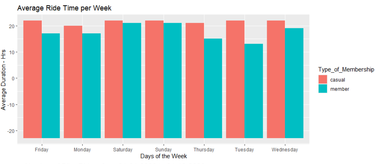
```
#Average Ride Time per week
bike_ride %>% ggplot(aes(x = weekday, y = hour,fill = Type_of_Membership)) + geom_bar(position = "dodge",stat = "identity") + labs(title ="Average Ride Time per Week",x = "Days of the Week",y = "Average Duration - Hrs")

#Number of Bike Riders from 01-01-2020 to 31-12-2020
bike_ride %>% ggplot() + geom_bar(mapping = aes(x = Type_of_Membership,fill = Type_of_Membership))+ labs(title = "Number of Bike Riders from 01-01-2020 to 31-12-2020",x = "Riders",y = "Total No. of Riders")

#Which Bike Works the Most?
bike_ride %>% ggplot() + geom_bar(mapping = aes(x = Type_of_Membership,fill = Type_of_Membership),position = "dodge") + facet_wrap(~rideable_type) + labs(title ="Which Bike Works the Most?",x = "Type of Bike",y = "Number of Rentals")

#Rides per Month
bike_ride %>% ggplot() + geom_bar(mapping = aes(x = month,fill = Type_of_Membership),position = "dodge") + scale_x_continuous(breaks = seq_along(month.name),labels = month.name) + labs(title = "Rides per Month",x = "Month",y = "Total Number of Riders")

#Trip Duration (Mins)
bike_ride %>% ggplot() + geom_histogram(mapping = aes(x = trip_duration),filter(bike_ride,bike_ride$trip_duration <120)) + facet_wrap(~Type_of_Membership) + labs(x = "Trip Duration (Mins)", y = "Total No. of Riders")

#Riders according to Seasons
bike_ride %>% ggplot() + geom_bar(mapping = aes(x=season,fill = Type_of_Membership),position = "dodge") + labs(title = "Riders according to Seasons",x = "Seasons", y = "Total Number of Riders")

#Number of bike ride week wise
bike_ride %>% ggplot() + geom_bar(mapping = aes(x = start_hour,fill = Type_of_Membership)) + labs(title = "Number of bike ride week wise",x = "Time - hrs", y = "Total Number of Riders" ) + facet_wrap(~weekday)

#Rides per Day of Week
bike_ride %>% ggplot() + geom_bar(mapping = aes(x = weekday,fill = Type_of_Membership),position = "dodge") + labs(title = "Rides per Day of Week",x = "Day of Week",y = "Total Number of Riders")
```

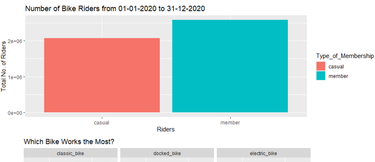
In order to answer our first business question, it would be beneficial to plot a few of our observations revolving around:

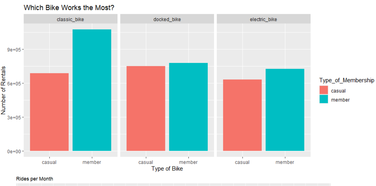
1. How do casual and members use their bikes differently throughout the week
2. Peak hours of bike usage between casual and annual members
3. Bike usage throughout the year
4. The average trip duration between casual and annual members
5. Most popular bike among casual and annual members
6. Which season casual and annual members love the most.

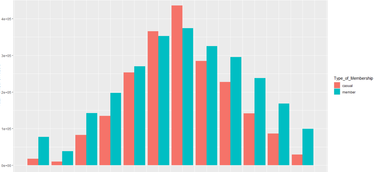
5) Share

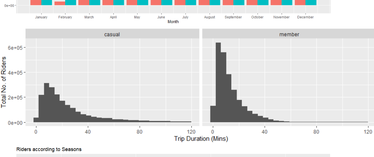
The fifth phase is to Share data. We need to create visualizations to share your findings. We can create a Dashboard with Power BI or Tableau. I decided to skip this step for the sake of time as I want to report using R Markdown.

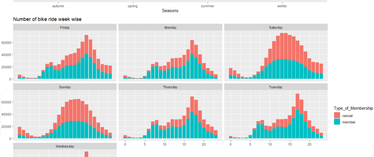












6) Act

The sixth phase of data analysis is to use every insight we learnt to solve the problem. We have to provide our stakeholders information that can help them to decide.

1. We can clearly see a peak in casual riders on a few occasions. On the weekends as well as in the months of June, July & August, we should prioritize marketing.
2. We should advertise promotions on the weekends and in the months of June, July & August whereby current casual members would be able to upgrade to annual members at a discount.
3. Another way to promote bike is to reduce greenhouse gas emission if we use bike instead of car which thereby increases the healthy way of life.