CYCLISTIC-CASE STUDY

Google Data Analytics Capstone Project

Documentation

Introduction

This documentation is of a capstone project of Google Data Analytics specialization on a company named Cyclistic. Cyclistic is a company of bike sharing in Chicago. Their strategy is based on the flexibility of pricing plans they offer as well as the diversity of bikes, customers can choose the plan and bike that suits them.

- Cyclistic has 5824 bikes and 692 stations.
- They offer reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities.
- There are three types of pricing plans: single-ride passes, full-day passes, and annual memberships.
- Customers who purchase single-ride or full-day passes are referred to as casual riders.
- Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders, and now the task is to determine whether converting casual riders to members by using social media can increase its profit. That's why I as a junior analyst will use data to help my team to support or reject that decision.

I used the 6 steps recommended by the Google Data Analytics course: ask, prepare, process, analyze, share and act.

1. ASK

The questions were already mentioned in the case study

- How do annual members and casual riders use Cyclistic bikes differently
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

My imaginary marketing manager has assigned me the first question to answer: **How do annual members and casual riders use Cyclistic bikes differently?** So, the business task was to perform a descriptive analysis to identify the differences between casuals and members. By finding the areas of distinction the marketing team can know what to target if it's doable, any relevant opportunities, and possible obstacles.

I also defined my primary and secondary stakeholders and their expectations to focus on my goal during the entire analysis.

Cyclistic Executives (Primary Stakeholder) :
Their expectations are a clear definition of the problem, precise data visualizations, and recommendations.

Cyclistic Manager (Secondary Stakeholder) :
The cyclistic Manager named Lily Moreno has the following expectations:

- Validate Decisions.
- Documentation of the data as well as a description of the data source and quality of the data.
- A log containing modifications of data.
- Answer to the problem: How annual members and casual riders differ.

Cyclistic Marketing Analytics Team (Secondary Stakeholder) : They expected to validate decisions and discuss any issues encountered during the analysis.

2. PREPARE

For the **prepare** phase, we are supposed to collect the data needed for the analysis and they gave us the Cyclistic's historical trip data for the previous 12 months of the Cyclistic trip, you can find the link here.

The datasets are given as monthly based trip data in a .zip file. I downloaded the trips from **July 2021 to June 2022** in .zip files and extracted them into a folder that I named Case Study 1 as this was my first case study. Each file was a .csv file, so we were ready for the next phase which is **process**.

3. PROCESS

I used R programming for processing my data because it was too big to be processed in a spreadsheet or excel. The original dataset dimension when combined was 5900385 x 13.

I imported the libraries needed for working on data like data cleaning, data validation, data integrity, etc. I used the following libraries for working:

- tidyverse - For data import and reading the data.
- janitor - Cleaning and Exploring Data.
- lubridate - For dates and times.
- skimr - For summary and statistics.

```
# 1. Download the libraries needed

library(tidyverse)
library(janitor)
library(lubridate)
library(skimr)

# 2. Remove any data from the environment

rm(list=ls())
```

After importing the libraries, I loaded each of the twelve datasets by specifying its location and using the ***read.csv*** method, I named them “ds1” to “ds12” for “data frame” and inspected the column names of each to have a first look at the kind of data I have.

```
# 3. Reading data sets

ds1 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202107-divvy-tripdata.csv")
colnames(ds1)
ds2 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202108-divvy-tripdata.csv")
colnames(ds2)
ds3 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202109-divvy-tripdata.csv")
colnames(ds3)
ds4 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202110-divvy-tripdata.csv")
colnames(ds4)
ds5 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202111-divvy-tripdata.csv")
colnames(ds5)
ds6 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202112-divvy-tripdata.csv")
colnames(ds6)
```

```
ds7 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202201-divvy-tripdata.csv")
colnames(ds7)
ds8 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202202-divvy-tripdata.csv")
colnames(ds8)
ds9 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202203-divvy-tripdata.csv")
colnames(ds9)
ds10 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202204-divvy-tripdata.csv")
colnames(ds10)
ds11 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202205-divvy-tripdata.csv")
colnames(ds11)
ds12 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202206-divvy-tripdata.csv")
colnames(ds12)
```

Each **colnames** function execution gave me a result similar to this one:

```
> # 3. Reading data sets
>
> ds1 <- read.csv("C:/Users/Hp/Desktop/Case Study 1/CSV Files/202107-divvy-tripdata.csv")
> colnames(ds1)
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"           "start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"          "start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
>
```

Then I examined if there was any column type mismatch between the datasets, and R has a cool function for that: ***compare_df_cols.*** If there is a difference between data types of two same elements of different data frames then we get to know from this function.

```
# 4. Making sure that the the columns have the same type

compare_df_cols(ds1,ds2,ds3,ds4,ds5,ds6,ds7,ds8,ds9,ds10,ds11,ds12,return = "mismatch")
```

In the next step, I combined all the data frames into one dataframe and removed empty rows and columns if any.

```
# 5. Combining Data sets and removing empty rows and columns if any

bike_rides_global <- rbind(ds1,ds2,ds3,ds4, ds5,
ds6,ds7,ds8,ds9,ds10,ds11,ds12)
bike_rides_global <- janitor::remove_empty(bike_rides_global,which =
c("cols"))
bike_rides_global <- janitor::remove_empty(bike_rides_global,which =
c("rows"))
dim(bike_rides_global)
```

```
> # 5. Combining Data sets and removing empty rows and columns if any
>
> bike_rides_global <- rbind(ds1,ds2,ds3,ds4, ds5, ds6,ds7,ds8,ds9,ds10,ds11,ds12)
> dim(bike_rides_global)
[1] 5900385      13
> bike_rides_global <- janitor::remove_empty(bike_rides_global,which = c("cols"))
> bike_rides_global <- janitor::remove_empty(bike_rides_global,which = c("rows"))
> dim(bike_rides_global)
[1] 5900385      13
>
```

Since the number of rows did not change, it means that there was no empty row or column.

I also wanted to have a general idea about the dataset so I used the function **skim_without_charts**. This function shows the whole summary of the data.

```
# 6. Summary of the data frame

skim_without_charts(bike_rides_global)
```

When looking closely, I noticed that:

- Missing data: end_lat, end_lng are 99.9% complete.
- Inconsistency: 1294 start_station_name and 1316 end_station_name whereas we should have at max 692 station names. Also, the start and end station ids are 1184 and 1198 which is 71% more than what should be included. Some start trip dates/times are greater than the end trip date/time.

```
> # 6. Summary of the data frame
>
> skim_without_charts(bike_rides_global)
── Data Summary ────────────────────────
                           Values
Name                       bike_rides_global
Number of rows             5900385
Number of columns          13
_________________________
Column type frequency:
  character                9
  numeric                  4
_________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable      n_missing complete_rate min max  empty n_unique whitespace
1 ride_id                    0             1   8  16      0  5900378          0
2 rideable_type              0             1  11  13      0        3          0
3 started_at                 0             1  16  19      0  1600181          0
4 ended_at                   0             1  16  19      0  1603267          0
5 start_station_name         0             1   0  64 836018     1294          0
6 start_station_id           0             1   0  44 836015     1184          0
7 end_station_name           0             1   0  64 892103     1316          0
8 end_station_id             0             1   0  44 892103     1198          0
9 member_casual              0             1   6   6      0        2          0

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean     sd    p0   p25   p50   p75  p100
1 start_lat             0         1      41.9 0.0472  41.6  41.9  41.9  41.9  45.6
2 start_lng             0         1     -87.6 0.0308 -87.8 -87.7 -87.6 -87.6 -73.8
3 end_lat            5374         0.999  41.9 0.0473  41.4  41.9  41.9  41.9  42.2
4 end_lng            5374         0.999 -87.6 0.0304 -89.0 -87.7 -87.6 -87.6 -87.5
> 
```

For the missing data, since the dataset is huge, I decided to ignore the rows and remove them, however for the station names and ids inconsistency, since I set to myself a week of work, I did not find in that limited time a solution for it. If you have any recommendations, I would be happy to read your comments!

After that, I converted the fields *started_at* and *ended_at* into dates, and created new fields that would be helpful later on :

- start_hour, end_hour: the hours on which the trip started and ended.
- start_date, end_date: the dates without time on which the trip started and ended.
- ride_length_hours, ride_length_mins: the trip duration in hours, and in minutes.
- day_of_week: the day in which the trip started in numbers where Sunday is 1 and Saturday is 7.
- day_of_week_letter: the day in which the trip started in letters.

```
# 7. Processing datetime

bike_rides_global$started_at <-
lubridate::ymd_hms(bike_rides_global$started_at)
bike_rides_global$ended_at <-
lubridate::ymd_hms(bike_rides_global$ended_at)

# 8. Creating start and end hour fields

bike_rides_global$start_hour <-
lubridate::hour(bike_rides_global$started_at)
bike_rides_global$end_hour <- lubridate::hour(bike_rides_global$ended_at)

# 9. Creating ride_length field

bike_rides_global$ride_length_hours <-
difftime(bike_rides_global$ended_at,bike_rides_global$started_at,units="ho
urs")
bike_rides_global$ride_length_mins <-
difftime(bike_rides_global$ended_at,bike_rides_global$started_at,units="mi
ns")

# 10. Creating day_of_the_week fields

bike_rides_global$day_of_week_letter <-
lubridate::wday(bike_rides_global$started_at,abbr = TRUE,label = TRUE)
bike_rides_global$day_of_week_number <-
lubridate::wday(bike_rides_global$started_at)

# 11. summary of data

skim_without_charts(bike_rides_global)
```

It's important to make another summary after adding new fields, to be sure there are no surprises, and indeed thanks to that I found that some ride lengths were negative, which means that the date time in the started_at is bigger than the date time in ended at, so I decided to remove them. But first I deleted NA values, duplicated rows, then those negative ride lengths.

```
# 12. Removing Na

bike_rides_global_no_na  <- drop_na(bike_rides_global)
rm(bike_rides_global)

# 13. Removing Duplicates

bike_rides_global_no_na <- distinct(bike_rides_global_no_na)

# 14. Removing negative ride length and quality check rows

bike_rides_global_no_na_length_correct <- bike_rides_global_no_na %>%
filter(ride_length_mins>0)
rm(bike_rides_global_no_na)
rm(ds1,ds2,ds3,ds4,ds5,ds6,ds7,ds8,ds9,ds10,ds11,ds12)

## summary of data

skim_without_charts(bike_rides_global_no_na_length_correct)
```

After each big step, I always rerun the skim_without_charts function

After cleaning

- The original dataset dimension was (5900385 x 13) while the new dataset dimension is (5852016 x 19).
- The completion rate of all columns is 100%.
- One issue is that there are still 1294 station names and 1184 station ids.
-

Finally, I saved that data frame in a .csv file named **"divvy_dataset_cleaned.csv"**:

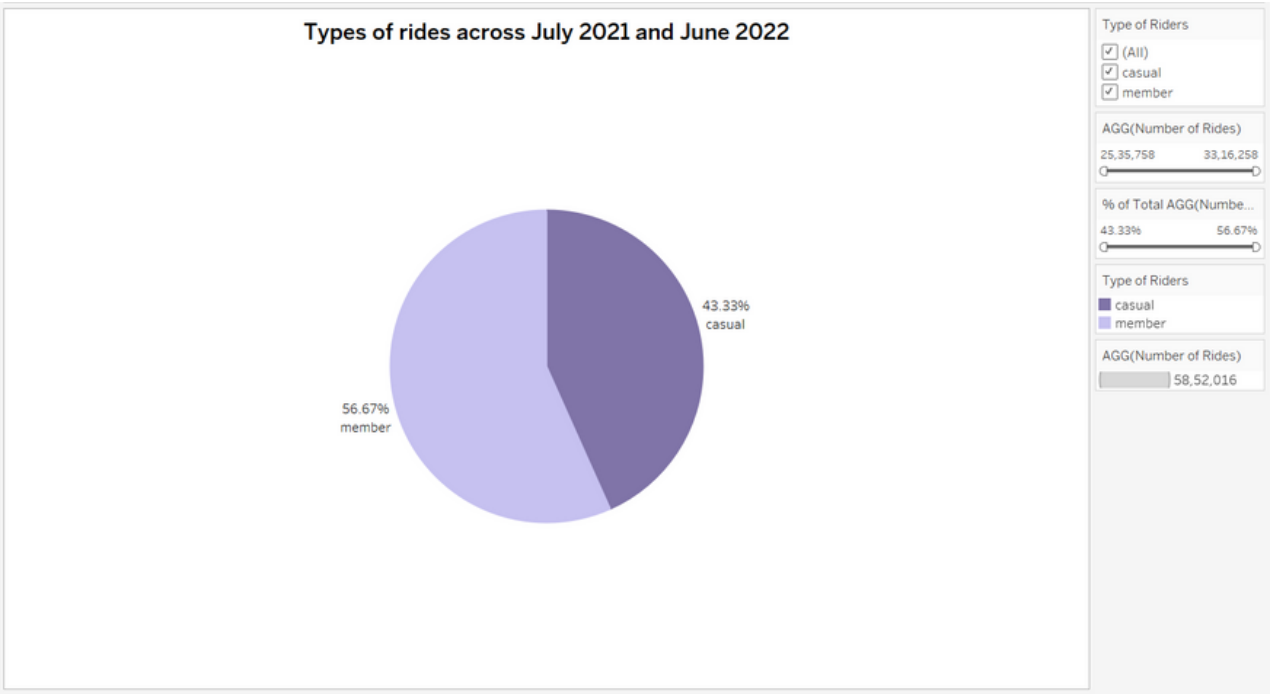
```
## Saved the final data frame into csv file by using below command
  write_csv(bike_rides_global_no_na_length_correct,"divvy_dataset_cleaned.c
sv")
```

4. ANALYZE

I decided to use tableau to visualize and analyze my data, it's simple and doesn't require code. We can also create similar graphs using R but I was more interested to try tableau this time. In order to be more organized, I prepared my plan first in a spreadsheet with the description of the things I wanted to explore while looking at my data so that things go faster and smoother.

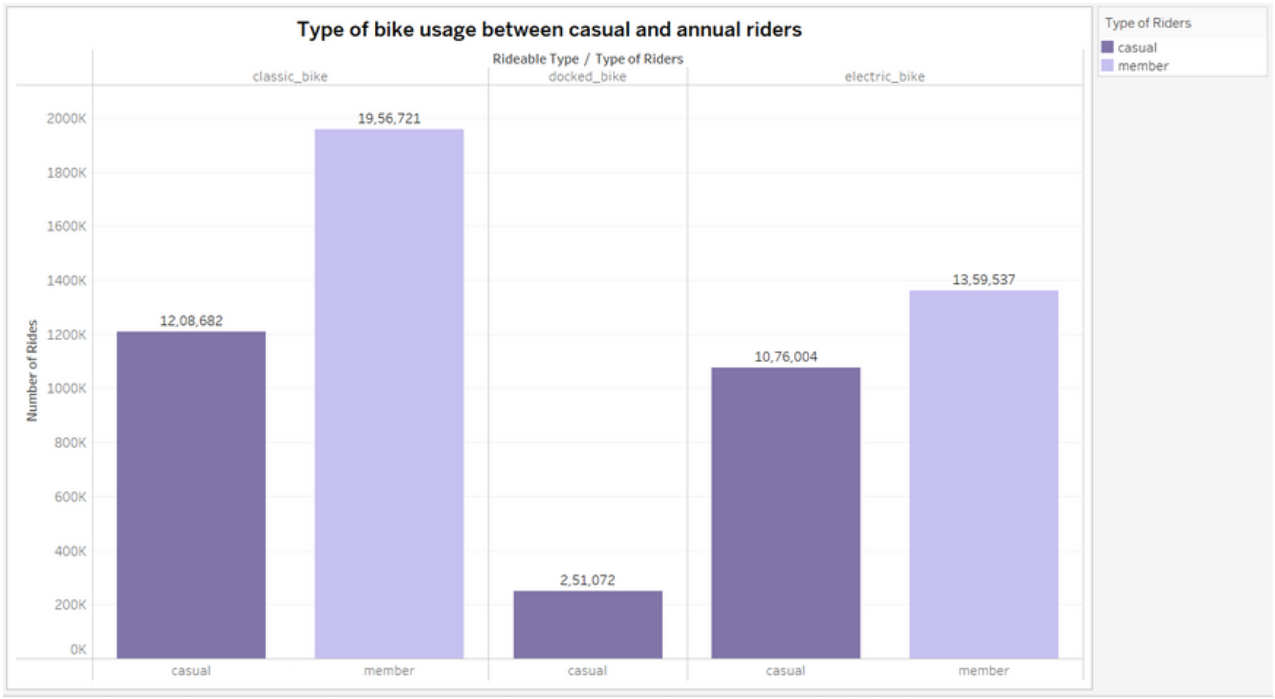
Following are the things that I explored and analyzed:

Type of rides distribution



This pie shows us the percentage of casual and member trips during **July 2021 and June 2022** in Chicago. You can see that the number of member trips is slightly bigger than casual trips which means that member riders do ride more than casual riders.

Bike usage



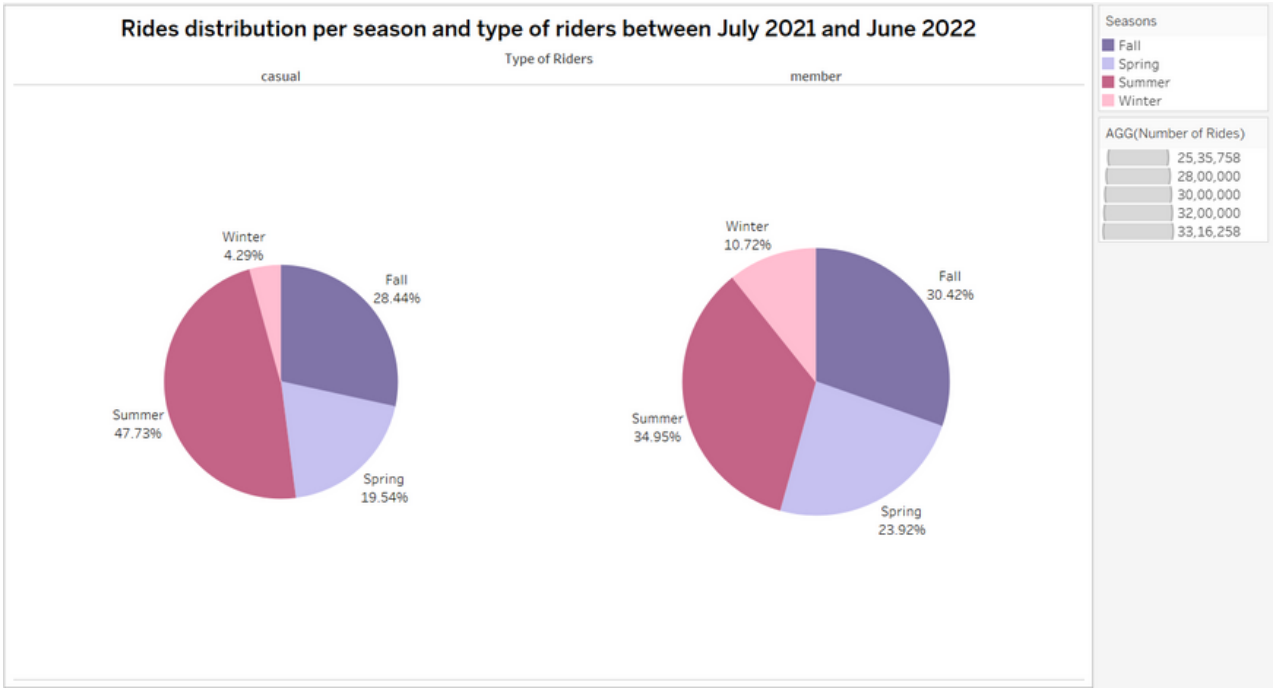
Among the different available bikes which Cyclistic offers, the dataset included these three types:

- Docked bikes
- Classic bikes
- Electric bikes

This bar graph represents the number of rides per bike type for casual and member riders. It is clear that classic bikes are the most popular followed by electric bikes then docked bikes and we can see those member riders didn't even prefer to ride docked bikes. But, there is no evidence or reason behind the preference of riders in terms of the type of bike they use.

Note that during analysis you want to explore any possible areas, you might find something, and you might not, it is worth removing any aspect, so even if we didn’t find a difference yet at least it’s information worth knowing.

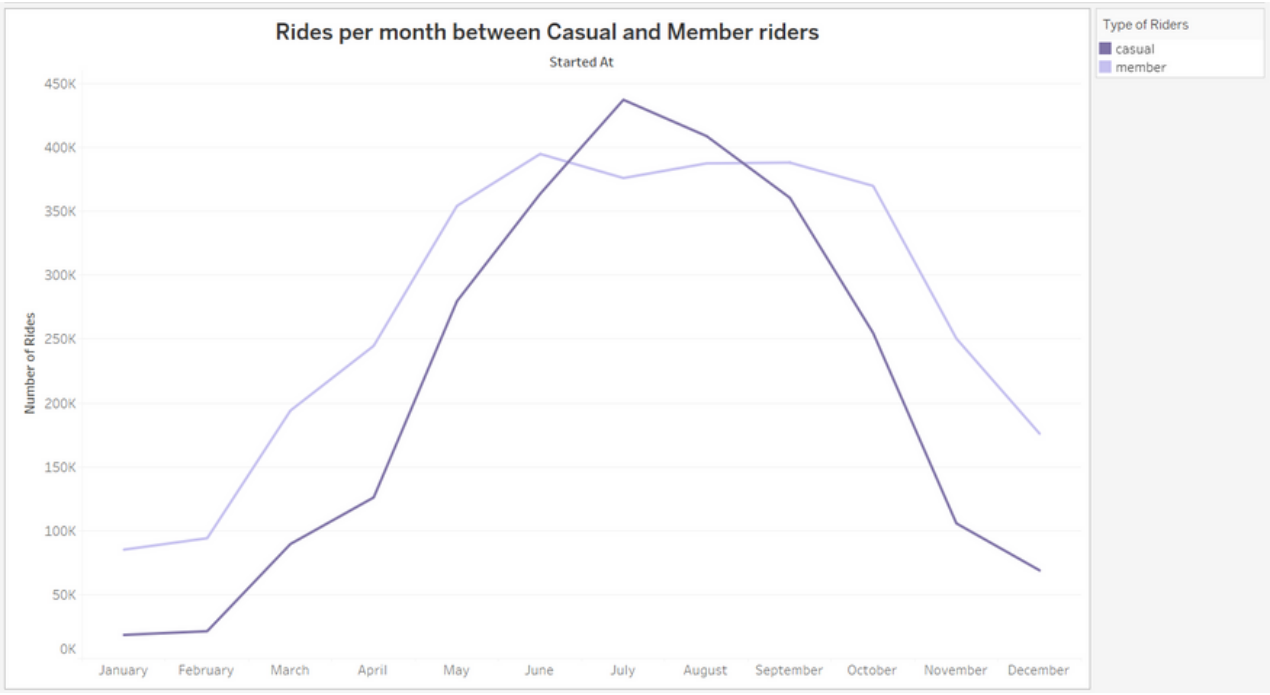
Season riding tendencies



So, I looked at how seasons might affect the frequency of riding of each group and I learned that both groups prefer summer and ride the most during that season, also during winter members ride 2 times more than casuals.

Rides per month

I actually wanted to look more into it, and this time I wanted to see the number of trips during each month:



From this, I observed that July is the month in which casual riders ride most and June is the month for member riders peak riding. We can observe that the peak starts from January and rise till June-July month then starts dropping til December.

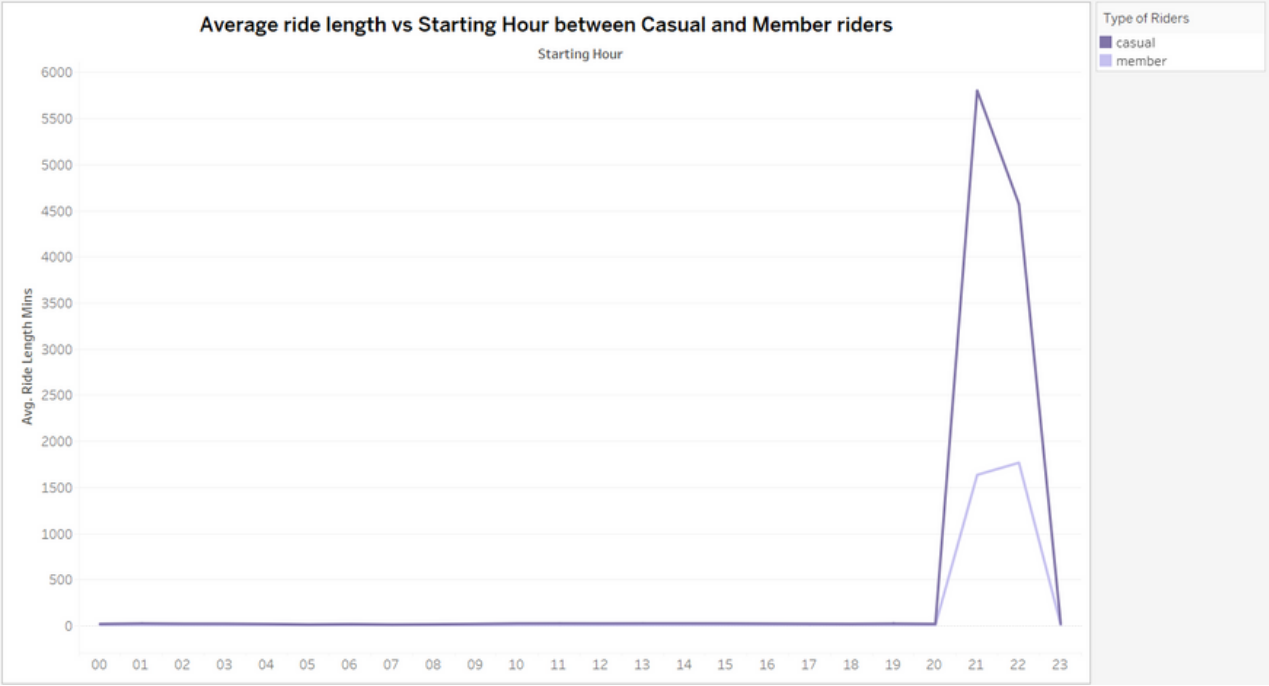
Favourite ride hours

I saw that seasons (winter/rain) affect the frequency of rides, after that I wanted to know if there is a particular hour where they ride the most, furthermore, I divided the trips per workdays and weekends.

From this graph, it seems that people ride mostly at 9-10 pm and people ride more on workdays as compared to weekends.

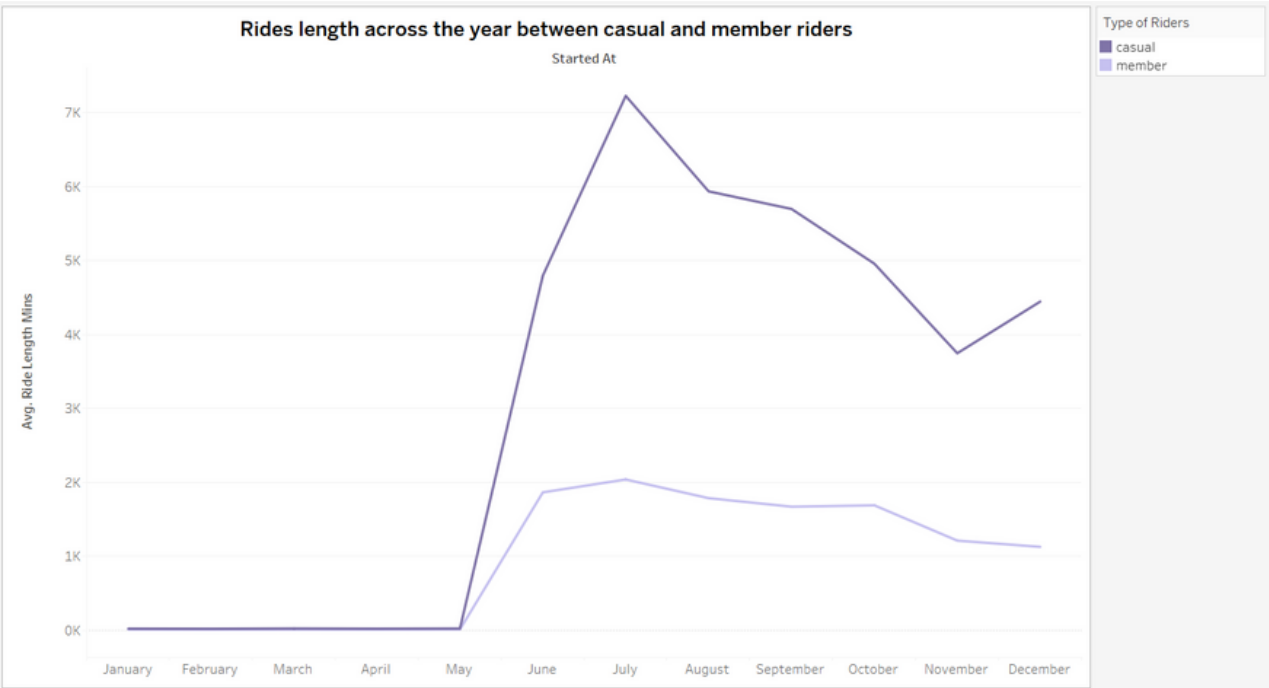
Member riders also ride slightly more at 8-9 am and 5 pm as compared to casual riders. It can meant as member riders may use their bikes for going to work while casual riders use them for leisure.

Ride lengths per hour



This graph also concludes that people ride mostly between 9-10 pm and also cover more distance during these hours. Also, member rides more than casual riders during these hours.

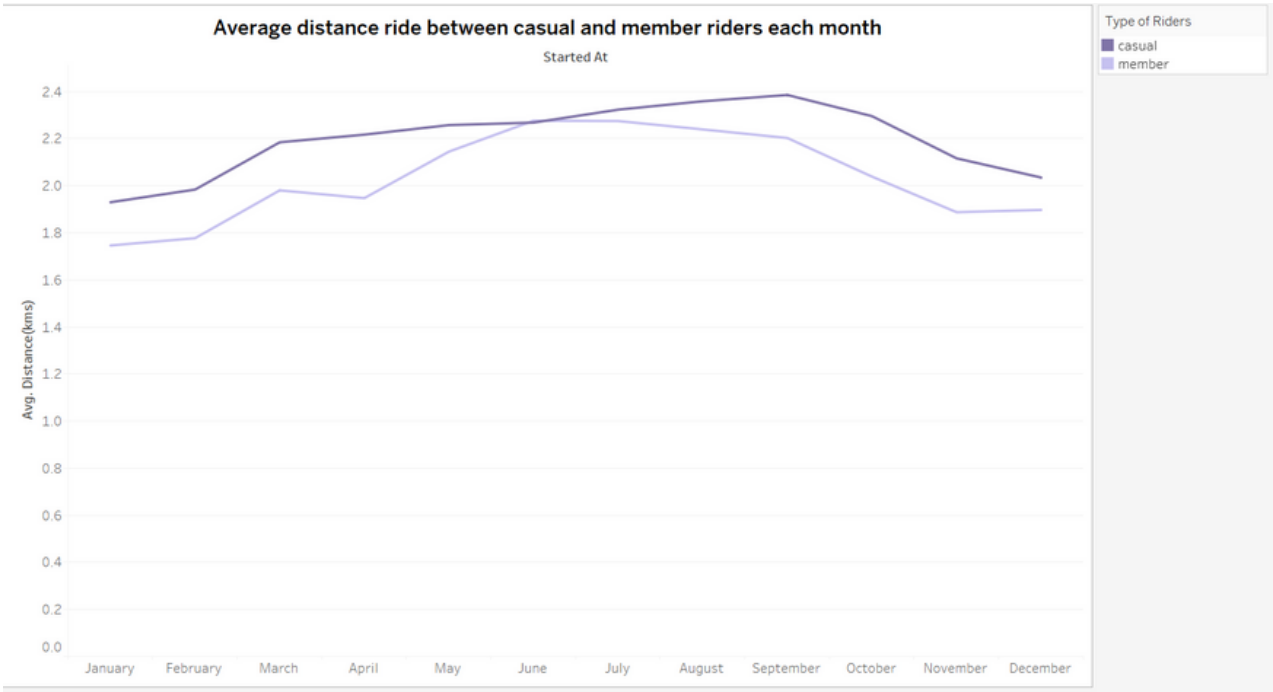
Ride length per year



This graph shows that the average time increases suddenly from May till July and then starts decreasing after July till December.

The same is true if we checked per hour or day or week. But why is that? We can't tell the reason with current data.

Rides average distance across the year



This graph shows that the average of member riders is around 2 km while 1.8 km for casual riders. This graph also tells that the average increases from January till September and then drops for casual members while the average of member riders increases from January till June and then decreases. This graph again didn't tell the reason why is this happening. Why is the average of casual more than member riders?

5. SHARE

To share my findings, I created a presentation using Canva. You can find it at this link.
I also provided a dashboard following this link.
I also shred my presentation link.

6. ACT

Key findings summary

Our data is good enough to make some assumptions but we can't jump to any conclusion, there is a high chance that the story is bigger than that, hence we cannot approve nor reject the decision to convert casuals to members. In my opinion, until we fully grasp our users, it's more dangerous than safe to jump to any conclusion.

Next steps

So, what are our options? If cyclistic can afford to be more careful, it is better to conduct further analysis so we can zoom in. Conducting further exploration is the safest option that can lead to a good marketing strategy, but it requires more time and more resources. If not, we still can use our key findings to start a marketing strategy while leveraging the risks at stake. Proceeding using the current key findings is risky and unpredictable but doesn't require time or resources.

1. Further Exploration

If we went with conducting further exploration, which is most probably the safest course of action, the analytical team will have to do a diagnostic analysis (Redefine the problem) to find the root cause of what makes each group different, because we know that there are differences between them but we don't know why they exist. The data previously collected is still useful however we need more data that concerns:

- Financial information of users
- Qualitative data about user preferences and reasons

- More details about each trip (eg: trips per user)
- Other demographic data such as their age, sex, and clinical conditions because it might be that some of them are on diet to lose weight and some are athletic. Then the analytical team would have to process to the analysis phase and eventually share the new discoveries in order to take a decision.

For instance, let's imagine this scenario where Yash is one of our clients. His annual income tells us that he is comfortable with having a membership, he probably rides for losing body fat and gain muscles, however, he has a daily or single pass. We can target this type of rider by promoting the benefits of consistent riding and how consistency is important.

But what if most of the casual riders are people who cannot afford the annual membership? Suppose for example that Joshua is also one of the Cyclistic clients, he really wants to lose weight but his income is an obstacle, he can only purchase single or daily passes and rides twice a week for example. In this case, no matter how good the strategy of Cyclistic is, it is impossible that their objective will succeed.

2. Act based on current findings

Now, if the company can't afford to reiterate the process for collecting more information for any reason, we still have some options:

- We saw that summer trips make almost half of the entire year, so we can for example make a special plan exclusively for summer, the price won't be as high as a yearly plan but it can benefit more than daily or single pass, for the rest of the year, we can also offer weekend pass.
- Another alternative is to promote the benefits of riding for people cyclists like Yash and offer some gifts to encourage them to subscribe to an annual membership.

However, we should be careful about this, note that there are some risks if we are not careful enough about converting casuals to members. Imagine a scenario where we have 80% of cyclists are members and they all ride frequently, it may be possible that Yash took the last bike near the station then a member who uses it for work and expects to find a bike in that exact station won't find any, so Cyclistic might deal with Bikes shortage and angry clients, we might lose their trust if we don't make sure that Bikes will suffice, furthermore, suppose now a situation where a station doesn't have anymore dock, with the unlimited number of docks, there is no guarantee that this scenario won't happen.

Conclusion

- We discovered some differences between casuals and members but were unable to explain why they are different.
- More data collection and analysis are required to solve our problem and identify marketing opportunities.
- The best next phase is to conduct a profound analysis to find the root cause for the group difference. You can still proceed to the next phase while taking the utmost precaution for any possible issue.

I hope I was clear enough, If you have any question, feel free to ask.

Contact Me:

- Linkedin
- GitHub
- Gmail - pranshu2105j@gmail.com

BY - PRANSHU JAIN