Cyclistic-Analysis

H. Mahmoud

Background

Cyclistic, a bike sharing company in Chicago, launched a successful bike-share offering in 2016. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its 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*.

Business Task

The **business task** of the Data Analysis team is **to analyze the historical bike trip data to understand how annual members and casual riders differ and identify possible trends and habits that would influence buying annual memberships.**

Questions to explore:

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

Data Source

As I mentioned previously, the data source for this task is the monthly historical trip data for the year 2022. Each dataset is downloadable from https://divvy-tripdata.s3.amazonaws.com/index.html as a zipped folder containing the (.csv) 'comma-separated values' file. The data is made available to the public under this license.

After downloading the files, I went ahead and loaded the data from the csv file to the work space as follows:

```
tripdata <- read_csv("202201-divvy-tripdata.csv")
```

Limitations:

- The data does not contain any customer records. This makes it impossible to explore the riding frequency of different types of riders and analyze trends. However, it is understandable that this data would be omitted for privacy concerns.

- The data does not include the ride distance. This could be a possible limitation because Longitude and Latitude values do not much in terms of riding behavior.

Cleaning 'Cyclistic' Dataset using R

I will be applying this process to data sets from 12 months starting from Jan-2022 till Dec-2022. The data is difficult to clean in Excel & Google Sheets, due to it being very large in size and will cause your spreadsheet to crash at some point. This is why I decided to do this process in Rstudio instead.

After downloading the files, I went ahead and loaded the data from the csv file to the work space as follows:

```
    tripdata <- read_csv("202201-divvy-tripdata.csv")
```

Because, the datasets are so large, I wasn't able to combine them all into one table. Instead, I tried to group them into quarters. It worked for Q1, Q2 & Q4, but crashed when I tried to group data for months (7, 8 & 9). I used the rbind() to join these tables, and since I wasn't able to join all of them, I used command prompt on my windows to join the multiple .csv files into one after the cleaning process.

```
  trips_Q1 <- rbind(tripdata1, tripdata2, tripdata3)
```

Cleaning Process

For each data set I apply this cleaning process to, I make sure to take note of the number or rows before and after.

1. Started by installing and loading the right packages:

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


    library(tidyverse)
    library(janitor)
    library(lubridate)
    library(hms)
    library(dplyr)
```

2. Created a new variable with the name (trip_duration).

```
    tripdata <- tripdata %>%
      mutate(trip_duration = as_hms(difftime(ended_at,started_at)),.keep =
"all")
```

3. Check for inconsistencies in timings by filtering for negative values.

```
    View(tripdata %>% filter(trip_duration < 0))
```

In this dataset, there were no negative 'trip_duration' values. However, in other datasets there were instances where this could occur. In those cases, I applied the following code to discard

that data, as I wasn't sure what the source of those errors were and didn't want those data points to skew my analysis.

```
    tripdata <- tripdata[tripdata$trip_duration >= 0,]`
```

4. Check for duplicates in ride_id:

```
    View(tripdata %>% distinct(ride_id,.keep_all = TRUE))
```

This will view the distinct rows filtered by (ride_id), comparing the number of observations here and in the original data frame will indicate whether or not there are repeated entries.

There weren't any duplicate values for any dataset that I used for this case study

5. I decided to delete all rows with NULL or NA values in all the datasets. Most NAs are in variables that indicate station name and id, and sometimes, latitude and longitude of stations. There is no way of retrieving those missed values, it is better to go ahead and remove them all together.

```
    tripdata <- drop_na(tripdata)

insert table of cleaning
```

6. Create more new variables… These will be calculated from existing variables as follows:

start_day : Day of the week the ride started on:

```
  tripdata <- tripdata %>%
    mutate(start_day = weekdays(tripdata$started_at))
```

start_year

```
  tripdata <- tripdata %>%
    mutate(start_year = year(tripdata$started_at))
```

start_month

```
  tripdata <- tripdata %>%
    mutate(start_month = months.POSIXt(tripdata$started_at, abbreviate =
FALSE))
```

ride_time : This is a variable that I added to categorize ride times into Morning, Afternoon, Evening and Night as follows.

```
  tripdata <- tripdata %>%
    mutate(ride_time = case_when(hour(started_at) >= 5 &
                                 hour(started_at) < 12 ~ 'Morning',
hour(started_at) >=12 &
                                 hour(started_at) <17 ~ 'Afrernoon',
hour(started_at) >=17 &
                                 hour(started_at) <21 ~ 'Evening',
hour(started_at) >= 21 &
```

```
                                hour(started_at) <=23~ 'Night', hour(started_at)
>=0 &
                                hour(started_at) <5 ~ 'Night'))
```

7. After adding all the variables I need for Analysis, I deleted the ones I won't be using. For this step I saved the data frame under a new name 'tripdata1'; the added number will indicate to me that this is the first cleaned set.

Romoving the variables I don't need:

```
  tripdata1 <- select(tripdata, -c("started_at", "ended_at"))
```

8. Saved and exported the file as CSV using the write.csv(). "dataset_name" & "dataset_file_name.csv" are modifiable.

```
  write.csv(dataset_name, "dataset_file_name.csv", row.names = FALSE)
```

This marks the end of the cleaning process I applied for this case study.

Data Analysis and Visualization

Summary

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

Visualizations and Key Findings

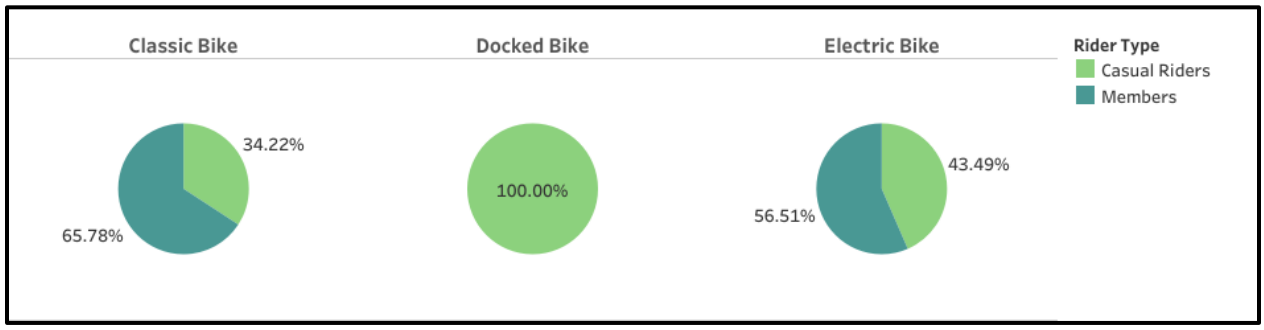
At this stage, it is clearly evident that casual riders and members each have different riding habits, which I'm going to explore further.

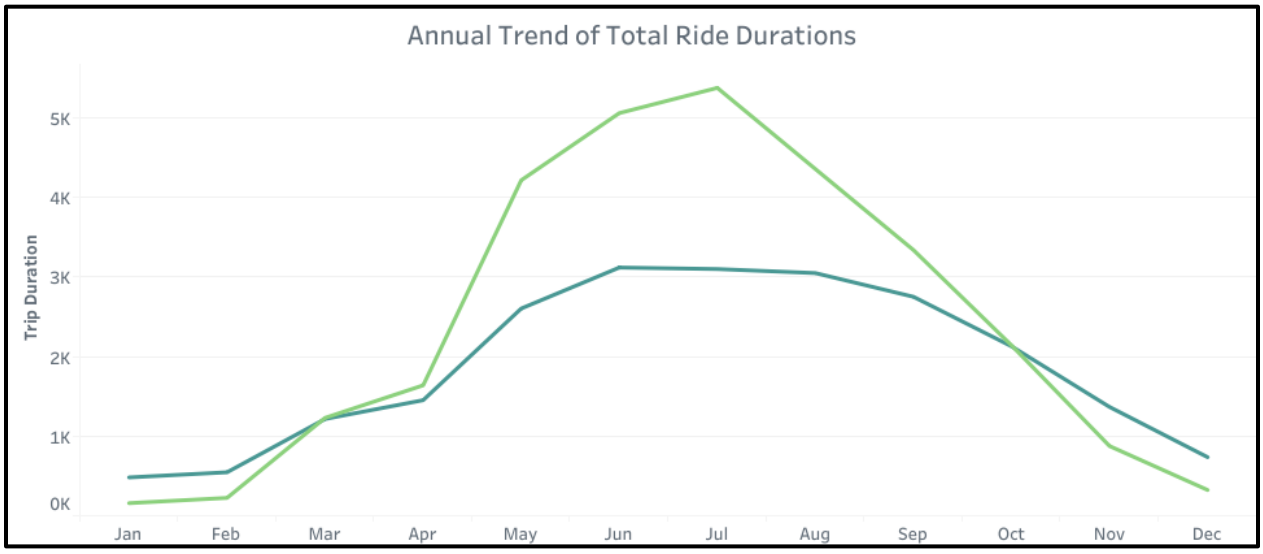
- Although members start more rides overall than casual riders, casual riders tend to use the bikes to take longer rides, especially in the 2nd and 3rd quarter, when the weather is warmer. It is also clear that all riders generally tend to use the bikes more and take more rides when the weather is warmer.

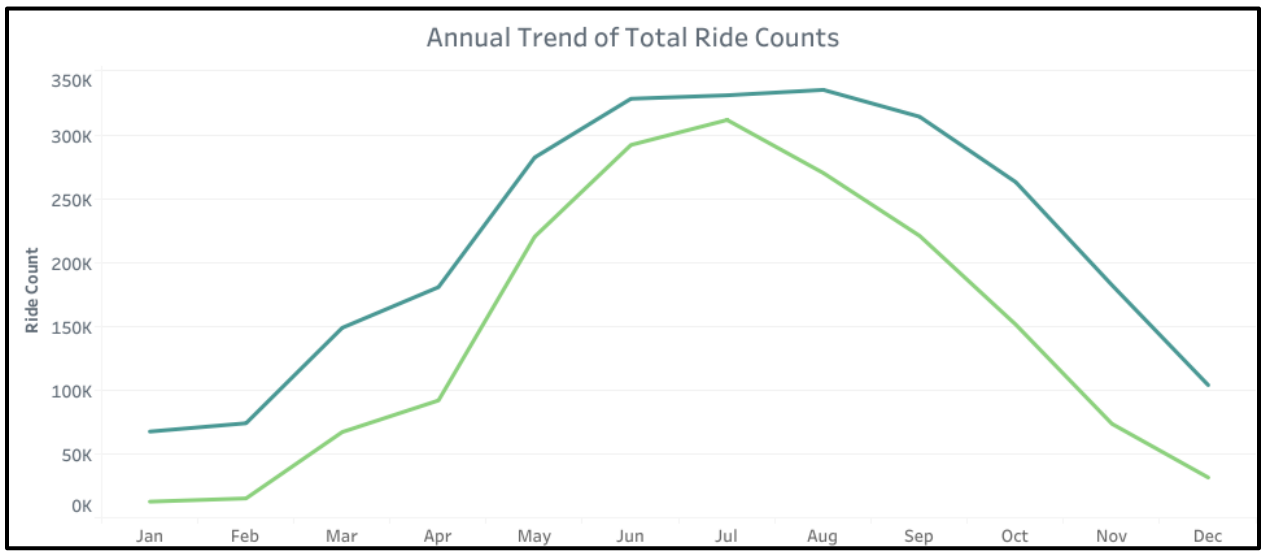
- Members use classic bikes significantly more than casual riders, whereas, casual riders use docked bikes 100% of the time.

- Most common stations used by casual riders are 'Streeter Dr & Grand Ave' and 'DuSable Lakeshore & Ave' significantly more than other stations and the highest riding frequency being on the weekends.

- Casual riders take more rides over the weekend. On the other hand, members take more rides during week days. This difference in behavior indicates that members may prefer to use the bikes to commute to work. Members also take significantly more morning rides.


Classic Bike
Docked Bike
Electric Bike
Rider Type
Casual Riders
Members
34.22%
65.78%
100.00%
43.49%
56.51%


Annual Trend of Total Ride Durations
Trip Duration
5K
4K
3K
2K
1K
0K
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


Annual Trend of Total Ride Counts
Ride Count
350K
300K
250K
200K
150K
100K
50K
0K
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

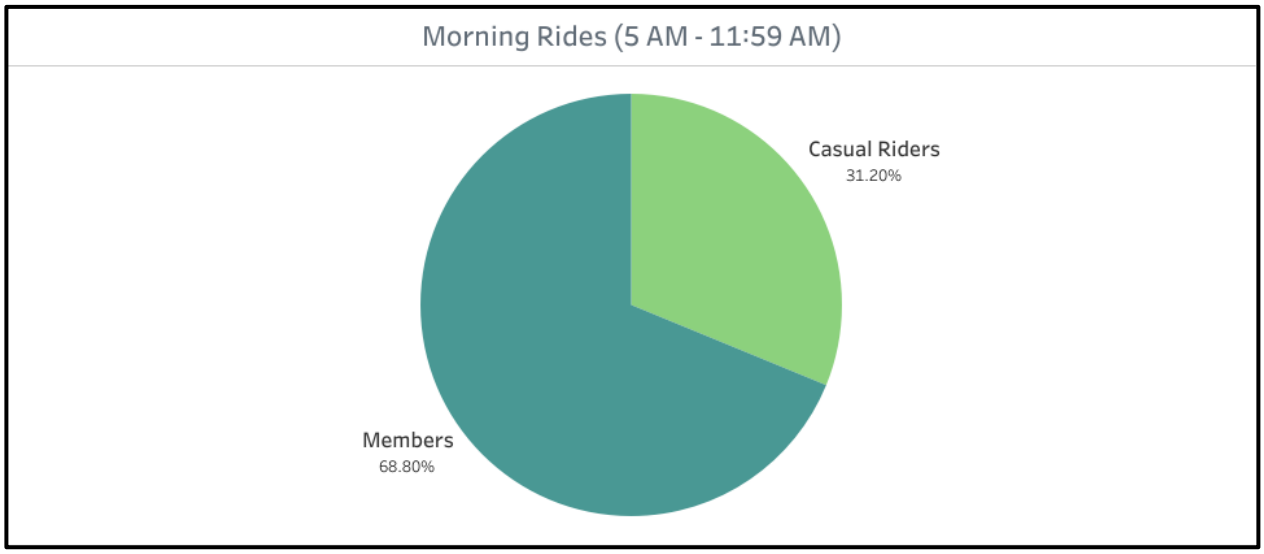


Tableau Dashboard Link:
https://public.tableau.com/views/Cyclistic2022Viz/StationTraffic?:language=en-US&:display_count=n&:origin=viz_share_link

Recommendation

Casual riders are leisure riders, are mostly active on the weekends and take longer rides. Offering them memberships that suit their riding habits can persuade them to join. My recommendation is a membership plan that offers a discounted rate or perks for longer rides. Through social media, collaborations and sponsorships of leisure events and weekend activities in the city can help with brand awareness and promote memberships for goers of such activities. Also, an app can be launched where members can easily find the nearest charging stations if they are riding an electric bike, or the nearest docking stations.