CYCLISTIC BIKE-SHARE CASE STUDY

Oct 2023

SCENARIO

This case study is about Cyclistic, a fictional bike-share company that provides bicycles to customers through single rides, full day passes or annual subscriptions. The director of marketing believes that for the company to grow an increase in the number of annual members is required, for which the marketing team has decided to come up with a strategy that will convert casual riders to annual riders. The job is to work along with the marketing team to gather information on how customers make use of their bikes differently with the purpose of identifying trends and generating insights.

ABOUT THE COMPANY

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked 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.

STEP 1: ASK

- Business task: Analyze trip data from the first half of 2023 to develop strategies oriented to the conversion of casual riders to annual riders.
- Primary stakeholders: the director of marketing (Lily Moreno) and the executive team.
- Secondary stakeholders: the marketing analytics team

STEP 2: PREPARE

The dataset is comprised of six comma-separated value (CSV) files that represent the first half of 2023's bike rides. Proper naming conventions were already being followed so none of the names were altered. Excel didn't seem like a viable option due to their large size, so R was chosen instead for its capabilities at handling large amounts of data.

First and foremost, certain libraries were loaded into R:

```
library(tidyverse)
library(tidyr)
library(lubridate)
library(DescTools)
```

Next, the individual CSV files were imported then combined into a single data frame:

```
df1 <- read.csv("202301-divvy-tripdata.csv", na.strings = c("","NA"))
df2 <- read.csv("202302-divvy-tripdata.csv", na.strings = c("","NA"))
df3 <- read.csv("202303-divvy-tripdata.csv", na.strings = c("","NA"))
df4 <- read.csv("202304-divvy-tripdata.csv", na.strings = c("","NA"))
df5 <- read.csv("202305-divvy-tripdata.csv", na.strings = c("","NA"))
df6 <- read.csv("202306-divvy-tripdata.csv", na.strings = c("","NA"))
biketrip_df <- rbind(df1, df2, df3, df4, df5, df6)
```

The codes returned a data frame with a total of 2,390,459 entries and 13 variables. After getting a quick overview of its contents it soon became apparent that some columns were missing information (*start_station_name, start_station_id & end_station_name & end_station_id*). It was deemed necessary by the director of marketing to remove such cases, for which the following code was implemented:

```
biketrip_clean <- biketrip_df %>%
  drop_na()
```

Removing cases with missing information reduced the number of observations to 1,820,473. The R.O.C.C.C method was then implemented to evaluate the credibility of the data:

- Reliable – Data was provided by Motivate International Inc. under this license.
- Original – Data comes from the City of Chicago's Divvy bicycle sharing service.
- Comprehensive – Data contains all of the information necessary to fulfill the business task.
- Current – Data was less than a year old when this study was made.
- Cited – Data can be found here.

STEP 3: PROCESS

Sample size wasn't calculated because all of the cases were considered for this study. The following steps were taken in order to ensure data was clean and ready for analysis:

First, started at & ended at variables were reformatted to date/time format:

```
biketrip_clean$started_at <- ymd_hms(biketrip_clean$started_at)
biketrip_clean$ended_at <- ymd_hms(biketrip_clean$ended_at)
```

Next, new variables that contain the date, weekday & month were created respectively:

```
biketrip_clean$start_date <- as.Date(biketrip_clean$started_at)
biketrip_clean$day_of_week <- weekday(biketrip_clean$started_at)
biketrip_clean$month <- month(biketrip_clean$started_at)
```

Consequently, the trip duration was calculated and stored in a variable called ride length:

```
biketrip_clean <- biketrip_clean %>%
  mutate(ride_length = ended_at - started_at)
```

After sorting and filtering these new variables, it was found that a total of 129 cases in the ride length variable returned negative values, for which they were dropped after consulting the director of marketing. The code used to drop these columns is the one that follows:

```
biketrip_clean <- biketrip_clean %>%
    filter(ride_length > 0)
```

From there on, the columns that weren't going to be used in the analysis were dropped:

```
biketrip_clean = subset(biketrip_clean, select = -c(ride_id, start_station
_id, end_station_id, start_lat, start_lng, end_lat, end_lng))
```

Lastly, the processed dataset got exported as a new .csv file which was later loaded onto R:

```
write_csv(biketrip_clean, "2023-divvy-tripdata_cleaned.csv")
biketrip_final <- read_csv("2023-divvy-tripdata_cleaned.csv")
```

STEP 4: ANALYZE

Descriptive analysis was performed using R for the purpose of obtaining the mode, number of trips, longest trip duration and the average trip duration. Some additional data aggregation techniques were implemented to summarize more data points grouped by dates, rider types and bike types.

- **Descriptive analysis of main variables**

```
analysis_df <- biketrip_final %>%
  group_by(member_casual) %>%
  summarize(
    mean_ride_length = mean(ride_length),
    max_ride_length = max(ride_length),
    mode_day_of_week = Mode(day_of_week),
    count_of_trips = n()) %>%
   mutate(max_hours = hour(seconds_to_period(max_ride_length)),
          max_minutes = minute(seconds_to_period(max_ride_length)),
          mean_hours = hour(seconds_to_period(mean_ride_length)),
          mean_minutes = minute(seconds_to_period(mean_ride_length)))
analysis_df %>%
  unite(mean_ride_length, mean_hours, mean_minutes, sep = ":") %>%
  unite(max_ride_length, max_hours, max_minutes, sep = ":")
analysis_df = subset(analysis_df, select = -c(max_hours, max_minutes, mean
_hours, mean_minutes))
write_csv(analysis_df, "2023-divvy-analysis_df.csv")
```

- **Total number of trips by biketype**

```
biketype_trip_count <- biketrip_final %>%
  group_by(member_casual, rideable_type) %>%
  summarize(count_of_trips = n())
write_csv(biketype_trip_count, "2023-divvy-biketype_trip_count.csv")
```

- **Station summary: member/casual**

```
station_analysis <- biketrip_final %>%
  mutate(station = start_station_name) %>%
  group_by(start_station_name, member_casual) %>%
  summarize(count_of_trips = n()) %>%
  arrange(desc(count_of_trips))
write_csv(station_analysis, "2023-divvy-station_analysis.csv")
```

- **Ride length: day of week**

```
ridelength_day_of_week <- biketrip_final %>%
  mutate(day_of_week = factor(day_of_week, levels = c("Sunday", "Monday",
"Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))) %>%
  group_by(member_casual, day_of_week) %>%
  summarize(
    mean_ride_length = mean(ride_length),
    count_of_trips = n())
write_csv(ridelength_day_of_week, "2023-divvy-ridelength_day_of_week.csv")
```

- **Ride length: month**

```
ridelength_month <- biketrip_final %>%
  mutate(month_name = month.name[month]) %>%
  mutate(month_name = factor(month_name, levels = c("January", "February",
 "March", "April", "May", "June"))) %>%
  group_by(member_casual, month_name) %>%
  summarize(
    mean_ride_length = mean(ride_length),
    count_of_trips = n())
write_csv(ridelength_month, "2023-divvy-ridelength_month.csv")
```

STEP 5: SHARE

Tableau was used to create compelling visualizations which were later included in a PowerPoint presentation created with the purpose of delivering key insights to stakeholders. The presentation can be downloaded by clicking on this link. In summary, the analysis supports the following conclusions:

- More than half of the trips analyzed come from members.

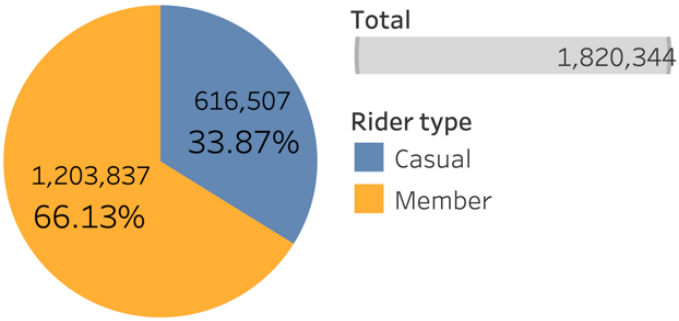

Figure 1: Total number of trips by rider type

- Average trip duration for casual riders is almost two times bigger than members.

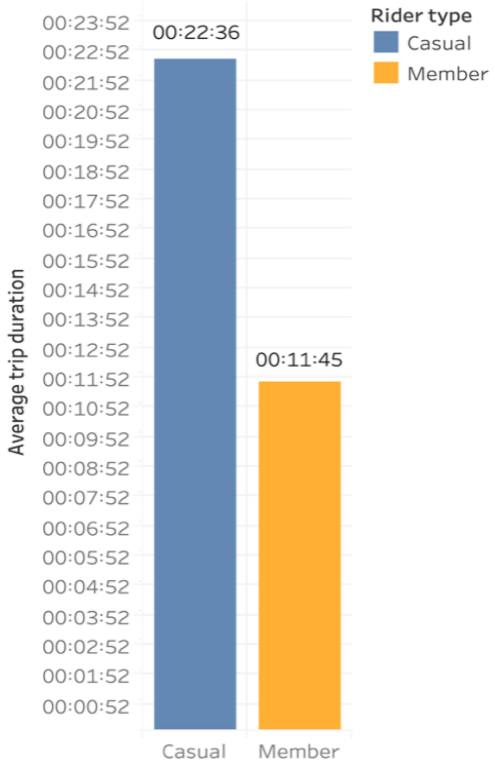

Figure 2: Average trip duration by rider type

- Members ride frequently during weekdays, while weekends are more popular between casuals.

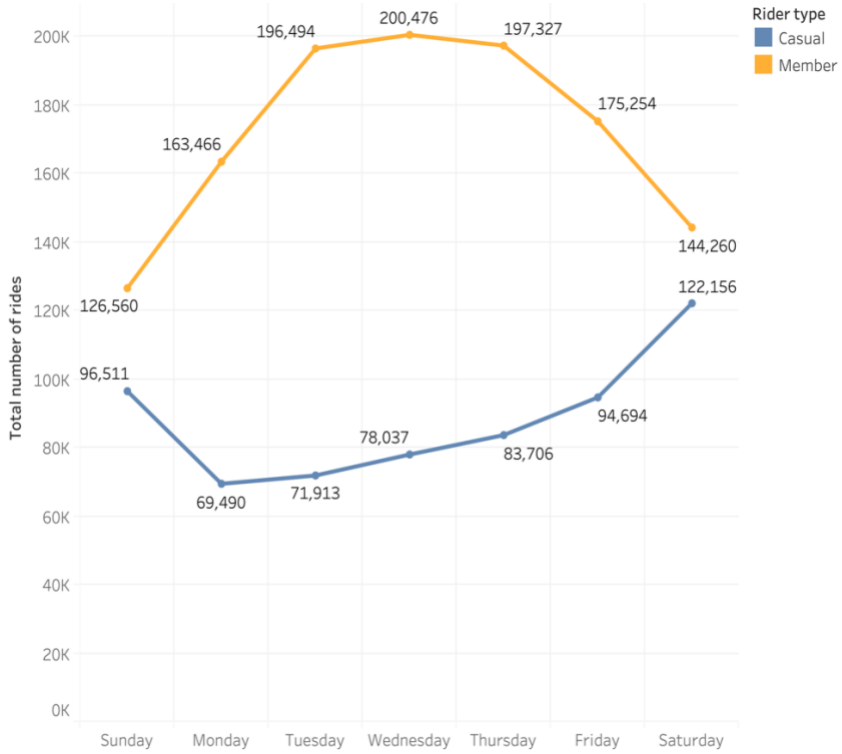

Figure 3: Total number of weekly rides

- On average, casuals ride for longer during weekends, but for members it stays about the same.

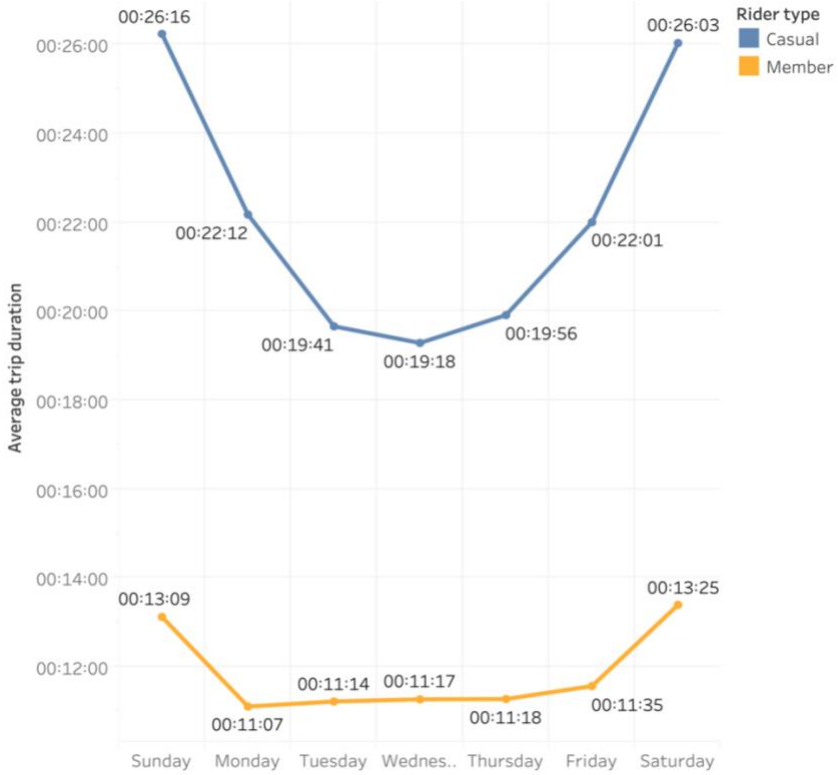

Figure 4: Average trip duration measured by day of the week.

- Rides increase exponentially in the months that follow.

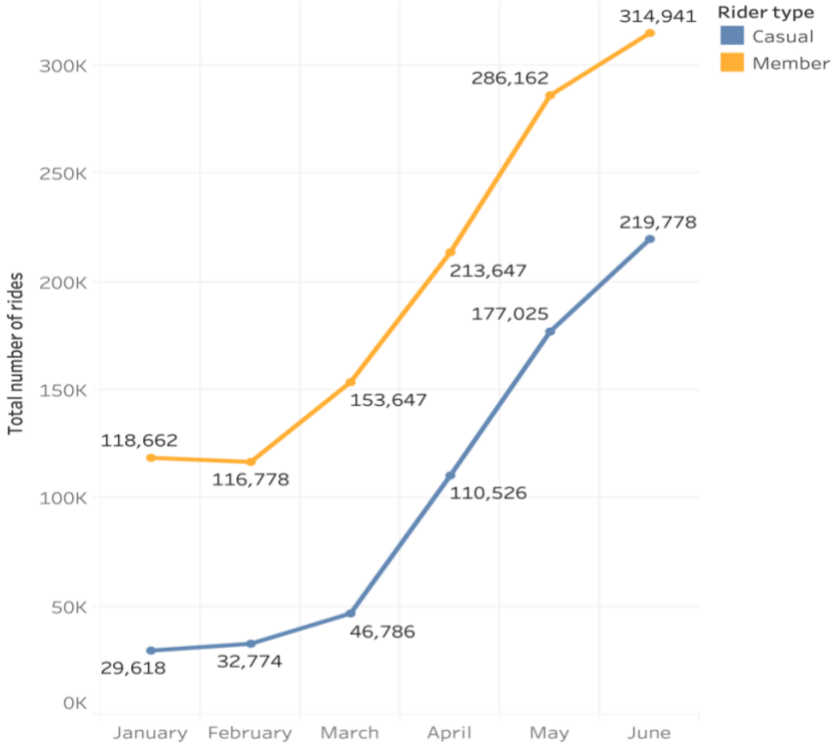

Figure 5: Total number of monthly rides

- Trip duration increases monthly for both rider types, but it's more significant for casual riders.

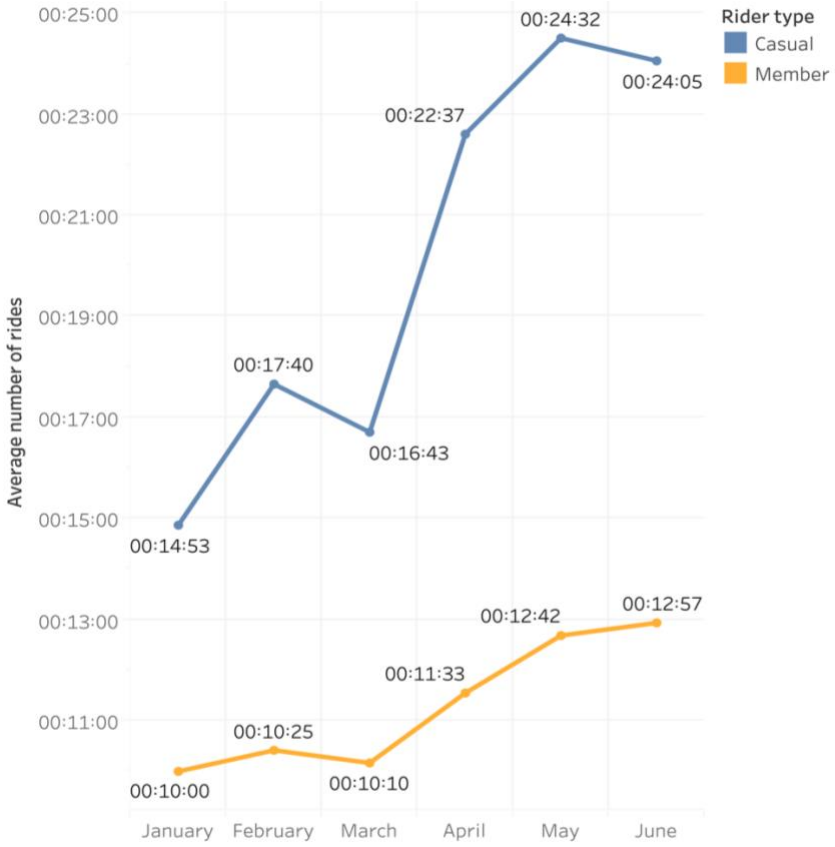

Figure 6: Average trip duration measured by month.

STEP 6: ACT

Based on the conclusions, the following recommendations were offered with the purpose of incrementing the number of annual members through membership conversion:

- Introduce a referral program by assigning each member a unique code and offering them a discount when a casual rider uses it to sign up for a membership.
- Launch riding packages on the busiest months of the year that lead to membership conversion if a differential fee is paid.
- Target ads to casual riders at the most popular sations during the weekends.
- Offer trial periods while collecting payment information upfront to incentivize casual riders into membership conversion by granting them the benefits of an annual subscription for a week.