Cyclistic Case Study

Lucas Argeles

5/19/2021

Ask

What is the problem that needs to be solved?

How do annual members and casual riders use Cyclistic bikes differently?

Business Task

Design marketing strategies aimed at converting casual riders into annual members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Findings will be used to find trends in order to convert more casual riders into annual members as part of a marketing strategy.

Stakeholders

- Lily Moreno: Director of Marketing
- Cyclistic marketing analytics team
- Cyclistic executive team

Prepare

Location and Storage

Data is provided by Motivate International Inc. and stored on AWS and Google Cloud Storage. More information on Motivate International can be found here.

Organization and Structure

All files are originally in comma-separated values (.CSV) and are organized into 15 columns:

Column Title	Column Variable
Ride ID	ride_id
Ride Type	rideable_type
Start Time	started_at
End Time	ended_at
Starting Station Name	start_station_name
Starting Station ID	start_station_id
Ending Station Name	end_station_name
Ending Station ID	end_station_id
Starting Station Latitude	start_lat
Starting Station Longitude	start_lng
Ending Station Latitude	end_lat
Ending Station Longitude	end_lng
Membership Type	member_casual

Credibility
Data is provided for public use and serves as a framework for this case study. More information can be found under this license to show data is reliable, original, comprehensive, current, and cited.

Process

Import necessary packages and load libraries for analysis.

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

library(tidyverse)
library(ggplot2)
library(skimr)
library(janitor)
library(readr)
library(lubridate)
```

Using BigQuery, an extra column was added in order to track activity throughout the week by including which day of the week the ride took place.

```
## Documentation of SQL queries. Day of the week for each ride was added.

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.april2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.may2020_trips`
```

```
# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.june2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.july2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.august2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.september2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.october2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.november2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.december2020_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.january2021_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.february2021_trips`

# SELECT *,  EXTRACT(DAYOFWEEK from started_at) AS day_of_week
# FROM `oval-flow-286322-309905.bqtest.march2021_trips`
```

Import tables.

```
april2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/april2020_updated.csv")
may2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/may2020_updated.csv")
june2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/june2020_updated.csv")
july2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/july2020_updated.csv")
august2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/august2020_updated.csv")
september2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/september2020_updated.cs
october2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/october2020_updated.csv")
november2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/november2020_updated.csv"
december2020_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/december2020_updated.csv"
january2021_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/january2021_updated.csv")
february2021_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/february2021_updated.csv"
march2021_updated <- read_csv("Documents/Datasets/Cyclistic/months_updated/march2021_updated.csv")
```

Column types are then changed to be consistent throughout all tables. There were mismatched column types for December 2020, January 2021, February 2021, and March 2021 tables. The start/end ID columns were set as “character” types rather than “numeric” values.

```
december2020_updated <- mutate(december2020_updated, start_station_id = as.numeric(start_station_id),
                               end_station_id = as.numeric(end_station_id))
january2021_updated <- mutate(january2021_updated, start_station_id = as.numeric(start_station_id),
                              end_station_id = as.numeric(end_station_id))
february2021_updated <- mutate(february2021_updated, start_station_id = as.numeric(start_station_id),
                    end_station_id = as.numeric(end_station_id))
march2021_updated <- mutate(march2021_updated, start_station_id = as.numeric(start_station_id),
                    end_station_id = as.numeric(end_station_id))
```

Analyze

Aggregate all data into one table for analysis.

```
# This combines all data into a data frame representing the year.
ag_rides <- bind_rows(april2020_updated,
                      may2020_updated,
                      june2020_updated,
                      july2020_updated,
                      august2020_updated,
                      september2020_updated,
                      october2020_updated,
                      november2020_updated,
                      december2020_updated,
                      january2021_updated,
                      february2021_updated,
                      march2021_updated
                      )
```

Adjust date format for later analysis.

```
ag_rides$date <- as.Date(ag_rides$started_at)
ag_rides$month <- format(as.Date(ag_rides$date), "%m")
ag_rides$day <- format(as.Date(ag_rides$date), "%d")
ag_rides$year <- format(as.Date(ag_rides$date), "%Y")
ag_rides$day_of_week <- format(as.Date(ag_rides$date), "%A")
```

Remove unnecessary columns for faster processing.

```
ag_rides <- ag_rides %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

Create new column with ride length.

```
ag_rides$ride_length <- as.numeric(difftime(ag_rides$ended_at,ag_rides$started_at))
# units will be in seconds

ag_rides$ride_length <- transform(ag_rides$ride_length/60)
# converts to minutes
```

Remove null values.

```
ag_rides_cleaned <- na.omit(ag_rides)
# Data changes from 3,489,748 observations to 1,835,164 due to
# null values posing as an obstacle for analysis
```

Remove negative ride lengths.

```
ag_rides_final <- ag_rides_cleaned[ag_rides_cleaned$ride_length > 0, ]
ag_rides_final$ride_len <- as.numeric(unlist(ag_rides_final$ride_length))
ag_rides_final <- ag_rides_final %>%
  select(-c(ride_length))
```

Share

Observe general statistics of finalized table.

```
summary(ag_rides_final)
# Mean ride: 33.14 minutes
# Median ride length: 17.17 minutes
```

Look at mean and median of casual riders versus members.

```
aggregate(ride_len ~member_casual, data = ag_rides_final, mean)
# casual    51.86898
# member    17.73425
aggregate(ride_len ~member_casual, data = ag_rides_final, median)
# casual    24.16667
# member    13.25000
```

Most popular day of the week.

```r
my_mode <- function(x) {        # Create mode function
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}
my_mode(ag_rides_final$day_of_week)
# [1] "Saturday"
```

Casual vs. Member activity.

```r
# Order days of the week
ag_rides_final$day_of_week <- ordered(ag_rides_final$day_of_week, levels=c("Sunday", "Monday", "Tuesday


# TABLE 1
ag_rides_final %>%
  group_by(member_casual, day_of_week) %>%
  summarise(rides = n()) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title="Casual vs. Member Rider Daily Activity") +
  ylab("Number of Rides") +
  xlab("Day of Week")

# Casual riders ride more on Fridays, Saturdays, and Sundays
# Members ride  more on weekends as well though difference weekdays versus weekends is not as drastic


# TABLE 2
aggregate(ride_len ~member_casual + day_of_week, data = ag_rides_final, mean)
# Show results in table using summarise function
ag_rides_final %>%
  group_by(member_casual, day_of_week) %>%
  summarise(avg_dur = mean(ride_len)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = avg_dur, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Casual vs. Member Daily Average Duration") +
  ylab("Average Duration (minutes)") +
  xlab("Day of Week")
# casual riders ride about 2.5-3x longer than members
```

Rider activity throughout the year.

```r
# TABLE 3
ggplot(ag_rides_final, aes(date, color = member_casual)) +
  geom_freqpoly() +
  labs(title = "Bike Activity Throughout the Year") +
  ylab("# of Rides (1e+05 = 100,000)") +
```

```
  xlab("Day of the Week")
# Rider activity picks up during the warmer seasons and drops during colder
# seasons due to freezing temperatures in the Chicago area
```

Rider activity by bike type.

```
ag_rides_final %>%
  group_by(rideable_type, day_of_week) %>%
  summarise(rides = n()) %>%
  arrange(rideable_type, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = rides, fill = rideable_type)) +
  geom_col(position = "dodge2") +
  labs(title = "Count of Rides by Day and Bike Type") +
  ylab("# of Rides (1e+05 = 100,000)") +
  xlab("Day of the Week")
# Most common bike type used is the docked bike

ag_rides_final %>%
  group_by(rideable_type, day_of_week) %>%
  summarise(avg_dur = mean(ride_len)) %>%
  arrange(rideable_type, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = avg_dur, fill = rideable_type)) +
  geom_col(position = "dodge") +
  labs(title = "Average Ride Duration by Day and Bike Type") +
  ylab("Average Duration (min)") +
  xlab("Day of the Week")
# Riders ride docked bikes for twice as long as electric and classic bikes
```

Observations and findings:

- The average ride for all riders is 33.1 minutes and the median ride is 17.7 minutes.
- The day of the week with the most activity is Saturday.
- Casual riders ride more on Fridays, Saturdays, and Sundays.
- Members ride more on weekends as well though the difference between weekdays and weekends is not as drastic.
- Casual riders ride about 2.5-3x longer than members.
- Rider activity picks up during the warmer seasons and drops during colder seasons due to freezing temperatures in the Chicago area.
- The most common bike type used is the docked bike compared to the electric and classic bikes.
- Riders use docked bikes for twice as long compared to electric and classic bikes.

Act

The original objective was to look at trends in order to design marketing strategies aimed at converting casual riders into annual members.

Recommendations based on analysis:

1. Create an annual weekend subscriptions aimed at casual riders.
    - Casual riders use bikes for leisure (Table 1 & 2); an annual weekend plan would allow unlimited bike use on Fridays, Saturdays, and Sundays. A lower annual rate would entice more casual customers to commit to an annual pass.
2. Run promotions mid-March when temperatures warm up and rider activity increases.
    - Chicago temperatures reach under 32 degrees throughout most of winter. Data shows bike activity is low during the winter (Table 3). By having a strong digital presence through ads during warm seasons, riders are more likely to pay attention to special announcements upon re-using bikes early Spring.
3. Offer discounted annual membership in October when temperatures drop in order to guarantee users for the following year.
    - Chicago temperatures reach under 32 degrees throughout most of winter. Data shows bike activity is low when temperature are below freezing (Table 3).

In what ways do members and casual riders use Cyclistic bikes differently?
Member riders use the bikes more during weekdays than casual riders. It is safe to assume that this is due to members using the bikes for their daily commutes to work compared to casual riders who use the bikes more than members as shown in the graph "Casual vs. Member Rider Daily Activity." Furthermore, casual riders use the bikes for longer on weekends. It is inferred that casual riders use bikes for more leisure activities.

Why would casual riders buy Cyclistic annual memberships?
If casual riders are more avid weekends users, an annual membership may help them save money if they ride for a long period of time and regularly use it on weekends. If an annual weekend pass is made available, casual riders are more likely to switch over to an annual subscription rather than a pay-as-you-go plan.

How can Cyclistic use digital media to influence casual riders to become members?
By marketing an annual weekend pass to casual riders, Cyclistic can increase its annual subscriptions users. Referring back to the graph "Bike Activity Throughout the Year," ride activity is low during the winter. By promoting all digital ads in the spring when temperatures warm up and rider activity increases, casual riders will be more enticed to switch to an annual subscription.