Case Study: How Does a Bike-Share Navigate Speedy Success?

Lingyu Tan

Dec 9, 2021

Background (Ask)

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.

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.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Three questions will guide the future marketing program:

1. **How do annual members and casual riders use Cyclistic bikes differently?**
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

In this report, I will try to answer **the first question** with the following deliverables thus helping Cyclistic design marketing strategies:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top three recommendations based on your analysis

Data Source (Prepare)

The datasets which are used in this case study can be accessed here. The datasets include previous 12 months of historical trip data (2020/12 - 2021/11) of Cyclistic, a fictional company. The data has been made available by Motivate International Inc. under this license.

Setting up Environment

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Defining Functions

```
swap <- function(x, y){
  temp <- x
  x <- y
  y <- temp
}
```

Loading Datasets

```
### Load datasets
data <- read_csv("Data/202012-divvy-tripdata.csv")
```

```
## Rows: 131573 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# data_202012 <- read_csv("Data/202012-divvy-tripdata.csv")
# data_202101 <- read_csv("Data/202101-divvy-tripdata.csv")
# data_202102 <- read_csv("Data/202102-divvy-tripdata.csv")
# data_202103 <- read_csv("Data/202103-divvy-tripdata.csv")
# data_202104 <- read_csv("Data/202104-divvy-tripdata.csv")
# data_202105 <- read_csv("Data/202105-divvy-tripdata.csv")
# data_202106 <- read_csv("Data/202106-divvy-tripdata.csv")
# data_202107 <- read_csv("Data/202107-divvy-tripdata.csv")
# data_202108 <- read_csv("Data/202108-divvy-tripdata.csv")
# data_202109 <- read_csv("Data/202109-divvy-tripdata.csv")
# data_202110 <- read_csv("Data/202110-divvy-tripdata.csv")
# data_202111 <- read_csv("Data/202111-divvy-tripdata.csv")
#
# ### Combine datasets
# data <- rbind(data_202012, data_202101, data_202102, data_202103,
#               data_202104, data_202105, data_202106, data_202107,
#               data_202108, data_202109, data_202110, data_202111)
# rm(data_202012, data_202101, data_202102, data_202103,
#    data_202104, data_202105, data_202106, data_202107,
#    data_202108, data_202109, data_202110, data_202111)
```

Exploratory Description

```
glimpse(data)
```

```
## Rows: 131,573
## Columns: 13
## $ ride_id            <chr> "70B6A9A437D4C30D", "158A465D4E74C54A", "5262016E0F~
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "~
## $ started_at         <dttm> 2020-12-27 12:44:29, 2020-12-18 17:37:15, 2020-12-~
## $ ended_at           <dttm> 2020-12-27 12:55:06, 2020-12-18 17:44:19, 2020-12-~
## $ start_station_name <chr> "Aberdeen St & Jackson Blvd", NA, NA, NA, NA, NA, N~
## $ start_station_id   <chr> "13157", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", NA, NA, NA, NA, NA, NA~
## $ end_station_id     <chr> "TA1306000003", NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.87773, 41.93000, 41.91000, 41.92000, 41.80000, 4~
## $ start_lng          <dbl> -87.65479, -87.70000, -87.69000, -87.70000, -87.590~
## $ end_lat            <dbl> 41.88872, 41.91000, 41.93000, 41.91000, 41.80000, 4~
## $ end_lng            <dbl> -87.64445, -87.70000, -87.70000, -87.70000, -87.590~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

In this datasets, there are overall 5,479,096 observations and 13 variables, including 7 character, 4 double, 2 date-time.

Cleaning Data

First we need to check if there is duplicated data in the dataset.

```
sum(duplicated(data$ride_id))
```

```
## [1] 0
```

We see that `ride_id` seems to be a unique identifier without repetition thus there seems to be no duplicates. We can check all the records without `ride_id` to see what happens.

```
### There is no duplicated rows so the sum is zero
# sum(duplicated(data[, -1]))
# = 502

# data_no_id <- data[, -1]
# dup <- duplicated(data_no_id)
# dup_data <- data[dup, ]
# dup_data[order(dup_data$started_at), ]

# nrow(unique(data[, -1]))
# = 5478594
```

There are 502 out of 5,479,096 records are duplicates without considering `ride_id`. We can have a look at one duplicate below:

```
data %>% filter(started_at == "2020-12-01 14:17:37 UTC")
```

```
## # A tibble: 2 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 0DF824~ docked_bike   2020-12-01 14:17:37 2020-12-01 14:26:31 Theater on the ~
## 2 6C8319~ docked_bike   2020-12-01 14:17:37 2020-12-01 14:26:31 Theater on the ~
## # ... with 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

We clearly see that all records of these two observations are identical except `ride_id`. It is possible that a couple start their journey at the same time/place and end simultaneously. Therefore, we decide not to remove these potential "duplicates".

In this report we are not interested in the information related to the locations