Google Analytics Capstone Project: Cyclistic

Erin Hart

2022-07-14

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Objectives of Analysis

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?

Preparing the data

Data is from Divvy-Trip data. For this project we are using the last 12 most recent months. I have downloaded the data to a folder on my desk top for future use in RStudio. I want to make sure the data is ROCCC- Reliable, Original, Comprehensive, Current, and Cited.

There are some problems with the data. Many records have missing start and end station names and IDs. I did find a Station name and ID list, but the date is from 2013 and the IDs are out of date so it is not usable. Disregarding the data with missing station names and IDs will take away about 20% of the original data. There is a possibility that valuable insight might have been gained if data was complete. For our purposes we will only analyze the data that is completely filled in. But in a future project I would request more information about station names and IDs to pull in more of the original data.

Processing the data

Call Libraries used for this process

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(tidyr)
library(skimr)
library(janitor)
```

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Load the data into RStudio

```
trips_2021_06<-read.csv("202106-divvy-tripdata.csv")
trips_2021_07<-read.csv("202107-divvy-tripdata.csv")
trips_2021_08<-read.csv("202108-divvy-tripdata.csv")
trips_2021_09<-read.csv("202109-divvy-tripdata.csv")
trips_2021_10<-read.csv("202110-divvy-tripdata.csv")
trips_2021_11<-read.csv("202111-divvy-tripdata.csv")
trips_2021_12<-read.csv("202112-divvy-tripdata.csv")
trips_2022_01<-read.csv("202201-divvy-tripdata.csv")
trips_2022_02<-read.csv("202202-divvy-tripdata.csv")
trips_2022_03<-read.csv("202203-divvy-tripdata.csv")
trips_2022_04<-read.csv("202204-divvy-tripdata.csv")
trips_2022_05<-read.csv("202205-divvy-tripdata.csv")
```

Merge data into a single file to work from Before we combine Lets make sure the data is able to be integrated and check that column names compatible and data types are compatible.

```
compare_df_cols(trips_2021_06, trips_2021_07, trips_2021_08, trips_2021_09, trips_2021_10, trips_2021_1
```

```
##          column_name trips_2021_06 trips_2021_07 trips_2021_08 trips_2021_09
## 1            end_lat       numeric       numeric       numeric       numeric
## 2            end_lng       numeric       numeric       numeric       numeric
## 3      end_station_id     character     character     character     character
## 4    end_station_name     character     character     character     character
## 5            ended_at     character     character     character     character
## 6       member_casual     character     character     character     character
## 7             ride_id     character     character     character     character
## 8       rideable_type     character     character     character     character
## 9           start_lat       numeric       numeric       numeric       numeric
## 10          start_lng       numeric       numeric       numeric       numeric
## 11   start_station_id     character     character     character     character
## 12 start_station_name     character     character     character     character
## 13         started_at     character     character     character     character
##    trips_2021_10 trips_2021_11 trips_2021_12 trips_2022_01 trips_2022_02
## 1        numeric       numeric       numeric       numeric       numeric
## 2        numeric       numeric       numeric       numeric       numeric
## 3      character     character     character     character     character
## 4      character     character     character     character     character
## 5      character     character     character     character     character
## 6      character     character     character     character     character
## 7      character     character     character     character     character
## 8      character     character     character     character     character
## 9        numeric       numeric       numeric       numeric       numeric
## 10       numeric       numeric       numeric       numeric       numeric
## 11     character     character     character     character     character
## 12     character     character     character     character     character
## 13     character     character     character     character     character
##    trips_2022_03 trips_2022_04 trips_2022_05
## 1        numeric       numeric       numeric
## 2        numeric       numeric       numeric
## 3      character     character     character
## 4      character     character     character
## 5      character     character     character
## 6      character     character     character
## 7      character     character     character
## 8      character     character     character
## 9        numeric       numeric       numeric
## 10       numeric       numeric       numeric
## 11     character     character     character
## 12     character     character     character
## 13     character     character     character
```

Looks good. Lets combine data with rbind and rename alltrips:

```
alltrips <- rbind(trips_2021_06, trips_2021_07, trips_2021_08, trips_2021_09, trips_2021_10, trips_2021_
```

Clean the Data

First lets remove duplicates based on the ride_id.

```
alltrips <- distinct(alltrips, ride_id, .keep_all=TRUE)
```

Now we have a lot of blank station names and station IDs that account for about 20% of the data. Unfortunately we do no have the right information to update these records and for our purposes we should delete them and work to analyze the only completed data. I'm going to start by changing the blank fields to NA and then omitting the NAs.

```r
alltrips2 <- na_if(alltrips,"") #NA in place of blanks
```

```r
alltrips3 <- na.omit(alltrips2) #Removes all NA rows
```

Next we are going to change the "started_at" and "ended_at" from character strings to date/time.

```r
alltrips3$started_at = strptime(alltrips3$started_at, format = "%Y-%m-%d %H:%M:%S", tz = "UTC")
alltrips3$ended_at = strptime(alltrips3$ended_at, format = "%Y-%m-%d %H:%M:%S", tz = "UTC")
```

Now lets check that:

```r
str(alltrips3)
```

```
## 'data.frame':    4667299 obs. of  13 variables:
##  $ ride_id           : chr  "0D904FEC5F84A538" "C4185F300D6B552B" "60F97090AC85F55E" "FBC7B1F0160AA3
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXlt, format: "2021-06-04 07:29:18" "2021-06-23 08:39:36" ...
##  $ ended_at          : POSIXlt, format: "2021-06-04 07:45:34" "2021-06-23 08:41:37" ...
##  $ start_station_name: chr  "Orleans St & Elm St" "Desplaines St & Kinzie St" "Clark St & Grace St"
##  $ start_station_id  : chr  "TA1306000006" "TA1306000003" "TA1307000127" "KA1503000043" ...
##  $ end_station_name  : chr  "Orleans St & Elm St" "Kingsbury St & Kinzie St" "Clark St & Leland Ave"
##  $ end_station_id    : chr  "TA1306000006" "KA1503000043" "TA1309000014" "TA1306000003" ...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  - attr(*, "na.action")= 'omit' Named int [1:1193477] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:1193477] "1" "2" "3" "4" ...
```

There is a start time and an end time for each ride. Lets create a column called trip_duration.

```r
alltrips4 <- mutate(alltrips3, trip_duration = ended_at - started_at)
```

Now we can filter this to see if there are any trips that are under 60 seconds. These trips are either errors in the data or false starts

```r
alltrips5 <- filter(alltrips4, trip_duration < 60 )
```

Ok. Lets go ahead and delete those trips.

```r
alltrips5 <- filter(alltrips4, trip_duration >60)
```

We can do the same for trips over 24 hours or 86400 seconds as those bikes will have been flagged as stolen.

```
alltrips6 <- filter(alltrips5, trip_duration <86400)
```

Lets also create a column for what day of the week these trips started on.

```
alltrips6$week_day <- weekdays(alltrips6$started_at)
```

Looks really good! We have deleted 1,259,304 rows because they were either duplicates, incomplete, false starts, or stolen bikes. But lets arrange it by the start time of each ride and name it the final copy.

```
Cyclistic_data <- arrange(alltrips6, started_at)
```

Exporting the Cleaned Data

I'm going to be working in Tableau to create some visualizations of the data. So I'll save this and I can export it as a .csv file.

```
write_csv(Cyclistic_data, "C:\\Users\\Erin Hart\\Desktop\\Cyclistic_data.csv")
```