Case1_Case_Study_Road+Map

Luis Rincones

2021-07-02

Case Study 1: How Does a Bike-Share Navigate Speedy Success?

This document will describe the work done using the google analytics process. **APPASA** (Ask, Prepare, Process, Analyze, Share and Act)

ASK

The first phase of the APPASA is Ask

The problem I am trying to solve:
How do annual members and casual riders use Cyclistic bikes differently?

How can my insights drive business decisions?

The differences in the different variables and their relations, would give key information for the difference between members user (bought an annual plan) and casual user did not
The results of my answers, should generate useful information to guide the future Marketing Program

Identify the business task
Analyze the data available their relations, to discover the story of the differences between members and casual users

Consider key stakeholders

- Director of Marketing
- My manager responsible for campaigns development and initiatives to promote the bike share program
- The Cyclistic marketing analytics team, I am part of this team
- Cyclistic executive team, the decision makers to approve or disapprove the recommendations

PREPARE

The second phase of phase of the APPASA is Prepare

Data Source
The data is in the divvy website " https://divvy-tripdata.s3.amazonaws.com/index.html " The data is organized in
* Monthly Trip data, latest 14 months * Data before 2020 for Stations, trips aggregated in two quarters, and some in one quarter

The data is credible, it is not secondhand information and similar data has been used before

Each trip is anonymized and includes:
* Trip start day and time
* Trip end day and time
* Trip start station
* Trip end station
* Rider type

The data has been processed to remove trips that are taken by staff as they service and inspect the system; and any trips that were below 60 seconds in length (potentially false starts or users trying to re-dock a bike to ensure it was secure).

The data fulfills the ROCCC criteria
* Reliable
* Original
* Comprehensive
* Current
* Cited

The Data is provided by the Data Owner from their systems, first party information. This takes care of Reliable and Original.

Comprehensive. The columns describe the trip origin, destination, starting and ending times, geo location and membership

Current, the latest file is May 2021

Cited a quick google search gave several mentions among others for "data.cityofchicago" and "gov.publicstuff"

The data is provided by the company in the site
https://www.divvybikes.com/system-data
according to the license https://www.divvybikes.com/data-license-agreement

The data is:
Accurate the records reflect a unique trip
Complete (it provides the columns indicated)
Consistent among the 14 files I will use, the source is trustable, and it has been used previously

The data provided per trip and rider type will describe the trips, I need to discover the differences that matters for my question
I only found a few records where the start date is later than the end date

I created a R Project(Case_1) in my local documents file and use its directory to store the data

- The data is organized by month, the data it is consistent with the month indicated by the name
- The records describe every trip and allows to extract the differences
- An initial review showed the variables format and content were consistent
- Dates fields to be managed with lubridate functions

A description of all data sources used

- One file per month from April 2020 to May 2021
- Each file with 13 columns will use 6 to work with.
- Columns indicated below
- **ride_id**
- rideable_type
- **started_at**
- **ended_at**
- start_station_name
- **start_station_id**
- end_station_name
- **end_station_id**
- start_lat
- start_lng
- end_lat
- end_lng
- **member_casual**

The six columns in bold will be used
The four geo location columns (lat and long) will not be used to analyze differences
The rideable type is a constant value, will not be used
The name for the two stations will not be used, the stations id suffice

The third phase of phase of the APPASA is Process

What tools are you choosing and why? I am using the R programming language with RStudio

I made a temporary copy(temp1) of each original file using the six columns indicated above
A second temporary copy(temp2) of as the working file. Any issue I have temp1 to recover
Read the zipped csv files, cleaned and manipulate each then merged to the unique file to work with
Use the str, head and tail function to look at the format and content
To define the cleaning, data manipulation, did a review and short analysis of one month data, the findings were used for the final coding with all the files
Checked for negative durations, converted the dates using lubridate
Create a column for duration, two column with weekday for start and end and two columns for start and end hour
This code is included in a Markdown file for sharing

ANALYZE

The fourth phase of phase of the APPASA is Analyze

The data to work in the analysis was a single file with more than four million records
The format for dates, duration and hours was done using lubridate
The duration in seconds is a large number even when expressed in other time measures, for my analysis is just a number to measure how long each trip was

The analysis showed that the members have more trips than casuals, however time used by casuals was greater than members, it is expected since casuals are expected to have more non-working time. Only in the 6am to 8am hours the members have more duration than casuals. The answer to the question
How does annual members and casual riders use Cyclistic bikes differently?
Is indicated in the charts done using ggplot and tableau.

My results are in the Rmardown document Case1_prettydoc

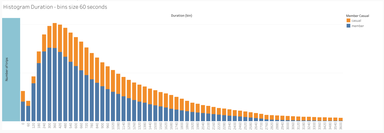
SHARE

The fifth phase of phase of the APPASA is Share

I was able to answer the question, using simple visualizations comparing the number of trips and durations, for members vs casual, summary with basic statistics to complement the story showed with the visualizations.

The amount of bicycle used by the casuals in greater than members except for the early hour (6 to 8 am), members start biking to work

I use simple visualizations, bar charts and histograms to facilitate the communications with the different stakeholders. The charts elaborated with ggplot and tableau are available for the different stakeholders. Example



ACT

The sixth and final phase of phase of the APPASA is Act

My final conclusion is that the differences among members and casuals do exist

In order to complement the information for the marketing campaign to gather more members from the casual. I suggest the following considerations.

How is the revenue distributed among the different plans
Do the registration requirements for members allow to register visitors?
How the change of membership will translate in revenues

Regarding Chicago Visitors and their length of stay
I found the following information searching in Google

Information from the enjoyillinois.com in the 2016 LEISURE VISITOR PROFILE FOR THE STATE OF ILLINOIS AND THE CITY OF CHICAGO
Page 82 in Chart 65 Segment: 2016 Leisure Stays (%)

indicated the visit length:
* day stay 51%
* 1 day stay 16%
* 2 day stay 15%
* 3 day stay 8%
* 4 to 7 days 8%
* more than 8 days 2%

My recommendations
The differences by themselves are just a consideration for a marketing campaign to increase membership, other considerations should be analyzed

Among other considerations:
Analyze the feasibility for new plans alternatives that appeal for the visitors and with an increase in revenues?
For example two day plans?