Created by: Michael Grue
Creation date: 9/18/2024
Last edited: 9/18/2024

Case study: Bike-share

Ask

Cyclistic is a fictional bike-sharing company that owns more than 5,800 bicycles and over 600 docking stations in the city of Chicago, IL. Since their launch in 2016, Cyclistic has focused their marketing strategy on boosting general awareness and engaging broad consumer segments. In order to appeal to the largest possible consumer base, Cyclistic has pricing plans for single-rides, day passes, and annual memberships. The annual memberships have been found to be much more profitable to the company vs. the casual single & day passes. Instead of trying to sell annual memberships to all-new customers, the Cyclistic marketing team has decided to develop a marketing campaign to convert their casual riders into annual members.

Business Task

In order to create a successful marketing campaign, the marketing team first needs to understand how the annual members and the casual riders use the Cyclistic bikes differently. Cyclistic's trip data for the last 12 months will be analyzed to find out "**what distinguishes an annual member from a casual user?**".

Prepare

About the data source

Since Cyclistic is a fictional company, data from an actual bikeshare company "Divvy" will be used in its place.

The bikeshare data came from this Divvy site. Per this Data License Agreement, the Divvy bikeshare data is collected by "Lyft Bikes and Scooters", is owned by the City of Chicago, and is available for

public use. In order to independently verify the reliability of the data, I was able to confirm that Divvy is owned by the City of Chicago via their website (www.chicago.gov), and I was able to navigate to the data directly via the Divvy website (www.divvybikes.com). The data has been regularly updated each month and is current as of August 2024. This dataset has already been stripped of personally identifiable information (addresses, credit card info, etc.) so there is little concern about privacy issues.

Satisfied that I was working with an official dataset that is reliable, original, comprehensive, current, cited, and licensed for public use, I continued with the analysis.

Preparing the data for analysis

Since April 2020, the bikeshare data has been stored as monthly CSV files. We are primarily interested in the different uses between members and casual riders so we will focus on data from the last 12 months. Including older data may skew the results since the usage of the bikes has likely changed since the company's inception in 2013.

Initial data download and storage:

1. Downloaded the 12 orig csv files to a personal, unshared folder on Google Drive.
2. Stored files in an orig csv file folder to retain the original datasets.

Excel:

1. All 12 files were opened and saved as individual Excel spreadsheets.
2. Opened all 12 files with excel (had to select "Don't Convert") and saved them in a new working excel folder. This took quite a while to open, not convert, and save as 12 excel files.
3. The size of the monthly files varied from 14MB to 73MB. In order to quickly analyze the structure, I focused the initial EDA on the smallest file (Jan 2024).
4. The Jan 2024 file has a total of 13 columns and 144,874 rows.
5. The columns that will not likely be needed for analysis include the [*ride_id*] (each row is a unique ride after all), [*start_station_id*] and [*end_station_id*] (I don't suspect that I will need this info for my analysis). The other columns may prove useful in establishing where, when, what type of bike was rented, and what type of user rented the bike).
6. [*rideable_type*] is a column that distinguishes if a bike is an electric bike or a classic bike. There are no blank rows in this column for this month.
7. [*member_casual*] is a column that categorizes the rider. There are no blank rows in this column for this month.
8. This quick summary was performed on **one** of the twelve monthly datasets from the past year. Even though it was the smallest dataset, it still took minutes for Excel just to open the file, not to mention performing basic sorting and filtering functions.
9. Conclusion: It would be inefficient to perform such a summary on the eleven larger files, and merging the CSV files into one CSV file would crash Excel when attempting to open it. Another tool must be used.

BigQuery:

1. I attempted to load the CSV files into Google's BigQuery to analyze them with SQL, but the 100MB size limit per file prevented half of the files from loading.

2. I next attempted to create a Google Cloud Storage bucket to load the data into (which could hold that much data and then be consumed by BigQuery), however I would have had to pay Google to host my large datasets. I'm hoping to avoid paying money to process and analyze this dataset… if possible.
3. Conclusion: Unless I want to pay to have Google host my large CSV files, another tool must be used.

RStudio (desktop):

1. I used the **cmd** prompt code below to create one combined CSV file from the twelve original files. This made it easier to import the data into RStudio.
    a. *C:\Path\to\csv_files> copy *.csv combined-divvy-tripdata.csv*
2. Using the combined CSV file, I loaded the data into a dataframe using readr's function read_csv().
3. It looks like RStudio can handle the size of the data, let's explore and clean up the data!

Process

The data processing can be found in the file '*Process-divvy-tripdata.html*' in this repository on GitHub. This document details how the Divvy bike-sharing data was loaded into R, inspected with various Exploratory Data Analysis (EDA) functions, processed via the removal of unneeded data, and readied for analysis through newly calculated fields.

Analysis

The data analyzing can be found in the file '*Analyze-divvy-tripdata.html*' in this repository on GitHub.

1. TODO when analyzing: examine the distribution of the number of rides and the month of ridership. See if some months had lower ridership.
2. See if can correlate storms to low # of rides.
3. TODO: look at the pct_miss of the full dataset and see what percentage didn't get docked? and add this to the analysis document: Per the above summary, we can see that xxxxx% of rides didn't have an end_station_name. What would cause a ride to not have an end station recorded? Could it be that the bike never made it back to a docking station? xxxxx% seems high for a bike to just be abandoned, maybe a ride only gets an end station if it was able to be successfully docked. If a docking station was full, and the rider left the bike near the station, the ride may have received an ending lat and long, but not a station name. Checking with those responsible for data collection may help answer these questions.

Appendix

This is a case study for the Capstone project for the Google Data Analyst Certificate on Coursera. While the company "Cyclistic" is fictional, the data analyzed is from an actual bike-share company - Divvy Bikes.