Project: Wrangle and Analyze Data

By Utibe Edet Bassey

Introduction

This report gives a summary of the steps taken to complete the Udacity Wrangle and Analyze Data project. The steps below were given by Udacity for guidance:

Project Steps Overview

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing and visualizing data

Step 6: Reporting

- your data wrangling efforts
- your data analyses and visualization

Step 1: Gathering data

This project involved gathering 3 datasets using 3 different methods, manually, programmatically and web scrapping using Twitter API. The twitter-archive-enhanced.csv was downloaded manually from the Udacity classroom, the image-predictions.tsv dataset was downloaded programmatically using the Requests library and the URL provided in the classroom while the Twitter API dataset was to be gotten using Tweepy. Due to the denied elevated access from Twitter the tweet-json.txt file provided by Udacity was used.

Step 2: Assessing data

The following specifications were provided by Udacity for this Wrangle and Analyze data project

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- The requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned.

After gathering the required data, visual and programmatic assessments were made on the 3 datasets and the following Quality and Tidiness issues were detected:

Quality issues

Twitter_archive_enhanced.csv

1. Incorrect datatypes for timestamp column (str instead of timestamp).
2. Dog stages not properly captured, 10 doggos,26 puppers, and 8 puppos were not captured.
3. 'None' values for unavailable values for doggo, floofer, pupper, and puppo.
4. Dog names Eazy, O, his, my, None(35), None(72) instead of Eazy-E, O'Malley, Quizno, Zoey, Howard, and Martha.
5. Dog names shouldn't be'actually','unacceptable','all','old','infuriating','by','life', 'space', 'a', 'the', 'quite', 'one', 'not', 'O', 'my', 'just', 'light', 'his', 'an', 'this', and 'None'.
6. rating_numerator column not properly extracted from text column (only integer part extracted).
7. Confusing ratings for both rating_numerator and rating_denominator from text column in index(313, 342, 516, 784,1068,1165, 1202, 1662, and 2335), multiple ratelike expressions like 'account started 11/15/15' at index[342], '9/11 search dog' at index[784, 1068], and 'she smiles 24/7' at index[516]).

image-predictions.tsv

1. Complex column names.

Twitter_archive_enhanced.csv

1. Retweets not needed.
2. Rows without images not needed.
3. Many irrelevant columns.

Tidiness issues

1. doggo, floofer, pupper, and puppo columns in twitter-archive-enhanced dataset should be one column instead of 4.
2. favorite_count and retweet_count columns in df_twit_api (json.txt) should be part of the Twitter_archive_enhanced dataset.

Step 3: Cleaning data

Copies of the original pieces of data were made prior to cleaning. All the issues identified in the assessing data phase were successfully cleaned using Python and pandas after which a tidy master dataset with the necessary pieces of gathered data was created.

Step 4: Storing data

The cleaned dataset was stored as 'twitter_archive_master.csv' using the pandas dataframe to_csv function.

Step 5: Analyzing and visualizing data

A new column year was created from the timestamp column and analysis and visualizations were made on the dataset.

Step 6: Reporting data

The wrangling efforts made, analysis, and visualizations are well documented.