Report: Wrangling Efforts

Introduction

In this report, the steps of data wrangling process and the efforts of data wrangling are documented. Three pieces of data are gathered using three different methods. The datasets are then checked for cleanliness, trimmed, and cleaned for analysis. Each steps are documented and each cleaning decisions are justified. This report is framed as an internal document.

In this project, three datasets are explored.

- **Enhanced Twitter Archive**
- **Image Predictions File**
- **Additional Data via the Twitter API**

Steps of Data Wrangling

The steps of data wrangling process are as follows:

- Step 1: Gathering data
- Step 2: Assessing data
- Step 3: Cleaning data

Gathering Data

Using Python and its libraries, data is gathered from a variety of sources. The method required to gather each data is different. The first step of gathering data is importing various Python libraries.

- pandas and numpy are required for the overall data wrangling and analysis
- matplotlib to produce visualizations
- requests and os libraries are imported to download image_predictions.tsv file programmatically
- time, json, tweepy, and OAuthHandler are imported to gather data for tweet_json.txt file

Enhanced Twitter Archive

The twitter_archive_enhanced.csv file is downloaded manually from the given project resources. It is then uploaded to a Jupyter Notebook and read into a pandas DataFrame (named twitter).

Image Predictions File

The image_predictions.tsv file was hosted on Udacity's servers and it is downloaded programmatically using the Requests library and a given url. The file is saved on the computer and opened in wb or write binary mode. It is then read into a pandas DataFrame (named image).

Additional Data via the Twitter API

Each tweet's retweet count and favorite count are gathered. Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API for each tweet's JSON data is queried with Python's Tweepy library. Then, each tweet's entire set of JSON data is stored in a file called tweet_json.txt file. The file is then read line by line into a pandas DataFrame (named tweet_list).

Twitter API keys, secrets, and tokens were required for quering data from Twitter API. These keys are not included in project submission.

Assessing Data

After gathering all three pieces of data, visual and programmatic assessments are performed for quality (content issues) and tidiness (structural issues). In this process, 11 quality issues and 5 tidiness issues are detected and documented. Two types of assesments are conducted:

- **Visual assessment:** three of the tables - twitter, image, and tweet_list - are displayed to their entirety to get acquainted with the meaning of the datasets. Additionally, twitter-archive-enhanced.csv and image-predictions.tsv files are assessed in Google Sheets and tweet_json.txt file in text editor Atom.

- **Programmatic assessment:** pandas' functions and methods are used to assess the data.

Cleaning Data

In this step, data cleaning is performed to solve the issues that were documented in the assessing section. The original datasets are copied to before performing any cleaning. All the necessary cleaning are done on the copied datasets. It prevents losing any original data while coding. In the cleaning process, a define-code-test framework is followed and every step is documented clearly. Finally, every piece of cleaned data is merged to create a high quality and tidy master pandas dataframe.

Define

Define section describes how each problematic data is cleaned using pandas functions and methods.

Code

In the coding section, necessary functions and methods are applied to solve the quality and tidiness issues of the three datasets. After that, the twitter_clean, image_clean, and tweet_list_clean tables are merged to create a final table (named df_master).

Test

In this step, the codes written in the code section are tested to check if all issues are solved. The final look of the master dataframe is also checked. The master dataset is then stored to a csv file

(named twitter_archive_master.csv). The data is now ready for conducting analyses and visualizations.

In [ ]: