Josh King
Data Wrangling Project

Introduction

This project's focus was the gathering and cleaning of data related to the WeRateDogs Twitter account using a number of Python libraries. WeRateDogs provides numerical ratings for posted images of dogs (on a score out of 10 with the numerator generally higher than 10). This provides a unique opportunity to investigate which factors result in more popular metrics for the posted dogs (via rating, # of retweets, # of likes, etc.). The code used, along with explanatory comments, can be found in the wrangle_act.ipynb file included with this document. A snapshot, however, of the process and end-results of the project are described below.

Gathering of Data

The data utilized in this project was compiled from the following 3 sources:

- **twitter_archive_enhanced.csv** - A pre-prepared file including information about individual WeRateDogs tweets (such as the tweet's text, the WeRateDog rating, etc.).
- **Image_predictions.tsv** - A file downloaded programmatically from a provided link on a Udacity server using the 'requests' Python library. This file housed the predicted entity type in the image associated with each tweet and whether or not it was a dog (determined via a machine learning process).
- **A Twitter API data pull** - Using Twitter's API and the 'tweepy' Python library, we were able to pull two further metrics about each tweet-- the number of likes and retweets. The information pulled was matched from the twitter API using the tweet id's provided by the twitter_archive_enhanced.csv file and stored in a new file called **tweet_json.txt**.

The prepared csv and tsv files were opened using Pandas and stored to a dataframe. The tweet data from the API data pull was parsed using the 'json' Python library and then also stored to a Pandas dataframe.

Data Assessment

The data was inspected for both quality and tidiness concerns using a number of visual and programmatic techniques. Visually techniques such as head(), tail(), and sample() were used. The data was also explored programmatically with methods like describe(), info(), value_counts(), as well as other methods used to produce operations such as looking for duplicates, abnormal values, and checking data types.

Data Cleaning

The data issues identified as well as the means of cleaning each issue are identified below:

- Quality Issues
    - Not all tweets were original - Tweets that were replies or retweets were dropped in all dataframes.
    - Not all tweets were about dogs - Tweets identified to not be about dogs using the image predictions dataframe were dropped across all others.
    - Not all tweets are still available - Some tweets could not be downloaded as they were found to be deleted during the API pull and as such were dropped from all data frames.
    - The csv file's date did not import as the right data type - This was changed to date-time.
    - A number of tweets had denominators that weren't 10 - These were changed to 10 since we know this to be true of all WeRateDogs ratings.
    - A number of tweets had abnormally high numerators - These were inspected manually to determine whether they should be dropped (from all dataframes) or manually repaired.
    - The 'source' column in the csv file included unnecessary html strings - These were parsed out using a string splitter.
    - The image predictions database had an inconsistent capitalization scheme - These were set to all lower case.
    - Erroneous names existed in the name column of the csv file - These were replaced with "Unknown".
- Tidiness Issues
    - The API data did not need to be its own dataframe since it included metrics about individual tweets, so it was merged with the dataframe from the csv file.
    - The twitter archive csv file had multiple columns for the dog stage type which were merged into a new column ('dog_stage') to be more tidy.
    - Unnecessary columns now blank (such as those regarding retweets or replies) were dropped.
    - All data needed for analysis was combined into one master dataframe and exported as **twitter_archive_master.csv**.

Data Analysis & Results

The resultant twitter_archive_master.csv was analyzed for which dogs evaluated over three domains-- dog stage, dog type, or dog name -- received the highest ratings, number of likes, or number of retweets. Generally the 'dogg' stage was found to be somewhat more popular overall, the dog type varied in popularity depending on the metric viewed, and the distribution of top dog names was almost flat.