We Rate Dogs Project: Data Wrangling

Table of Contents

Data Gathering

The data used for analysis was gathered from 3 different sources using different techniques.

- The Enhanced Twitter data was read directly using the pandas read_csv function. The data was provided and downloaded manually.
- The Image Prediction Tab Separated file. This file was downloaded programmatically using the requests library. It was also read using the read_csv function with the delimited specified as tab.
- The third dataset, tweet_json.txt was collected from Twitter using Tweepy. The data collected using the tweet id in the enhanced twitter data into the text file. The text data was read, line by line into a json file which was then converted to a dataframe, json_df

The Enhanced Twitter data was used as the main source of data. In addition, 2 columns: Retweet Counts and Favorite Count were added from Twitter data, json_df. From the Image Prediction data, the predicted dog breed was also added after wrangling.

Accessing Data

The data was explored visually and programmatically for possible issues/inconsistencies in the data. Upon accessing the data, some quality issues and tidiness issues were found.

All the issues identified (Quality and Tidiness) were cleaned in the data cleaning phase.

Data Cleaning

The data cleaning stage consists of 3 sub-stages: Define, Code and Test. For each of the issues identified above, the Define-Code-Test method was used as a procedure for cleaning the dataset.

Quality Issue 1: The timestamp column is not in the correct format.

The data was converted to timestamp using regular expressions and the python datetime method. Upon cleansing, the few rows of the column is checked.

Quality Issue 2: There are some invalid names in the dataset.

There were some invalid names such as “A” and “None” contained in the dataset. These names were removed and replaced with missing values.

Quality Issue 3: The source column is enclosed in html tags.

The source column contained strings enclosed in html tags which were removed using string operations.

Quality Issue 4: The data should not contain retweets/replies.

The rows of the data containing retweets and replies were removed because the focus is on original tweets.

Quality Issue 5: All Retweet and Reply related columns are dropped.

The retweet and reply related columns were also removed since the data is focusing on original tweets.

Quality Issue 6: The Expanded urls contains 3 missing values.

The rows containing missing values in the expanded url column were removed.

Quality Issue 7: The Retweeted and Favourite columns should be added to the dataset.

The retweet and favorite counts in the json dataframe were combined to the twitter archive dataframe.

Quality Issue 8: Some images were not correctly classified as dogs.

The image prediction data contained 3 possible predictions of dogs breeds. However, some of the predictions were not dogs. For cases where all predictions are not dogs, the brredPredict column was replaced with np.nan.

Tidiness Issue 1: Combine Numerator Rating and Denominator Rating to one Rating column

The rating columns were combined into one column where the denominator is uniform, over 10. Some numerators were above 10, which is not invalid in this case.

Tidiness Issue 2: The dog stages are spread across several columns.

The 4 dog stages in separate columns were combined into one column using string replace and concatenation. For columns with no stage recorded; they were replaced with np.nan.

Data Storing

The data collected from the 3 sources specified above were combined into one dataframe and saved as a CSV file named: *“twitter_archive_master.csv”*