Wrangle Report

The dataset to be wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering data

1. Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. The csv file provided is loaded using pandas dataframe.

2. Image Predictions File

This file image_predictions.tsv is hosted on Udacity's servers and is downloaded programmatically using the Requests library. It is then stored as a tsv file, which is loaded using pandas dataframe

3. Additional Data via the Twitter API

Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. Load the data to a datframe, we need only the tweet_id, retweet count and favorite count.

Assessing data

Here the data that was gathered is assessed both visually and programmatically for quality and tidiness issues. Analysis requires only the tweets with original ratings with images, no retweets and reply tweets to be retained.

Quality: issues with content. Low quality data is also known as dirty data.

Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data.

The following assessments were made:

Quality

`df_twitter` table

- 181 Retweeted data
- 78 reply tweets
- Source column with HTML tags
- Expanded URLS have 59 missing values
- Timestamp in string format
- Numerator and denominator rating in integer format
- Numerators with decimal values, need to see how many of them are retweets.
- Denominator rating other than value 10

`df_image` table

- Only 2075 tweets have image
- Entries where "p1_dog" has a False value
- The "p1" and "p1_conf" column names are not explanatory

Tidiness

- In df_twitter table dog's stage dogoo, floofer, pupper, puppo is spread in different columns
- df_image , df_info_json needs to be joined with archive table df_twitter
- columns associated with retweets and reply tweets
- p2,p3 related columns is not of use, as we take most confident predication

Cleaning

At first, we create a copy of the three data frames.

For cleaning the data, we have followed the following process:

- Define: convert our assessments into defined cleaning tasks.
- Code: convert those definitions to code and run that code.
- Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

After cleaning these dataframes were combined to a single datset. And made sure that this master dataframe didn’t have any retweets, reply tweets and tweets with no images.

Further insights were derived from this master dataframe using statistical and visual analysis.