Jee Naughton
Wrangle_Report
Jan 2019

This Wrangling project consists of three parts namely:

Gathering

There are 3 sets of data to assess and conclude the results. They are:

• twitter-archive-enhanced.csv. I read this file into my dataframe ultimately called “df1_clean”.

• This file was downloaded programmatically from the tsv_url = https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
It contains the image prediction file. It is called ‘image-predictions.tsv’. They took the top 3 dog breeds and run into algorithms to predict what the images that ran through it and recorded the success of predicting the correct umage of the dogs. I named this file df2_clean.

• The last data set was taken from Twitter API using the python tweepy library. This was later saved as a JSON file with UTF-8 encoding. I later named this read file as tweet_likes_df. In order to get this file, I had to create an account with Twitter which enabled me to obtain a consumer_key, consumer_secret, access_token and access_secret. Then, I would plug-in these codes into the program to run it.

Assessing

There were a number of missing data in some columns of these files. But before they could be deleted, all of these 3 files needed to be merged with a join by tweet_id.

There were several empty records in ‘in_reply_to_status’,’in_reply_to_user_id’,’retweeted_status_id’,retweeted_status_id’,retweeted_status_timestmp’.

The rating_numerator and rating_denominator had errors in them. There were considered outliers.The denominators had numbers other than 10 and numerators had some very large numbers

There were 4 kinds of dogs (pupper', 'puppo', 'doggo', 'floofer') which were in 4 separate columns. They needed to be consolidatd to one.

The number of image_prediction records and the twitter-archive records did not match. There were less records in the image prediction than the archive account. We don’t need records that do not have images.

Timestamp needed to be converted to datetime format.

Some of the dog name were erroneous. Example ‘a’,’such’, etc etc

All 3 files, twitter-archive-enhance.csv,JSon file, image-prediction files needed to be consolidated.

There were some null values in tweet_ids

Cleaning

After the files were merged, removal of rempty records from retweeted_status_id

Deletion columns of in_reply_to_status_id ,in_reply_to_user_id, retweet_status_id, retweeted_status_user_id , and retweeted_status_timestamp occurred because we do not need them.

Timestamp was converted to datetime format

Deleted any records that didn't have any images in them

Consolidated 4 unnecessary dog kinds (pupper', 'puppo', 'doggo', 'floofer') to 1 column named 'dog_kind' and deleted the original 4 columns

Deleted tweet_ids that were null.

There are several issues with the rating_numerator and rating_denominator. Namely, the rating_numerator should not be more than 30 (because we will be calculating the ratings which is numerator/denominator and it should not be more than 3). And the rating_denominator must always be consistently be 10. Those numerators and denominators were identified and the numerators and denominators were rectified accordingly. Denominators were converted to 10 and numerators were assessed to be no more than 30. Another column was 'rating' was created which contained the results of numerator divided by denominator. This would than, be used for graphs.

Store (even though it's not part of wrangling, I thought I would include it for a tight fit in the process)

The data from df1_clean was ultimatlely saved as 'twitter_archive_master.csv''.

The data that generated the JSON file was saved as 'tweet_json.txt'

The image-predictions.tsv was also uploaded to Udacity github for evaluation.