Data Wrangle_Report

Obtaining data

Three different datasets for this project are image-predictions.tsv, tweet-json.txt, and twitter-archive-enhanced.csv. The twitter-archive-enhanced.csv was directly downloaded and uploaded to the jupyter workspace before loading into a pandas dataframe in the notebook. The image_predictions.tsv was downloaded using the requests library in the Jupyter notebook and loaded with read_csv(). This is an ordinary csv-like file with tab separators instead of comma separators. The tweet_json.txt file was used from additional materials for Twitter found in the project space as I could not obtain Twitter API credentials. This file was loaded into jupyter notebooks as a text file, read line by line and loaded each line into a json object using json.loads(line).

Assessing data

Different methods of assessing the data for quality and tidiness issues were used; namely: visual assessment and programmatical assessment.

Cleaning data

The spotted issues were then cleaned and also tested. Each copy of the three data set was made for the purpose of cleaning.

Issue 1 (Quality issues): There is no consistency on the names of dogs, others starts with a capital letter and other with small letters. Image_predictions file.

I have converted the first characters of the names to uppercase in image_predictions dataframe.

Issue 2 (Quality issues): Retweets on the tweeterarchive dataframe which are not required.

I have deleted rows with retweets, it was 181 rows of retweeted_status_id.

Issue 3 (Quality issues): There are jpg_url which appears twice on the image_predictions file.

I deleted rows with duplicated links under jpg_url column.

Issue 4 (Quality issues): "None" have been used where there should be missing values, and these missing values cannot be detected with a code.

I replaced None with an empty space.

Issue 5 (Tidiness issues): Dog stages (doggo, floofer, pupper, and puppo) has been divided into four columns instead on one column on twitterarchive dataframe.
I made one column for dog stages and named it DogStages.

Issue 6 (Quality issues): There are dogs which have been classified under two stages, e.g doggo and pupper (doggopupper).

I dropped rows with dogs classified not as either pupper, doggo, puppo, or floofer.

Issue 7 (Quality issues): Dog breed names are separated by underscore, e.g German_Shepherd.

I replaced underscores with an empty space on column p1, p2, and p3 of image_predictions data frame.

Issue 8 (Quality issues): There are dogs which are not falling into any category.

I checked dogs which are not classified. Dogs which are not classified are 84.6% of the total number of dog, this number is high, therefore, non-classified dogs will not be dropped.

Issue 9 (Quality issues): Reply_to_status_id and Reply_to_user_id on the tweeterarchive dataframe which are not required.

I deleted rows with reply on status id and Reply_to_user_id.

Issue 10 (Quality issues): There are predictions which are not giving dog breed.

This were predictions and will be treated as such, it is expected that some named will not be dog names.

From this project I have learnt that data does not always come clean, and there are steps which needs to be followed before getting data to a stage where it can be useful. The process of preparing data is called data wrangling. This process involves gathering data, assessing data, and cleaning data. When cleaning data it is important to start by checking missing values on the dataset.