WE RATE DOGS DATA WRANGLING PROJECT BY CHINEDU UZORUE

The data wrangling process is an important aspect of the data analysis process. This is because it forms the input to the actual data analysis. Hence, any low-quality or messy data gathered here will not serve the intended purpose of deriving insightful information from the data.

I carried out data wrangling on the Weratedogs datasets provided by Udacity for the Data analysis Nanodegree.

The process typically involves 3 parts:

1. Gather data
2. Assess data
3. Clean data

1. Data Gathering

For this project, data gathering involved 3 parts:

- Direct file download using the provided link (twitter-archive-enhanced.csv)
- Use the **requests** library to download the tweet image prediction (image_predictions.tsv) and
- Using the Twitter tweepy API (I was not able to use this method as my Twitter Developer Portal application was not approved, so I had to use the included tweet-json.txt file).

2. Assessing the data

After gathering the data, visual and programmatic assessment were next.

I visually assessed the data using a spreadsheet application to look for some data quality issues and tidyness issues. For the programmatic assessment, I made use of the pandas library's standard methods like describe(), info(), and sample() to assess the three (3) datasets. The following issues were identified:

Quality Issues:

General

1. The tweet_id column type is integer instead of string (object type)

Image Prediction file (img_prd)

1. There are only **2075** entries instead of **2356** (missing values issue)
2. In the **"p1", "p2", and "p3"** columns, the names are separated by an underscore "_" instead of a whitespace (validity issue)
3. In the **"p1", "p2", and "p3"** columns, some breed names begin with Upper case letters whereas some begin with lower case letters (consistency issue)

Twitter API file (twitter_data)

1. There are only **2354** values instead of **2356** (missing values issue)

Twitter archive file (tw_archive)

1. Some names are invalid e.g "a", "an", "None", "incredibly", "quite" etc
2. (a). Row with tweet_id **"666287406224695000"** shows a rating_numerator of 1 and rating_denominator of 2, however on careful

observation, the text states that the rating is 9/10, so rating_numerator should be 9 and rating_denominator 10
(b). Row with tweet_id "835246439529840000" shows a rating_numerator of 960 and rating_denominator of 0 i.e rating is 960/0, this is invalid, as stated in the text column

3. The timestamp column has trailing "+0000"
4. The timestamp column type is string instead of datetime
5. There are **181** tweets that are retweets and **78** replies to tweets that need to be deleted

Tidiness issues

1. Four redundant columns in tw_archive dataframe - doggo, floofer, pupper, and puppo - This goes against the second condition for tidy data (every column must be a variable)
2. The 3 dataframes are all related but separated which is an example of messy/untidy data as documented in the tidy data documentation - *A single observational unit is stored in multiple tables.*

3. Cleaning the data

After identifying the data quality and tidyness issues, the next step was to make copies of the datasets before cleaning. The cleaning process involved the use of some pandas methods as well (e.g melt() method). After having cleaned copies of the datasets, I merged them together and performed some analysis on the final dataframe.