PROJECT: WRANGLING AND ANALYSE DATA

The main aim of this project was to wrangle data to produce high-quality and tidy data which would eventually be analyzed to produce at least 3 insights and 1 visualization. Wrangling data consists of three steps: gathering, assessing, and cleaning. In this project, data was gathered from three different sources, in three different formats, the individual datasets were assessed and 11 quality issues and 2 tidiness issues were found. The issues found were cleaned, and the individual datasets were merged into a single clean dataset.

Gathering Data

The data gathering step included**:**

- Loading **a file on hand, *twitter_archive_enhanced.csv,*** the tweet archive of Twitter user @dog_rates (or WeRateDogs) provided on Udacity into a pandas dataframe called **weratedogs_df**.

- Next data was downloaded from the **internet** using the **requests library** and stored in a file called ***image_predictions.tsv*.** The file was opened and the data loaded into a pandas dataframe called **image_predictions.**

- Lastly, data was downloaded from Twitter through **Twitter API**. Using the tweet IDs from the WeRateDogs Twitter archive, the Twitter API was queried for each tweet's JSON data using Python's **tweepy library** and each tweet's entire set of JSON data was stored in a file called ***tweet-json.txt***. This file was opened and loaded into a pandas dataframe called **twitter_api**.

Assessing Data

• This **weratedogs_df** dataset consisted of features such as the tweet ID, retweet observations (retweeted_status_id, retweeted_status_user_id, and retweeted_timestamp), the rating for each dog (rating_numerator and rating_denominator), and dog stages (doggo, puppo, pupper, and floofer) amongst others. It had a total of 2356 rows and 17 columns.

• **image_predictions** dataset included features such as three image predictions (p1, p2, and p3) along with the confidence of the predictions (p1_conf, p2_conf, p3_conf), and features which returned True/False if each prediction was for a dog (p1_dog, p2_dog, p3_dog). It also included the tweet IDs, image URL, and the image number corresponding to the most confident prediction (this had numbers 1 - 4). It had a total of 2075 rows and 12 columns.

• The **twitter_api** dataset features were the tweet IDs from weratedogs_df, along with the retweet and favorite (likes) counts for each tweet ID. This dataframe had a total of 2327 rows and 3 columns.

The individual dataframes were assessed **visually** and **programmatically** for quality and tidiness issues, and the following issues were documented.

Under **quality issues, for** weratedogs_df:

• There were "None" values in the dog stages columns (doggo, floofer, pupper, puppo) instead of null values.

• Some rows had both doggo and pupper dog stages.

• Some rows had both doggo and puppo dog stages.

• Some rows were retweets.

• The tweet_id column data type was integer.

• The data type for the timestamp column was string.

• The data types for in_reply_to_status_id and in_reply_to_user_id columns were float.

• Some rating_denominator values were greater than 10.

For image_predictions:

• The tweet_id column was also integer.

• Some img_num rows had values of 4, but there were only three predictions (p1, p2, and p3) in the table.

• Some rows present were not for dogs. That is, the most confident prediction concluded the images in those rows were not dog images.

For **tidiness issues,** in weratedogs_df**:**

• The various dog stages (doggo, floofer, pupper, puppo) were column names.

• Retweet observations (retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp) which are separate were found in the table.

Cleaning Data

First copies were made of each dataset, and the various issues documented above were cleaned beginning with the first three quality issues to make addressing the first tidiness issue easier. Next, the second tidiness issue was cleaned, followed by the other quality issues. Afterwards, the three datasets were merged to form a single clean master dataset.