Wrangle Report

As I was wrangling the data from the 3 different data sources containing tweet information, there were lots of quality issues that needed to be sorted out in order to analyse the information.

I have numbered them to highlight them individually.

Image Prediction file

The first was the Image Prediction file that first had to be downloaded from a web page into a CSV using the requests.get() method and then opened with the pandas .read_csv() method to make the first dataframe. The data was mostly clean(especially compared with the others) but:

1) The columns were not very clear in meaning. There were columns such as 'p1' and 'p1_conf', and that aren't clear at first glance. P1 is the Algorithms Number One Prediction for which dog is in the image. 'Conf' is the confidence level of that prediction. So, I renamed them 'Prediction1/2/3' and 'pred1/2/3_confidence' to make it clearer/understandable.

Twitter Enhanced Archive File

For the Twitter Enhanced Archive File:

2) There were many nulls making empty columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp')So I dropped the columns from the dataframe.

3) There was also the issue of many dogs stages ('pupper, 'doggo', 'puppo', and 'floofer') names not being filled in in the columns. I could only see that when the data was visualised. They were not discovered as Null values because they had a string 'None' in there instead. And as most were not filled in, I have decided to disregard them from my analysis.

Tweet JSON File

For the Tweet JSON file, there were many more quality issues.

4) The 'display_text_range' column had the data in a list of two integers, with the first always being 0 and the second being the number of characters. So to analyse this, I made it into two columns and only kept the second integer value as this was what is needed to analyse the character lengths.
5) There were also nested dictionaries as values in the columns such as 'entities', 'extended_entites' and 'users'. I got rid of the 'entities' and 'extended_entites' columns as I didn't want them, but the 'user' column could be interesting so I 'flattened' that, turning it into its own dataframe temporarily and then matching it back with the tweets dataframe.
6) There were also lots of outlier values in the favourite count and retweet count which could affect the average values in our analysis, and also potentially suggests mistakes have been entered for the extreme values.

Master Tweet Archive

I merged the dataframes on the tweet_id column as this was the unique identifier that linked all the records and ended up with the final datafame containing all the information. On further examination:

7) There were lots of unwanted/not-wanted columns which I then removed.
8) While going to look for trends with the type of dog (based off the number one prediction) with the amount of retweets and favourites, I also noticed that for the image algorithm's prediction, it was guessing things other than dogs which we want (e.g. 'ice lolly', 'toyshop' and 'seatbelt'), which is a data quality issue for us.

<u>Tidiness Issue's</u>

1) In the Twitter_JSON file, we also find there are different units of observation in the same table, i.e. information for both Tweets and for Users, with the user information held within a nested dictionary inside the 'User' Column which is a data tidiness issue.
2) We can see that the attributes related to the tweet itself needed for the analysis a such as the counting retweets and favourites data is held in different dataframes which must be merged at a later point for analysis.