Wrangle Report

It is worth mentioning that the steps to cleaning the data are not in numerical order as this was not necessarily the easiest way of representing the steps before.

Starting with Quality 1 I have looked here to drop columns that contained all NaN values as these would not obviously be useful in any further data.

Step 2 saw dogs with names ‘a’ or none being replaced with Null values as it would still be useful to visualise these later on.

Step 3 saw rows showing retweets being dropped as per the requirements of the project as these were not to be included.

Quality 5. And Tidiness 2 next (sorry for the confusion) saw me using the astype() function in order to address some simple format changes in the tweet IDs of each table to string. This was needed to do at this stage so that any later joins could be run. I also used a pd.to_datetime(df[‘col’]).dt.date to split the full timestamp into both a date and time column in order to make this more manageable to read with the naked eye.

Step 4 saw a quality issue with expanded URLs that seemed to have more null values than other columns. After inspection, I noticed that these had no corresponding images and could therefore be retweeted/deleted. In order to mitigate these I completed a left join from the general df and the images df. While doing this I also joined the resulting table to the share count table with a left join.

For the next steps, I made sure to create a column of the dog types as in the uncleaned DF this was spread across 4 columns and a master table will most efficiently show this data within one column. This was done using a simple

dftext.str.extract method. I also addressed my first quality point in this process by representing this in a column with a decimal percentage, e.g. 87% = 0.87. (Later Analysis showed decimal percentages to be the most effective for statistical analysis)
Next, within steps 7 and 8 I ran the p1.sre.replace to replace any _'s with spaces in the dog breed column and also shows the prediction accuracy as a percentage as to me this would be easier to visually assess when looking at the table or any average prediction values later in my work.
Finally, I completed a simple column rename on the columns that existed within the data frame to more readable headers and also rearranged the columns in an order that I found to have the most use, starting first with tweet id. (What can be considered the primary key of the dataset and therefore the value used in any earlier joins)