Wrangle_report

In this report, I will briefly discuss the data wrangle procedure and the thought behind some of the steps.

After the initial assessment through both visual and programming, I have first tackled the tidiness of the tables. As the required report was on the original tweets only, all the retweets and replies from df_tweet_archive and df_json were deleted. For the latter, quoted tweets are removed as well.

Columns containing this information (retweet_id etc.) are subsequently dropped. Next, I have reduced four columns that described four stages of a dog ('doggo', 'floofer', 'pupper','puppo') into one column and replaced the 'None' entries to NaN. Some rows have multiple stages of dog, and for these, I have separated each term with a comma. This completed the tidying up of the tweet_archive table. For the df_json table, irrelevant tweets were dropped first, as mentioned previously. Selected information (retweet and favourite counts) were then joined to the tweet_archive table. For this inner join was used to make sure that there are no rows with missing values. Lastly, for the image_prediction table, I have added an extra column with the correctly predicted breed of dog. There are entries with many 'True' value for the breed, in which case I have picked the one with the highest confidence.

Completing the tidiness issues, I then moved on to clean some of the quality issues in tweet_archive table. Firstly, I have changed the data type of the timestamp column to DateTime. In the next step, I have sliced out redundant URL information included in the text column, as the extended URL column has already existed. Then I worked on the 'name' column. Tweets without the name of a dog were listed as None. However, there are cases in which some word (such as 'this', 'a', 'quite' etc.) were wrongly listed as a name. All of these entries were changed to NaN. There was also a name O'Malley listed as 'O', and this was also fixed. I have then tackled on 'rating_numerator/denominator' columns, which had several different quality issues. The initial extraction of rating from the tweet did not work when the text included fraction notation (9/11, 4/20 referring the date, etc.). Using a regular expression, I have extracted the correct rating and updated the respective entries. The next problem was that some rate had decimals and these were not reflected correctly in numerator column (e.g. 9.75 /10 would be listed as 75/10). I have firstly changed the datatype of the numerator column to be a float. Therefore there would be no data type mismatch when updating these values. Regular expression was once again utilised to extract and correct the rating. Finally, there were tweets rating multiple dogs which gave ratings that were not out of 10, but out 80 for eight dogs, for example.

I have created a new 'rating' column by dividing numerator and denominator to solve this issue. I then have checked for outliers, which were: a tweet without a rating, rating on a rapper Snoop Dog (rated 420/10) and independence day dog (rated 1776/10). These entries were removed, additionally 'rating_numeartor' and 'rating_denominator' were dropped.

Finally, I worked on the image prediction table. I first capitalise and removed underscore and replaced with a space. Then I have added the predicted dog breed from the image prediction table to the tweet_archive table and created tweet_archive_master.