WRANGLE REPORT

This briefly describes my wrangling efforts

The datasets used for this wrangling project were twitter_archive, api_df, and img_df datasets. After data gathering process, I assessed the datasets visually and programmatically. During my visual assessment of the twitter_archive dataset, I found various quality issues and tidiness issues. The rating_denominator column had erroneous data as some of the tweets had denominator ratings below or above the standard rating of 10, this had to be rectified. I also found out that the rating_numerator column contained outlier ratings, such as a rating of 960. The most frequent numerator ratings were between 7 and 15. During my programmatic assessment, I used the describe method to have an idea of the numerator statistics. It was found that the mean rating of the rating_numerator column is 13.1 while the 75th percentile was a rating of 12, therefore, outlier numerator ratings had to be discarded. With the use of programmatic assessment, I intended to find out why some ratings were too high to be true. I found out that some tweets had numerator ratings in decimal number form and instead of returning that decimal number, only the numbers after the decimal point was extracted as the tweet image numerator rating. I also found out that some images contained multiple dogs, which multiplied the rating by the number of dogs in the image. All these had to be rectified to have a cleaner rating_numerator column.

I used the .info () method on twitter_archive_clean dataframe to find out the columns with null values and their data types. The timestamp column was found to have the wrong data type. 5 columns had null values but these columns were easily dropped because they weren't needed for my analysis. Two tidiness issues were also found in the twiter_archive dataframe. By visually inspecting the dataframe, I discovered that 4 columns (doggo, puppo, pupper, and floofer columns) were actually representing the same thing, which is the dog stage name. These 4 variables are meant to be a single column which I eventually named dog_stage_name. Also by visual assessment, I realised that the retweet_count and favorit_count columns in api_df are meant to be a part of the twitter_archive. All of these were taken care of during the merging.

The other two dataframes (api_df_clean and img_df_clean) had no null values but they both had a lesser number of rows when compared with the twitter_archive_clean dataframe. This suggested that the twitter_archive dataframe had some rows with tweets identified by their tweet_id that weren't present in the other dataframes. These rows will eventually be dropped after all datasets were merged together to form the twitter_archive_master dataframe. These rows were dropped because they were only a small part of the overall dataset which means they wouldn't affect my analysis.

Before merging, I had also visually assessed the img_df_clean and api_df_clean dataframe. I found no quality or tidiness issues in the api_df but the img_df had two issues. The first issue was that the p1, p2, and p3 columns which represented the predicted dog's breed have some names starting with lower case and some with upper case. The second issue was that there was no column that stated the prediction result, that is, whether the tweet image is a dog or not. Both issues were rectified in the cleaning stages.