wrangle_report

February 17, 2019

0.1 Wrangle Report

After collecting the data for this project and assessing it, a list was generated for each data set for appropriate cleaning steps.

WeRateDogs Twitter Archive A number of features didn't really add anything to the data, or worse - could actually take away from the analysis, as is the case with retweet observations. The columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestampe', 'source', and 'URL' have been dropped. Some feature data types have been adjusted, 'timestampe' was changed to a datetime dtype and 'doggo', 'floofer', 'puppo', and 'pupper' were changed to catagorical. The 'text' feature name was changed to 'description'. And lastly the rating columns were normalized with any observations containing a zero in either the numberator or denominator were dropped.

Tweet Image Predictions 'p1', 'p2', and 'p3': I don't care what the top three predictions for dog type were, I only really care which was the actual breed choosen in the end, which in theory will always be the first dog listed (highest probability or first in list if misidentified at first). There was some formatting for capitol letters in the dog breed name. Changed the '_' and '-' in breed names.'p1_dog', 'p2_dog', and 'p3_dog' offered absolutely nothing so the were dropped. All mislabeled breeds like 'dishwasher' etc were removed and consolidated the breeds into a single column of the highest probability, if all three predictions in the row were not dogs, the observation was dropped. The 'img_num' feature was renamed and changed to catagorical.

Twitter API Query There wasn't a lot of cleaning activities for this data frame. The data was complete (all observations were filled out), using the describe() function, none of the values were terribly out of place at first glance, the features are all appropriate data types, and there are no duplicate tweet ID numbers (key for merging). The only thing "cleaned" in this dataframe was renaming the feature 'id' to 'tweet_id' to match the other data sets during merging.