Data Wrangling Report

Dana Cody

Gathering the Data

There were three data sources that I gathered from; a csv file that was given from the start, a tsv file that I had downloaded programmatically, and json data that I had queried twitter's API for. The csv file (**twitter-archive-enhanced.csv**) was originally @dog_rates twitter archive, but it had been slightly wrangled. The tsv file (**image_predictions.tsv**) contained dog breed predictions for each tweet with a picture as generated from a neural network. The data from twitter's API was extracted, and written to a text file (**tweet_json.txt**) in json format. This txt file really just contained each tweet's tweet id, favorite count, and retweet count.

Assessing the Data

I used pandas .describe() , .value_counts() .sample() and .info() mainly to assess the data. I didn't quite know what 'one' quality or tidiness issue was so I had to make some executive decisions. The issues I found with the data were as follows:

Quality Issues

1. In the tweet_archive dataframe, the following columns are of the wrong data type: tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and timestamp.
2. In the additional_tweet_data dataframe, and doggo_pred tweet_id is also of the wrong data type.
3. There are some entries that are not pictures of dogs, typically rated less than 10.
4. There are a few tweets that have been deleted. I.e missing API data
5. We are missing data for the dog breed predictions.
6. Some names in the name columns can't possibly be dog names. (a, the, an, this...)
7. We only want original ratings. So we don't want tweets replying to another tweet.
8. The following variables are of the wrong data type: type, p1, p1_dog, p2, p2_dog, p3, p3_dog
9. There are some missing dog types in our dataset.

Tidiness Issues

1. The additional_tweet_data dataframe, the tweet_archive dataframe, and the doggo_pred dataframe should all be one dataframe.
2. The doggo, pupper, puppo, and floofer columns are possible values, and should all in one column.

Cleaning the Data

The actual cleaning of the data was probably the most time consuming part of the entire project but also again the most helpful. Having to clean a dataset with little to no guidance on how we perform these tasks was fun and challenging. I think there were definitely some tasks where I made it more complicated than it had to be. There were certain tasks where I struggled to really fix the issue, (issue 3) but it would take way too much time to really fix them accurately.