Data Wrangling

Data Wrangling effort for this project was done in the following three parts:

- Data Gathering
- Data Accessing
- Data Cleaning

Data Gathering

We gathered three different data frames from various sources. The first data frame was given to us in the form of a csv file. This data frame has relevant information about the status of the dog and the ratings. Next, we gathered a data frame by sending a request to a website. The response was written in a tsv file. A tsv file is very similar to csv with an exception of using the delimiter as tab instead of a comma. Finally, queried Twitter API for each tweet in the Twitter archive and save JSON in a text file. Using relevant information in the json file in order to create the third data frame containing retweets and favorite count. We use the dataframe, and save it as a text file which can be later used for reading the data frame.

Data Accessing

For the first given data frame twitter_archive_enhanced, there are multiple attributes for the status of the dog. On careful visual inspection, it is noted that the text column has information about the relevant status of the dog. Some records may have multiple status, but we extracted the correct status(es) from the text attribute. Moreover, on visual assessment it was inferred that their were some names of dogs like 'a','an','the' and more similar words which aren't names and are listed in lower case. We need to remove the records of the retweets and replies. The denominator with value as 10 can be kept to maintain consistency, and easy for feature engineering. Some values in the numerator rating seemed like outliers, and required case by case analysis. For some of the cases, the rating was provided in the text. The second data frame on image prediction has attributes with prediction of the dogs breed, and the confidence of the prediction. The highest prediction of dog breed seems to be a significant attribute. And finally, the tweet_data drataframe's user_favorite attribute looks redundant and does not provide us with any useful insight. Lastly, made sure the tweet_id for each data frame is not duplicating.

Data Cleaning

We created fresh copies of the three data frames in question. We added a feature 'status' which extracted the correct status of the dog, and removed the four attributes, making our dataframe tidy. This feature is converted to a category data type as it has 5 categorical variables. Removing all names of dogs which were't names, and replacing them with Null Values. Removed the records of the retweets and replies. Feature engineered a new column which calculates the rating for each dog using the numerator and denominator. Some incorrect values were hardcoded, as they were extracted from the text. In the image prediction data frame, only focussed on the highest correct prediction for dog, removing information about less useful predictions. Also, converted this attribute into a category which will be later used in analysis. Removing the records for the retweets and replies. Dropping multiple columns in the three dataframes as it does not provide any useful information for our analysis and visualizations. Finally, merging all three data frames based on the tweet_ids, and saved it in a csv file.