Wrangle Report

The dataset used is WeRateDogs from twitter. WeRateDogs is a twitter account that rate's people's dogs with a humorous comment about the dog.

The very first step was to gather data from 3 different sources. File twitter_archive_enhanced.csv was provided by Udacity. This file has tweet information like tweet text, dog name, dog stage. There are 2356 records in this file, only those records which has rating for the dogs.

Next from file image_predictions.tsv. The images of dog were ran through a neural network that can classify breeds of dogs.

And querying twitter API for each tweet's JSON data using Python's Tweepy library to fetch each tweets retweet_count and favorite_count. Storing each tweet's entire set of JSON data in a file called tweet_json.txt. The greatest challenge was to learn tweepy library. It was hard to make breakthrough that the result of a function which fetches the tweet information from twitter for a tweet_id has to be json serializable by adding a small piece of code.

Then reading all the files into pandas dataframe and then assessing the data visually and programmatically. Visual assessment led to some quality and tidiness issues. Here the most difficult part was to correct the wrongfully recorded dog names. First to identify the pattern that the recorded dog name is correct or not. And then identifying the pattern in text to search for correct dog name. And then there were some records where the text didn't have any dog name in the tweet. Then searching for those tweets and updated corresponding dog names to None.

For rating numerator, there were few records where the numerator was in decimals but the value was incorrectly recorded with only digits in decimal place. Finding those tweets with decimal places and then updating the numerator with correct values. And there was a tweet which had a fraction which was not the rating for dog, this appeared in the statistical description, in min val for denominator column, searched for that specific record and updated numerator and denominator with correct values.

There was a column timestamp for recording tweet's date and time. Both the date and time were in the same column separated with a space. I extracted only the date part from the column and stored it in a separate column for analysis part.

There were 4 columns for 4 stages of dog. Merged those 4 columns into 1 column stage and assigned values appropriately.

Converting the column datatypes appropriately where ever required. And renaming the column id to tweet_id in one of the dataframes.

Deleting the retweets from all the dataframes. Dropping the unwanted columns from all the dataframes.

And then finally merged all the clean dataframes, and saved all the dataframes along with the merged one to new files.

This completes the wrangling part of the dataset.