Wrangle Report

The data looked at in this project is from the Twitter page, WeRateDogs. The wrangling done on this project goes through the 3 steps of: Gathering, assessing, and then cleaning. The data was gathered from 3 sources:

1. The Enhanced Archive given to us which has basic information about the tweets.
2. Querying the Twitter API using Tweepy to collect additional information about the given tweets.
3. An Image prediction neural network tool used to predict the type of dog breed which is stored on the Udacity servers.

Gather

The standard Python libraries, NumPy, Pandas and matplotlib are imported and used to do the data analysis.

The Enhanced Twitter Archive is given as a CSV file and read in using pandas read_csv function.

To gather data from the Twitter API we create a twitter developer account and once that is complete, we are given 4 codes that give us permission to access the Twitter database. These 4 codes are the consumer key, consumer secret, access token, and access secret. These codes are passed onto Tweepy which is then used to query twitter and search based on the tweet ID given to us in the Enhanced Archive List. The data that is returned is stored as a JSON file as this is a quite easy to use format to read data. This portion takes a long time because Twitter has a rate limit on how many requests you can make per 15 minutes. From the 5000+ tweets that WeRateDogs has we pulled 2331 tweets which have been run through the image prediction tool.

Finally, the Image Prediction file is gathered by using the “requests” library and downloaded using a URL link given. The resulting file downloaded is a tab separated file (TSV), so it is read in through pandas read_csv function with the separator as “\t”

Assess

Starting with the Enhanced Archive the data is assessed visually and then programmatically. We get familiarized with the data by exploring the columns in the data frame, checking for columns with missing data, duplicated data, data types used, and looking for quality and tidiness issues.

We do the same for the downloaded Tweepy data and realize we have a lot more columns to explore. We then decide which columns would be interesting to analyze and consider which ones to drop.

Lastly the same is done for the image prediction file, and it is worth noting this is already a clean file. We just need to change the datatype for the tweet ID’s.

Clean

Before we clean, we create copies of the data frames created so that we have a safe place to restart from should there be any issues. We begin cleaning the issues assessed and do it one table at a time starting with the enhanced archive. The columns with incorrect datatypes are corrected, the undesired columns are dropped and the dog stage columns are combined into one column.

Next from the Tweepy data the dog rating, and dog name information is extracted and cleaned. The cleaned results are visually assessed again and further cleaned until satisfied.

The image Prediction file was mostly clean and only the data type for the tweet_id was changed.

Finally the 3 data frames were combined into 1 for tidiness and saved as a master data frame and CSV file.