Wrangle Report

Overview:

In this wrangle project, we will be exploring WeRateDogs Twitter database. The database is a Twitter database of postings of people's dogs, with information like ratings, name, and dog stage. We will be combining that database with information on retweet counts and favorite counts. In addition, we will be using an image prediction database that takes the image of the dog posting and predicts the breed of the dog. The three database will be combine to analyze information on pet dogs' community on Twitter.

Data Wrangling:

Gathering:

The first database that I looked is the 'twitter-archive-enhanced.csv'. This file was given and could be downloaded directly using Panda's 'read_csv' function. The next database is the image-prediction database. This will be gathered using the 'requests' function that downloads the data via the internet. The last database is the retweet / favorite count. This will be access through Twitter API. Since, I don't have an account and do not want to establish one, I accessed the data through 'read_json' function.

Assessing:

This database will be reference as ‘dog’ dataframe. To get a quick overview of the dataframe, I used commands like list, info, describe, value_counts, and duplicated. This helps me get a visual understanding of the database and provides some basic information. Using a visual analysis of the dataframe, I noticed a few quality issues that needs to be addressed.

The first visual quality issue that showed was that there was an inconsistent use of rating systems. Some rate the dog out of 10. Others use rating out of 50, 80, etc. We need to convert all that to a consistent value like out of 10.

Next, there are several columns that deals with retweets and replies. Since we don’t desire these information, we will need to filter all the retweets and replies and delete them. We need to delete those columns afterwards with other columns that are needed like ‘source’ and ‘expanded-urls’.

Visually, I could also see that many of the name column contains the value of ‘None’. This needs to be address. Either it needs to be deleted or changed to ‘NaN’.

After visual, I used programmatic assessment to assess the database. Through the ‘info’ function, I noticed that the ‘timestamp’ column is a string format instead of ‘datetime’ format.

The next issue that I noticed through ‘value_counts’ function is that many of the dog’s name are not actually names. They have names like ‘a’ or ‘an’. After more research, I noticed these are not original tweets and retweets or other information that we didn’t want. These rows will be dropped. There are other rows that are not original tweets through additional researching.

As for tidiness issues, the ‘doggo’, ‘floofer’, ‘pupper’, and ‘puppo’ columns should all be under one column and those columns needs to be dropped. I will also need to combine the image database and merge it with the dog database. The same goes for the image prediction database.