Data Wrangling Report

Gathering Data:

The data I wrangled, analyzed and visualized was collected form tweet archive of Twitter account WeRatingDogs. This account rates people's dogs with a humorous comment about the dog. The data separated into three files **Enhanced Twitter Archive, Image Predictions File, and tweet json** which provided by Udacity.

Enhanced Twitter Archive:

The WeRateDogs Twitter archive contains basic of the 5000+ tweets data where tweet_id is the unique identifier for the data set. This file contains columns which were extracted programmatically: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo).

Image Predictions File:

This file contain of the result of apply neural network on each image in the WeRateDogs Twitter archive to classify breeds of dogs. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

tweet json:

This is the file where you store all the tweets extracted using twitter API I extracted all the tweets attributes such as favorite count , retweet count ....etc. The data was reprinted in JSON format and stored as text file encoded using UTF-8.

Assessing Data:

The imported data needed to be assessed. This step done by using .info , .describe to discover the data and the parts need to be cleaned. During the assessing stage I have across the following quality and tidiness issues:

Quality:

- Incorrect and misspelled dog names

- The name column has values don't seem to be a name like 'a','an', and 'his'

- The numerator need to be convert to float type and drop the unusual values in denominator
- Calculating the over all rating
- Timestamp, tweet_id need to be convert to datetime object and str respectively
- Drop columns that aren't needed for analysis
- Exclude any tweet that is a retweet
- Entries in p1,p2 and p3 have inconsistent capitalization
- Change id name in df_tweet to be tweet_id to match other two tables

Tidiness:

- Join the 3 tables
- Make all dog stages in one column

Data Cleaning:

During cleaning I solved problems that I outlined in assessing part. Having these steps written down helped me a lot to make cleaning data more easy and took less time than before.