DATA WRANGLING PROJECT

Data Source: We Rate Dogs Twitter Page Data
Tool Used: Jupyter Notebook, Spreadsheet(Google Doc)
Aim & Objectives:

- Wrangle twitter data using the following process
    - Gathering
    - Assessing
    - Cleaning
- Storing and Exploring the cleaned data
- Reporting the Analysis

GATHERING DATA PHASE

Firstly, i had to gather data from the following sources for the We Rate Dogs Twitter Project;

- The WeRateDogs twitter archive, the 'twitter_archive_enhanced.csv' file was provided by udacity to all learners like myself available for download manually containing basic tweet data for all 5000+ of their tweets spanning up until 2017.
- The ImagePrediction File: This has classification of dog breeds which was obtained by running the archive data through a neural network and was provided for us learners also to get easily.
- Twitter API to gather retweet count, favorite counts and several other metrics i find valuable from the WeRateDogs tweet_id using the tweepy library

ASSESSING DATA PHASE

Here i visually and programmatically assessed my various gotten data sources for Quality and Tidiness issues, here are what i found to be worked on during pre-processing;

Quality Issues

1. **archives table**
    - Keep original ratings (no retweets) that have images
    - We should drop all columns not needed for our analysis
    - Make corrections to the Incorrect data types in these columns
    - Correct all numerators with decimals
    - Correct all non-name characters in the name column
    - lets try to identify all dog_growth_level represented as 'None'
    - drop unused columns after the above
    - Source column is in HTML-formatted string not a normal stringError
    - get the standard unit dog ratings
2. **image_predictions table**
    - Erroneous data type (tweet_id) convert to string
    - Missing images (only 2075 counts out of possible 2356)
3. **Twitter API extract table**
    - Erroneous data type (tweet_id) convert to string
    - Missing tweets (only 2327 counts out of possible 2356)

Tidiness Issues

1. **archives table**
   - doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "dog_growth_level"
2. **image_predictions table**
   - Image predictions table should be merged to twitter archive table
3. **Twitter API extract table**
   - Twitter API table should be merged to twitter archive table.

CLEANING DATA PHASE

I then moved on in cleaning up the data as respectively itemized in the assessing section using the **Define**, **Code** and **Test** method of cleaning. We can see this in the Wrangle_act.ipynb file attached to this file.