Project: Wrangling and Analyze

Data We Rate Dogs Data.

Data Gathering Phase: I started the project by downloading the ‘twitter-archive-enhanced.csv’ file manually. Then, programmatically from Udacity's server used the Requests library to download the tweet image prediction (image_predictions.tsv). Next, I wrote it into image_predictions.tsv.

I was unable to use the tweeter api because i couldn't get approval so i used the code provided by udacity to query twitter Api and download the last file. tweet_json.txt: This is the resulting data from twitter_api.py. Then read the tweet_json.txt file line by line into a panda DataFrame with (at minimum) tweet ID, retweet count, and favorite count."

Assessing Data: In this section, I used both visual assessment and programmatic assessment to assess the data.

- **Visual assessment**: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.
- **Programmatic assessment:** pandas' functions and/or methods are used to assess the data.

Quality issues:

Twitter_archive Table:

Issues	Solution
1. keep only original tweets	1. use drop method to drop unnecessary columns
2. incorrect datatype in some columns(tweet_id, timestamp)	2. change the columns to appropriate data types.
3. Error in dog names	3. Remove incorrect names
4. missing values in some columns	4. drop empty data entry
5. source column is in html format	5. extract data from the column

6. invalid values in tweet_source column	6. remove invalid entries

Image_prediction table:

Issues	Solution
1. invalid data type in id column	1. change column data type
2.	2.

Tweet_data table:

Issues	Solution
1. invalid data types in id column	1. change column data type

Tidiness issues:

Issues	Solution
1. in Twitter_archive table, 4 different columns(doggo, floofer, pupper and puppo) are the same and should be melted into a column with value name as dog_life_cycle.	1. Melt the columns
2. twitter_archive table, image_prediction table, tweet_data merged together	2. merge the tables