**Wrangle and Analyze Data**

**Introduction**

Goal of this project is to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. The dataset that will be wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

**Gathering Data**

Data was gathered from multiple sources:

1. The WeRateDogs Twitter archive: twitter_archive_enhanced.csv provided for this project
2. The tweet image predictions: what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and will be downloaded programmatically.
3. Each tweet's entire set of JSON data in a file called tweet_json.txt aquired using tweet IDs in the WeRateDogs Twitter archive

**Assess**

Lets start with the twitter archive loaded into dataframe

Initial observation

1. Total 2356 tweets
2. Incomplete columns:
    - in_reply_to_status_id
    - in_reply_to_user_id
    - retweeted_status_id
    - retweeted_status_user_id
    - retweeted_status_timestamp
    - expanded_urls
3. Timestame and retweeted_status_timestamp is object not of type DateTime
4. Source column has html tags present
5. Name Column has invalid names such as None
6. The last 4 columns doggo, floofer, pupper and puppo can instead be a single categorical column
7. Rating numerator and denominator don’t seem to fall in valid range of values since their max values are 1776 and 170

Next, lets look at the Tweet image predictions dataframe

Initial Observation:

This dataframe looks complete for the most part and has the correct datatypes for each column. Key here would be how we use it.

1. Inconsistency, the prediction columns p1, p2, p3 do not follow similar case format and are separated by an underscore.
2. This data can be combined with tweeter archive dataframe

There are 5 areas of focus:

Completeness

1. Incomplete columns:
    - in_reply_to_status_id
    - in_reply_to_user_id
    - retweeted_status_id
    - retweeted_status_user_id
    - retweeted_status_timestamp
    - expanded_urls
2. Tweet_id should be string not int

Validity

1. Dog name column has some missing data filled with None string.
2. Retweets are repetition of original tweet and must be removed for analysis.
3. Dog Prediction table, some predictions are non-dog names and those should be removed for analysis

Accuracy

1. All timestamps should be DateTime type.
2. Rating numerator and Denominator min and max values seem odd. Needs to be standardized.

Consistency

1. Tweet_id should be string among all tables
2. Rating denominator should only be one value, 10 in this case
3. Remove html tags from source column to match others
4. Dog prediction table, columns p1, p2, p3 do not follow the same format and case

Tidiness

1. Replace last four columns with a category type

2. Combine the dog predictions table with the archive table as they part of dog rating information

**Clean**

Tidiness Issue 1

*Define*: Consolidating dog type to 1 column

Dog types are: doggo', 'floofer', 'pupper', 'puppo'

These columns were originally created by extracting from the text associated with each tweet

*Code*: Extract the dog type from each individual column and populate the new dog_type column

*Test*: Verify if new column is created and matches the previous individual dog type columns

Tidiness Issue 2

*Define*: Drop unwanted columns from df_img_preds and merge with df_twitter_clean

*Code*: Use pandas merge function on tweet_id column

*Test*: Sample the dataframe and ensure all columns look correctly merged

Quality Issue 1

*Define*: Convert tweet_id from integer to string

*Code*: Use astype() function to set a new type to existing column

*Test*: Review dataframe information for datatype

Quality Issue 2

*Define*: Remove retweets

*Code*: We achieve this by removing any column where retweet status id is not null. Find all indices where retweeted_status_id is NaN

*Test*: We will verify the DataFrame information to check if all retweeted variables non-null count should be 0

Quality Issue 3

*Define*: Remove unwanted columns

Since we removed all rows which were retweets, we can remove columns relating to retweet as well as others which dont think are relevant

*Code*: A this point we can drop these columns as they serve no purpose.

*Test*: Lets take a look at the dataframe once again and what we are left with now

Quality Issue 4

*Define*: Convert timestamp to datetime format

*Code*: Use pandas to_datatime() function

*Test*: lets take a look at the dataframe information if the change has been applied. we expect to see the DateTime as the datatype for timestamp column

Quality Issue 5

*Define*: Transform dog names to match same case preference

*Code*: We use regular expression to capitalize the first letter in all names

*Test*: Verify a random sample to check if the names appear correct now

Quality Issue 6

*Define*: Remove HTML tags from source column

*Code*: Use regular expression to extract the html tag content

*Test*: Verify the dataframe sorce column

Quality Issue 7

*Define*: Fix ratings numerator and denominator

*Code*: First lets change the data format for numerator and denominator to be a float

*Test*: Verify numerator and denominator values in the rows affected

Quality Issue 8

*Define*: Create a rating column using the numerator and denominator so it is easy to compare one dog against another irrespective of their rating scale

*Code*: We divide numerator with denominator and assign to new column

*Test*: Sample the dataframe and view the new column

Quality Issue 9

*Define*: Add column dog_breed to predictions data frame based on predictions

*Code*: Consolidate the predictions and their repective probablitiy to one dog_breed column

*Test*: Sample the dataframe as well its summary information to look at the newly created column