Data Wrangling Twitter *We Rate Dogs*

Data Gathering

The data for this project came from 3 sources

- Data gathered from twitter using tweepy (tweet_json.txt)
- Data set provided by Udacity (twitter-archive-enhanced.csv)
- Predictions data programmatically downloaded (image_predictions.tsv)

Findings & Solutions

Data Exclusion Criteria

For this project we are expected to exclude.

- Retweets
- Tweets with no images
- Reply tweets

Data Quality Issues

1.	There are 3 separate tables.	We fix this by merging our data into one table using “twitter_id”.
2.	Timestamp datatype values are typed as object.	We change this to datetime.
3.	Re-tweets are included in the dataset these are identified by “RT @”	We query the text field in our merged data for “RT @” and drop those columns.

4.	Replies are included in the dataset.	Identified in “reply_to_status_id” Query the table if “reply_to_status_id”not null then drop rows.
5.	Dog classification name values are “NONE” should be NaN.	We run a for loop to identify the instance of “NONE” and we change to NaN.
6.	Redundant columns (doggo, puppo etc).	We Place the values into a new one dimensional column called “dog_type” and drop redundant columns.
7.	Missing fields for newly created dog type field.	This is not worth fixing, noticing that we may build inaccurate data. We observe that the name could potentially be contained within the text field and mined. However the risks of building false data are too high.
8.	Dog names are missing in the name column.	We fix this by running a for loop and populating our name column based on observable criteria.
9.	Various tweets without images	We query “jpg_url” in our merged data and exclude the null fields

Other Notable Quality Issues

- Erroneous datatypes
- Missing info in expanded_urls
- Some names are not names just random data
- Some rating values are in the 100's when I assume it's max 10
- For the predictions table the fields are not that human readable, we could change them but I decided not to.
- The predictions columns values are stored as TRUE & FALSE should be stored as INT64 using 0's and 1's to determine TRUE & FLASE.

Tidy Data

1. *After gathering the data we merge it into one dataset using outer join based on "Tweet_id" as to not lose data.*

    - First we check our separate dataset shapes.
    - Then we merge using an outer join based on unique tweet_id.

2. *We then create an SQL database and store our cleaned data.*