Wrangling Report for Project 2: Wrangle and Analyze Data

Goal

Our goal is to wrangle and create interesting analysis from WeRateDogs's twitter data.

The Data

For this project I used the following datasets:
1. **Enhanced Twitter Archive**: This contains WeRateDogs' tweets with ratings, dog names, dog stages.
2. **Image Predictions**: This is a table of predicted breed of dogs from each image in WeRateDogs' tweets
3. **Tweet_json**: This datasets contains the retweet and favorite count for each tweet in the Enhanced Twitter Archive.

Step 1: Gathering the data

I directly downloaded the Enhanced Twitter archive and stored in the root of the project directory as *twitter_archive_enhanced.csv.* I used requests to programmatically download the image predictions from this link and saved it as *image-predictions.tsv.* I was unable to get approved so I couldn't use the Twitter's API so I downloaded the tweet_json.txt from Udacity.

Step 2: Assessing the data

I used both visual and programmatic methods to asses the datas and found the following issues in the datasets:
For quality issues, I found:

1. **Enhanced Twitter Archive** : Incorrect data types for timestamp, it is in string while it should be in datetime
2. **Enhanced Twitter Archive** : Expanded_url in the twitter archive has duplicate urls
3. **Enhanced Twitter Archive** : names are incorrect, names with None, a, an, the and so on should be Null
4. **Enhanced Twitter Archive** : Ratings with decimals did not get extracted properly therefore rating_numerator might be incorrect
5. **Enhanced Twitter Archive** : incorrect rating_denominator like those with 2,11 should be 10, scores with 20, 7 etc should be NaN
6. **Enhanced Twitter Archive** : we don't need retweets just the original rating
7. **Twitter JSON** : id_str and quoted_status_id_str should be string
8. **Twitter JSON** : language should be categorical
9. **Twitter JSON** : possibly_sensitive and possibly_sensitive_appealable should be boolean
10. **Image Prediction** : Breeds are lowercase and uppercase and words were seperated by underscore

For tidiness, I found the following issues:

1. **Enhanced Twitter Archive** : Url in text should be split to text and shortened_url
2. **Enhanced Twitter Archive** : Four dog stages should be a melted into one
3. Remove any columns with retweet and reply information in **Enhanced Twitter Archive** and **Twitter JSON**
4. **Enhanced Twitter Archive,** have the same tweets and refer to the same thing.

Step 3: Cleaning the data

- To the tackle the issue of retweets in the data I retained rows with retweeted_status_id as null.
- I split the url in the text column in the enhanced Twitter Archive to a seperate column called shortened_url
- Combined the dog stages into a single column
- I replaced names that are in lowercase with NaN since all of them were not names
- I removed duplicates urls in expanded_url
- Converted all the errorenous datatypes in the datasets to the correct ones
- Extracted the rating from the text using regex
- Harmonized the breed names so they are all in lowercase and removed underscores
- Removed columns that had retweet and reply values and combined the datasets into one dataset

Step 4: Storing the data

I stored the dataset into a csv file with the name *twitter_archive_master.csv*