Wrangling Report

Abstract:

The project was about gathering data from different resources related to the We Rate Dogs twitter account.

Data Gathering:

There were a couple of data sources that I used to gather the data from:

1. The `twitter_archive_enhanced.csv` file by Udacity. it includes a pre-scrapped thousands of tweets from the @dog_rates, each observation (row) has a tweet_id, the source and the content of the tweet, rating of the dog, the name of the dog and finally the stage of the dog that is one of doggo, floofer, pupper or puppo.

   I used python programming language and one of it's packages called Pandas to use the file in my workspace.

2. The Image_predictions.tsv file on the internet. It includes the prediction of the dog breed given the image of the dog. And it uses the power of neural networks and the deep learning to classify the images and get the breed of the dog.

   I used a Python package called Request to fetch the file from the internet and add it to the workspace.

3. The Twitter REST API of the tweets, I used it to fetch some more information about each tweet given in the twitter_archive file, like the number of followers, number of favorites, etc.

Data Assessing :

I started looking at the data visually and programmatically then found the following issues:

Quality

1. useless columns for the anaylsis in twitter_archive like in_reply_to_status_id,in_reply_to_user_id ,retweeted_status_user_id,retweeted_status_timestamp.
2. 181 records have a retweeted_status_id, these will need to be exluded from the twitter_archive table.
3. many wrong invalid breed names in twitter_archive .
4. Consistency issue with the source column because it includes the wholoe <a> tag while it has to be just the URL in twitter_archive.
5. 11 Tweets in twitter_archive ain't available anymore while fetching their data them from the API.
6. retweets in tweet_json.
7. the timestamp data type in twitter_archive .
8. Remove the underscore between the words in image_predictions.

9. tweet_id data type should be a string in both image_predictions and twitter_archive.
10. the rating is int in twitter_archive.
11. the rating_denominator is less or greater than 10 in twitter_archive.
12. 57 URL errors in the tweet_json

Tidiness

1. the stages of the Dogs looks awful in twitter_archive.
2. the col names in tweet_json weren't matching the other tables but that was fixed while retrieving the JSON .
3. the Tables have to be combined for the analysis.
4. the Breeds in image_predictions is ordered from left to right, hence if p1 is big and True then no need to the others.

Data Cleaning:

I finally started working on the issues using Python programming language and it's package Pandas.And the final result was a dataframe (a new csv file) with no data types errors, more clear file appropriate for the next phase of analysis.