We rate dogs

The goal of this project is to wrangle WeRateDogs Twitter data to come up with fascinating analyses and visualizations. There were three datasets used i.e. twitter archive, image predictions file and additional data extracted using Twitter API

To begin with our data wrangling the following steps were followed:

1. Data Gathering

    - Read the twitter_archive_enhanced.csv onto our notebook
    - Used the Requests library to download the tweet image prediction (image_predictions.tsv).
    - Used the Tweepy library to query additional data via the Twitter API (tweet_json.txt). We then had to write a code that read the json file line by line.

2. Assessing data

For this part we used two methods which is visual assessment and programmatic assessment. Visual assessment involved scanning through the data and spotting abnormalities. Programmatic assessment involved using code to assess errors that we might have missed during visual scanning

These are the issues that were spotted.

Quality issues

1. The time_stamp has an incorrect data type
2. Some row are retweets
3. tweet id datatype is incorrect
4. There are rating numerators which are abnormally higher
5. There are rating denominators which are greater than 10
6. Some columns have a lot of missing entries and won't be needed for my analyses i.e. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
7. some dog names have been incorrectly read
8. There is an unnecessary long link for the source column
9. some image predictions are not dog breeds
10. Change the column name p1 to Dog breed

Tidiness issues

1. The duggo, puppo, floof columns should be one column
2. merge the all the three datasets to one

3. Cleaning data

The last part of data wrangling was cleaning all the quality and tidiness issues. The cleaned data was later saved to a master CSV file twitter_archive_master.csv