Data Wrangling the Twitter Archive of WeRateDogs

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. In this project, this twitter data of WeRateDogs is wrangled to create interesting analysis.

Data Gathering:

The data for this project has been gathered from three sources:

1. **Downloadable csv file in Udacity resources**. This file has been filtered for tweets with ratings only from WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. It is downloaded manually from Udacity's resources page.

2. **TSV file containing image predictions** of dogs for each tweet which is hosted on Udacity's server (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv). This file has been downloaded programmatically using requests library

3. **JSON data on each tweet** that will be queried from Twitter's API using the Tweepy library. From this additional information on tweets like retweet count and favourite ("like") count has been gathered

After gathering data from three sources, there were three dataframes with necessary data.

1. tweets_csv : Dataframe with basic tweet data from enhanced WeRateDogs Twitter Archive
2. image_predictions: Dataframe with image predictions data for each tweet
3. tweets_api: Dataframe with additional tweet data extracted using Twitter API

Data Assessment:

After gathering data, visual and programmatic assessment was done on the three dataframes to check for quality and tidiness issues.

From the assessment, the following issues have been identified:

Quality issues

1. 181 tweets in tweets_csv are retweeted tweets
2. 51 tweet have nulls in expanded_urls column of tweets_csv, implying these tweets have no image
3. Some tweets in tweets_csv has rating_denominator greater than 10. Some of these tweets have multiple dogs and hence the denominator is number of dogs* 10. However some have wrong values for rating. (ex: date considered as rating)
4. 442 tweets in tweets_csv have rating less than 1. Some of these are not pictures of dogs. Some of them have incorrectly extracted ratings. There are also tweets about dogs with rating less than 1
5. timestamp column in tweets_csv should be a datetime object
6. Some of dog names (names column in tweets_csv) are unusual like 'a', 'quite'
7. tweet_id in all three dataframes should be a string instead of int
8. The breeds of dogs in image_predictions (p1, p2, p3) columns have inconsistent format. Some start with capital letter while other don't. All of them should be changed to lowercase format
9. The jpg_url column in image_predictions is redundant and can be dropped

Tidiness issues

1. Dog stage is one variable but has four columns
2. There is only observational unit: tweets. However we have 3 dataframes
3. Both the text and url have been stored in the same column (text) in tweets_csv

Cleaning

Based on the assessment, the following cleaning was done on the dataset.

1. Converted the timestamp column to a datetime object and the tweets_id column to string
2. Joined all three datframes into one dataframe using the tweet_id column and dropped redundant columns
3. Filtered the dataframe to remove all the retweets and tweets without any images
4. Combined the four columns:doggo, floofer, pupper, puppo into one column: dog_stage
5. Removed the url from the text column so that it contains only the text of the tweet
6. Converted all the strings in p1, p2, p3 columns to lowercase format
7. Replaced all unusual dog names like a, quite with None

8. Corrected the ratings for tweets having multiple numerator/denominator format
9. Removed the tweets with ratings less than 1 to remove tweets without dogs. Here it has to be acknowledged that some of the proper dogs tweets would also be removed
10. Divided the numerator and denominator of tweets having multiple dogs with number of dogs

The final clean dataframe has been stored as twitter_archive_master.csv.