@dog_rates Twitter Wrangling

Introduction:

The objective of this project is to perform data wrangling on twitter data of @dog_rates twitter account. This report consists a detailed explanation of how the data was gathered, assessed and cleaned. The @dog_rates account rates peoples' dogs in a humorous way. The contents of this report are divided in three parts as mentioned previously.

Gathering Data:

The data was gathered from following sources:

1. Twitter Archive File: 'twitter-archive-enhanced.csv'
2. Server: 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'
3. Twitter API

Source 1 – Twitter Archive:

This file ('twitter-archive-enhanced.csv') was provided by the instructor, it is the twitter archive of the account @dog_rates. The file contains tweet-data (*tweet_id, timestamp, tweet_url, text, rating_numerator, rating_denominator, etc.)* as it stands on August 2017. The contents of the file were loaded into a ***pandas.DataFrame***.

Source 2 – Udacity Server:

Requests to the server were made using the url – 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'. A ***get*** request was made to the server using python's ***requests*** module. The contents of the response were written to a file ***image-prediction.txt***. The file data was then read into a ***pandas.DataFrame***. The dataset contains data like *image_url, three dog_breed predictions, confidence-level of the predictions, etc.*

Source 3 – Twitter API:

A twitter developer account was made, Twitter API can be accessed only if one has a developer account. All the ***tweet_ids*** from source 1 were used to make ***get_status*** requests through a library called ***tweepy***. The json response for each tweet was written to a file named ***tweet_json.txt*** using python's ***json*** library. Its contents were then loaded into a ***pandas.DataFrame***. It contains *tweet_id, retweet_count, favorite_count* for each tweet the account has made until August 2017.

Assessing Data:

- **Twitter Archive Data-Set:** Visual analysis was performed and then summary of the data-set was printed. The summary depicted that the data-set contains 2356 rows and 17 columns. Certain columns like *retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp* were deemed as extraneous data. Some of the *tweet_ids* were actually retweets such rows were also deemed peripheral. Upon inspecting the data-types it was ascertained that the data-type of *rating_numerator* is integer which should have been float. Upon further assessment some values of *rating_denominator* were found to be more than ten, *text* columns of such rows were examined, it was evident that these ratings were actually used to rate more than one dog. While examining the *text* column the inaccuracy of some ratings due presence of dates like 9/11 and store names 7/11 in the text column was pointed out. After assessing the *name* column, invalid data like use of articles like ‘a’, ‘an’, ‘the’ instead of name and use of None instead of null values was found. It was also noted that some dogs were assigned multiple slang categories.

- **Image Predictions Data-Set:** After performing visual assessment and creating a summary it was noted that this data-set contains 2074 rows and 12 columns. Some column names like *p1, p1_dog, p1_conf* were non-descriptive and confusing. The *jpg_url* column had a couple of png image links along with 66 duplicate rows.

- **Twitter API Data-Set:** Data-set summary was inspected. The data-set had 2356 rows and 3 columns. Out of these only 2331 rows were non-null, the rest of the rows were classified as missing data. Upon inspecting the data-types it was found that the tweet_ids were floating point numbers instead of integers.

Cleaning:

- **Twitter Archive Data-Set:** A copy of the data-set was created, and cleaning operations were performed on the copy. Extraneous columns were dropped, rows that contained retweet data were also dropped, ***rating_numerator*** and ***rating_denominator*** columns were cleaned by extracting correct data from the ***text*** column, the rows with ***rating_denominator*** greater than 10 were cleaned by dividing the ***rating_numerator*** and ***rating_denominator*** with appropriate numbers. Then both the columns were merged into a single column ***rating***. Another column, ***rating_num*** was created that stores the floating point rating for the purpose of analyses. The column ***name*** was purged for wrong names and “None” values were replaced with ***numpy.NaN***. The dogs that were categorised into more than one slang category were assigned null value, also the multiple category columns were merged into a single column called ***slang_category***.

- **Image Predictions Data-Set:** A copy of the data-set was created, on which the cleaning operations were performed. The column names were confusing so they were change to ***prediction1***, ***prediction1_confidence***, ***prediction1_iscorrect*** and so on. Duplicated rows from the ***url*** column were dropped. All the prediction related columns were merged into two columns ***predicted_breed*** and ***prediction_confidence***. The breed names were cleaned by replacing underscores with whitespace.

- **Twitter API Data-Set:** The column name ***id*** changed to ***tweet_id***, then changed the data type of the same column to integer. Finally all three data-sets were merged on their ***tweet_id***, this merged data-set is saved in a file named ***twitter_archive_master.csv***