Udacity Data Engineer Nanodegree – Twitter We Rate Dogs Dataset – Wrangle Report

Gathering Data

In this step, the We Rate Dogs Twitter data is gather from the following sources

- **Twitter Archive Enhanced Dataset CSV file**
- **Images Predictions TSV File**
- **Twitter Additional Data File JSON File**
    - This file was supposed to the downloaded via Twitter API using Tweepy library but code failed while Twitter API settings authentication failed and so I had to raise ticket with Knowledge support team, Mentor has advised me to use the 'tweet-json.txt' dataset and continue the module
    - ### Ticket Reference - https://knowledge.udacity.com/questions/986996

I have loaded the 3 above source files into the three separate dataframe in Python using Pandas library.

1. df_twitter_archive_enhanced_dataset
2. df_twitter_image_predictions_tsv
3. df_tweet_json_file

Note – I have continued the further into my local PC environment where Python Anaconda package is installed because the Udacity cloud workspace was getting disconnected frequently.

Assessing Data

Once the data is gathered and loaded in the respective dataframe, I have analyzed and assessed the data using the functions for each of the data frames.

I have found the following issues

Twitter Archive Enhanced Dataset

1. Timestamp field is string object.
2. The data columns - tweet_id,in_reply_to_status_id,in_reply_to_user_id etc. should be an object data type but currently they of data types - integer and float in the original dataset.
3. There are only 181 retweets - columns - retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.We need only original "tweets" not "retweets".
4. There are "in_reply_to_status_id" and "in_reply_to_user_id" columns which are extra as we need only "original tweets" column in our dataset.
5. There are 109 inconsistent names with small letters with no valid names,also there are NaN values in the dataset in the "name" column should be rename to 'None'.
6. There is one record with rating denominator with the value Zero.
7. There are rating denominator values greater than 10, should be removed.
8. There are only 4 types of values in the source column which can be simplified further to remove complexity. Twitter for iPhone, Vine - Make a Scene, Twitter Web Client, TweetDeck

9. There are 59 number of tweets which are having missing expanded URL's

Twitter Image Prediction Dataset

1. There are 324 rows in the Twitter Image Prediction Dataset dataframe which didn't predict anything for three of dog type, best way to predict the dog breed is to combine it with master dataset.

Tidiness issues

1. The 4 Columns doggo, floofer, pupper, and puppo - should be merged into the one column as this doesn't comply "tidy data" rule.
2. Json Tweet text file dataset should be union with 'Twitter Enhanced Archive' Dataset to get the full view.
3. The data columns in the Json Tweet text file dataset should be reduced to the "tweet_id", "recount_tweet", "favourite_count", extra columns must be removed.

Cleaning Data

All issues identified in the assess phase are successfully cleaned (if possible) using Python and pandas, and include the cleaning tasks required.

a) Merging columns - "doggo", "floofer","pupper","puppo" into single column called - "dog_breed_stage"
b) Dropping the columns - 'doggo', 'floofer', 'pupper', 'puppo' which are no longer required
c) Checking the dataset to confirm if columns are clubbed properly.
d) Merging the Twitter Enhaced dataset with the Image Predictions dataset after cleaning
e) Renaming the "id" column to the "tweet_id" in the tweet json file dataset so that it can be further used for the merge in the next step

Storing and Acting on Wrangled Data

A tidy master dataset "**df_final_master_merged"** gathered, assessed, and cleaned master dataset(s) to a CSV file with the name - **"twitter_archive_master.csv".**