Reprort : Data Wrangling

Goal: Wrangle and analyze *WeRateDogs* Twitter data.

1. Gathering the Data:

- **Enhanced Twitter Archive:** I used the *WeRateDogs* Twitter archive to study twitter data on dog ratings. The data was provided in a file called : **twitter_archive_enhanced.csv**

- **Tweet Image Predictions:** The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet was also provided in a file: image_predictions.tsv. This data was downloaded using the **Requests** library from the Udacity servers.

- **Tweet JSON data :** Using the tweet IDs in the WeRateDogs Twitter archive, I used Python's **Tweepy** library to query the Twitter API for each tweet's JSON data.
  - In order to query the twitter API, I created the application in my Twitter account and got my API keys and Token.
  - Then I queried the API and stored each tweet's entire set of JSON data in a file called tweet_json.txt file which could then be read into a pandas DataFrame.
  - To install Tweepy in Anaconda: I did a pip install tweepy in Anaconda. Since some of the tweet Id's were deleted I used the "try" command before I read the API. It took about 30-40 minutes to get the entire JSON data.

2. Assessing the Data:

Quality issues in the data:

A. twitter_enhanced.csv:

1. Large amount of Missing data for some variables such as: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp
2. Twitter_id is int but we will not be using it. So change it to string. Do it after the merge as its better to use the int format when merging.
3. Incorrect ratings or data: Examined the tweets and found a number of incorrect ratings.

   - The decimal ratings are extracted incorrectly.
   - Some have double ratings. It could be because of texts which look like ratings, like, 9/11 or 7/11 and these are not correct. The rating which occurs at the end of the text should be taken.
   - Some double ratings happened, because the user updated the rating. He thought it should be some rating but gave some other rating at the end. The one at the end should be taken.
   - Some have pics which are not dogs. These can be identified by "We only rate dogs".

B. df_json:

1. Some variables have all null values and need to be investigated further: geo, contributors and coordinates.
2. Some variables have lots of missing values: in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id,in_reply_to_user_id_str, quoted_status, quoted_status_id,quoted_status_id_str, retweeted_status
3. The user column has no valid information and has duplicated texts.
4. Json has 11 duplicated twitter Id's.

C. df_image:

1. Inconsistency: In the Image dataframe, some names start with Capital letters, others don't.

Tidy issues:

A. df_tweets :
   1. Since the dog can have one of the 4 categories, the "doggo", "floofer"..etc columns need to be row values instead, otherwise the information is just redundant.
   2. The source column has lot of unuseful text which can be removed and only the useful part retained, i.e. the actual source, such as iphone, vine etc.

B. General:
   3. Many columns are duplicates between df_tweets and the df_json data, and the two dataframes need to be merged and the duplicate columns dropped.

3. Cleaning:

JSON data:

- Make a copy of the json data to clean
- Drop columns which have un-useful and a lot of missing data:
  - *df_json_clean.dropna(axis=1,thresh=79,inplace=True)*
- Use only dates up-to Aug 1st 2017.
  - *df_json_clean[(df_json_clean['created_at']<datetime.date(2017,8,2))]*
- Drop all rows which have retweeted status = True
- Drop columns with unuseful information and lot of missing data:
  - *['display_text_range','entities','extended_entities','id_str','possibly_sensitive',*
    - *'possibly_sensitive_appealable', 'source','user']*
- Drop the duplicated twitter Id's in json data.

Tweets data:

- Drop columns with unuseful information: *'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'*
- Incorrect ratings extracted from Texts. Fixed them. (See section on Quality Issues). Then drop the original columns ratings_numerator and ratings_denominator and fill them with the fixed ratings wherever applicable.
- Some columns are common between the json and tweets data. Drop the common columns and merge the json and tweets data.
- Merge the image data as well.

- The dog categories have redundant information, because only one of the categories is possible. Transform the dog categories from columns to row information and drop the original columns. Merge new category column with the old data frame.

Image data :

- Capitalize the names of the dogs so as to have standard name format.

The final data frame is called ‘df_fin’.