Data Wrangling

I have made an account on twitter. But I couldn't obtain an acceptance to access to the twitter API as a developer. So I used the instructions listed in Udacity Project guidelines in case of not having a twitter account.

Gathering the data:

1) I downloaded the files: twitter-archive-enhanced.csv and tweet-json.zip. manually. And imported it into the jupyter notebook.
2) Programmatically downloaded tweet-json.zip file
3) Reading the files into csv files, json, and tsv for twitter-archive-enhanced.csv, tweet-json, respectively

Data Assessment:

1) Visual assessment:

By exploring the twitter-archive-enhanced.csv file manually using excel spread sheet, I found

Quality issues:

1. The 'source' data column has url which contained unneeded tags like <a at the beginning of the url and also at its end.
2. **There are several missing values.**
3. **tweeter_id values are of type int. It should be converted to string type as they are not required in mathematical operations**
4. **Columns name should be modified to be more descriptive (ex: P1, P1_conf,..etc)**

2) Programmatic Assessment

<u>Quality issues:</u>

4) column 'time_stamp' in the arc_df is of type string so it should be changed to datetime type.

5) There are data related to retweets in arc_df. They are considered duplications and should be removed.

5) There are some tweets have no image information.(number of records in arc_df is 2365 however there are only 2075records in image_predections.

6) columns retweet_count and favorite_count sholud be added to arc_df 6) the data in column 'source' has unneeded symbols and HTML tags. The URLs should be extracted.

7) The data in column 'name' of arc_df starts with small characters. It should start with capital characters.

8) The Ratings data were incorrectly extracted. (there are errors in rating_numerator and rating_denominator). So there is a need to re-extract the correct rating data from column 'full_text' that exists in API.

9) Data that is not about dogs should be excluded.(removing non-dog tweets)

Tidiness Issues:

1) The column 'id' in twitter Api (twit_json dataframe) is related to column 'tweet_id' in arc_df. So it can be changed to 'tweet_id' to simplify understanding and dealing with the data

2) columns retweet_count and favorite_count should be added to arc_df

3) the columns 'doggo' , 'floofer', 'pupper', and 'puppo' in arc_df can be merged together in one column called 'stage'

4) p1_conf can be added to the arc_df (after renaming).