Data Wrangling Report

By Areej AlZahrani

References:

Reading from Json file:
https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/
JsonDecoder problem: https://stackoverflow.com/a/48154834
Helped in cleaning:
https://stackoverflow.com/questions/31511997/pandas-dataframe-replace-all-values-in-a-column-based-on-condition

This project is about data wrangling. The goal is to wrangle the data and then drive some insights based on the wrangling. I would like to clarify that only tweets before August 1st, 2017 will be considered in this project since beyond that data we don't have results from the prediction Algorithm. The main steps are:

1. Gathering:

Will gather three data files. 1: WeRateDogs, Twitter archive (This file is available at hand from the supporting materials) 2: get the predictions data (Will use the Request library to send HTTP request and get HTTP response. After that, write the response text to a file named with the last part in the url). 3. tweet-json.txt (unfortunately I couldn't get access to Twitter API so I will use the data file in the supporting materials)

2. Assessing:

Assessing each gathered dataframe visually and programmatically for quality and tidiness issues. I will assess each dataframe separately and document the detected issues. Later, in the assessing summary section I will document eight quality issues and two tidiness issues which will be cleansed in the Cleaning section.

Summary:

Quality:

1. Inaccurate names in name column at archived tweets data.
2. Inaccurate ratings at archived tweets data.
3. Source column should be categorized into 4 categories.
4. In df_arch columns: 'retweeted_status_user_id','retweeted_status_id', 'retweeted_status_timestamp', 'in_reply_to_user_id', 'expanded_urls' are not valuable or meaningful
5. In df_arch columns: timestamp should be of type datetime
6. There are not original ratings (retweets)

7. Invalid data: during the visually assessing some tweets are invalid. Like they are news or ratings for non-dogs.
8. in_reply_to_status_id column name is not clear and should store info about weather the tweet is a reply or not.

Tidiness:

1. Merge the retweets and favorites columns to the tweet archive table, joining on given tweet_id.
2. Replace doggo floofer pupper puppo columns with one column called stage.

3. Cleaning

Cleaning includes merging individual pieces of data according to the rules of tidy data.

Quality:

9. Drop retweeted tweets.
10. Drop useless columns:

a. 'retweeted_status_user_id','retweeted_status_id', 'retweeted_status_timestamp', 'in_reply_to_user_id', 'expanded_urls'

11. Use in_reply_to_status_id column as indicator if the tweet is a reply. and rename it to is_reply. 1 if true and 0 is false.
12. Categorize the source column to have values: 'iPhone', 'Vine', 'Web', and 'TweetDeck'
13. Extract the correct name if exist. I noticed that names that start with small letters are not correct. So, will use patterns to extract to the correct names or will handle them manually. name patterns encountered:

'named (REAL_NAME). '
'name is (REAL_NAME). '

14. Remove the invalid tweets

tweets that are not about dog ratings or with overwrited rating

#832645525019123713 it is a news tweet, to be removed
#832088576586297345 not a dog rating tweet, to be removed
#810984652412424192 this dog has no rating, to be removed
#686035780142297088 to be removed
#684222868335505415 overwrited by 684225744407494656, to be removed
#682808988178739200 not dog rating tweet, to be removed
#746906459439529985 not dog rating tweet, to be removed
#835152434251116546 to be removed

15. Extract the correct rating. Based on rating_denominators not equal to 10 we have two issues to be taking care of:

a. Having two fractions at the same tweet or decimal value in the numerator led to incorrect ratings ==> so will re-extract the rating using faindall lambda. The regular expression will return last fraction in the text which is the correct rating. Also, will take care if the numerator is a decimal value.

b. Multiple dogs ratings: the tweet is for multiple dogs so the rating is multiplied by number of dogs ==> divide the rating, add column # of dogs

16. In df_arch column timestamp should be of type datetime.

Tidiness:

3. Merge the retweets and favorites columns to the tweet archive table, joining on given tweet_id.

4. Move the first prediction to the tweet arch table.

5. Move the text column from tweet archive table to the prediction table.

6. Replace doggo floofer pupper puppo columns with one column called stage.