Wrangle report

Introduction

This report details the wrangling done on the Project 4 of the Udacity NanoDegree Data Analyst course.
The wrangling activities were split into gathering, assessing and cleaning.

Summary:

The data is about twitters o dogs and score for them. Out of 2.3K input records, after cleaning, we got to 1.7K records. During the activity I had in mind other people that would work on the analysis of the data, where I wanted to provide the most simple but complete file as possible to ease and minimize the learning curve for those people. The major task that brought simplicity and agility aws combining the 3 data sources onto a single file, which reduced the complexity and distractions, allowing me to improve the cleaning process and think clearly.

Also, after the end of the cleaning process, a new file was created and store, which can be sent around for analysis, without the need of running this whole Jupyter notebook by other people.

Below you have more details about the process, it includes also some analysis using the cleaned datah.

Gathering:

The data was collected from three datasets, obtained as following:

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing:

In a jupyter notebook I assessed and provide the findings below.

That was an iterative process, where some other cleaning and was identified during that process itself.

The findings:

Quality or 'q' : completeness, validity, accuracy, consistency issues
Tidiness or 't': structural issues. strucuture datasets to facilitate analysis

Generic findings

0.1-(t) combine the 3 datasets into "df_master", that later will be saved to twitter-archive-master.csv file

0.2-(t) remove unnecessary columns identified as not relevant for the analysis (iterative process)

1. twitter-archive-enhanced.csv file findings

1.1-(t) create a stage column and add dog stages (doggo, puppo, pupper, floofer), and delete the 4 do type columns

1.2-(q) some "expanded" URL fields have duplicated info. remove the duplicated data.

1.3-(q) some rows do not have a tweet ID, and can't be used/matched with data sets 2 and 3

1.4-(q) some denominators are different than 10 (smaller or bigger). This could be set to 10, but actually this column could even be deleted since it has no use. Delete it.

1.5-(q) - Remove tweets that are retweets (rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp).

1.6-(q) - Remove replies as they are not original tweets (field in_reply_to_status_id), similar as item 1.5 above.

2. The tweet image predictions.tsv findings

2.1-(q) If all predictions are FALSE, drop the row, as it seems to be something else than a dog.

2.2-(q) p1 shall be used if true, if not use p2 or p3, in this order. create a new

2.3-(t) create a new row to store the preferred prediction ('predicted_dog') + score ('predicted_dog_score')

2.3.1-(t) drop p1,p2,p3 rows and its scores.

2.4-(q) Some dog names have the "_" character. Remove it and add a space instead.

2.5-(q) on dog names, put first letter in upper-case.

3. Twitter API & JSON (stored into tweet_json.txt) findings

3.1-(q) Remove columns not relevant when creating the df_3_tweets_from_API dataset (kept id, favorite count and retweet, retweeded_status count only) - done in Gathering data step

3.2-(q) Remote 'retweet' column, as none was retweeded

Storing:

After the cleaning was done, the final dataset was store in the twitter_archive_master.csv file and it contains 1659 rows.