WRANGLE REPORT - WE RATE DOGS

Introduction

This present report describes the wrangling efforts involved in completing the “WeRateDogs” second project as part of Udacity’s Data Analysis Nanodegree program.
The Data Wrangling Report process consists in:

1. Gathering
2. Assessing
3. Cleaning

1. Gathering Data

Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one testing a different way of obtaining a dataset.

The first was to download a file manually and be able to open a csv file. In this case the file was called “*twitter_archive_enhanced.csv*”.
The second was to be able to download a file programmatically using Python Requests library. The file contained image predictions on the breed of the dog coming from a neural network on some of the tweets already downloaded in the archive file. The file was in tsv format and tested your ability to open this type of file successfully.

The final dataset tested your ability to query Twitter’s API and use a Python library called Tweepy to obtain further data on the tweets in the archive file using the tweet id. An copy was saved in csv format of the data frame created (for my own case, I have used the saved csv).

Details:

- twitter_archive_enhanced.csv;
- image_predictions.tsv;
- tweet_json.txt

2. Assessing Data

The three saved data frames were then assessed visually inside a jupyter notebook with pandas and because the datasets were not too large, a copy of each was exported into one Excel workbook. This allowed quick scanning through the rows and use of filters to identify areas for more detailed investigation. Following this a programmatic assessment was made inside jupyter with pandas using the following functions, df.info(), df.head(), df.sample(5), df.value_counts(),etc.

The datasets were accessed under two criteria, quality and tidiness. When an issue was detected it was documented under one of these two criteria.

Details:

2.1. Quality issues

1. Merging/Associating the all 3 dataframes into 1;

   1.1. 'tweet_id' instead of id in order to merge with the other tables.

2. Erroneous datatypes (tweet_id, timestamp, source, rating_denominator, dog_stage, rating_numerator, p1_dog, p2_dog, etc)

   2.1. Extra: "tw_df" has 2356 rows while the "img_predictions" has only 2075 rows, probably due to retweets and missing photos;

   2.2. Extra: "tw_df" contains 181 retweets and 78 replies which not needed.

3. The 'name' column contain inconsistencies like lower case senses and others;

4. Renaming "Alitta" instead of "a" as dog name;

5. Texts not readable in source columns;

6. Some texts contain floof and still have 'None' as their dog stage value;

7. Cleaning up and Dropping columns of merged dataframe.

3. Cleaning Data

The final step in the wrangling process is cleaning the data for quality and tidiness issues. The cleaning followed the standard process of define, code and test for each of the issues and they were tackled in a logicalorder, which is reflected in the numbering order in the “*wrangle_act.ipynb*” notebook and closely followed standard practice of cleaning missing data first, then cleaning for tidiness and finally quality.

Details:

3.1. Tidiness issues

1. Between the 3 dataframes we have, only 1 should suffice to incorporate the data we need.

2. Not needed columns (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, and in_reply_to_user_id

3. We have 3 predictions for the dog breed in the ″img_predictions″ but maybe the best prediction will suffice to reduce the size of the final dataframe.

4. Merging/Associating the Image predictions, tweet count and twitter archive tables to form a single table;