WeRateDogs Wrangle Report

The goal of this report is to briefly describe the data wrangling efforts that have been made to accomplish the 'WeRateDogs: Wrangle and Analyze Data' project. The 3 steps in which data wrangling has been carried out are: Gathering, Assessing and Cleaning.

Gathering

Three different files had to be obtained in various ways:

- **twitter-archive-enhanced.csv:** This document was delivered directly by Udacity and downloaded manually. It contains the following columns: tweet id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls , rating_numerator and rating_denominator. The name assigned to this dataset in the script is df_twitter.

- **image_predictions.tsv:** This file is located on a Udacity server and was downloaded through its url using the request library. It contains information related to a neural network classifier of dog breeds. Its columns are: tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf and p3_dog. The name assigned to this dataset in the script is: df_predictions.

- **tweet_json.txt:** This information was obtained by communicating with the Twitter AP using the library tweepy. It shows the quantification of retweets and favourites for each tweet. Its columns are: tweet_id, retweet_count and favorite_count. The name assigned to this dataset in the script is: df_api.

Assessing

During the gathering step, the pandas and numpy libraries have been mostly used. Finally, a list has been made with some of the issues contained in the datasets, differentiating between quality and tidiness issues. All quality issues that have been assessed correspond to the dataset associated with the ‘twitter-archive-enhanced.csv’ file.

Quality issues

1. Lines with in_reply not null are replies
2. Lines with retweeted status not null are retweets
3. Names very, quite, incredibly, infuriating and his are not dogs
4. Nan names dogs with name None, a, one, ...
5. Timestamp column is a string
6. There should be an extra column showing the actual rating without differences in the denominators

7. Missing expanded urls (not possible to solve)
8. Tweet id 765395769549590528 name is Zoey
9. Tweet_id 878604707211726852 name is Martha
10. Source column has html tags

Tidiness issues

1. df_twitter: Columns doggo, ploofer, pupper, puppo should be only one
2. df_twitter and df_api should be only one table

Cleaning

Firstly, a copy of all the dataframes has been made to clean up the data, always being able to access the originals. All problems found have been cleaned up following a three-step process: defining, coding and testing.

In the defining stage, the assessments have been converted into cleaning tasks by writing short guides on how to do it. In the coding stage, the definitions have been translated into code. Finally, in the testing stage, it has been checked that the written code has worked correctly.