Data Wrangling report

In this project, I revised three dataset that contains detailed transactions from WeRateDogs twitter account, who rate pictures of dogs and share it with their followers along this social network. Those datasets has some problems related with the quality of data and how the structure of the dataset store the information, so those files need to be loaded into a Jupyter notebook and apply the wrangling process before the data could be ready for analysis and visualization.

As i mentioned, there is 3 dataset that contains the data we want to use:

- Tweet dataset, from WeRateDogs accounts, that has been already processed by Udacity instructors.
- Complementary information of tweets, captured by API requests, made with Python tweepy library.
- A dataset that contains a prediction of dog breed from the images poste on WeRateDogs account.

1. Gather the data

First of all, I loaded the tweets and the prediction datasets, and makes an effort to capture the json data through the API mentioned before, but besides that tweepy API works fine, Twitter has some criteria for obtains this information, and limits how many information you can gather this way, so I finally loaded the tweet_json.txt file that already contains the data needed for this project. The implementation of a function that makes this capture is already on the `wrangle_act` Jupyter notebook, but the call for this is commented so has no effect when we'are running the whole notebook.

The others 2 files has been loaded with traditional pd.read_csv function, so no further details needs to be documented.

2. Assess

The assess process needs to be done programmatically, so I wrote a buch of functions and code to make visual and statistic inspection of the data. With the df.info() and df.describe() functions, I have a kickstart to begin analyzing some variables, change some names to human-readable ones, and delete columns that has almost null values on it. I write down the issues that had been found, to keep it in mind for the clean process.

3. Clean

Once the assess process shows me what part of the dataset has some problems, i write the specific code to solve this issues, like renaming some columns, convert some values (like "source" "timestamp" column). This process needs to be done on all three of the datasets, before trying to make a merge, because some of the columns are repeated, key value (id) has differents names, so it's has been neccessary to clean and make those changes, before the final merge of all the three

datasets. When I finally had the final and cleaned dataset, had been saved as local csv file, to make it ready to use for analysis and visualization.