Wrangle Report

With the initial dataset with 2356 tweets from the Twitter account @WeRateDogs, I was able to use the Twitter API to get more data for the further analysis, and also use the provided breed predictions for most of the tweets given. But, before doing the data analysis, it needed some wrangling steps to ensure the data quality and tidiness. The first thing was to assess the data, and with visual and programmatic methods try to identify major problems in the dataset. 11 issues were detected regarding *data quality*, and 4 issues regarding *tidiness*.

The first issue detected was that there were retweets included in the dataset, but we should consider only original ratings from the account, therefore all the retweets should be removed from the dataset. This was done by filtering all the rows and removing the ones with data related to retweets (IDs, timestamps and user) and/or with text starting with “RT @”.

The next issue was that there were some tweets without pictures, and therefore, they were not useful for the study, as we were interested only in the ones with dog ratings (with pictures). To solve this problem, I removed from the master dataset (all tweets) the ones that were not in the secondary dataset that had the predicted breed, because not having a prediction means the tweet did not have a picture.

Then, the next issue was that the master dataset had too many columns with useless information. Although it was not a big issue, having too many columns that would not be used in the analysis was making it harder to deal with the dataset, especially regarding the visual analysis of the data. To solve this, I dropped the columns from the dataframe, maintaining only the ones needed to continue with the study.

Also, during the import of datasets to *Python* using *pandas*, the tools did not correctly recognize the data types of some columns, like date and like counts. The correction for this was basically changing the data types for the correct ones, allowing better programmatic functions to treat it.

In the dataset, some columns like *source* and *text* were in a raw format, without any treatment to remove additional information that is not useful. To solve this problem, I extracted the text from them using regular expressions and splitting, and then storing

the clean data again in the columns. Some columns like *dog_stage* in the dataset had null values stored as the string “None”, which could make it difficult to use automated functions in those columns. I solved this by replacing those strings by the correct *NaN* (null) value.

Another problem identified in the dataset is that most of the tweets did not have any information about the stage of the dog. Unfortunately, as this data should be provided, there was nothing I could do to recover or treat this, leaving the dataset with many null values.

Regarding the ratings given, the main issues identified were that there was a zero denominator (mathematical indetermination), and also different bases for the rating, which would make the comparison between tweets harder. To solve this, I cleaned those tweets with invalid ratings, and manually parsed the ones with other ratings because most of them were problems in the parser that created the dataset.

After all the assessing and cleaning, the master dataset went from 2356 to 2057 entries, almost 13% of reduction on total data, but now with a clean and tidy dataset, ready for doing analysis and visualizations to understand better the tweets and ratings given by the Twitter account @WeRateDogs.