Wrangle report

By: Luis Enrique García Orozco

Description

In this document I briefly describe the overall strategy I used to complete to wrangle data relative to the "WeRateDogs" twitter account tweets.

Gathering

When I gathered the data there were no big complications. However, all the skills gained through this section of the course were applied:

I manually downloaded file "twitter_archive_enhanced.csv", programmatically downloading "image_predictions.tsv" using the requests library, **d**ownloading and later store in a text file a set of JSON data into "tweet_json.txt file" using the tweet IDs in the WeRateDogs Twitter archive and Tweepy library.

Later, I was able to read all this archives and open them in the form of a Python DataFrame, which was useful to later asses each table.

Assesing the Data

When assessing the data the main strategy was to first create a summary of what the dataframe contained. Pd.DataFrame.info function was very useful in this step, as it gives us the number of rows and columns and also the name, type and number of null values contained in each column.

Whit that info I was able to get an idea of the tidiness level of the table (for example, when I saw 4 columns each assigned to one dog stage) and to focus in looking if some column of interest contained null values where it should not have, or contained values when it should have null values (for example, the column retweeted_status_id in our twitter_archive_enhanced.csv should be empty, cause we do not want to analyze retweets).

The next step was to see some rows of the table with the functions pd.DataFrame.head and pd.DataFrame.tail, this allowed me to catch some of the data quality issues present in the table and document them. When I suspected that there might be an error in the data I also used the pd.DataFrame.column_name.value_counts() function, which automatically gives you a list of the values contained in the column and the number of occurrences for each (this was useful for example when exploring the "name" column in the twitter_archive_enhanced.csv).

Cleaning the Data

When all the quality and tidiness issues detected were documented, the final step of the data wrangling was to clean them. The process was mainly trial and error, because not all the solutions worked perfectly and some tiny exceptions might apply.

At the end of this step we already had a “master” dataframe, which compiled all the previously different tables into a single one.