Mathieu Rella

Wrangle Act Report

We Rate Dogs

UDACITY | **Data Analyst Nanodegree**

The final project of the data wrangling part of udacity's nanodegree data analyst was to analyze the twitter we rate dogs account through the use of different sources and then deliver visualizations and insights.

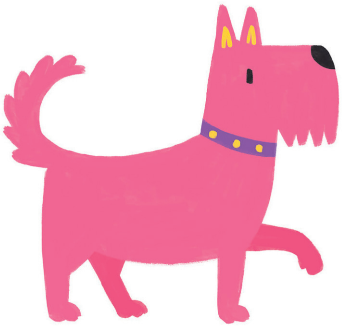
I. Data Gathering

The information was gathered from 3 different sources as follows:

- **Enhanced Twitter Archive** file directly accessible from the jupiter notebook, you just had to read it it was saved as twit arch.

- **Image Predictions** file accessible and downloadable through a udacity link save as img_pred, this file contains the dog breed predictions according to the images linked to the tweet.

- **Additional data accessible by the twitter API** by matching twit_arch tweet ids to the api to retrieve retweet and favorite using the tweepy library, this file has been saved as tweet_metrics.



II. Data Assesment & Cleaning

In the assesment phase of the different dataframes, I first wanted to look for visual inconsistencies and then confirm and find more errors programmatically.

In **twit_arch** dataframes, the first visually visible inconsistency was the missing values for the following columns:

- in_reply_to_status_id.
- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp

Those one were deleted

Then the html tags in the source columns represented the second inconsistency to be cleaned up.

Finally the 4 last columns did not meet the criteria of tydinness

Programmatically it turned out that some values were missing for the expanded urls column and sometimes they were duplicated, they've been deleted.
The second inconsistency concerned the names, some had no name (None) and others were strange as (a, an ...) the common point was that they all contained lowercase letters

The timestamp have the wrong data types (object instead of datetime)

Finally, although it was clearly marked not to correct the scores in the introductory section of the project, some of the ratings in the rating_numerator and rating_denominator columns were biased,
Some of the rating in the original tweet were not the same as in the original tweet, but they have been corrected.
Sometimes the denominator was greater than 10, they were removed...

In **img_pred** dataframes :

Visually The names of dog breeds are a mixture of upper case, lower case and underscore letters to standardize the breed name I decided to replace the underscore by a space and the first letter of each word in upper case.

Finally the columns following the p1 do not respect the rules of tydinness. in this case I decided to keep only the race and status of the best predictions.

Programmaticallywe notice that this datafram is less populated with less records than the twit_arch df

some races don't exist in real life like the scorpion, the mailbox, the pillow... (since these are predictions we'll leave them as they are.)

tweet_metrics because I chose the metrics I was interested in beforehand and didn't need to be cleaned up

Once cleaned up these last 2 tables were merged with twit arch under the name twitter_archive_master . Here I noticed that the tweet_id didn't match completely between all the tables, so I delete the row with missing values.

The new table twitter_archive_master was then stored in csv format to allow an analysis in analysis_act