Author: *Marcello Victorino*

Data Wrangling Report

This part of the project consisted in **Gathering** data from 3 different sources and pre-storing them locally, and then **Assessing** and **Cleaning** as necessary (*Quality* and *Tidiness* issues).

Once cleaned, the data was saved locally as its "final" version, making sure each table contains information regarding a specific subject.

Gathering

- **Twitter Archived Enhanced**: this `.csv` file was made available and easily incorporated using the `pd.read_csv` function
- **Dog Breed**: the data was available online in a specific `url`, requiring to access it via the `requests` library and writing it to a local file. Since this data is *tabular separated*, it was incorporated via the `pd.read_csv` function, specifying the parameter `sep='\t'`, to read from tabular structure. The pre-version of this file was saved as **dog_breed.csv**
- **Twitter Specific Data**: this data was accessed via the *Twitter API*, using the wrapper library `tweepy` to easy authentication and information retrieval. With the unique *Tweet ID* for each observation, it was possible to extract additional data from each tweet by parsing the *JSON* format returned by the API (using the `json` package). More specifically, we obtained the following:
    - *Date of Creation*: allowing to extract further data such as `year`, `month`, `day of week`, and `hour of day`;
    - *Retweet Count*: amount of times a specific tweet was shared by others;
    - *Favorite Count*: number of times a specific tweet was "liked" by users

Assessing & Cleaning

Tidiness Issues

1. **Twitter Archived Enhanced**: multiple stage categories (puppo, fluffer etc.) in separate columns -> should be consolidated in a single `category` variable

2. Overall tidiness, the existing 3 dataset should be rearranged into 2 tables:
    - **Tweet data**: tweet_id, creation time, retweet count, favorite count, text, rating
    - **Dog data**: tweet_id, breed, name, category, jpg_url

Author: *Marcello Victorino*

Quality Issues

Twitter Archived Enhanced:

1. `tweet_id` stored as integer -> should be stored as string
2. Many columns filled with missing value or just useless data -> remove 7 variables ['in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'retweeted_status_id', retweeted_status_user_id', retweeted_status_timestamp', 'expanded_urls']
3. Replaced 'None' entries with `np.nan`
4. `rate` stored as *float* -> should be stored as integer
5. Remove entries with `rate` higher than 50 (refer to odd ratings or jokes... Snoop Dog rapper)
6. Missing values in Name column (there is nothing we can do. Some tweets do not have the dog's name)

Dog Breed

7. Remove irrelevant columns -> ['p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog']
8. Remove entries with Confidence Score lower than 59% (totally inaccurate breed predictions)
9. Remove entries not related to Dogs
10. `tweet_id` stored as integer -> should be stored as string

Scraped Twitter API

11. `tweet_id` stored as integer -> should be stored as string