Good Doggos: Wrangling Process

*by **Ana Júlia Bellini***

10 Dec 2021

Introduction

This project aims to put students' data wrangling skills to the test by cleaning three datasets on @WeRateDogs – a Twitter page – tweets spanning from X to Y. All of the data comes from a variety of sources, such as APIs and external servers.

Throughout the following sections, we'll bring details on the issues found and how each wrangling process step was conducted.

Data Sources

Our final dataset contains information extracted from three sources, listed below:

- **twitter-archive-enhanced.csv** – a CSV file with the Tweet content, ratings, stages, timestamps, and other information; imported from a local file.
- **Twitter API** – using my particular credentials, we've gathered favourite and retweet counts for each Tweet contained in the previous CSV file.
- **image-predictions.tsv** – a TSV file containing predictions over each dog's breed, coming from an AI algorithm. Imported from Udacity's server via an HTTP request.

To gather this data, the following libraries were used:

- pandas;
- numpy;
- json;
- io;
- re;
- tweepy.

Identified Issues

After programmatic and visual assessment performed mainly with Pandas and Numpy, there were **10 quality issues** and **2 tidiness issues** found in the data, described as follows:

Quality

1. A considerable number of columns that don't bring much value to our analyses (e.g. `expanded_url` in Twitter archive);
2. There are some Retweets among the posts that must be dropped, as per Udacity's request;
3. There are also Tweets older than August 1st, 2017 that are to be dropped as well;
4. Breed names start with uppercase and its words are separated by underscores, instead of actual white spaces;
5. Not all Tweets from the main archive are in the other two files, since some of them seem to have been deleted;
6. Rows with invalid column names (e.g. "a", "the");

7. Rows with ratings not conforming to the standard system of some score out of 10 (e.g. 24/7);
8. Some Tweets mention two dog stages;
9. Columns with the wrong datatype (e.g. `timestamp` should be `datetime`).
10. Dog breed information could be within any of three different columns, depending on the prediction's confidence level

Tidiness

1. The "dog stage" information (doggo, pupper, puppo, floofer) is spread across four other columns, when it's part of a single categorical variable, violating the "each variable forms a column" rule;
2. We currently have three different tables containing information on Tweets, when it should be all part of a single set, not abiding by the "each type of observational unit forms a table" rule;

Cleaning

Each of the aforementioned issues could be solved by **filtering the DataFrame based on certain complex conditions**, along with some special Pandas methods, such as:

- `cumsum`;
- `agg`;
- `get_loc`;
- `apply`;
- `to_datetime`;

And many others.

To get a better idea of what we're working on when treating each issue, the Define-Code-Test framework was used.

After performing all cleaning steps, we were left with a master dataset made of the following columns:

- `tweet_id`;
- `timestamp`;
- `text`;
- `rating_numerator`;
- `rating_denominator`;
- `name`;
- `retweet_count`;
- `favorite_count`;
- `stage`;
- `breed`.