We rate dogs – wrangle report

By Petra Schneckenburger

In this project the task was to gather several data belonging to the twitter site **‚We rate dogs'**, to examine the data for quality and tidiness, to clean them and at last perform some analyzes on them.

Data gathering:

There have been three initial datasets:

- Twitter_archive_enhanced.csv was downloaded manually
- Image_prediction.tsv was downloaded programmatically
- each tweet's retweet count and favorite count (and other items of interest)

The latter should have been downloaded via the twitter API.
Twitter didn't accept my applying for a developer account, so I had to use the dataset provided on the project page. I inspected the json_txt with an online json formatter https://jsonformatter.curiousconcept.com to get better insights about the structure and to decide, which other information than the counts might be of interest. Since most relevant information in that dataset was already given in the twitter_archive file, I decided to only use retweet count and favorite count besides the id.

All three datasets were opend as dataframes for assessing, cleaning and analyzing.

Assessing the data:

I assessed the data visually with Microsoft Excel and visually and programmatically in the provided jupyter notebook.

The Twitter **archive** dataframe contained detailed information about 2356 tweets, including timestamp, urls, the complete feed text and from the text extracted values as numerators and denominators of the ratings, dog names and dog stages.
The image **prediction** dataframe contained data from a neural network model trained to detect dog breeds on images. It consists of images of 2075 twitter feeds and three prediction groups, each consisting of a column with predicted dog (or other detected things/animals), the confidence and if a dog is detected.
The dataframe containing the **counts** for favorite and retweet consisted of data from 2354 tweets.

I could identify several minor issues, as undescriptive column names or not fitting datatypes, but also more severe issues like one variable spread over several columns (dog stage), and wrong values extracted from the text (rating values, dog names), and not identified dogs by the breed prediction model. There also have been retweets in the archive dataframe.
The found issues are listed in detail in the jupyter notebook.

Cleaning the data:

The different cleaning steps were performed in the jupyter notebook. Since the cleaning process is documented in detail there, I here want to address only some of the issues:

- The extraction of data from the tweet text has its limitations in identifying the correct values, there sometimes have been numerator values with decimals, or other expressions in the text that have been captuered instead. Detected errors have been corrected manually.
- The extraction of dog names out of the tweet text also had limitations in capturing. There have been captured other words like ‘a’, ‘an’… that then were replaced to ‘None’. In several cases, there still are names in the text, which could have been extracted manually, due to time limitations I left them as they are.
- The prediction model had some limitations in detecting dogs. There are some misses, where there are dogs but weren’t recognized by the model. I didn’t find any false positives. Since I found no way to discriminate between missed dogs and correct recognized other items, I decided to leave the data as they are.

The complete cleaning process resulted in a **master** dataframe with 1968 tweets used for further analyzing.