Wrangle Report

by Gerrit Tombrink

This “Wrangle and Analyze Data” project is part of the Data Analyst Nanodegree, created by Udacity. The goal of this project is to wrangle, analyse, and visualize 'WeRateDogs' Twitter data in combination with additional prediction and Twitter datasets, downloaded via a URL and a Twitter API.
The first imported .csv file of 'WeRateDogs' Twitter archive contains tweet data for all 5000+ of their tweets. The second imported .tsv file was programmatically downloaded via an URL using the Python “requests” library and saved in a new folder. This file contains a table full of image predictions about the breeds of dogs, which based on the images of the 'WeRateDogs' Twitter archive. Based on the tweet IDs of the 'WeRateDogs' Twitter archive each additional information of the Tweets was programmatically downloaded via the Twitter API using the Python “tweepy” library and saved as a .txt file. The “JSON” library helped to read and download all the tweets within this third dataset correctly. In particular, the collection of data from the Twitter API was a time-consuming task. To download the tweets it took more than half an hour.
After the gathering process, each dataset was visually and programmatically evaluated. During this assessing process, twelve quality issues and two tidiness issues were considered, and documented for the three complete datasets:

Quality Issues

- Some dog names within the column "name" are misspelt or missing. We can use the dog names within the text of the column "text" as a second reference.
- 181 retweeted records (see column "retweeted_status_id", "retweeted_status_user_id" and "retweeted_status_timestamp") should be deleted from the dataset.
- A wrong data type of the column "timestamp". This column should be converted to a datetime object.
- Missing values within the columns "expanded_urls", "in_reply_to_status_id" and "in_reply_to_user_id".
- We need to make the spelling of the dog breeds within the column "p1" consistent.
- We need to make the spelling of the dog breeds within the column "p2" consistent.
- We need to make the spelling of the dog breeds within the column "p3" consistent.
- We have to delete the _ between each word of the columns "p1", "p2" and "p3".
- We have to round the values of the columns "p1_conf", "p2_conf" and "p3_conf".
- 168 retweeted records (see column "retweeted_status") should be deleted from the dataset.
- The columns "contributors", "coordinates", "extended_entities", "geo", "in_reply_to_screen_name", "in_reply_to_status_id", "in_reply_to_status_id_str", "in_reply_to_user_id", "in_reply_to_user_id_str", "place", "possibly_sensitive", "possibly_sensitive_appealable", "quoted_status", "quoted_status_id", "quoted_status_id_str" and

"quoted_status_permalink" should be deleted from the dataset, because they have to many missing values.
- Change the name of the column "id" into "tweet_id" to make the dataset consistent.

Tidyness Issues

- Merge all four dog stages of the columns "doggo", "floofer", "pupper" and "puppo" into one stage column.
- Join the second (predic_data) and third (tweet_data) dataset into the first dataset (twitter_data).

Before the following cleaning processes, all the datasets were copied (.copy) to protect raw data. The cleaning process consists of three parts: 1) the definition of the cleaning code, 2) the programming part, which cleans the dataset, and 3) the programming test of the cleaned dataset. Several Python and Pandas methods, such as .replace, .to_datetime, .fillna, .join, .title, .round, .drop, and .rename were used to clean the dataset. Afterwards, all the datasets were merged with an inner join (.merge) and saved as a .csv file (to_csv).
It should be noted that the cleaning of raw data is the most time-consuming process, depending on the quality requirements. In particular, through these data-cleaning processes, the datasets gain in value and increase the validity of analyses.