Wrangle Report
WeRateDogs

In this project, WeRateDogs twitter account was analysed. As the name indicates, the tweets describe dogs' characteristics such as name, numerator and denominator as part of rating, timestamp, source, stage of development. On top of that, extra data was provided and collected to expand the analysis. The table of content in the beginning of the jupyter notebook contains all the steps implemented: data wrangling, storing clean data, analysis and visualization, and conclusion. The very first step, just like in most data analysis projects, is data wrangling that aims to prepare the data to feed into plots and analytical functions.

Data wrangling consists of three main parts: gathering, assessment, cleaning. First of all, we need to gather data from various sources. In this project, there were three sources.

1) twitter-archive-enhanced.csv provided by Udacity

The dataset contains information derived from the tweets directly. Read_csv function was used to store the data in a dataframe format (df_archive).

2) image_predictions.tsv also provided by Udacity

This dataset is a result of predictions on images of the tweets with breed and probability of correctness. The data frame was created via the requests library (df_images).

3) tweet_json.txt as the result of the requests library

This dataset was supposed to be gathered with the help of a Twitter developer account, however it was problematic to open one due to travelling across the countries and not getting access by Twitter. Thus, a txt file provided by Udacity was used. If a twitter account was used and a tweepy api was utilised, then we would have received the same txt file. A code snippet was copied from Udacity guide that describes how to get data via tweepy api wa added. The resulting 'data' was stored in 'json_text_udacity.txt'. The resulting data frame was obtained from another file 'json_text.txt'. This was necessary since json_text_udacity.txt is blank, because the streamed data via tweepy api is not correct due to unset twitter account (df_additional_info).

As a result of the gathering step, 3 data frames were created. These data frames were to be assessed next.

First, a visual assessment was performed and some issues could be already seen. Then, a programmatic assessment helped to finalize all issues. The issues were divided into 2 categories: quality and tidiness. Overall, 8 quality and 3 tidiness issues were found across all 3 data frames. They were documented as observations.

The next step is cleaning. Before starting to manipulate the data, copies were created for each data frame. For quality issues: missing data was fixed - deleted the columns with mostly NaN records, since they are useless for analysis, removed rows with ambiguous data such as meaningless names and null values, fixed data types and column names for consistency, changed breed names to all lower case to maintain consistency, extracted and stored useful parts from predictions set and dropped redundant columns afterwards. For tidiness: squeezed the data of 4 stages into one column and dropped the messy columns, a similar manipulation was performed with breed and confidence, and in the end, api data was merged with the archive data frame.
In the end all 3 data frames were merged into one master data frame.

In the next step, the cleaned data frame was transferred into csv and stored locally. This dataframe (twitter_archive_master.csv) was further analysed and visualised.