Wrangle Report

Introduction

The purpose of this project was to prove certain skills in the data wrangling process. This process consists of gathering, assessing and cleaning data. The available data is taken from the twitter account @dog_rates (WeRateDogs) which rates pictures of dogs in a funny way. The wrangling process is done in a jupyter notebook whereas in this report the process is described.

Content

1. Gathering
2. Assessing
3. Cleaning

Gathering

Gathering the data depends on the source where the data comes from. In this case the available data was obtained from three different sources:

- Twitter_archive: This dataset was provided by Udacity in the csv format. After reading it with the pandas command "read_csv" the data is stored in a dataframe called *df_archive*.
- Image predictions: This .tsv file needs to be downloaded manually by getting the response from a given webpage via the requests library. After that the content of the response is written in a .tsv file named *image_predictions.tsv*. Because the data is separated in tabs you can read it like a .csv file, but the "sep"-value in the read_csv command should be set to "\t". This dataframe is called *df_pred*.
- Tweepy data: With the tweepy library further information like the number of retweets or favorites can be gathered. After setting the keys and tokens of the account, I run through every tweet_id in *df_archive* and queried every tweets json data. The gathered data is written in a file called *tweet_json.txt* and read with the pandas command read_json. It is saved in the dataframe *df_tweepy*.

Assessing

After gathering the data the next important step is to assess the data. The gathered data is observed by quality or tidiness issues. These issues were obtained visually by investigating the different csv files in Excel and programmatically by using different python commands. This commands give information whether there are duplicates in the dataset or other obvious issues. In this project at least 8 quality and 2 tidiness issues should be found. Quality issues refer to the content whereas tidiness issues refer to the structure of the data.

Cleaning

In this process the previous found issues need to be cleaned. After making a copy of each dataframe, every single dataframe was cleaned by the Define, Code, Test process. Therefore each cleaning process is defined in words, programmatically coded and tested afterwards. In the end the dataframes are joined together with the necessary columns for better analysis and visualization of the data.