Data Wrangling – Final Project

For this project I worked with a database from Twitter called @Dog_rates or WeRateDogs. The whole wrangling process was performed in 3 steps. However, I iterated through these steps as I was performing them. These steps are defined as:

1. Data gathering
2. Data assessment
3. Data cleaning

Data Gathering

I used data from three different sources: A csv file, a tsv file and the twitter API.

- The csv file was stored locally in the directory of the project
- The tsv file was downloaded programmatically from the Udacity servers
- I used the library Tweepy to download further information

The Twitter API was the most challenging part, as I was asked to download the whole JSON file and store each file as a line in a text editor. Furthermore, the twitter API let you make up to 900 calls every 15 min – that is why I had to build a timer in my code that make 850 calls every 15 min. The stored JSON data had to be transformed to a DataFrame for us to be able to work with it. So, I had to extract all the relevant information from the JSON data and store it in a DataFrame.

At the end of this step we had loaded three DataFrames to Python: df_main, df_pred, df_counts. All three ready to be assessed

Data Assessment

In this step I identified all the “issues” with the data that had to be corrected in the last step. I looked for two types of issues: tidiness and quality. As expected, I found some issue regarding tidiness and a lot of issues regarding quality. More details can be found in the Jupyter Notebook “wrangel_act”. At the end of this chapter you will find a list with all the issues I found.

Data Cleaning

In this step I approach all the issues identified in the previous step. This include things like changing the datatype of variables but also creating new DataFrames that are better suited for the analysis (tidy data). The Data Cleaning process was well documented in the Jupyter Notebook mentioned before. All efforts have a description (“define”), code with further illustration, and testing.

By the end of the data wrangling process I saved all cleaned DataFrames as a csv file in my local directory and use these for further analysis of the data.