# Wrangle Report

Wrangling the data in this project was quite tough, but I think it is one of the most useful ones in the Data Analyst Nanodegree Program. The instruments used here, especially to gather data will be a big leverage for (private) projects in the future.

The task of this project was to gather data from 3 different sources:

- **archive:** An exclusive dataset from the Twitter user WeRateDogs that contains basic tweet data (tweed ID, timestamp, text, etc.) in the form of a csv-file
- **images:** A tsv-file that predicted the breed of the dogs in the tweet images with a convolutional neural network
- **twitter:** All tweets from Twitters database that have the same tweet ID than the data in the exclusive WeRateDogs database.

Gathering “archive” and “images” was no big problem since the first was given and the second was automatically downloaded via Python’s “Requests” library with a short code. However, quering the the twitter database via an API (Pythons “tweepy” package) was the hardest challenge in this project. Unfortunately the tweepy package didn’t offer much insight. Only by spending several hours researching on this subject to figure out the final code to store each tweet’s data in an entire JSON dataset, which in a next step was read line by line to create a pandas dataframe.

After gathering those three datasets I cleaned quality and tidiness issues. I began by addressing the tidiness issues. I transformed the columns the dog_stages from a wide to the tidy long format and I left joined the other two dataframes on the images dataframe to create a master_dataframe. By doing the latter as one of the first steps a didn’t have to make copies of the original dataframes because following cleaning steps were performed on the master-dataframe. To improve dataquality I handled wrong datatypes, missing data and mislabeled information mainly in the “archive” dataset. Some quality issues were already fixed by left joining on the images dataset and some by removing retweets.

I ignored some quality “issues” that in my opinion were not worth cleaning or where I wasn’t sure about the meaning. The images dataframe for instance contained a column named “img_num” but I couldn’t find out the meaning of it or whether there was a meaning. Another example is the numerator and denominator of the WeRateDogs Rating that sometimes varied to 165 and 150 but most times ranged between 8-14 and 10 respectively. I could have omitted these observations but I didn’t because I think they are no errors but they are part of the unique rating system. However, for my analysis I omitted these instances to get better results.