My wrangling process contained three major steps, which were collecting, assessing, and cleaning data. There were three areas of data I needed to collect to form the structure of my analysis. Two of them were handed to me which did not require too much efforts. However, the third part of data, which needed to be acquired from Twitter API , took significant attention and time since I needed to construct some codes the gain an access to it.

After collecting these data, I converted all three datasets to individual data frames. Next, I needed to assess each data frame and spot the issues embedded in each of them. The data frame that contained archived information was the one that had the most issues. There were seven quality issues and one tidiness issue after my assessment. The quality issues included having missing values in many columns which were not significant in terms of contributing to our conclusion, such as “in_reply_to_status_id” and “retweeted_status_user_id”, etc. Also, some rows were not original tweets. Instead, they were retweets or replies to others’ tweets which we did not need. Another big issue was that the rating system did not have a standard numerator, which indicated that the comparison between each tweet could be difficult to interpret. Other minor issues were associated with inconsistent data types, spelling format, and invalid dog names. A minor quality issues of inconsistent capitalization format occurred in the image data frame as well. In terms of tidiness issues, one major problem was that all three data frames contained information targeted to the same topic, which was tweets about dogs. There was no data that demonstrated different topics, such as description of dogs that needed additional observational units. Thus, we should construct only one data frame instead of three. Another tidiness issue was that in the archived data frame, the columns, “doggo”, “floofer”, “pupper”, and “puppo” were too scattered which could be consolidated to convey a simple idea associated with the “growing stages” of dogs. Additionally, some columns in the image data frame were too redundant, such as “p1”, “p2”, and “p3”, etc. which simply were measures of the dog breeds.

The next step, where the magic happened, was the data cleaning. Based on each problem I spotted earlier, I came up with individual step to resolve each of them. For example, we should not have rows that were retweets or replies because we only wanted original comments. Therefore, I deleted those rows. If there were columns that contained too many missing values and I couldn’t fill them with reasonable information, I got rid of them as well. For format issues, I converted timestamp to datetime, converted growing stages to categorical data, and I converted user IDs to objects for data frame merging. To standardize the rating system, I set all denominators to 10 and transformed corresponding numerators for further comparison between the ratings. Last but not least, I merged all the information within each data frame to one main data frame for the ease of further analysis.

The final data frame was clean and concise with information that only mattered to reach important conclusions for my project. The data frame wass finally left with 1971 rows for each column and there were no null values within each column.