Of the three parts of data wrangling, data gathering was the most technical one to do I found. Gathering data from a csv file was easy enough. Just import the pandas library and you're ready to go. The second file needed in this project was a tsv file located located on a server. To get this file, the requests library was needed to download the file programmatically. The result was read into a pandas dataframe for easy use later. The last piece of information we needed had to be gathered using the Twitter API. After connecting to the API using the tweepy library, the json data was collected in an txt and again read into a dataframe. After double checking I had al the information from the three different sources above, the next part of data wrangling was ready to commence.

The data which was gathered previously needs to be assessed. This started with just asking for some general information about the three datasets, the different variables and their datatype and some statistical measures when useful. I started with the first variable in the twitter archive dataframe and worked my way to the right of the dataframe. The same was done with the other two dataframes, one containing image predictions of the dog breed, linked to a certain tweet id and the other one was the additional information gathered for every tweet using the twitter api. Different issues were found, both quality issues and tidiness issues. Some tweets no images associated with them, some had invalid dog names or invalid denominators. Different columns had the wrong datatypes or had a lot of empty values. These were the quality issues. As for tidiness issues, some columns needed to be collapsed into one, obviously the different dataframes had to be merged at some point. Also, there was too much information in the dataframes, this needed to be addressed if something useful had to come out of this data wrangling process.

Next was the time to clean the above mentioned mess. But first a copy was made of every dataframe, just to be sure. First the wrong datatypes we're fixed in all three dataframes. Then some columns from the twitter archive we're deleted as they were not needed in this project. Next, we had four columns that needed to be collapsed into one, so that's what happened. The result was one column displaying the same information as the four columns, which we're deleted after the procedure was done. Apparently an ampersand was not displayed correctly in same tweet. So a replacement function was used to remove al the wrong displays of the ampersand and insert the correct normal one. The following issue which was fixed were the wrong dog names, first the wrong names had to be selected through different procedures, which were then replaced with the correct if it could be found. Also some variables were removed from the image prediction dataframe as they were just distracting and unnecessary. At last the three dataframes were merged using a lef join in both cases and save as a new named dataframe to a csv file.