WeRateDogs @dog_rates is a Twitter account that rates people's dogs with a humorous comment about the dog. Each tweet contains image(s) and text with the name, rating and other comments about the dog.

In this project we need to analyze the archive of WeRateDogs. In order to do that we need to move through the analysis steps (Gathering, Assessing and Cleaning).

First of all we should gather the data since we have no data! The twitter-archive-enhanced.csv file was given by Udacity which contains most of the data needed for each tweet (tweet ID, timestamp, text, name, rating,etc...). The second data was the image prediction file that was given as a link, so I downloaded the file porgramatically using python and pandas. This file contains prediction results of three different algorithms that predict the image of the tweet with % of confidence and the breed of the dog. Third data was collected through the Twitter API after creating a developer account on twitter. This data Contained the number of retweets and favorites for each tweet.

For sure the collected Data wasn't clean enough for analysis, so I started with the assessing phase.
Through Assessing, we categorize the issues under quality and tidiness. Some of the issues were detected through Visual assessment like name of the columns in the prediction table, and prediction dog names were of lower and upper case. But for sure not all the issues can be detected by the eye, so I go to Programmatic assessment and detected several issues, like wrong datatypes and many null values.

If the assessing phase were clear, it will be easy to start with cleaning the issues. For example I found a tweet with rating=960 but in fact it was a reply not a tweet, so we drop its row, rather than dropping another columns related to the retweets and replacing 4 columns of the dog stage with 1 categorical column. For sure after each cleaning code you should run a test code to check if your issue was cleaned or no.
After finishing these 3 steps successfully I stored the cleaned data into new CSV files to start my analysis.