Data Wrangling

The data I've wrangled was about the twitter archive of @dog_rates acount, also known as WeRateDogs that rates people's dogs with a humorous comment about the dog. The wrangling process has been done in 3 steps:

1- Data gathering

In this part I wrangle data from 3 different resources (csv file, file on a given URL, text file of JSON objects gathered using tweeter API). The difficulty was in dealing with different format, especially JSON objects, as it has many nested list and objects.

2- Assessment

This is the process of documenting the Tidiness and Quality issues using visual and programmatic assessment.

The visual assessment has a power more than expected, a good understanding of the data helps on identifying the problems that might exist on the data. This is a good guide of what code should we write in the programmatic assessment.

Programmatic assessment is done using pandas. Pandas has many useful methods that facilitate the process of assessment.

This part was the easiest but it's important, as documenting the issues via assessment make the cleaning process organized and helped on deciding what to do next.

3- Cleaning:

This is the process of finding solution to the issues that is documented in the assessment step. This is the hardest part and it uses some complicated functions, to break things down, it's done in three steps for each issue, Define (define what actions should be done to solve the issue, it's like pseudocode), Code, and Test. One important thing is to make a copy of each dataframe before cleaning. I've faced many problems while cleaning and if I didn't make a copy, I will waste my time in running all the cells from the beginning. Another difficulty is which to clean first, the issues are dependent, and some issues couldn't be solved till other issue is solved, this was a challenge for me.

Data Wrangling takes too much time but the result is interesting and help to make many interesting conclusions.