

# Wrangling Report

**Note: All the details will be mentioned in this document is in the notebook, where you can see the code after the definition.**

## 1. Gathering

In the gathering process, I collected three datasets.

- a CSV File
- a TSV File
- a data from tweepy API

## 2. Assessing

the assessment process was a little trickier, at first, I tried to assess the data visually, where I first noticed that the denominators are not all out of 10! using tweet IDs, I opened some random links and I found that I have to run through them all. I wrote a small script to open all of them in my browser. and I divided them into three categories:

- unrelated or corrupted data. I dropped them.
- more than one dog, so he rated them out of their number multiplied by 10. for example. 7 dogs are rated out of 70. it may be sarcastic when you see it on twitter. but trust me it is not when you have to clean the data.
- more than rating-like numbers. such as dates or 24/7.

Then I moved on the columns themselves, as usual. when importing dates from a CSV or a TSV file. they are strings. I converted them into datetime in order to improve insights accuracy and specificity.

some IDs were miscategorized as floats, I converted them into strings

That is about the quality issues, how about the tidiness?

The data was tidier than more I thought! and merging it into a single DataFrame did not need any work more than renaming a column. the one that was quite challenging that the last four columns (doggo, floofer, pupper and puppo) in the DataFrame that extracted from the CSV file. I decided that it must be melted into a single column

## 3. Cleaning

Applying the three stages of cleaning: defining - coding - testing. the cleaning process was tricky, sometimes hard. but it never became messy! for example: when dealing with the non-consistency in ratings. I collected each similar IDs in a single data structure (list, dictionary when needed) and iterated through them to change them. changing the datatypes was not any hard at all. but in some columns that have too much missing data it was annoying to look for an existing data to check it.

My philosophy in cleaning is choosing a series from the DataFrame. then I make a dummy variable: foo, bar or baz. - which I did not include in the notebook at all- then I try with it until I reach the cleanest solution. then I rename it into the original DataFrame name