Data wrangling report

Introduction:

In this project, we will wrangle and analyze data from (@WeRateDogs) Twitter account. Where there is over 2500 post that rate dogs based on certain criteria. This report will contain all the details on each step starting with Gather the data step and ending with Clean the data step.

Gather:

Gathering data is the first step of wrangling data and is considered the longest one, since collecting data is not always easy.
However, in this project we gathered our data from three different sources:

*1- **Twitter archive file**: we gathered the data manually by downloading the file ('twitter-archive-enhanced.csv') and then upload it to the Jupyter notebook stage.*

*2- **Image prediction file**: we gathered the data programmatically by importing the Request library.*

*3- **Tweets information file**: This file supposed to be gathered by querying the Twitter API but because I didn't get access from Twitter, I downloaded it manually, It was provided by Udacity in the classroom.*

Asses:

In this step what I have done basically is inspecting the data trying to figure out Quality or Tideness issues.
Here is some of them :
1- Inappropriate datatypes in certain columns. (Quality)
2- Labeling null values on the name column as None instead of leaving it as NaN. (Quality)

3- Rating denominator out of 0, this is logically unacceptable. (Quality)

4- There are four observations that form columns (doggo, pupper, floofer, puppo). (Tidiness)

5- We can combine all tables into the main dataset Twitter archive table, it`d be tidier. (No need to split the dataset into many tables, since they all relevant). (Tidiness)

Clean:

In this step and before we actually cleaned the data we created a copy of the original data to work upon it. Working on a copy is very important because sometimes we need to recover the original data. After that, we cleaned the data that we have gathered by applying certain codes that suit our issues.