Wrangle Report

Gathering phase

1) First of all in this project we have **three main datasets** that are needed to fulfil our analysis.
   a. The **Twitter_enhanced_archive.csv** which is provided by Udacity and it has the needed data of the tweets of **WeRateDogs** account on twitter.

   b. The second dataset was the image prediction data set which we downloaded it programmatically and then put it in a DataFrame.

   c. The twitter API dataset, I didn’t have the credentials so I worked from the data provided by Udacity and also put it in another DataFrame.

Assessing phase

1) The data was assessed visually first by viewing them on Jupyter Notebook and also on Excel
2) Then The data was assessed programmatically and defined the data which needs to be cleaned in the next phase
   a. The Quality issues were addressed first for each data set
   b. Then Tidiness issues were addressed

Cleaning Phase

Quality issues was cleaned first, and then tidiness issues.

1) First I made a copy of all the data sets
2) I deleted the replies and retweets from the archive data (please note that the tweets without images will be dropped once the image data set is merged to this one)
3) I then dropped the unneeded columns in archive data
4) Deleting ('\n') from some values in text column using str.replace()
5) Replacing the “None” values to first “” (empty cells), then replaced to Nan after step no. 10
6) Fixed some ratings manually and deleted extreme ones
7) Deleted duplicates in image prediction data
8) Reshaped the image prediction data
9) Dropped the unneeded columns in image prediction data
10) Then fixed the Tidiness issue in archive data which is dog classification columns
11) Then merged the two other data sets to master one