WRANGLING THE DATA REPORT:

The data wrangling part for this project has been performed in three steps, gathering the data, accessing the data, cleaning the data

Gathering:

For this project I used data sets

1) The we rate dog twitter enhanced by Udacity
2) The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.
3) Data extracted from twitter like the retweet count, favorite count.

Assessing the data:

There are several type of issues related to the data sets present. Few of them were Quality issues, few of them were tidiness issues.

I started assessing the quality issues first,

The first data set the twitter enhanced has several issues within itself. There were names for the dogs that have been entered wrongly, Time stamp column associated with the data set is given as string but it was originally supposed to be a datetime data type. There were several missing values in the data set that were missing for example: in_reply_to_status_id','retweeted_status_user_id','retweeted_status_timestamp' ,'in_reply_to_user_id','in_reply_to_status_id','retweeted_status_id' has lots of missing values and they are to be dropped.

The text field has several unnecessary characters that are to be omitted and the source column had links associated with it.

In the predictions data set some breed names have the first letter lowercase in p1, p2, p3 columns. All the above mentioned issues are to be cleaned.

Tidiness issues:

The date and time column should be separated into two columns. We can also create a new column for the ratings ratio, i.e, numerator divided by the denominator.
Dog stages are split into four different columns they can in whole be combined into one.

In the predictions data set p1,p2,p3 has several dog species they can be condensed into a single column and we can perform analysis on that.

Finally merge the three data frames into a single and after that we can check if any duplicates are present and if any missing values are present and drop them.

The clean data frame is obtained and the data wrangling process is completed but at any time if new data is added we can follow any of the steps data gathering, accessing and cleaning can be done at any stage .