Wrangle Report

This project is aimed as an introduction to all the activities involved in data wrangling. The dataset used for this project is the tweet archive of a Twitter account known as "WeRateDogs". This account rate images of dogs belonging to other people using a humorous approach.

The dataset needed for this project are to be got from three different sources. These datasets are then evaluated and clean. After cleaning, the dataset is then merged and analysed. The first part of the data, which is the twitter archive for "We Rate Dogs" was provided in the instruction page for this project. I only downloaded and then uploaded in this Jupyter notebook for analysis. The second part of the data, which is the "tweet image predictions" is available on the Udacity server as a "tab separated value (tsv)" file. We are required to download this data programmatically using the "request" library. The third and final part of the data is the retweet count and the favourite (like) count of each of the tweets in the dataset. This was supposed to be got using the Twitter API. However, since the API is not open source and because of the difficulty involved in getting Twitter's permission for this, the needed data was provided in a text file. The text file is then read line by line into a dataframe. After collecting the needed data from all three sources, they are then assessed, cleaned and processed for the analysis.

For this project, we are instructed to list out at least 8 quality issues and 2 tidiness issues observed in our data. To do this, we can either visually assess the data or programmatically access the data. Some of the issues can be identified by just looking at the data, using a word processing application like Microsoft Word, or we run some codes to check for issues like duplicates, missing values, etc. Some of the quality issues observed includes: datatype of the “timestamp” column needs to be changed to "datetime", the dog names are not in capital letters, there are lots of columns with 'none' as the entry, etc. Tidiness issues observed are: data in all the tables must be combined to form just one table, some of the tweets represents more than one dog.

```
In [ ]:
```