# WRANGLE REPORT

Data wrangling simply refers to the process of gathering, assessing and cleaning data to be able to make sense out of it. There are three main processes involved in data wrangling. These include gathering the data, assessing the data and finally cleaning the data.

In the project, we rate dogs' data was gathered from three different sources before it was assessed for quality and tidiness issues after which these issues were resolved by cleaning.

In the gathering process, data is acquired or collected. The first set of data which is the twitter archive was downloaded manually from a link and saved to a workspace. The second part of the data which is the image predictions data that contains a neural network for dog breed prediction was downloaded programmatically with the request's library. Finally, the last piece of data was obtained from the twitter API in json format.

The next step is the assessing process which refers to looking through the data to identify issues that will make the data difficult to work with. This can be done both visually or programmatically. After gathering the data, each dataset had some issues that needed to be resolved. These issues can be classified as quality or tidiness issues.

For the twitter archive dataset, the following quality issues were observed;

1. Missing values in 'in_reply_to_status_id', 'in_reply_to_user_id', retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' and 'expanded_urls'
2. 'in_reply_to_status_id' and 'in_reply_to_user_id' should be integers
3. Timestamp should have had a datetime attribute
4. Dog stages had None instead of NaN
5. Dog names had None instead of NaN

For the image predictions dataset, the following quality issues were observed;

6. There were 2356 rows in archive and 2075 rows in images
7. Some images were missing

For the twitter API dataset, the following quality issues were observed;

8. Duplicate rows in data that was obtained from API
9. Retweets and Favorites should have been integers

The following tidiness issues were also observed;

1. Tables should be combined into a single master dataframe
2. One column could be created for the dog stages

The final step in the wrangling process is the cleaning step. The above issues observed in the three datasets were resolved using manual and programmatic processes after the three datasets were combined into a single master dataframe. These steps included removing the unnecessary columns, changing the datatype of the timestamp to datetime, replacing None with

NaN, combining the dataframes and removing duplicates and empty rows and finally converting retweets and favorites datatypes to integers. In solving the tidiness issues, all three dataframes were merged into a single master dataframe and a new dog stage column was created to contain the values of the different stages of the dogs.

In the end, the master dataframe was tidy and very easy to work with. Visualizations were then made from insights observed in the data.