# WRANGLING REPORT

## INTRODUCTION

In this section of the wrangling report, I'll briefly go over how I uncovered all the data quality and tidiness problems in the dataset and how I went about fixing them.

## BODY

After loading all three datasets required for this project, I proceeded to identify data quality and tidiness issues, where I found the following:

- I investigated the top 5 rows of the datasets using both programmatic and manual assessment. I noticed many columns in the data which contained missing values, while some missing values were misrepresented, such as "None".

I solved these issues using the pandas replace() method to replace all instances of the "None" value with N/A in the dataset.

- By going through the details of the loaded Twitter archived data using programmatic assessment, I noticed the following issues:
    - Tweet id is loaded as integer type instead of string, and this happens because tweet id contains bunch of numbers.
    - There are only 78 values in the in_reply_to_status_id and in_reply_to_user_id columns.
    - The timestamp column is a string, instead of a datetime object.

  I solved the erroneous data type using both astype() and pandas.to_datetime() functions to change both the tweet id and timestamp column to their accurate format. Since the analysis did not require using retweeted and replies to tweets, I removed all the retweeted and replied tweets from the dataset.

- Then, I checked the descriptive and analytical summary of the dataset. This is where I identify the following issues:
    - Abnormal rating's numerator and denominator.
    - Url at the end of each tweet.
    - Unknown dog names, e.g., "such", "the", "this".
    - Inconsistent dog breed names, i.e., separated with underscore instead of space, and some breed types were capitalized while others were lowercase.

  After checking the ratings given to some dogs in different tweets, I noticed that some rating numerator and denominator were wrongly represented, e.g., 13.5/10 is represented as 5/10. I tried using a regular expression pattern to match the ratings in the text, which works perfectly for fractional and integer ratings, but some tweets have multiple ratings, which leads the regular expression to match the wrong ones, which I couldn't help but remove all ratings that are above or below the normal rating interval. In order to solve the issue of the unknown url at the end of each tweet, I replaced all urls in the tweet text with empty strings, and this solves the issue. I noticed that all dogs with unknown names have the same pattern, i.e., lowercase, so I dropped all dog names with lowercase. I solved the issue of inconsistent breed type by replacing underscore with space and also capitalizing all dogs' breeds.

- Looking at the dataset, I found that dog stages were spread across multiple columns, which I identified as a tidiness issue.

I resolved this by using the pandas melt() method to place the column names and values into two different columns where I removed the dog stage without a value identified as N/A. So, I merged the new dog stage back to the original data frame using the tweet id in the dataset.

To deal with the tidiness issue, I merged the three data frames based on the common tweet id found in the data set.

## CONCLUSION

A significant number of data observations were lost during the data cleaning step. The entirety of the dataset is clean and devoid of poor data quality issues. All issues were adequately addressed, and the project notebook is well-documented.