Data Wrangling Report: WeRateDogs

In [ ]:
```
To wrangle the data successsfully, I followed these stages:
```

Gather

The data used for the analysis was gathered from three sources. The first source was a CSV file provided by Udacity. I had to download the file by clicking on the link included in the Project instruction, then read the content into a dataframe called `twitter_archive`. The second source was a TSV file also provided by Udacity and to be downloaded programmatically using the `requests` library and the provided URL. The file was then read into a dataframe called `twitter_img_prediction`. The third source required getting the data from the Twitter API. To achieve this, I had to create a developer's account with Twitter by providing adequate information about the intended project. After my account was approved, I was provided the API keys,Token and used `Tweepy` to query Twitter API for the archived tweets using the `tweet_ids` provided in the CSV file. Each tweet was read into a new line and saved in a TEXT file. I then, read each tweet line by line using the `json` library to create a dataframe called `twitter_new_data`.

Assess

To assess the data, I performed two types of assessment; visual and programmatic. For the visual assessment, I used `Google spreadsheets` to manually scrutinize the data and I used some `Pandas` functions to assess the data programmatically. Overally, I noticed some issues related to quality, tidiness, and missing data. Some of the issues include missing data in the `twitter_new_data` probably due to some tweets being deleted, incorrect extraction of dog names, irrelevant columns, and incorrect datatypes. These issues were adequately documented.

Clean

Before cleaning the data, I had to create copies of each dataframe to prevent tampering with the original dataframe. Then I started with the Quality issues before proceeding to Tidiness and Missing Data Issues. For some of the quality issues, I deleted tweets that were actually retweets, converted datatypes, renamed a column, etc. For tidiness issues, I represented the dog stage as a column instead of having several columns to represent this information. The tables were merged as they contain related information and did not need to be in several tables. Generally, most of the cleaning process was in the `twitter_archive_copy` dataframe. After cleaning the data, the cleaned data was saved to a CSV file called `twitter_archive_master`.