# Project: Wrangling the Dog Rate twitter account + other files

## Gather

In this part of the project, I had to gather the information from 3 different types of sources, first the *twitter-archive-enhanced.csv* which was ready to downloaded it and then uploaded in the workspace. The second file needed to be downloaded from a specific URL and programmatically.

The last file was the hard one, because through the Twitter API I needed to make a request with the twitters ID that already had from the *twitter-archive-enhanced.csv* file, then gather the response from the Twitter API and save it in a Json file. In this part I had to create a Twitter Developer account in order to get the keys to making the request.

## Asses

In this part I decide to use Excel to visually asses the dataset that I gather in the previous part, it helps to review all the datasets.

Also, I use the info() function to observe the datatypes and the null values in some columns.

Another function that I use was describe() which statistically describe each dataset, which this function I see some issues with mean in the numerator and denominator from the rating.

After assessing the datasets, I came up with 9 quality and 2 tidiness issues:

## Quality Issues

### Twitter-archive

- There are retweets in the dataset
- Wrongly extracted rating components (numerator and denominator)
- Wrongly extracted the name of the dog
- The timestamp should be in date format and dog_stage should be categorical, plus side *tweet_id*, *in_reply_to_status_id*, and *in_reply_to_status_id* should be strings.
- In some cases, the dog stage is missing
- Tweets with no images
- The 'text' column is show contracted

### Image-predictions

- The dog’s breeds are merge by a _ in the cases of compound names
- In the case of the dog breeds, some of them are in lower cases

### Tidiness Issues

- Each dog stage is a column in the *twitter_archive* dataset
- We can easily merge *df_twitterApi* and *image_prediction* with *twitter_archive* in order to get one master table

## Clean

In this part I decided to start with the tidiness issues because they were only 2 issues, so for the joint of the 3 dataset in one master dataset I used the merge function, then for the *dog_stages* columns I extract the word puppo|pupper|floofer|doggo from the text column, according the case, and saved it in the new *dog_stage* column, then drop the puppo|pupper|floofer|doggo columns.

For the quality issues I started in order as I found them in the assessment stage, in the project detail it said that we shouldn't consider the retweets so I only save the data that didn't have a value in the *retweets_status_id* column.

I had 2 quality issues that took me some time to fix them, the first one is the numerator and denominator from the dog rating, eventually, I came up with the correct regex so I can find the correct value.

In the case with the name of the dog, I had 3 scenarios that they were: when the name of the doge was around the word 'named', around the word 'name is' and in the case that there was no name.

For the rest of the quality's issues, I used different function that I saw on the cleaning part of this course.

After finishing with the cleaning part, I saved the dataset into a csv file with the name of twitter_archive_master.csv