# Investigate a dataset (weratedogs)

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

## Gathering

In this project firstly I imported pandas and numpy to my workspace, then I directly downloaded the Twitter archive data named (twitter_archive_enhanced.csv) from my udacity classroom, then I read the csv file into my notebook. The second file needed for this project was the image predictions tsv which I programmatically downloaded into my workspace as instructed in the classroom using requests. The last gathering of data was to use the tweepy library to secure additional data via the twitter API.

## Assessing

The 'twitter_archive_enhanced.csv' had a total of 17 columns of which 4 columns were floats, 3 integers, 10 strings and 2356 rows . The image predictions tsv had 2075 rows and 12 columns of which 3 were booleans, 3 floats, 2 integers and 4 strings. The 'tweet-json.txt' had 2354 rows and 3 columns, all integers.Then i made copies of the three data frames and joined them into a single dataframe which I named joined_df which had a total of 2356 rows and 32 columns.
We were told to detect and document at least 8 quality issues and 2 tidiness issues and use both visual and programmatic assessment to access the data.
The issues i listed were;

### Quality issues

1. Irrelevant columns: dropping irrelevant columns because they are not needed
2. removing retweet rows that have non-empty
3. Drop retweet columns

4. Correct columns with wrong data types
5. invalid data: name column has invalid dog names
6. source column needs to be cleaned to make data clearer
7. some data in rating were not extracted correctly and datatype issue
8. Missing values: drop columns with nan values

## Cleaning

To solve the quality issues listed above, I first went ahead to drop columns which I felt were irrelevant to my analysis. such as ‘in_reply_to_status_id’, ‘in_reply_to_user_id’, ‘expanded_urls’ etc which reduced the number of columns to 18.  Secondly, I removed retweet rows that have non empty rows which reduced the tweet_id from 2356 rows to 2175 rows which lead to the third issue which was to drop the retweet columns, the columns were now empty so i had to drop them. For the fourth issue, i corrected the columns with wrong data types (I changed the timestamp column to a datetime using pd.to_datetime and also the ‘tweet_id’ column to a string using the .astype() method). The fifth issue was solved by using code to fish out the strings in small letters out of the name column to get the invalid dog names such as a,an,just,unacceptable,etc. For the sixth quality issue I extracted the source from the string using code also. Another issue was in the rating columns(‘rating_numerator’ and ‘rating_denominator’) some values were not extracted properly, so regex was used to extract the correct rating from the text column, then i also changed the data type of the rating columns to floats. Then lastly i dropped nan values and all columns in the joined_df data frame became a total of 344 rows each.

### Tidiness issues

1. Merging the three data frames into one
2. Join "doggo", "floofer", "pupper", "puppo" columns into one