# Gathering

Three distinct methods were required for gathering the data.

### Twitter Archive Data

This was easily the most straightforward, as the data was provided by Udacity. It was a simply matter of opening it up.

### Images Data

Gathering the images data was slightly more challenging, but still fairly simply. Udacity provided a link to the file, which I was able to download programmatically.

### Twitter Data

Using the Tweepy API, I was able to download twitter data and store it as a json file.

# Assessing

I completed a visual inspection of each dataframe to identify obvious abnormalities. For example, it was immediately obvious that there was a tidiness issue with the four dog stages (doggo, pupper, *etc*) broken out into separate columns.

However, due to the volume of data, programmatic assessment yielded a far more thorough understanding of the quality and tidiness issues. Particularly effective functions were:

- info()
- value_counts()

# Cleaning

This data set requires substantial time and effort to clean. While cleaned critical quality and tidiness issues, many remain. For example, in the *images* table, there are dozens of inaccurate dog breeds. In fact, many are not dog breeds at all! They are often things like furniture or sea animals.

Additionally, there are some quality  issues that are hard to assess. For example, there are many ratings less than 10. Are these typos? Or are these intentional? A score of less than 10

appears to be in conflict with the spirit of the account. To determine accuracy, it would at least require reading through the texts and looking through the images. Or better yet, contacting with the author of the tweets, which would be nearly impossible. Assuming one could find the author, the author would then need to remember the intentions behind hundreds of old tweets.

# Reflection

Wrangling data is incredibly burdensome. It requires a huge amount of time, patience, and technique. I should have left far more than a single day for this activity. I found myself growing frustrated and fatigued, particularly over writing and reading the tweet data. Furthermore, downloading large amounts of data simply takes time.

It was shocking the imbalance between the time spent cleaning and the time spent analyzing. While I did not measure it, I suspect it was nearly 10:1 - 10 hours spent wrangling and 1 hour spent analyzing. Granted, I was so exhausted by the end of wrangling, I did not put my best foot forward in the analysis. I simply wanted to complete the project on time and go to bed!

I have a great appreciation for the challenge AND importance of data wrangling.