**Introduction**

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user [**@dog\_rates**](https://twitter.com/dog_rates), also known as [**WeRateDogs**](https://en.wikipedia.org/wiki/WeRateDogs). WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "[**they're good dogs Brent**](http://knowyourmeme.com/memes/theyre-good-dogs-brent)." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs [**downloaded their Twitter archive**](https://support.twitter.com/articles/20170160) and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

**Project Motivation**

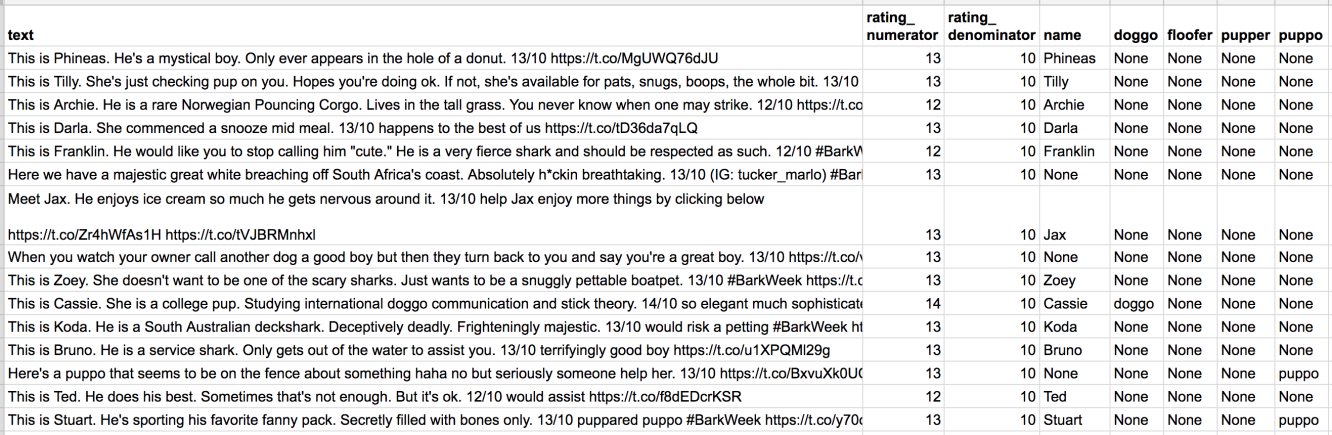
**Context**

Your goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "*Wow!*"-worthy analyses and visualizations.

**The Data**

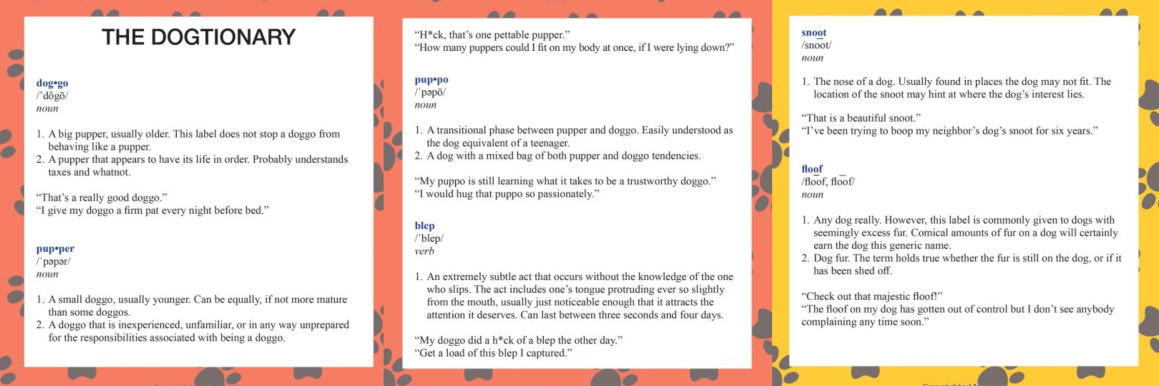
**Enhanced Twitter Archive**

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

**[[](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)**

***[The extracted data from each tweet's text](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)***

I extracted this data programmatically, but I didn't do a very good job. The ratings probably aren't all correct. Same goes for the dog names and probably dog stages (see below for more information on these) too. You'll need to assess and clean these columns if you want to use them for analysis and visualization.

**[[](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)**

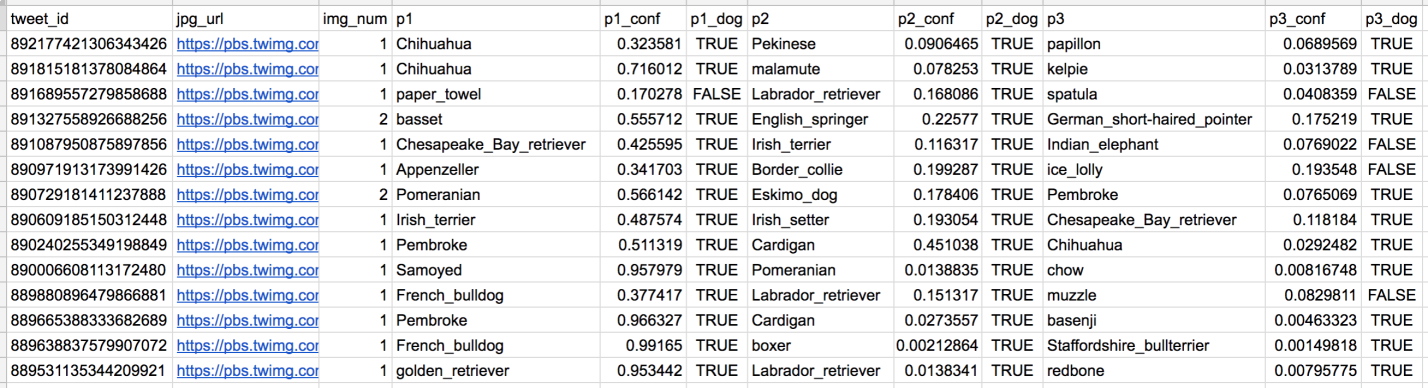
***[The Dogtionary explains the various stages of dog: doggo, pupper, puppo, and floof(er) (via the](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)***[***#WeRateDogs book***](https://www.amazon.com/WeRateDogs-Most-Hilarious-Adorable-Youve/dp/1510717145)*on Amazon)*

**Additional Data via the Twitter API**

Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

**Image Predictions File**

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a [**neural network**](https://www.youtube.com/watch?v=2-Ol7ZB0MmU) that can classify breeds of dogs\*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

**[[](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)**

***[Tweet image prediction data](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)***

So for the last row in that table:

* tweet\_id is the last part of the tweet URL after "*status/*" → [**https://twitter.com/dog\_rates/status/889531135344209921**](https://twitter.com/dog_rates/status/889531135344209921)
* p1 is the algorithm's #1 prediction for the image in the tweet → **golden retriever**
* p1\_conf is how confident the algorithm is in its #1 prediction → **95%**
* p1\_dog is whether or not the #1 prediction is a breed of dog → **TRUE**
* p2 is the algorithm's second most likely prediction → **Labrador retriever**
* p2\_conf is how confident the algorithm is in its #2 prediction → **1%**
* p2\_dog is whether or not the #2 prediction is a breed of dog → **TRUE**
* etc.

And the #1 prediction for the image in that tweet was spot on:

**[[](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)**

***[A golden retriever named Stuart](https://classroom.udacity.com/nanodegrees/nd025-indsc/parts/9fc59674-5b65-480f-81cc-ce6178c15acc/modules/8294e967-eaa8-437f-b8af-4ca08baa76b5/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576)***

So that's all fun and good. But all of this additional data will need to be gathered, assessed, and cleaned. This is where you come in.

**Key Points**

Key points to keep in mind when data wrangling for this project:

* You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
* Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
* Cleaning includes merging individual pieces of data according to the rules of [**tidy data**](https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html).
* The fact that the rating numerators are greater than the denominators does not need to be cleaned. This [**unique rating system**](http://knowyourmeme.com/memes/theyre-good-dogs-brent) is a big part of the popularity of WeRateDogs.
* You do *not* need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.