Data Wrangling Overview:

1. Gather - Gather up the various datasets.
2. Assess - Assess the quality and tidiness issues with the data sets.
3. Clean - Tidy up the data sets, removing bad data, combining into a single set.
4. Store - Store the cleaned and combined dataset to a new file.
5. Analyze and Visualize - Analyze the data and support the analysis with good visualizations of the data.

# Gather

I started by gathering the various datasets. This went well for the Enhanced Twitter Archive and the Image Predictions, but the instructions for using the twitter API make no mention of the free access having been stopped, or curtailed so much that this step could not be done as laid out. I wasted far too many hours trying to figure it out. I finally went with the alternate method and downloaded the provided additional data.

# Assess & Clean

I assessed the three data sets using both visual inspection and programmatic assessment. Through this process I found a number of issues that would need to be addressed. They, and their resolutions are listed below.

## Tidiness

1. three data sources
2. "Stage" data spread across multiple columns

The separate data files were joined on the tweet\_id field and saving to a master file.

The puppo, pupper, doggo and floofer columns were merged to a single stage column. Then all five columns were dropped as useless busy work, as they were so poorly extracted, and so many tweets actually contain two or more of those, and the "floofer" descriptor is of a physical characteristic whereas the other three relate to the age of the dog. So, should the original wrangle\_act.ipynb file be lost, but this one somehow preserved, don't bother with the "Stages", just delete them.

## Quality

1. Posts w/o photos were removed from the data set.
2. Not original posts, that is replies & retweets, were also removed from the data set.
3. rating\_numerator was changed to a Float and the value was re-extracted from the text data.
4. name values were extracted poorly turning "O'Malley" into "O", this was corrected with a one-off replacement.
5. "None" being used as a name - these were changed to "none" then delt processed further in the next step.
6. name values that are not proper nouns, just words from the post. These values were all dropped from the dataset, replaced with NaN.
7. retweets in the image prediction data, these were not merged into the master file, as the retweet records had already been removed from the main data file before the merge, and so the tweet\_id could not be matched.
8. image predictions that did not predict a dog were dropped from the data set.
9. image predictions missing dogs that were actually in the images. There was nothing to be done about this issue in this class. These records were dropped from the data set.

# Store

The cleaned data sets were merged into a single set (as mentioned above) and it was saved as twitter\_archive\_master.csv as instructed.

# Analyze and Visualize

I analyzed the data and came up with three insights plus a bonus. I also created three visuals, using seaborn for the second and third:

1. A pie chart showing the top ten dog breeds from the data set.
2. A histogram showing the number of tweets per week
3. A scatter plot with trendline showing the popularity of the tweets over time and the popularity trend.

The "Tweets Per Week" and "Tweet Love by Tweet Age" make understanding the data much easier.

A bonus insight was included at the end with a brief discussion of the viral nature of tweets and observations on how more than one photo, or even better a video clip could be used to draw more attention.