Project Information- In Project 4, I worked on gathering data from various sources, assessing the data that I gathered, cleaning the data after assessing the data, and finally analyzings and finding insights into the data. I will describe each process in more detail down below

Data Gathering- The first set of data came from archive.csv which was downloaded from the Udacity website. This data was then put into a dataframe. The second set of data came from the internet and was brought in using the requests library. This is the source of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. This tab separated file was then put into a dataframe. The third set of data came from the tweet-json.txt file. The information was in JSON format, then put into a dataframe.

Data Assessment- I assessed the data from these sources using both visual and programmatic assessment. After doing both of these, I found issues with the data from the gathered sources.

The issues with the archive data (weRateDogsDF) included the following: some of the names in the 'name' column were not actual names (there were entries that were None and some others words that were not names), there were unneeded html tags in the source column and only the important information outside the tags were needed, some of the entries were retweets (which were not needed), there was an instance where the denominator was zero in one of the rows, there were some duplicate urls in the expanded urls column, and the dog stages were in separate columns when they could be combined into one column.

The issues with the image predictions data included the following: there were predictions that were not dogs, there was an unnecessary column called img_num which could be removed, some of the dog breeds in the p1,p2,p3 did not have consistent capitalization, there were also duplicate jpg image urls.

Data Cleaning: After finding all these issues, I dealt with them by dropping unnecessary columns, removing duplicates, removing certain entries/rows based on a certain condition, combining some columns into one. Finally, I combined all the data into a master table.

Data Analysis/Insights: Some of the insights that sought to find included the following:

1. What is the most common dog name
2. Which dog stages had the highest average rating.
3. Which dog had the most favorites and retweets.