We Rate Dogs

Importing Libraries:
We begin by importing our necessary libraries Pandas for data wrangling and analysis, Requests will allow us to download our files programmatically and Tweepy for accessing the Twitter API.

Gathering Data:
We now move on to the gathering phase where will be acquiring our data through different means. The first file is a local csv file called twitter_archive_enhanced using panda's read csv function we create a data frame called tac. The next step dataset we want is located on the web, so we use the requests library to get the data from the url and then we write the content to a local tsv file which we then read into a pandas data frame called img_pred. The last portion of our data comes from the use of the tweepy library in conjunction with the twitter API. Using a for loop we are able to grab the favorite count and retweet count for the tweet ids we have in our dataset. While looping through the list of tweet ids we are writing the favorite count and retweet count on a single line, separated by a space before creating a line break. Once all of our information has been downloaded we write our information into a pandas data frame for later use.

Assessing:
Once we gather our data from the appropriate sources we are able to move on the assessing phase, this phase is quite important, we must now review our data both visually and programmatically to determine what we need to clean in order to produce quality analysis. As someone who has yet to become a seasoned professional, I like to start with the obvious methods in order to get a feel for the data. This includes missing rows, duplicates, erroneous data types that way I can gradually build momentum while trying to uncover more and more complex issues. During the assessing phase the focus is on creating awareness of the issues and not actually correcting them until we get into the cleaning phase. We were able to uncover 14 Quality issues and 2 tidiness issues.

Clean:
The Cleaning phase was the final phase in our wrangling effort. Using the define, code, test methodology we would isolate issues that we identified in the assess phase. We would clearly identify what the issues was and how we would solve it. We would then move on to problem solving and then testing to ensure we got the outcome we were looking for. One example of an issue that was repeated across multiple data sets is the tweet_id, it was interpreted as an int, however it needed to be converted as a string to accurately reflect the fact that it was a unique identifier for each tweet. Aside from the data transformations we had to perform, there was also a lot of "fat trimming". It doesn't make sense to keep data that we would not be using in our analysis, so we dropped multiple unnecessary columns. We then consolidated all of our data sets into a single one.