Gathering Data

The WeRateDogs Twitter data is a robust and challenging set of data that, when cleaned, could provide some interesting insights. The first step in wrangling this data was gathering the data. There were three separate data files needed to obtain all the necessary information and each data file was obtained through different methods

- Loading a saved csv file – for twitter_archive_enhanced data
- Downloading a file from the internet – for tweet_image_predictions data
- Utilizing Twitter's API and working with JSON to extract data – for retweet and favorite count data

Assessing Data

Once the data files have been successfully gathered, it was time to assess the data. I started off by reviewed the structure of each of the data tables and seeing exactly what information each table contained (and did not contain). As I was reviewed each table and came across different issues, I logged each issue and categorized them as either a Quality issue or Tidiness issue to revisit during the Cleaning stage. In this step I found a handful of issues such as tweet_ids listed in one table but not present in another or not present when looked up in the API. Another example of an issue/action item was merging data were necessary and consolidating columns in an effort to make the master data more tidy.

Cleaning Data

The last step in this data wrangling exercise was to clean the data I have gathered and assessed. This is where all the issues and action items noted in the Assess step were acted upon. Here, a variety of Pandas and dataframe methods were used to merge, melt, drop parts of each data file to form a universal master file which could be used for insights. A number of quality issues, such as inaccurate dog ratings, names, and styles were also addressed in this step. Once all the issues I noted were resolved, I saved the updated dataframe to a csv file.