My wrangling efforts started by gathering the data required. I read in twitter_archive_enhanced.csv as df1, image_predictions.tsv as df2, and tweet_json.txt as df3. Unfortunately, I was unable to use the tweepy library myself to query the data from the Twitter API.

I started assessing df1 by verifying the data was retrieved correctly. I then opened twitter_archive_enhanced.csv file in excel and did a visual assessment within the document. In Jupyter Notebooks I then used .info, .describe, and .duplicated to find any useful information. Here I found some quality issues that needed to be corrected. I plotted a boxplot of the rating numerator and found there were a few outliers that might cause inconsistencies when analyzing the dataset. I assessed df2 by using .info, .describe, and .duplicated. A visual analysis of the data showed some quality and tidiness issues. I assessed df3 by using .info, .describe, and .duplicated. A visual analysis of the data showed some tidiness issues.

I identified 8 quality issues, and 2 tidiness issues as follows:

- Incorrect score for 960/0 in df1
- Rating for Bella (5/10) incorrect, need to allow for decimals in df1
- (doggo, floofer, pupper, puppo) data types in df1
- Ratings out of higher than 10 in df1
- Outliers in df1
- Non-dogs in df2
- Retweets in df1
- Timestamp datatype in df1
- Dog breeds have mix of capital and lowercase in df2
- Do not need url or retweet status in df3

I made copies of each of the datasets and labeled them with _clean at the end of each name.

Issue 1 was cleaned by using the index number to change the numerator and denominator to the correct rating.

Issue 2 required multiple changes. The rating in excel used a decimal point and from inspecting datatypes in df1 info, rating_numerator only allow integers. The datatype of rating_numerator was changed to float to allow decimal points, then the row of “Bella” was changed to the correct rating.

Issue 3 involved changing the datatype to one that would make readability better and data analysis more streamlined. I used a loop to change each of the “None” values to spaces so when the datatype was changed to bool, the none values would be represented as False. To increase readability, the datatype was then changed to int to show the values as 1s and 0s.

Issue 4 was cleaned by filtering the dataframe by only including those rows that has denominators of 10.

Issue 5 was cleaned by filtering the dataframe by only including those rows that had numerators equal or less than 70. This made the box plot much more readable.

Issue 6 was identified by visually inspecting df2. Using the p1, p2, and p3 values of True the dataframe was filtered to only include images of dogs.

Issue 7 was cleaned by removing any rows that had non-null values in the retweet status id column.

Issue 8 was cleaned by changing the datatype of timestamp from object to datetime.

Issue 9 was cleaned by changing all the dog types to lowercase for easier reference to the type of dogs.

Issue 10 was cleaned by removing unnecessary columns and merging the 3 dataframes into one master dataframe.