In this report we will be discussing the wrangling efforts made in response to the dataset provided by the Twitter user WeRateDogs. This report is divided into the following sections: Gathering Data, Assessing Data, Cleaning Data, and Conclusions; with each section going further into detail on what took place in each section.

Gathering Data:

In order to start the data wrangling process, there were 3 pieces of data that had to be gathered in different methods:

1. The WeRateDogs Twitter archive
   - This file was manually downloaded as a .csv file and read into our Jupyter Notebook using the .read_csv() function.
2. The tweet image predictions
   - This was a .tsv file and it was taken from the internet programmatically using the Requests Library.
3. Additional data from the Twitter API
   - This data was taken using Tweepy and converted into a JSON file that was then imported into the Jupyter Notebook. Due to not having a Twitter account I was unable to do this and had to rely on old code that was given to me by instructors.

Assessing Data:

After gathering the data it was time to assess the data for tidiness and quality of the data. Quality refers to data content and tidiness refers to structure of the data. Assessment was done both visually and programmatically and concerns with cleanliness were documented for the next step. Concerns were sorted by dataset table and were documented below:

Quality:
archive

1. Non null values in name instead of null.
2. Invalid name values in name (a, an, etc.)
3. Non null values in dog types instead of null.
4. Column tweet_id is type int instead of type string.
5. Column timestamp is string instead of type datetime.
6. Contains retweets.

image_predictions

1. Column tweed_id is type int instead of type string.
2. Duplicate strings in jpg_url column.
3. Dog breeds listed in p1, p2, p3 columns are inconsistent with capitalization.

tweet_json

1. Column tweet_id is type int instead of type string.

Tidiness issues

1. 3 tables with only 1 observational unit.
2. Unneccesary columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) due to no retweets being included.
3. Columns doggo, puppo, pupper, and floofer should be merged into one column, dog type.

This was just some of the concerns identified for these datasets. There were others but for brevity these were the ones that were focused on and cleaned.

Cleaning Data:

This is the last step in the data wrangling process where concerns identified in the previous step get fixed or “cleaned.” To complete this step cleaning methods were done programmatically due to the quantity of errors, but manual cleaning could be done in one-off cases. Each issue was listed out and followed the Define, Code, Test framework where we define what needed to be changed, wrote code to complete the change and then tested to see if our code fixed the issue before storing our final product in a new, clean csv file. Due to not identifying all issues in the datasets this would be the first time through assess and cleaning process, and it would need to be continued in an iterative process until all issues are identified and cleaned.

Conclusion:

Data wrangling is an essential skill to have as a data analyst. Thirteen issues with the data were identified and those were just those that were to be fixed the first time through the process. Without data wrangling analyzing and visualizing your data would be virtually impossible because you wouldn’t be able to verify whether the data was clean or not making any conclusions useless.