The data wrangling for this project includes gathering, assessing and cleaning data. Data was gathered from three sources. This includes 1. twitter-archive-enhanced.csv, which is gathered directly from memory, 2. image_predictions.tsv which is stored on a server and pulled via HTTP using a Python request object, and 3. a text file tweet_json.txt generated by accessing the Twitter API and writing each tweet as a JSON object to file. image-archive-enhanced.csv and image_predictions.tsv were able to be read directly into a Pandas DataFrame. tweet_json.txt is a text file, so had to be loaded as a JSON object. Since Python's json library can only read one object at a time, and since each object has to be flattened to make sense in a Pandas DataFrame, the text file consisting of many JSON objects (tweets) must be read in and flattened one line at a time, and then appended onto a DataFrame.

The most notable things during the assessment phase are that the rating_numerator is inconsistent. Through visual inspection in a spreadsheet software, it can be seen that the algorithm that was used to extract ratings from each tweet missed some ratings. These rows were dropped during cleaning.

Similarly, some of the names were missed by the algorithm when extracting names from tweets. These rows were dropped during cleaning.

Many tweets from this archived dataset no longer exist and so reproducing the analysis would be dependent on having the archived data and the 3 files used, and so would not be able to be reproduced directly from Twitter. This does not affect our analysis, but possible reproducibility in the future.

In order to make it easier to merge datasets and perform analysis, it would be helpful to have matching ‘id’ data types and column names. The actual Twitter data in 3 accessed via the API includes an ‘id_str’, and since calculations are not being performed on ‘id’, it makes sense to convert ‘id’ in 1 and 2 to str data type and rename all columns referencing ‘id’ to ‘id’.

There were no particular issues during the cleaning phase as this was all straight-forward.