This project's objective is to collect WeRateDogs Twitter data and use it to provide insightful and reliable analyses and visualizations. Although the Twitter archive is fantastic, it only includes the most fundamental tweet data. To create stunning and deserving analysis and visualizations, more data needs to be gathered, assessed, and cleaned. This project aims to manipulate WeRateDogs Twitter data to produce engaging and reliable analyses and infographics.

I had to collect three different pieces of data and load it to my local working space in three different ways. Firstly, I loaded all necessary libraries needed to wrangle my dataset, downloaded the first data file ‘twitter-archive-enhanced.csv’ manually then loaded the second and third datasets, ‘image-predictions.tsv’ and ‘tweet-json.txt’ respectively, programmatically.

I was able to assess the data visually and programmatically, whereby programmatically being more efficient, I was able to deduce data quality errors which were as follows:

1. All retweets should be removed from the dataset

2. Missing values in columns doggo, floofer, pupper,puppo are represented as None and should be converted to NaN

3. In the image predictions dataset some p1 values start with small letters while others start with a capital letter, so for convenience all will start with capital letters.

4. In the image predictions dataset some p2 values start with small letters while others start with a capital letter, so for convenience all will start with capital letters.

5. Some dog names start with small letters thus will be converted to capital.

6. Some dog names such as this and unacceptable do not make any sense thus will be replaced with No_Name from the twitter archive dataset

7. Predictions P1, P2 and P3 in the image prediction dataset will be formatted such that spacing will be whitespace rather than '_'

8. Timestamp for the tweetarchive will be converted to datetime format for better manipulation.

I was able to spot some data tidiness issues according to me and were:

1. The columns 'doggo', 'floofer', 'pupper' and 'puppo', are to be merged to dog stages.

2. The columns rating_numerator and rating_denominator can be merged to a single column for better manipulation.

I was able to use different pandas, methods and functions to help me clean these issues using the define, code and test framework, which helped me perform this cleaning process easier. In the end I merged all these cleaned data into one csv file.