Reporting: wrangle_report

1. Gathering Data

First and foremost wrangling is divided into 3 steps which are gathering, assessing and cleaning. Before a data analyst can start working on the data, it has to be gathered from many different sources. In this case 3 datasets were gathered for the wrangling exercise. The first dataset was the WeRateDogs Twitter archive which was available to download and load into python. The second dataset was the tweet image prediction which was downloaded using the Requests library, the data was pulled from HTML using the BeautifulSoup library which was then written to a file and loaded into python by setting the separator parameter as tabs since the file which was generated is a tab separated values file. The third dataset was obtained from Twitter using the Tweepy library via the Twitter API. When the Twitter API was queried the data came out in JSON format and it was saved to a text file.

2. Assessing Data

Before the data is cleaned, it has to be assessed. There are two ways of assessing data which are manual or programmatic. Manual assessment is done for small datasets and this can be done by opening csv or excel files and checking them to see if there are any issues or loading the dataset into python and displaying it. It is not sustainable to use manual assessment for large datasets and this is where programmatic assessment comes in since it is fast and can be scaled to be used for larger datasets. Assessing data is just taking note of the issues with the dataset. All the issues discovered are listed below:

2.1. Quality issues

1. There is some missing data in the twitter archive dataset
2. There are retweets included in the twitter archive dataset
3. There are replies included in the twitter archive dataset
4. Some of the tweets are not about dogs
5. Some of the rating denominators are not out of 10
6. Some of the rating numerators are outliers
7. Timestamp is of type object
8. The source column for the twitter archive dataset is in html format
9. Some records in one dataset are not found in other datasets
10. There are original tweets without expanded urls in the twitter archive dataset

2.2. Tidiness issues

1. The twitter archive dataset should have one column for dog stage

2. Text and the url are in the same column and need to be separated, the rating needs to be removed

3. Cleaning Data

After the issues have been noted down then the cleaning beginning begins. The cleaning is just simply fixing the issues that have been stated in the assessment stage. The cleaning process can be divided into 3 steps which are first defining how the cleaning is going to be done, secondly doing the actual cleaning by writing to code to fix the issues and lastly testing to check whether the issue has been resolved. The issues that were stated above were resolved by removing any missing values, removing retweets and favorites, converting the timestamp column to a datetime variable, extracting text from html, removing tweets which are not about dogs, combining the four columns about dog stages into one column and separating text and urls. After all the issues had been resolved the 3 datasets were merged to remove any records about tweets which are not common in all 3 datasets. The final dataframe was then saved into a csv file which was then loaded into python to do the analyzing and visualization.