In this project, the data wrangling includes three major parts: gathering data, assessing data and cleaning data.

To gather the data, firstly the twitter_archive_enhanced.csv is downloaded manually. The image_predictions.tsv is downloaded programmingly through its URL. Then, Tweepy was used to query Twitter's API for additional data which includes retweet count and favorite count.

To assess the data, the following python tools are used firstly to get the basic information of the datasets: .info(), .describe(), .value_counts(), .sample(). Through the information obtained by these tools and also reading the spreadsheets, the following quality issues and tidiness issues are observed.

Quality issues:

tweet_archive:

- In the 'name' column, some names are actually not real names.
- Wrong data type: 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'timestamp' and 'retweeted_status_timestamp'.
- Some expanded_urls are missing.
- There are some 'rating_denominator' which are not 10.
- There are some 'rating_numerator' which are less than 10.
- Exclude any tweet that is a retweet.

tweet_predict:

- Wrong data type: 'tweet_id'.

additional_tweet:

- Wrong data type: 'tweet_id'.

Tidiness issues:

tweet_archive:

- retweet columns not needed
- one variable in four columns (doggo, floofer, pupper, and puppo)
- Then, three data frames tweet_archive, tweet_predict, and additional_tweets should be one (combined table) since all tables' entries are each describing one tweet.

To clean the data, firstly copy the original dataframes and conduct the cleaning on the copied dataframes. For each issue summarized above, three steps are performed: ‘Define’, ‘Code’ and ‘Test’. ‘Define’ refers to define the issue and the proposed method to fix it. ‘Code’ is the specific python code. ‘Test’ refers to testing to see it the issue has been fixed by the code.

In the 'name' column, some names are actually not real names. The wrong names are either lower case or 'None'. Therefore, loop through the names and use the regex expression to further extract the name. The names which are not included in the tweet will be replaced by 'None'.

There are some wrong data type: 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'timestamp' and 'retweeted_status_timestamp'. 'tweet_id' should be string; 'timestamp' should be datatime; The rest should be int64;

Some expanded_urls are missing. Add the expanded_urls according to tweet_id row by row.

There are some 'rating_denominator' which are not 10. We check it row by row and assign 10 to it if it's not 10.

There are some 'rating_numerator' which are less than 10.

Exclude any tweet that is a retweet. Find entries that are retweets by matching text pattern 'RT @' and drop entries from the table.

The tweet_id in tweet_predict has wrong datatype. Convert it to string.

The tweet_id in additional_tweet has wrong datatype. Convert it to string.

One variable in four columns (doggo, floofer, pupper, and puppo). Make one column for dog stage (doggo, floofer, pupper, and puppo) by saving value ('None' if no dog stage given). Also record if there are multiple dog stages, separating by a comma.

Drop retweet columns from each dataframe.

Three data frames tweet_archive, tweet_predict, and additional_tweets should be one (combined table) since all tables' entries are each describing one tweet.