DATA WRANGLING REPORT

I performed wrangling in three steps:

1. GATHERING:

I gathered the required data from three sources.

First, I loaded a twitter_archive_enhanced.csv file from the Udacity website and read it into a data frame called x_archive.

Second, I used the requests library to download programmatically the TSV file provided by Udacity located at a URL. Then, I read this file into a data frame called predictions.

Third, I downloaded and read the JSON file provided by Udacity line by line to create a data frame tweet_data with tweet_id, favorite_count, retweet_count, retweet status and URL.

2. ASSESSMENT:

I performed the visual and programmatic assessment.

In x_archive and predictions, I noticed incorrect values for name, ratings and dog types, there were inconsistencies in capitalisation in prediction columns. I also noticed that there are records where no prediction is ‘dog’ and columns with missing values and duplicated values. Several datatypes needed to be corrected. The main issue was that the data frames contained multiple columns for several similar variables. Tweet_data looked clean.

3. CLEANING

First, I created copies of the data frames. In the cleaning process, for each task I followed the steps: define, code, and test the result.

I addressed and solved tidiness and quality issues from the assessment section and then prepared the final master data frame, and stored it in a CSV file. The most difficult and time-consuming part for me in the process was merging the columns in predictions.

As a final step of the wrangling process, I saved the clean data frames in CSV files.