Background:
For the purpose of this data wrangling, there are three datasets provided. The aim is to bring in the data into a jupyter notebook, assess the data and define,code and test for the cleaning processes.

Data Wrangling

Dataset 1:
The twitter_archives was provided as a CSV file and was brought in using the pandas read_csv() function.
To assess the data, I used the shape(), sample() and info() functions to understand how the data look, get the number of rows and columns, identify columns not important to the analysis goal and also know the important information like columns that have missing values and columns with inappropriate data type.
For the cleaning, I made a copy of the data using the copy() function and then firstly investigated the columns with missing values by using isna() and value_counts() functions. I deleted rows that contain retweets. I proceeded to drop columns who have the majority of their values empty and also columns I identified as not necessary for analysis. I proceeded to extract the data from the timestamp using the split() function and changed the datatype using to_datetime(). To combine the doggo,floofer,pupper and puppo columns into one, I used the melt() function and then proceeded to remove duplicates using drop_duplicate(). After all the coding, I used info() to check if the final state of the data fits the purpose.

Dataset 2:
Due to the technical constraints with using the Twitter API, I worked with the tweet_json text file provided. This text file was imported into Jupyter Notebook using the read_json() function and setting the parameter lines=True.
To assess the data, I used the shape(), sample() and info() functions to understand how the data look, get the number of rows and columns, identify columns not important to the analysis goal and also know the important information like columns that have missing values and columns with inappropriate data type.
For cleaning, I made a copy of the data using the copy() function and then checked for the number of missing values in each column using isna() and sum() functions and proceeded to drop selected columns using drop(). I proceeded to select columns useful for my analysis and passed it into a new dataframe tj_final.

Dataset 3:
To bring in the image prediction data, I used the requests.get(), open(), write() function to get the data downloaded from the provided url and then use read_csv() to read in the data to the Jupyter notebook.
To assess the data, I used the shape(), head() and info() functions to understand how the data look, get the number of rows and columns and check for missing values and inappropriate data types.

For cleaning, I used the astype() to change the datatype of tweet_id to str and img_num to category data type. I also renamed the column headers using rename().

After individual cleaning of the datasets, I used the merge() function to combine the three datasets into one final dataframe and write it to my PC using to_csv() function.