Wrangle Report

This report is a summary of the steps I performed to complete this project.

1) Gathering Data:

- Because I had trouble gaining access to a Twitter Developer account, I used the steps provided to access the Twitter data without the account
- I uploaded the 'twitter-archive-enhanced.csv', 'tweet-json.txt' and 'image-predictions.tsv' files into 3 DataFrames: 'twitter_df', 'json_df' and 'images_df'

2) Assessing Data:

a) Visual Assessment:

This assessment was done by using Excel and scrolling through the DataFrames in Jupyter Notebook.

Quality Issues:

In 'twitter_df':

- Some dogs' names are "a", "actually", "all", "an", "by", "his", "life", "space", "such", "the", "this", "unacceptable" or "very". They should be converted to no name (I can check for any name that is lowercase and see if it is a valid dog name)
- There are denominators that do not equal to 10 and some of them are wrong ratings
- Some numerators have exaggerated numbers
- Retweets and replies should be removed from the dataset
- Change "None" to null for columns doggo, floofer, pupper and puppo

In 'images_df':

- Assign null values to p1, p1_conf, p2, p2_conf, p3 and p3_conf with values of p1_dog, p2_dog, p3_dog = False
- Remove p1_dog, p2_dog and p3_dog columns
- The dog type columns should be consistent in the format (all lowercase)
- Change column names to be more representative

Tidiness Issues:

In 'twitter_df':

- Doggo, floofer, pupper and puppo columns could be merged into one column

In 'json_df':

- The columns can be merged with twitter-archived-enhanced.csv via the tweet_id

In 'images_df':

- The columns can be merged with twitter-archived-enhanced.csv via the tweet_id

b) Programmatic Assessment:

Quality Issues:

In 'twitter_df':

- tweet_id should be converted to string
- timestamp and retweeted_status_timestamp should be converted to datetime
- There are two extra rows that are not in the other two files. Check if any ID does not have a photo, likes or retweets

In 'json_df':

- tweet_id should be converted to string
- In 'images_df':
- tweet_id should be converted to string

3) Cleaning Data:

- Copies of the 3 DataFrames were created to clean the data on
- The first step done was to convert the incorrect datatypes to more suitable datatypes
- After that, the retweets and replies were removed from the DataFrame and their corresponding columns were dropped
- The following step was to combine doggo, floofer, pupper and puppo columns into one column called classification, then drop the previous columns
- The next step was to fix the incorrect dog names (which were in lowercase letters or were named None). Regex was used to check if the name was in the text, otherwise the name was set to null
- Null values were set to p1, p1_conf, p2, p2_conf, p3 and p3_conf for p1_dog, p2_dog and p3_dog with values = False. p1_dog, p2_dog and p3_dog columns were then dropped
- Columns p1, p2 and p3 were converted to lowercase letters
- Column names were changed from p1, p2 and p3, p1_conf, p2_conf and p3_conf to prediction1, prediction2, prediction3, prediction1_conf, prediction2_conf and prediction3_conf
- Numerators and denominators with exaggerated and incorrect values were adjusted, and numerators with decimal values were rounded
- The three datasets were combined into one dataset called 'twitter_archive_master'
- Rows without images were removed

4) Storing Data:

- Data was successfully stored in a csv file called 'twitter_archive_master.csv'