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’