WRANGLE_REPORT

In this wrangling process, I started with gathering all the needed datasets. The first dataset (twitter_archived_enhanced.csv) was downloaded traditionally by clicking on the provided link and saving it to a local device, after which I loaded it into the Jupyter notebook I used for the wrangling process. The second dataset (image_predictions.tsv) was downloaded programmatically from Udacity 's servers using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and loaded it into the Jupyter notebook. The third dataset (tweet_json.txt) was queried from the Twitter API using the API query process and I needed to apply for permission to access their API due to privacy reasons. After I got permission, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library, I stored each tweet's entire set of JSON data in a 'tweet_json.txt' file then I read the .txt file line by line into a pandas DataFrame with the **tweet ID**, **retweet count,** and **favorite count** columns

I proceeded to the next wrangling process, 'assessing' all the datasets. I used two methods to perform this part which were visually and programmatically. I opened the twitter_archive_enhanced.csv file using MS Excel to visually assess it and opened the other two datasets using a Jupyter notebook. Next, I used pandas functions like '.info()', '.describe()', '.head()' etc. to detect and documented **eight (8) quality issues** and **two (2) tidiness issues**.

Then I moved to the last wrangling process, 'cleaning'. The first thing I did in this part was to make copies, which is a good cleaning practice so I still have the original data as a backup copy in case I need to reverse any wrong action or changes. I implemented the cleaning process by converting the documented quality and tidiness issues that needed to be addressed to code, I used pandas functions like '.drop()', to trim and arrange the dataset columns and rows to remove what was not needed and correct some errors.

Lastly, I merged the three cleaned datasets into one master dataset and stored it in a CSV file called 'twitter_archive_master.csv', using the '.to_csv()' pandas function.