Introduction

In this report, I will briefly go through the steps that I adapted through the data wrangling processes: Gathering, Assessing and Cleaning data.

Gathering

- I downloaded the **twitter_archive_enhaced.csv** file from the classroom. Then, I used pandas method **pd.read_csv** to read the file in a dataframe

- I used the **requests** library to read **images_predictions.tsv** programmatically from given url in the classroom and read it into a second dataframe.

- Intialized the twitter api keys. Then, used this access to get the information about favorite_count and retweet_count of each tweet in the dataset and stored each tweet information in a **tweet_json.txt** file.

- Read the information needed for analysis from the tweets_json.txt file into a third dataframe separated from previous two.

Assessing

- Opened the first and second files in excel and assessed them visually and checked if there are any tidiness or single occurrence quality issues.

- Used **head()** to assess the dataframes visually.

- Used **info()** to assess the columns and datatypes in the dataframes.

- Used **duplicated()** to check for any duplicates.

- Used **shape** to check how the dimensions of the dataframes differ from each other

- Used **isna()** to check for NaN values.

- Used **Boolean indexing and Slicing** to check for certain rows in the dataset and check certain tweets photos and text that can help in the process of cleaning the dataset.

Cleaning

- First of all, I defined the issue and what I want to clean from the dataset by referring to my assessing results.

- Checked tweets photos, text and visited the tweets url to help me understand more about the issue.

- The previous step help me greatly in cleaning the dogs **names, ratings, classes and prediction results.**

- I used functions like **drop(), dropna(), replace(), pd.merge(), pd.wide_to_long()** in the cleaning process.

- Merged the three dataframes to have at last a complete dataframe with information about tweets ready for analysis and visualization.

- Cleaned all the issues programmatically and tested the results of the cleaning process by reassessing the issues with the same methods of assessing process.

- After cleaning, I stored the newly created dataframes in a new file formats as described in the project details.