DATA WRANGLING PROJECT

Data Gathering:

The project involved gathering three datasets: a csv file, a tsv file and a JSON file.

1st Dataset
The first csv file, "twitter-archive-enhanced.csv" was provided on Udacity classroom. The csv file was downloaded manually and read into python using pandas. The dataset had 2356 rows and 17 columns.

2nd Dataset
The second file, "image-predictions.tsv", was from a url. It was read into python using the request library. I had to create a new file and write into the file and saved as tsv file. The dataset has 2075 rows and 7 columns

3rd Dataset
The third data was supposed to be accessed and downloaded using twitter's API. It required access from twitter developmental team. I requested access from the team, I had to fill several forms on the purpose and nature of the analysis.
Unfortunately, I didn't get the access key or token, so I defaulted to manually downloading the JSON file(txt) from Udacity and extracting columns of interest from the dictionary. Data was extracted using a for loop, where I iterated over every line in the txt, collecting tweet id, retweet count, favorite count and followers count. All four columns were extracted into a dictionary called df_list. I subsequently converted df_list to a dataframe for exploration and data analysis.

Data Cleaning:

I identified some quality and tidiness issues with the dataset.

Quality Issues

Twitter_archive

Issues	Solutions
Wrong Datatype for tweet_id column (int)	Converted the datatype from int to str.
Wrong Datatype for timestamp column (int)	Converted the datatype from int to datetime.
Wrongly classified names for dogs like "a", "an"	Used a string match to convert strange name pattern to None.
Remove hyperlink from text	Removed http and text after http
Needed only original tweets	Deleted retweets by filtering the NaN of the retweeted_status_user_id
Columns with only missing values and little aid in analysis	Dropped the columns
Extracted the main source of tweet	Extracted only the last text before the closing tag to show main source
Correct numerators with decimals	Converted columns to float and extracted full numerators

Image_Predictions

Issues	Solutions
Wrong data type for tweet_id (int)	Converted the datatype from int to str.
No uniformed letter case for columns p1, p2, p3	Performed a string manipulation on columns

tweet_json

Issues	Solutions
Wrong data type for tweet_id (int)	Converted the datatype from int to str.

Tidiness

Issues	Solutions
Four different columns representing the stage of the dog	Concatenated the different columns into one column named stage
Data in 3 different tables	Merged all 3 tables using the tweet_id to have a master dataset
Rows without images	Dropping rows without images

Data Storing:

After all necessary cleaning and merging, dataset was saved into a csv file, called "twitter_archive_master.csv".