WRANGLING REPORT

The wrangling and analyze data project were focused on wrangling data from @WeRateDogs Twitter Account. The dataset was acquired from three sources and in different formats; I downloaded one file manually by clicking on a link, I downloaded another file programmatically using the Requests library and lastly, I queried Twitter's API for @WeRateDogs twitter data such as retweet count and favorite count. Then combined these pieces of data into one pandas DataFrame.

GATHERING DATA

I started by importing packages that were essential in this project such as numpy, pandas, matplotlib, tweepy, requests, os and json.

The first piece of data is the Twitter Archive data which was downloaded programmatically by Udacity and was made available via a link to be downloaded manually in a csv format. The data was read into a pandas data frame using read_csv method.

The second piece of data is the Image Predictions data provided in a URL which I programmatically downloaded using the Python Requests library. I started by creating a folder for the file then used the requests library to request for the file from a URL. Then I read the file into a pandas DataFrame.

The third piece of data was required because of missing key variables in the Twitter Archive data. I querried Twitter account @WeRateDogs for this additional information using the tweepy API. First and foremost, I applied for a Twitter developer account and requested for elevated access, I got the required consumer and access keys needed to query the API. I only needed tweet ids which I already had in Twitter Archive DataFrame, this was useful in pulling the data for the required tweets. I used a for loop for each tweet id in twitter archive to get the status, I also used a timer to check how long it took my code to run which was approximately 30 minutes. I wrote a code to open the Json file in read format and had to write the file line by line into a DataFrame.

ASSESSING DATA

I assessed the data visually and programmatically to scout for data quality tidiness issues. For Visual Assessment; I used Excel spreadsheet, also I printed out the pandas Dataframe checking visually for any issues which I documented.
For Programmatic Assessment; I used functions such as head, sample, info, describe to assess each dataframe individually and check for duplicates, check for nulls and incorrect datatypes. I found issues like; invalid data such as the name of a dog. I noticed a tidiness issue in the Twitter archive data, where dog stages (one variable) was represented in 4 columns violating the rule 'each variable is a column' for tidy data. Also, the retweets count, and favourite were on a different table other than the Twitter Archive which indicated a tidiness issue. I documented all data quality and tidiness issues I found in the three dataframes started my Data Cleaning efforts.

CLEAN DATA.

This stage had me getting my data from an Unstructured, untidy, and dirty format to structured and clean data. I dropped rows of data where Tweets were Retweets. I removed, invalid data, joined dog stages column into one and merged the three data frames as one. I also dropped columns which were not useful for my analysis and winded up by converting datatypes to the right format.

SAVE

After Cleaning, I saved the combined DataFrame into a CSV file for use for further analysis.