WRANGEL REPORT

GATHERING DATA

I imported nine python librae's which were

- Numpy for indexing through the rows and columns in the data frame
- Pandas for creating and storing the data in to a data frame of csv format
- Matplotlib for creating a simple visualization of variables in the data frame
- Seaborn for styling the matplotlib graph and setting its style to dark grid
- Request for making a get request from Udacity link to wrangle the data from the webpage
- Os for creating path to store saved data frame in my current directory
- Json for converting text format from string to json
- Sqlalchemy for storing data into database after wrangling
- Re for extracting data when cleaning

How I wrangled the tweeter API

Created a get request statement from the URL like of {'https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt' } and stored the content of the data into a text file called tweet_json.txt file in which a for loop was used to target each line of the text file which was then converted into json format thought the use of imported json library

I then created two list one for storing data of the targeted key and its values if found in the text file which was converted to json which I kept in a try block and for the other list for storing data which could not be found but placed in the except error block.

Which then I created a data frame with columns of names 'timestampe', 'tweet_id', 'tweet_text', 'tweet_image_url', 'source', 'retweet_count', 'favorite_count', 'follwers_count' then stored the newly created data frame in a CSV file called WeRateDog.csv

How I accessed the newly sored data frame

Accessing WeRateDog.csv without pandas doing a visual assessment on the data frame first viewing the data frame plainly in a spread sheet application which was excel with out pandas and with just few glimpse of the data frame could observer some visual assessment issues with the source column with an invalid format for source column and of which most the issues were from the **tweet_text** column of it containing multiples tables in just one single statements such as

- ratting of dogs in tweet text
- extra url links for images in text
- dog stage in tweet text
- inconsistent tweet text containing retweet instead of actual tweet

As Also for timestamped column visual assessment issues

- invalid time format for timestamp column
- time tweeted in time stamp
- day of the week in timestampe

Accessing WeRateDog.csv with pandas doing a programmatic assessment using pandas .info() method and found two programmatic issues which were

- inaccuare timestampe colunms data type
- inaccuare dog rating data type

Has for all assessment issues that were observed though visual assessment and programmatic assessment the were all programmatically cleaned and stored into a CSV file

called twitter_archive_master.csv and a database called twitter_archive_master.db with the use of Sqlalchemy.