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 one list one for storing data of the targeted key and its values if found in the text file which was converted to json which was placed in a try and 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 sorted data frames

Accessing WeRateDog.csv, image_preidiction.tsv , and twitter-archive-enhanced.csv with 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 Quality issues and then Tidiness Issues

Quality Issues

weRateDog_df Data Frame

- invalid time format for timestamp column
- time tweeted in time stamp
- day of the week in timestampe
- ratting of dogs in tweet text
- extra URL links for images in text
- dog stage in tweet text
- invalid format for source column
- inaccurate timestampe columns data type
- inaccurate dog rating data type

twitter_archive_enhanced Data Frame

- invalid format for source column
- invalid time format for timestamp column

Tidiness Issues

weRateDog_df Data Frame

- separate columns for tweet and retweet in weRateDog_df Data Frame
- names of dog in tweet text

twitter_archive_enhanced Data Frame

- dog stage in separate columns in twitter_archive_enhanced Data Frame

image_preidiction Data Frame

- correct prediction of dog type in multiple tables in image_preidiction Data Frame

Has for all Quality and Tidiness 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.