DECI L3 Data And AI Project

First:

I imported many libraries that is required to my analysis like pandas , numpy , malpotlib , requests , os

Second:

I created folder to take all files in it

Downloaded Files Using Requests and save them using os

Note:

Tweepy API(Twitter(X) API) Has Some Problems So I Downloaded The File From Udacity Classroom Using Requests , OS

Third:

Opening the first dataset its name is "twitter-archive-enhanced.csv" that has coulmns of 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'timestamp', 'source', 'text', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'doggo', 'floofer', 'pupper', 'puppo'

I found an unlogical problem in denominator column that some of rows doesnt have the same values of the rest so I assigned all of them to 10 Also Rating numerator column has sometimes vales that is greater than the denominator so I assigned the unlogical ones to the value of rating_denominator column

I found many null values in retweeted_status_id retweeted_status_user_id retweeted_status_timestamp doggo floofer pupper puppo

I dropped the expanded urls that has null values because we cant repair it i dropped doggo , floofer ,pupper ,puppo as most of it's rows is null and we don't have any refrence in the dataset abou it to repair it

I found many null values in these columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id. Since they indicate missing retweets or replies, I replaced nulls with 0 to represent "not a retweet/reply".

I replaced null values in the name column with "Unknown" because I can't know the correct name.

I converted the timestamp column to datetime format

Fourth: The Second Dataset that its name is "image-predictions.tsv" which columns names are 'created_at', 'id', 'id_str', 'full_text', 'truncated', 'display_text_range', 'entities', 'extended_entities', 'source', 'in_reply_to_status_id', 'in_reply_to_status_id_str', 'in_reply_to_user_id', 'in_reply_to_user_id_str', 'in_reply_to_screen_name', 'user', 'geo', 'coordinates', 'place', 'contributors', 'is_quote_status', 'retweet_count', 'favorite_count', 'favorited', 'retweeted', 'possibly_sensitive', 'possibly_sensitive_appealable', 'lang', 'retweeted_status', 'quoted_status_id', 'quoted_status_id_str', 'quoted_status'

I made a copy of it to clean it

I filled the null values in the place column with "Unknown" because location data was missing in many tweets and can't be detected.

I filled missing values in the retweet_count column with 0 because a missing count is likely equivalent to having no retweets.

I created a new column is_retweet by checking if retweeted_status is not null. This helps in quickly identifying whether a tweet is a retweet.

I created another column is_quote using quoted_status.notnull() to easily identify quote tweets.

I filled missing values in the retweeted_status column with False, indicating that these tweets are not retweets.

Also, I filled missing values in the quoted_status column with False to represent that these tweets are not quotes.

The quoted_status_id column had missing values, so I replaced them with 0 to indicate absence of a quote tweet reference.

I replaced null values in the quoted_status_id_str column with "None" because they refer to string IDs of quotes that don't exist.

For the geographical columns geo, coordinates, and place, I filled all missing values with "Unknown" because location data cannot be restored.

I replaced null values in the contributors column with "None" since contributor information is not present in most tweets.

Finally, I dropped these reply-related columns:

in_reply_to_status_id

in_reply_to_status_id_str

in_reply_to_user_id

in_reply_to_user_id_str

in_reply_to_screen_name

They were mostly null and not useful for this analysis.

Fifth: The third Dataset whose name is tweet-json.txt and columns names is 'tweet_id', 'jpg_url', 'img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'

but it was mostly cleaned

Sixth and last:

I renamed the column of id in each dataset to "id" to be able to merge the 3 datasets in one master using the id

I merged them into the master dataframe to be able to start my analysis

The Analysis Process is in the next pdf