WRANGLE REPORT

Data was collected from 'The WeRateDogs Twitter archive, Image Prediction File, Twitter API using the request library, pandas, tweepy accordingly. The dataset was programmatically and visually accessed using .info(), .head(), .tail(), .sample(), .describe(), .duplicated() methods and 14 quality , 2 tidiness issues were identified.

Proceeded to clean the dataset by copying the data using .copy() into a new variable, then I defined, code and test each process which is explained below;

- I filtered the table to only rows needed which excluded the retweet rows that had images because this were replies to original rows using np.isnan() method.
- I extracted the names with single letters in the name column of the clean twitter archive table using Reg Expression. I replaced it with None since some names had none already.
- I changed the rating columns (rating_numerator, rating_denominator) from int to float using .astype () method.
- When I visually assessed the text column for rating numerator column, there were ratings in decimal number that were not included which I extracted again using a for loop to the rows affected to extract the numbers which I also updated.
- I handled outlier by dropping after investigating the tweet using the tweet_id.
- Changed all rating denominator to 10 using '=' .
- Dropped all columns not needed for analysis.
- Renamed id in new_tweet table to tweet_id using .rename() method, to enable join the 3 table.
- Converted the tweet_id column in the 3 tables to string using .str()
- Converted the Date data type using pd.to_datetime()
- Handled missing images by dropping the rows affected.
- Removed links from the source column using .contain () and RegExp to get only the source without the html link.

the process included handling missing data in all table, changing datatypes of the tweet_id columns in all table, dropping columns not needed for analyses, filtering the data row to only rows needed