WERATEDOGS PROJECT WRANGLE_REPORT

The dataset used in this project is the tweet archive of Twitter user @dog_rates. This twitter account is also known as WeRateDogs and it rates people's dogs with a humorous comment about the dog. For this project, three datasets were gathered. The first dataset is the downloaded WeRateDog Twitter archive dataset, the second dataset is the Image Prediction dataset and the third dataset is an additional data via the Twitter API.

The major task of this project was data wrangling (and analyzing and visualizing) of the twitter archive dataset. In carrying out this task, the following steps were followed:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

The Twitter API data was gathered using the tweepy library while the Image Prediction TSV file dataset was gathered using the request library.

After the 3 datasets have been gathered, the data were thoroughly assessed visually using the .head() and through other external medium via Excel sheet. Programmatic assessment was also done using .info(), .describe(), .value_counts(). A number of observations were identified while assessing the data and these made the cleaning data stage a necessity.

During assessment, a number of quality and tidiness issues were identified as listed below:

Quality Issues

1. Change datatype for tweet_id to string in all the 3 tables

2. The ratings_denominator column value is 10 for most of the rows, we can drop the column to have just one denominator.

3. There are extraneous figures in ratings_numerator column in the archives table

4. Errors in dog names e.g a, an, actually etc.

5. There are hybrid dog stages as seen in some tweets.

6. Some columns are not needed and should be dropped.

7. The source column in the archives table has too long and not relevant details. This can be shortened.

8. The url link is also present in the image prediction table, so we can drop the url expanded column in the archive table.

9. The last five digit (+0000) in the Timestamp column is not needed. Change datatype to date-time.

Tidiness Issues

1. The following columns 'doggo’, ‘puppo’, ‘pupper’ and ‘floofer’ should be in a single column called dog_stage.

2. Twitter API and Image Prediction table should be merged with the Archive table.

The issues identified above were all addressed programmatically during the cleaning process.

To begin the cleaning process, I made a copy of the 3 datasets using .copy().

The tweet_id columns in the 3 tables were found to be of integer data type and this was corrected to string. Some extraneous figures were found in the rating numerators column which could be as a result of errors during extraction of the ratings values from the text column. This was corrected during the cleaning process.

Assessment shows errors in dog name column with most of them starting with small letters. All the affected rows were identified and later dropped. Issues of tidiness were also identified, especially were we had 4 different columns for dog stages. The 4 columns were merged into a single column titled dog_stage and the other 4 columns were deleted.

To tidy up our dataset, the 3 cleaned datasets were merged into a single dataset which was named **twitter_archive_master.csv** and was stored before analysis and visualization.