wrangle_report

October 29, 2022

0.1 Reporting: wrangle_report

0.1.1 Data Gathering

Three datasets was gathered for this project. I downloaded the **'twitter-archived-enhanced.csv'** dataset manually from the Udacity classroom, downloaded the **'image_prediction.tsv'** data programatically using the request library also from the link given in the Udacity classroom. The third **'tweet-json.txt'** dataset was to be downloaded using tweep query but I had issues getting my twitter verification request approved. So I resorted to used Json, operator and pandas libraries to download the data.

0.1.2 Assessing

I assessed the three datasets visually and progammaticall using pandas methods. Observation was made on quality and tidiness issues

0.1.3 Tidiness issues

- I observed that the html tag had tweet source.
- Timestamp column had both date and time in the same column and saved as object datatype.
- Tweet column had rating and links in the same column.
- All three datasets should be merged into one dataframe.

0.1.4 Quality Issues

- There were about 181 retweets and 78 replies in the twitter_archived_enhanced datasets, which were not part of the original tweet, so I removed them using the drop method.
- Rating should be float.
- Timestamp column with object datatype.
- Inconsistency in names across the three dataset.
- Irrelevant columnns.
- Inappropriate dog names like 'a', 'the'.

0.1.5 Cleaning

- To allow for changes in the datasets without affecting the original datasets, I copy the original datasets and added a suffic of _clean to the new dataframe created.

Tidiness

- Used Regular expression to extract the tweet sources from the html tag, as the html tag had links not needed.
- Timestamp column had both date and time, which I splitted into separate columns to enable me get more insights while analysing the data.
- I notice from programmatical assessment that the tweet column had rating and links which was not quite obvious, so I extracted the links and rating from the column.
- Merged all three datasets as it seems structurally okay for them to be in the same dataframe.

Quality

- There were about 181 retweets and 78 replies in the twitter_archived_enhanced datasets, which were not part of the original tweet, so I removed them using the drop method.
- Change rating datatype to float to allow for exploratory analysis.
- Timestamp column datatype was object so I change it to datetime to allow for futher analysis.
- Make all the names columns across the three datasets to be lowercase consistency.
- Drop columns that column not be needed for the analysis.
- Change inappropriate dog names to No_name for better recognition.

Storing

- Save the merged dataset into a single df_master dataframe as a CSV file named twitter_archive_master.csv