wrangle_report

November 15, 2022

1 Wrangle Report

1.0.1 By Arikanawasi Udoka

There are three different datasets. twitter-archive-enhanced.csv image-predictions.tsv tweet_json.txt

Pandas was used to convert the files into dataframes, with the `image-predictions.tsv` being downloaded programmatically first using the requests library before being read into a dataframe. The `tweet_json.txt` file was used from the provided resources due to the inability to open a developer account with Twitter.

Columns were dropped from all 3 dataframes, these dataframes were deemed unnecessary and cumbersome to the final result.

Assessing Data

Each data frame was programmatically checked, and no null rows were found. `tweet_json.txt` and `twitter-archive-enhance.csv` files had date columns, which were checked to confirm that no tweets were after August 1, 2017.

Some columns were checked using the unique() method to view what they contain.

The head() method was used to visually assess each data frame to further identity quality and/or tidiness issues

The following Issues were noticed and documented; Quality Issues `new_df_twitter` - `created_at` column conversion from datetime to date - `tweet_id` column conversion from int to object - `favorite_count` to int - `retweet_count` to int

`new_df_archive` - `timestamp` column conversion from datetime to date - fix `rating_numerator` outliers (outside the rating system) - fix `rating_denominator` outliers (outside the rating system) - change `rating_numerator` column conversion from float to int (this column had no floats prior to cleaning, after cleaning, these number turned to floats with no 0 behind the decimal point hence the conversion to int for stability.) - `tweet_id` to object

`new_df_image` - `tweet_id` to object - replace underscores in `p1` column with whitespace.

Next we found 2 tidiness issues they were; `new_df_archive` - melt doggo, floofer, pupper, puppo to change dataframe from wide to long

- `new_df_twitter`, `new_df_archive` and `new_df_image` tables to master table

After documenting, copies were created of each dataframe before commencing with cleaning and fixing. Finally, the master dataset was stored in the same directory with the file name `twitter_archive_master.csv`.

Analysis and Visualization

Here further cleaning was done to aid analysis, multiple rows in multiple columns were dropped, and the data set went from (8292, 13) to (203, 12). This new dataframe was then stored in the same directory with the file name `twitter_archive_master_cleaned.csv` which is what is used in the `act_report.pdf` file.

In conclusion, The dataset was cleaned extensively and insights were got from it.

In [ ]: