wrangle_report

September 11, 2022

0.1 Project Overview: Analysis of @dog_rates tweets (wrangle_report)

Tasked with analysing tweets about the ratings of dogs on the we_rate_dogs archive and get insights like; the most liked type of dogs, average ratings based on the dog types, or the type of dogs that's most tweeted about.

The dataset will be wrangled (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

The tweet archive of WeRateDogs contains basic tweet data for all 5000+ as it stood on the 1st of August, 2017.

0.1.1 Data Gathering

The datasets required for the analysis of the are located in 3 different modes; 1. csv dataset already downloaded 2. image prediction dataset stored in a url as a .tsv file 3. A json.txt dataset holding additional data pulled from tweeter API

- The csv file (tweeter-archive-enhanced.csv) was read direcltly into a dataframe using `pd.read_csv('tweeter-archive-enhanced.csv')`
- The image_prediction was requested and pulled into dataframe using and
- The tweet_json.txt was requested, read, and stored in a separate dataframe.

0.1.2 Assessing

Visual Assessment The data was visualized manually using microsoft excel to understand the data in a better way. Tidiness issues including poorly structured columns were caught here.

Programmatic Assessment .info(), .describe(), .sample(), etc. were used to assess the dataset and 8 quality issues were noted . - **Quality Issues**

1. Retweeted ratings in tweet archive
2. Some names are inappropriate in tweet archive. There are no such names like 'a', 'an', 'o', etc.
3. Some Numerators are suspicious and some Denominators are not equal to 10
4. Some of the columns in tweet_archive are not essential to analysis

5. 66 duplicated images in image_predictions
6. Split timestamp in twitter archive into three different columns (day, month, and year)
7. tweet_id should be a string rather than an int data type
8. retweeted_status_timestamp not needed for analysis

0.1.3 Data Cleaning

To clean the datasets, a copy of the datasets were made before attempting to correct the highlighted issues.

```
   tweet_archive_clean = we_rate_dogs.copy() image_pred_clean = img_pred.copy()
tweet_json_clean = tweets.copy()
```

Cleaning Quality Issues

tweet_archive

- The retweets and replies in the archived dataset were removed, inappropriate names were replaced with 'None'.
- Suspicious numerators were checked and wrong ones were duly corrected while denominators not equal to 10 were corrected.
- Columns not essential to the analysis were dropped
- The timestamp column was collapsed into the 'day', 'month', and 'year' columns
- The retweeted_status_timestamp was also deleted from the dataset

image_predictions

- Duplicated image urls were removed
- tweet_id data type was changed from integer to the more appropraite string type

tweet_json

- tweet_id data type was changed from integer to string type

Cleaning Tidiness Issues

- The doggo, floofer, pupper, and puppo columns were pivoted into a single column called 'dog_stage' because it is a structural issue to the analysis. Four columns are holding information that could be held in one
- **Finally**, the three datasets were merged into a single master dataset