Wragle_report

Introduction

This project sought to wrangle data from twitter user **@dog_rates** (WeRateDogs). It required that we gather datasets from three sources, assess and clean them, then produce insights and visualizations.

Data Gathering

The datasets gathered are:

1. **twitter_archive-enhanced.csv**: This was directly downloaded from the WeRateDogs twitter archive data.
2. **image-predictions.tsv**: The requests library was used to download this file.
3. **tweet_json.txt** : Tweepy library was used to query data from the twitter API, save it into this JSON file then read it into a pandas dataframe.

Assessing Data

The project required documenting at least **8 quality and 2 tidiness issues** from the datasets and analyzing a dataset with original ratings only.

The findings for assessing each dataset were:

Quality issues

1. Some of the tweets are retweets and replies.
2. Missing values (NaN) in the in_reply_to_(id, status) and retweet_status_(id, user_id, timestamp) and expanded_urls columns
3. Invalid dog names in the name columns e.g a, the, and some names are missing(none)
4. Misrepresentation of missing names in **name** column. (None instead of empty strings)
5. Misrepresentation of null values in floofer, doggo, pupper and puppo columns. (None instead of empty strings)
6. Incorrect datatype for timestamp and rating_numerator columns
7. Some rating values are not properly extracted from the tweet's text e.g 13.5 extracted as 5
8. **image-predictions table** :tweet_id is in ascending order & isn't consistent with other dataframes

Tidiness issues

1. archives_df: The doggo, floofer, pupper and puppo columns should form one column
2. All three dataset should be merged to one.

Cleaning Data

A copy of the original datasets was made from which the cleaning would be done.

The solutions to each issues were:

Quality Solutions

#1: All rows with retweets or replies were dropped.

#2: The columns with missing values were dropped.

#3: A new **names-list** (made of all valid names starting with capital letter) was created, replacing the original name column. Pandas' NaN assigned values replacing all invalid names.

#4: The misrepresentations in the name column were replaced with pandas' NaN.

#5: All values represented as **None** were replaced with an empty string..

#6: Converting the **timestamp** from object to datetime64[ns] and **rating_numerator** from int to float.

#7: Extracting rows with decimal tweet ratings and using row indices to assign the correct decimal rating to each row.

#8: Sorting the *tweet_id* column in descending order then dropping the original dataframe's index and resetting it to ascend from zero.

Tidiness Solutions

#1: Creating a new column (**dog_stage**) and filling it with corresponding values from the doggo, floofer, pupper and puppo columns.

Note: The above action created the following issues that required solving.

****** Multiple invalid dog stages e.g doggopupper.

****** Empty strings that appear as dog stages.

Solutions

#1.1: Replacing invald dog stages with the stage's suffix e.g replacing doggofloofer with floofer.

#1.2: Replacing empty strings with pandas' NaN.

#1.3: Dropping doggo, floofer and puppo columns.

#2: Merging the three dataframes into one (***clean_merge1***).

The merged dataset was saved to a CSV file named **"twitter_archive_master.csv".**

Conclusion

At Least eight quality and two tidiness issues were detected and documented as per the project requirements. However, there may have been more issues with the datasets needing to be addressed.

Some limitations encountered during cleaning are:

1: There is a lot of missing data (vital in our analysis) that cannot be acquired.

#2: During merging the dog stages, we are left with fewer dog_stage values than the average. This may have affected our analysis.

3: Assigning dog stages to some rows that weren't well represented (e.g "doggofloofer" instead of either "doggo" or "floofer") may have affected the analysis.