wrangle_report

July 12, 2022

0.1 Wrangle Report for WeRateDogs Project

0.1.1 The Project was in three different stages:

1. The Data Gathering Stage
2. The Assessing Stage
3. The Cleaning Stage.

Data Gathering Stage
In this data gathering stage, data was gathered from three different sources:

- The WeRateDogs Twitter archive, which was provided by our Udacity instructor. This archive dataset contains basic tweet data such as tweet ID, timestamp, text, ratings, etc, for all 5000+ of tweets.

- The image_predictions dataset. This file can be used to run every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs.

- Twitter API, which required using Python's Tweepy library to gather each tweet's retweet_count and favorite_count, using the tweet IDs in the WeRateDogs Twitter archive.

Assessing Stage
The Assessing stage was done **Visually and Programmatically**,
After assessing the datasets, both **Quality** and **Tidiness** issues were discovered.

For the **Quality issues**, the image_predictions data had about 2075 records, instead of 2356 records as found in the twitter_archive dataset, which could have resulted from retweets, replies and missing images. The image_predictions data also had 181 retweets, and we only needed original ratings. The dog names column also had incorrect names, and missing values which appeared as 'None'. The rating_denominator also had wrong values, and we just needed rating_denominator of 10. The rating_numerator also had invalid values. Timestamp and tweet_ids were also in the wrong format. The expanded_urls, reply and retweet columns in the twitter_archive also had many NaN values. The source column in the twitter_archive also had HTML codes attached to its values.

For the **Tidiness issues**, the dog_stages (doggo, floofer, pupper, puppo) should be in just one column. The breed and confidence columns in the image_predictions dataframe, and the name column in twitter_archive needed renaming for easy understanding. The reply and retweet columns needed to be dropped. The favorite_count and retweet_count from twitter API should have been part of the twitter_archive, so all three datasets should be merged as one.

Cleaning Stage

The cleaning stage followed the pattern of Issue-Define-Code-Test

Each of the **Quality and Tidiness issue** was listed and fixed in this cleaning stage.

The drop() function was used to drop reply columns, retweet columns and images that do not represent dogs. Incorrect dog names and 'None' values were replaced with NaN values. I selected only rating_denominator of 10, I dropped zero values and values greater than 20 in the rating_numerator. Timestamp was changed to datetime. Rows without images were dropped from jpg_url column. The extract() function was used to remove HTML codes from source column. The melt() function was used to combine dog stages into one column, and 'None', 'NaN' values were dropped. I used the rename function to rename dog name, dog breed and confidence columns. Datatypes were changed for tweet_ids, favorite_count and retweet_count using the astype() function. Finally, I merged all three dataset into one master dataframe, using the merge() function.

Conclusion.

The gathered, assessed, and cleaned dataset was stored into a CSV file named 'twitter_archive_master.csv', which can be used for future purposes. The importance of Data Wrangling as a data analyst cannot be overemphasized. It is a process that must be carried out in order to generate useful insights from any given dataset.

```
In [ ]:
```