wrangle_report

August 28, 2022

0.1 Reporting: wragle_report

This project involved wrangling and analyzing Twitter tweet archive of 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. Why? Because "they're good dogs Brent.". This archive contains basic tweet data for all 5000+ of their tweets as they stood on August 1, 2017.

The datasets used in this project are 3 parts: 1. The first is the main Tweeter tweet archive 2. The second is the Tweet Image Prediction 3. An addtional dataset that contained features like retweet count and favorite count

As expected, the datasets came with lots of issues that needed cleaning up. At least 8 issues were identified and fixed. The wrangling process began with twitter_archive_enhanced. Both visual and programatic approaches were used to identify problems with the datasets. The issues identified are categorized as Data Quality and Tidiness issues.

1. The first obvious issue were lots of NaNs in the twitter_archive_enhanced.csv dataset. This was noticed after running the sample() method a couple of times. These NaNs were eventually dropped.
2. Another obious issue was source column having some html data along with the actual/required data. These dara were sliced out since strings can be technically handled as lists in python.
3. Moving further to programmatically find more issues, it was seen that timestamp column on the contrary is of type string. This was fixed by converting it to pandas datetime type.
4. It was also noticed that some data points have None recorded instead of NaN. Pandas/python sees them differently. While the former is indeed a string value in pandas, the latter is actually "Not a Number" in which case means no record. None can therefore be misleading and represent a false record. This was fixed by changing all Nones to NaNs. This was done with dataframe replace method.
5. Although retweeted_status_id and retweeted_status_user_id were eventually deleted, it was notice that the the values were represented in exponentials which should not be so; this is because the columns are of float datatype. A proper recast to int should solve that.
6. It was also noticed that there were some erronous data in rating_denominator which are above the 10 denominator mark. This kind of error can help at the point of entry, collection or transfer. The issue was fixed by eliminating columns with erronous data.
7. A scrutiny on name column sort_values and value_count methods also reveal improper dog names. There were converted to NaNs likewise.

8. Expanded urls had lots of duplicate ulrs separated by commas. This was observed as a pattern. The data was splitted at comma and only one url returned. The comlumn was renamed as url which fits more
9. Inconsistent column naming was observed in tweet-json.txt dataset where tweet_id was give tha name id. Using the rename method solved that issue.
10. Another very vital issue was to ensure data worked on is not retweet data and with valid images. This was done by filtering out culprit columns.

The datasets were tidied up by: 1. merging doggo, floofer, pupper and puppo into categorical data in a column 2. merging all 3 datasets into for easy analysis

In all, the entire wrangling process was insightful and interesting though a bit challenging.

In [ ]: