Wrangling Report

Dataset ℹ

The dataset is about the tweet archive of a twitter user `WeRateDogs® / @dog_rates` which rates dogs based on their appearance, breed and their stage. This data contains 2356 record from 15 November, 2015 to 1st August, 2017.

Data Gathering

- **Twitter Archive CSV**

I downloaded `twitter-archive-enhanced.csv` file using the link provided in the Udacity Classroom, then assigning it into a dataframe named `df`

Link : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv

- **Image Prediction TSV**

I downloaded and saved the `image-predictions.tsv` file hosted on Udacity's servers programmatically using the Requests library provided by python. I assigned this file to a dataframe named `df_pred`

- **Twitter API**

Using Python's Tweepy library i was able to access the full data for each tweet. Referring to the tweet ids in the `twitter-archive-enhanced.csv` i stored each tweet's entire set of JSON data into `tweet_json.txt` then i converted it manually to json format `tweet_json.json`, The latter has been converted to a list of dictionaries and has been assigned to a dataframe named `df_json` with only tweet_id, retweet_count and favorite_count columns

Data Assessing

After merging all the three dataframes into a single dataframe called `df`, I managed easily to identify 8 quality issues and 3 tidiness some through visual assessment and others through programmatic assessment

- **Visual Assessment**

After showing 7 samples this is what i found :

- Missing values in multiple columns **(quality)**
- Unnecessary html tags in the source column **(quality)**
- Erraneous img_num datatype **(quality)**
- When removing duplicates retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp will remain empty therefore they should be dropped **(tidiness)**
- Dog stages are variables while they should be merged in one column to meet tidy data requirements **(tidiness)**

- **Programmatic Assessment**

Using pandas's various methods i was able to identify :

- rating_denominator contains values other than 10 **(quality)**
- Existence of retweets therefore there is duplicates **(quality)**
- Erraneous datatype in (favorite_count, retweet_count, p_dog and timestamp) columns **(quality)**
- Nulls represented as None in name column **(quality)**
- Name column contains multiple invalid dog names ('a', 'an', 'the') **(quality)**
- Dog stage columns contain multiple dog stage at once **(quality)**
- A new column dog_breed must be added based on the img_num and p1_dog **(tidiness)**

Data Cleaning

After identifying all the above issues, I started with creating a copy of the original dataframe named `df_cleaned` to perform all my cleaning procedures following the programmatic Data Cleaning process **Define, Code & Test**. I converted my observations from the

assess step into defined problems, translating these definitions to sophisticated code in order to fix these problems, Then i performed my testing codes.

Data Storing

Finally after cleaning, I saved the result which is a high-quality and tidy master pandas DataFrame to twitter_archive_master.csv file

References

- https://stackoverflow.com/questions/34126234/handling-json-file-with-python
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html?highlight=merge#pandas.DataFrame.merge
- http://seaborn.pydata.org/tutorial/categorical.html#:~:text=In%20seaborn%2C%20the%20barplot%20%28%29%20function%20operates
- https://www.tutorialspoint.com/how-to-write-text-above-the-bars-on-a-bar-plot-python-matplotlib#:~:text=Matplotlib%20Python%20Data%20Visualization%20To%20write%20text%20above,a%20set%20of%20subplots%20u
- https://www.geeksforgeeks.org/barplot-using-seaborn-in-python/

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js