WeRateDogs

Data Wrangling Report

Summary of wrangling efforts

1. Data Gathering:
    a. Twitter archive - a csv file provided by Udacity
    b. Image prediction for the WeRate Dog tweets - a tsv file provided by Udacity for students to download
    c. Original tweet data - each tweet was collected in JSON form using tweepy API and a Tweeter developer account. Then JSON data was parsed and saved in a pandas dataframe. There are also implementation details in the following section.
2. Assessment:
    a. Each data set was assessed to identify data quality and structural issues that need to be tidied up:
        i. Missing or obsolete data
        ii. Quality and uniformity of values in specific columns
        iii. Whether each observation forms a row or observation is spread among data sets
        iv. Whether each variable forms a column or spread among few columns
        v. Whether all data stored using an appropriate data type
3. Cleaning:
    a. Issues identified during assessment phase were thoroughly cleaned and well documented using define, code, test narrative to make the process intuitive for the reader
    b. Master data set df_master was created as a merge of three cleaned input data sources and stored as `twitter_archive_master.csv`

Implementation Notes

Gathering Twitter Data Using Tweepy

Note: This implementation avoids pauses in tweed data download using arrays of tweet ids. This was achieved by splitting the list of all tweet ids found in the provided twitter archive by list of 100 tweet ids. This allows downloading all tweet JSON data in one go.

The downside of this implementation is that downloading goes in a silent mode, and we cannot gather tweet ids which JSON is failing to download.

Nevertheless, those failed tweets would not be used in further analysis.

Merging Data Sets

Note: It was noticed during the merging phase that image prediction and WeRateDogs provided archive data sets both have orphans which were excluded from the final master data set. Orphans were assessed manually in Google Spreadsheets to ensure that no valuable data gets eliminated:

- 112 tweets in the image prediction data set were not found in the twitter archive. Visual assessment suggests those tweets are no longer available and could be deleted.
- 123 tweets in the twitter archive do not have image predictions. Visual assessment suggests those tweets contain videos and not images. That is a plausible reason why those tweets were not found in the image prediction data set.

Cleaning Issues Summary

Note: This summary is also provided on the supplied Jupyter notebook DataWranglingProject.ipynb

2.5.4 Data Cleaning

This is a summary of the issues noticed during the assessment stage that require cleaning.

2.5.4.1 Quality Issues

Image Predictions Table

1. p1, p2, p3 column values are in mixed case. Some are lower case some are capitalized. Underscore in these columns looks unnecessary
2. p1, p2, p3 column are non-descriptive

Twitter Archive CSV dataset

3. source column looks convoluted and its value can be simplified to value in href element. HTML markup is unnecessary for analysing data in this table.source can be simplified as listed value of:
    - Twitter for iPhone
    - Vine - Make a Scene
    - Twitter Web Client
    - TweetDeck
4. missing data in expanded_urls column: those are not original tweets with dog images and ratings. Those rows shall be dropped as they do not bring value for further analysis.
5. there are 181 retweets that shall be removed as this analysis required only original tweets with ratings
6. There are 78 records that are replies to other tweets. expanded_urls contains multiple repetitive urls. Only 23 replies have images in the image predictions table.
7. timestamp has string format, not date time format
8. doggo, floofer, pupper, puppo has None value instead of being empty. Some dogs have all these columns populated with None which means there are dogs without identified stage of their development. Further analysis will require a default value for dog development stage, and doggo fits as the most appropriate value.
9. floofer is misspelled and shall be floof
10. floofer shall have value as either True or False instead of many None and floofer values
11. name is not capitalized
12. name has None, a, an values that looks like omitted values
13. some values in name are not dog names
14. rating numerator and rating denominator have string format
15. Some ratings have decimals in tweet text and were not extracted correctly from the text column.
16. Dog Bretagne is a 9/11 hero who has an incorrect rating. It shall be 14 in numerator and 10 in denominator
17. there are tweets with 0 rating denominator. They are not tweets with dog images and ratings and shall be removed

Note: It looks like those will be outliers for the analysis as some raters got carried away. There is a dog with 1776/10 rating. This is not an issue for cleaning but rather a note for further analysis.

2.5.4.2 Tidiness issues

Twitter Archive CSV dataset

1. df_tw_archive: columns doggo, floofer, puppo, pupper represent 2 variables in 4 columns. Floof is not really a dog stage but rather an indicator if a dog has lots of fur or not
2. df_tw_archive: rating shall be a single number for analysis purposes defined as rating_numerator/rating_denomenator
3. df_tw_archive: omitted favorite count and retweet count will be merged from data collected via Tweepy
4. df_img_predictions: prediction shall be a single column with the most confidence coefficient where dog was predicted
5. df_img_predictions: image prediction data and tweet archive data are in separate data sets. Note: df_img_predictions has fewer records than tweeter archive so not all tweets in the twitter archive can be used in this analysis and only matching records shall stay