wrangle_report

June 26, 2022

0.1 Reporting: wrangle_report

For the purpose of this project which is basically wrangling and analyzing data from Twitter user, WeRateDogs, the data to be used were gathered from different sources. The first data (twitter_archive_enhanced.csv) was downloaded manually and uploaded to the workspace. The second dataset (image_predictions.tsv) was downloaded and saved programmatically from a provided url using the request library and its content written into the created 'image_predictions.tsv' file. The third dataset (twitter_api_data.csv) was scraped from twitter by querying its API. The tweets of each ids present in the twitter_archive_enhanced.csv data was collected via api.get_status() and written into a 'tweet_json.txt' file, each tweet on a new line. The content of this file was then read line by line, and tweet_id, favorite_count and retweet_count extracted from it. The extracted data was stored as 'twitter_api_data.csv'. All gathered datasets were then individually assessed both visually and pragmatically (.info(), .duplicated(), .head(), etc) for quality and tidiness issues. The issues detected were documented so they could be fixed. They were;

1. 1. twitter_archive: in_reply_to_status_id and in_reply_to_user_id columns have 78 values. Also, there are 181 values in the retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. These are not original ratings. Rows with non-null values were dropped. These columns were dropped too.

2. twitter_archive: Ratings with decimal values incorrectly extracted. Extract the correct rating numerators from the text column values using regular expression that matches and includes decimal values.

3. twitter_archive: datatypes of tweet_id, timestamp, rating_numerator and rating_denominator incorrect. Change datatypes to valid ones.

4. twitter_archive: rating_numerator contains ratings that aren't valid (in the actual tweet, these aren't the real ratings, and some have no ratings). Rows where the ratings were invalid were replaced with the valid ones. Then, rows where there are no ratings (we want only rows that have ratings) were removed.

5. twitter_archive: Some dog names in the name column are invalid. They have a pattern of starting with small letters. Invalid names were replaced with 'None'.

6. images: There are 3 predictions in this table. Just one with the highest confidence and a true dog prediction is enough. Also, the values aren't consistent with case type. A new column to extract dog_breed with a true value and highest confidence was created and values converted to title case.

7. images: The predictions and jpg_url columns are redundant. These columns were dropped.

8. images: tweet_id, img_num should be string datatype not int. Datatypes of these columns were changed to the valid type.

9. Twitter_archive: floofer, doggo, pupper and puppo should be in a column. These are values presented as variables. These four columns are supposed to be just a single column with the four titles as its values. To maneuver this, I concatenated values in the four columns, replaced 'none' with ' ' and then stripped the resulting values off excess whitespaces. I then replaced where the values are nothing('') with 'none' so that this dogs maintain their non-existent dog-type value. .value_counts() was used to confirm the list of values in the newly created column.

10. The 3 datsets should be in just a single dataset not separated since they all have a common column (tweet_id). This was achieved by using .merge() method to join the 3 dataframes on the tweet_id.

After cleaning and fixing all documented data issues, a single dataframe was arrived at and stored in a new file named 'twitter_archive_master.csv'.