wrangle_act

January 12, 2021

1 Gather

- using given twitter-archive-enhanced.csv
- dowloading image-predictions.tsv using requests
- using tweet_json.txt results becouse of auth issues with tweepy api

2 Assessing Data

The three saved data frames were first assessed programmatically in Jupyter Notebook with *pandas*, then visually in Excel/Google Sheets.

Several issues were detected and listed below:

Quality Issue (issues with content)

1. *df_WeRateDogs_Twitter_archive*:
   - 1.1 Only want original ratings (Delete the 181 retweets and 78 replies)
   - 1.2 Don't need those columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'img_num', 'expanded_urls' and 'jpg_url'
   - 1.3 All rating_denominator should be "10" and some rating_numerators are extreme values
   - 1.4 Since all the denominator is 10 after last step, we can get rid of rating_denominator column and change rating_numerators to 'rating'
   - 1.5 Many dog names are meesed up, such as "such" "a" "quite"
   - 1.6 timestamp have extra "+0000"
   - 1.7 timestamp's datatype should be converted to "datatime"
2. *df_WeRateDogs_predictions*:
   - 2.1 Remove "_" and capitalize the image predictions.(p1, p2, p3 column names)

Tidiness Issue (issues with structure)

0. Join 3 DataFrames.
1. *twitter_archive_df*:
   - 1.1 Dog stage's 4 variables: doggo, floofer, pupper, puppo should be in single column of categorical variable
   - 1.2 Dog stage have 'None' instead of np.nan

3. Cleaning Data

Tidiness Issues:

Issue 0: Inner join *df_WeRateDogs_Twitter_archive*, *df_WeRateDogs_predictions*, and *tweets_popularity*on tweet_id

Issue 1.1: Create *'dog_stage'* variable which is made by extracting the dog stage variables from the text column

Issue 1.2: Dog stage have 'None' and replace 'None' to np.nan

Issue 2.1: Use the ture prediction to fill in *dog_breed* column. If no ture prediction, fill in use np.nan

Quality Issues:

Issue 1.1: Select the rows from *twitter_archive_df* that *retweeted_status_id* and *in_reply_to_user_id* columns that is null

Issue 1.2: Remove columns: 1.*in_reply_to_status_id*, 2.*in_reply_to_user_id*, 3.*retweeted_status_id*, 4.*retweeted_status_user_id*, 5.*retweeted_status_timestamp*, 6.*img_num*

Issue 1.3: Drop rows where denominator of rating != 10 and where numerator rating >> 10 Issue 1.4: Drop

rating_denominator column

Issue 1.5: We find all the incorrect names have lowercase first letters. We will change those names to None, then change all the None to np.nan

Issue 1.6 &1.7: Use *str.strip* to remove "+0000" and use *pd.to_datetime* convert timestamp's datatype Issue 1.8: Use

regular expression and *Series.str.extract* to find real source between tags >

4. Storing Data

Store the clean df in CSV file with name using *.to_csv('twitter_archive_master.csv')*