wrangle_report

May 29, 2019

1 WeRateDogs - Twitter Data

1.0.1 1. Gathering Data

I looked at the instructions given by the Udacity team on how to gather data for this data wrangling analysis. - I downloaded the data which is a given CSV file and named as twitter-archive-enhanced.csv. - Next I created my twitter developer account and created the JSON file named tweet_json.txt by using the API. - Next I downloaded the file image predictions file which is in the tsv format. Once I had all the above three files, I created them into 3 different dataframes which are shown below. - archive_df - this is a dataset "twitter-archive-enhanced.csv" which was converted into a dataframe and gives information on basic tweet data. - tweets_df - This dataset will contain information like tweet_id, no of retweets and no of favorites etc., - img_df - This dataset will contain information about predictions about the image.

1.0.2 2. Assessing the Data

We have three dataframes: - tweets_df which has retweet and favorite counts - img_predictions_df which has the results of a neural network trying to identify dog breed in a tweet's picture - archive_df which has the tweet's text, rating, and dog category

1.0.3 Archive_df table

Quality

- Retweets need to be removed to avoid duplication in our analysis. This may be done by removing rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. When this step is correct, there should be a fewer number of non-empty tweet ids.
- Rating numerators have not been properly cleaned. The current pipeline captures incorrect values when rating numerators contain decimals.
- There are cases where there are multiple dog stages in a row.Need to handle these cases.
- Missing values in columns from in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- tweet id 835246439529840640 has a rating of denominator = 0
- weird names found for dogs - 'an','by','his','infuriating', 'just', 'life', 'light', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'
- timestamp and retweeted_status_timestamp should be datetime type instead of the object

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should not be float
- The columns which have missing values in doggo, floofer, pupper , puppo - has None instead of NaN
- We see that the information of text is truncated to 50 characters. Anything in excess is ellipsized
- This archive_df is like a main base table with the above attributes, there are some other attributes that are found to be there in other dataframes. #### Tidiness: ###
- Hence we need to join all other dataframes to get a final dataframe.
- Dog stages are found in multiple columns, hence we should find a way to club all these variables into single column. This will reduce the dimensionality of the dataframe

1.0.4 Tweets_info_df

Quality - tweets_df table

- 19 tweet ids information is Missing
- Retweets and Favorites needs to be joined to the archive_df table

1.0.5 img_df

Quality - tweets_df table

- Only 2075 tweets have images.
- Retweets and Favorites needs to be joined to the archive_df table

1.0.6 3. Cleaning

For cleaning all the 3 dataframes, Here are the steps I followed before after joining the dataframes. - Convert the datatype of "tweet_id" into string - Remove the retweets to avoid duplication in analysis - Create a universe dataset joining all the dataframes based on the tweet_id - Convert the dog stage or category into one column instead of the multiple columns - Rating numerators have not been properly cleaned. The current pipeline captures incorrect values when rating numerators contain decimals.Refetched the value from text - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id -- Convert all these into Object or string - retweeted_status_timestamp - Convert this variable into datetime format - We see that the information of text is truncated to 50 characters. Anything in excess is ellipsized. Let us increase the text format representation - Weird names found for dogs - 'infuriating', 'just', 'life', 'light', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'. Let us clean to ideal name by looking at the text. - retweeted_status_timestamp - has the null values , I dropped this variable

1.0.7 4. Store

I stored the final dataframe into csv file with name twitter_archive_master.csv with final data of 1990 rows and 26 columns