wrangle_report

September 11, 2022

0.0.1 WeRateDogs Wrangle Report

In this report I am outlining the steps I took gather and clean the datasets required fo this ananlysis.

0.1 Data Gathering

I gathered the datasets from three different sources: 1. The WeRateDogs twitter enhanced archive file was manually downloaded from the Udacity servers. 2. The tweet image prediction (image_predictions.tsv) file was downloaded programmatically using the requests library 3. Each tweet in the twitter enhanced archive file JSON data was downloaded by querying the Twitter API using Tweepy

These three files were then stored in 3 separate dataframes. twitter_archive_df,image_prediction_df and twitter_api_df.

0.2 Assesment and Cleaning of dataframes

I began assessing the dataframes by viewing each of the dataframes both visually and programmatically. I identified 8 quality issues and 2 tidiness issues as specified in the requirements for this analysis. I first convered the timestamp in the twitter_archive_df to datatime. I then converted the created_at column in my twitter_api_df to datetime as well. Next issue i tackled was formatting the text in the p1, p2 and p3 columns in the image_prediction_df. I also dropped the rows with no data in the expanded_urls column. The expanded_urls column also had some rows with multiple urls, so i handled that by only keeping the first url in the row and deleted the rest after th delimiter which was a comma sign. The expanded_urls column also had some duplicate values so those duplicates were dropped from the dataframe. Some text in the name column of the twitter_archive_df were in lowercase instead of uppercase, so that was handled by making all the strings in the column uppercase.The name collumn also had some invalid names so those names were removed as well.

I also improved the tidiness of the image_prediction_df by creating one prediction column instead of three based on the values and conditions of those three seperate columns. The same step was done for the fluffer puppo pupper and floofer columns in the twitter_archive_df. A new separate column Dog_Terms was created based on the vaules in the four separate columns.

Additional cleaning was done on the datasets by selecting only original tweets, this was achieved by selecting all rows with null values in the retweeted_status_id column and in_reply_to_status_id as specified in the requirements for this analysis. The source column was also worked on by extracting the string between the > and < /a> characters. The rating numerator and rating denominator columns were also cleaned by dropping the rows with

invalid data. For the numerator excessively high values were dropped and for the denominator column all other values except '10' were dropped. A new column rating was then created by dividing the rating numerator and rating denominator columns. All 0 values in the newly created Dog_Terms column were replaced by "No_Term". Then the columns in the twitter_archive_df not essential to my analysis were dropped. The columns dropped were 'doggo','floofer','pupper','puppo','in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted columns.

I also dropped non essential columns in the image_prediction_df dataset. Columns dropped were'p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog','jpg_url','img_num' columns.I also replaced the 0 values in the Prediction coulmn with "Not_Conclusive".

The id column in the twitter_api_df was renamed to tweet_id. All other columns in the twitter_api_df were dropped except the tweet_id,favorite_count and reweet_count columns.

After all the cleaning i merged the three dataframes by first merging the twitter_archive_df and twitter_api_df using a left join. I then merged this newly created dataframe with the image_prediction_df. After the merge I casted the favorite_count and retweet_count columns to int64. This new master_df was then saved to a csv file twitter_archive_master.csv'.