wrangle_report

September 6, 2022

0.0.1 Introduction

WeRateDogs is a twitter account that rates various dogs with a humorous comment about the dog. The dataset that we have wrangled and analyzed consists of basic tweet data for all 5000+ of their as they stood on 1st August, 2017. This report,I am going to outline the efforts in wrangling to obtain a clean dataset to be used for analysis.

0.0.2 Data Gathering

I gathered the dataset from the following three sources:

1. twitter_archive_enhanced.csv which was downloaded manually and uploaded in the Jupyter notebook.
2. image_predictions.tsv which was downloaded programmatically the Requests library.
3. Additional data from twitter API. An entire set of JSON tweets was downloaded by querying the twitter API using Tweepy library.

0.0.3 Data Assessing

The three datasets were assessed visually and programmatically for ensure that both quality and tidiness issues were identified. The following issues were identified:

A) Quality issues

i.twitter_archive_enhanced table has invalid names under 'name' column (a an, the,None)

ii. Retweets need to be dropped in twitter_archive_enhanced table (retweeted_status_id, retweeted_status_id, and retweeted_status_user_id)

iii. These columns need to be dropped (in_reply_to_status_id, in_reply_to_user_id)

iv. In image_predictions table, 'p1','p2', and 'p3' have entries beginning with lower case, others with underscore or dash.

v. Erroneous datatype for 'timestamp', 'doggo', 'floofer', 'pupper', and 'puppo' in the twitter_archive_enhanced table.

vi. In the image prediction dataset, there are 66 duplicated entries under the 'jpg_url' column.

vii. Missing data in the 'name' column in the twitter_archive_enhanced table.

viii. In image_predictions table, 'p1','p2', and 'p3' are not properly described., 'text' to 'tweet'.

B) Tidiness issues

i. In the twitter_archive_enhanced table, the dog 'stage' ('doggo', 'floofer', 'pupper','puppo')should be under one column.

ii. Need to merge the three tables (tweet_archive,image_predictions, and tweet_status)

0.0.4 Data Cleaning

In the twitter archive dataset, I began by dropping the following columns which were not of use in my analysis: 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'expanded_urls', 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp'. I also converted the timestamp datatype from object to datetime. This table contained invalid dog names under the 'name' column and some of the dog names' first letters were not capitalized. I replaced the invalid names with NaN using the np.nan function and the first letter of dog names were capitalized using str.capitalize() function. To create some tidiness in this table, column 'doggo', 'floofer', 'pupper', and 'puppo' were melted to create one column 'dog_stage' using the concat fucntion. The entries that had no do stage were replaced with NaN.

In the image predictions table, there were 66 duplicated entries in the 'jpg_url' column. The duplicates were dropped. In the same table columns 'p1', 'p2', and 'p3' had entries beginning with lower case and the columns were not properly described. The first letter for the entries were capitalized while the column names were replaced as follows: p1-prediction_1, p1_conf-confidence_1, p1_dog-dog_1, p2-prediction_2, p2_conf-confidence_2, p2_dog-dog_2, p3-prediction_3, p3_conf-confidence_3, p3_dog-dog_3. In the tweet_status table, I renamed the 'id' column to 'tweet_id'.

Lastly, I merged the three tables to obtained one table which I saved as twitter_archive_master.csv.