wrangle_report

May 25, 2020

1 Data Wrangling

1.1 Gathering

We download the provided WeRateDogs Twitter archive data. To get some other essential details about the tweets that are absent from the archive like the Number of Retweets and Favorites, we use the Tweepy library to gather the data based on the tweet ids in the archive. The data about the image predictions that was run through a neural network is downloaded using the Requests library from this link. These 3 datasets are wrangled and analyzed.

1.2 Assessing

We access the gathered datasets to find both Quality and Tidiness issues present in the dataset. ### The Provided Archive Data - archive_df

The archive data has 2356 rows and 17 columns. The source column contains full anchor tag which has 2 variables, that is, the name assigned to the source and the link for that source.

The expanded_urls column shows the links that form part of the tweets. A tweet that has an expanded_urls value indicated that at least it had an image attached to it. Some of the values were missing which meant that they didn't have images attached to the tweets. We only need tweets that had images for this particular analysis. find that there were duplicated values in some of the expanded_urls.

We only want to work with tweets originally tweeted by WeRateDogs and not Retweets of other user's tweets. We find these retweets by finding all tweets that contain 'RT @' in their text.

We also find some of the tweets to be updates to original tweets. These tweets include 'PUPDATE' in their text.

The names of the dogs extracted from the tweets included 55, 'a' and 745,'None' as values. These indicate some inaccuracy. We confirm this by looking at the text for the tweets with names of 'a' or 'None'. We observe that most of the names that were not captured were preceded by 'meet', 'named', 'name is' or 'This is'. We use this to find these tweets so that the names can be extracted again.

The various stages of the dogs are separated into their own columns. We also find that the dogs belonging to the 'floof' stage have been captured as either 'floof' or 'floofer'.

The names and dog stages are in a mix of Lowercase, Uppercase and Capitalized.

The dog rating was separated into numerator and denominator. A column that contains the full rating in the form or numerator/denominator is also needed.

We also do a programmatic assessment using several pandas functions and discover that some of the datatypes were erroneous.

1.2.1 Downloaded Tweet Data - tweet_df

The downloaded tweet dataset contained several columns but we interested in only the 'favorite_count' and 'retweet_count'. These columns would have to be added to the archive dataset.

1.2.2 Image Predictions dataset - image_predictions_df

The predicted breeds are in a mix of Lowercase, Uppercase and Capitalized. The predicted breeds which were truly dogs would have to be added to the archive data for this analysis.

1.3 Cleaning

We make a copy of all 3 datasets for the cleaning process. We drop the following columns from a copy of the archive data: * in_reply_to_status_id * in_reply_to_user_id * retweeted_status_id * retweeted_status_user_id * retweeted_status_timestamp

We make heavy usage of regex patterns to extract the information we need from the columns of interest.

- The source data was separated into source_url and source_name using regex.
- We extract the stage names from the text column and drop the 'doggo','floofer', 'pupper' and 'puppo' columns and store in a 'dog_stage' column.
- All the rows under the 'dog_stage' column are changed to lowercase.
- The 'floofer' dog_stage is also changed to 'floof'.
- We merge the 'favorite_count' and 'retweet_count' from the downloaded tweets to the tweet archive based on the tweet ids.
- The Retweets and PUPDATE(s) are removed by selecting only rows that do not contain either of these in their text column.
- We also remove tweets with null expanded_url values by selecting all the rows with non-null expanded_urls values.
- The duplicated in the expanded_urls are removed by; a) Create a list of the strings. b) Cast the resulting list to a set to remove the duplicates. c) Cast them back to a string joined by a comma.
- The names were extracted from the text column using the pattern observed during assessment for names which were initially stored as 'None' or 'a'. Not all of the text contained names of the dogs.
- All the names were converted to lowercase.
- A rating column was created with a combination of the numerator/denominator as a string.
- For the erroneous datatypes, the favorite_count and retweet_count were changed to int, timestamp to datetime and the dog_stage to category.
- The predicted breed names were also changed to lowercase.

- From the prediction dataset, we choose predictions that are dogs and have the highest confidence. In a case where none of the predicted names was a dog, we assign a NaN value under the newly created ‘predicted_breed’ column.
- The jpg_url and ‘predicted_breed’ columns from the image predictions dataset are merged with the archive dataset on the tweet_id into a final dataset named, ‘combined_clean’.