12, December, 2022

Wrangling Report- Twitter WeRateDogs Archive

This report discloses the wrangling efforts used to gather, assess and clean data regarding a twitter archive from the twitter user We Rate dogs.

A file called ‘twitter-archive-enhanced.csv’ was given to help start my analysis. This contained data such as the tweet-id for each tweet, the date and time of the posting of the tweet, as well as URL’s to the tweet and the text column which was the tweets themselves. Udacity started some data cleaning, by filtering through the text columns for Dog names, the type of dog, (Doggo, Floofer, Pupper, and Puppo) and the rating the dog received. The cleaning wasn’t perfect and needed to be fixed. I imported this into a dataframe in pandas.

Using the requests library, I was able to grab the image-predictions.tsv file which was a file containing predictions on what breed of dog was in the tweets photos. This was imported into a pandas dataframe.

The last piece of data that I needed was gathered using the Twitters API. Using the tweet ID from the twitter-archive-enhanced.csv, I iterated through each id and gathered info such as the favorite count and retweet count. This was saved to a tweet-json,.txt file, which I also imported into a dataframe in pandas.

To assess the data that I had, I first merged all the dataframes into one dataset to make it easier to view. I assed the data programmatically, as well as visually with google spreadsheets, to try and clean the dataset so it can be anaylzed.

The cleaning that I noticed that needed to be done, and was successfully completed are as follows: **In tidiness issues**

- 1. doggo, floofer, pupper, puppo could be put into 1 column with those options to pick inside the column.
-
- 2. Merge 3 Dataframes.
-

Quality Issues:

- 1. the rating of the dog inside the text column, if the rating ended in .5 (example 9.5/10), the rating numerator only took the number after the decimal point.
-
- 2. Some tweets contained in the dataset are retweets, which mean that they show up twice in the dataset.
-
- 3. Remove columns with all null values
-

- 4. Change columns data types to proper data types
-
- 5. Some tweets don't have photos to go with it, which doesn't get a prediction and also shouldn't get a rating for accuracy.
-
- 6. Majority of dogs are not categorized, could skim through the text with a list of strings likely to be in texts and categorize it.
-
- 7. Some dog names are not names. These seem to be indicated by just having lower case letters. removed rows that have only lowercase letters in the data.
-
- 8. Fix columns to be easily understood:
-    - p1 = first_prediction
-    - p1_conf = first_prediction_confidence
-    - p1_dog = first_prediction_is_dog
-    - p2 = second_prediction
-    - p2_conf = second_prediction_confidence
-    -p2_dog = second_prediction_is_dog
-

Lastly, I removed any columns that had all null values, and converted the date column to a datetime format for future analyzing. This may not have cleaned up all issues, but the dataset is now ready for analysis.