wrangle_report

February 15, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

1 Wrangle Report - February 2022

1.1 Data used for wrangling:

-Tweet archive from WeRateDogs
-Tweet data pulled from Twitter API
-Tweet images gathered from twitter

1.1.1 Quality issues

1. There are tweets that do not have images
2. Some dogs have invalid names, such as None, a, an
3. Timestamps should be datetime instead of float64
4. In the archive, the columns for doggo, floofer, pupper, and puppo should be combined into a single column
5. In the images, some of the breeds listed are objects
6. In images, some breeds are actually other animals
7. Archive contains retweets, which are not what we're looking for
8. In the columns for doggo, floofer, pupper, and puppo, if there isn't a value, it should be changed from None to NULL

1.1.2 Tidiness issues

1. There are unneccessary columns that have the value NaN
2. Image number in the images dataframe is unneccessary

1.2 Cleaning Data

1.2.1 Issue #1: archive contains retweets, which are not what we want in this dataset.

For our purposes we only need tweets, and not retweets.

Define: Remove retweets from archive_clean The retweets were removed by keeping everything that had datapoints related to retweets set to NULL

1.2.2 Issue #2: Some names listed are invalid, such as a, an, or None

While the program did well at picking out dog names from the data provided, it was not always accurate. Occasionally some dogs would have the name None, a, an, etc.

Define Change invalid names such as a or a to None, and then change None to NULL If dog names contained lower case letters, I changed them to NaN. Next I replaced all instances of "None" with NaN as well.

1.2.3 Issue #3: Datatype of timestamps is float64 when it should be datetime

Timestamp was stored as object. The correct datatype would be datetime.

Define Change datatype of timestamps to datetime instead of float64 Corrected the datatype to datetime.

1.2.4 Issue #4: The columns doggo, floofer, pupper, and puppo should be a single column

The four different classifications of dog should be combined in a single column, instead of four separate ones.

1.2.5 Issue #5: There are values of None in those columns. should be replaced with NaN

Some columns have the value of None. Instead the value should be NaN.

Define Combine the columns for doggo, floofer, pupper, and puppo into a single column and change any values of None to NaN The four columns were combined with the values of None being changed to NaN.

1.2.6 Issue #6: Some images are labled as things other than dogs

Some of the images were incorrectly labled as things like "shopping cart"

Define Remove rows where the images are labeled as not dogs. This handles both the issue where the images are labled as objects and the issue where they are labled as other animals. Rows in which the labels were not types of dogs were removed as there was not a simple way to correct the incorrect values.

1.2.7 Issue #7: Ratings should be combined into a single value

There is no need for the numerator and denominator to be a separate value, as they were not used for any math operations.

Define combine the ratings columns The two columns related to ratings were combined into one singular column.

1.2.8 Issue #8: Some tweets are replies to other tweets, these should be removed.

As was the case with retweets, replies were also not relevant for this project.

Define Remove tweet replies from archive_clean Any row with a value in the columns related to tweet replies that was not NaN was dropped.

1.2.9 Issue #9: There are unneccessary columns that are primarially NULL values

There were several columns that contained primarily NaN for values and had no relevance for this project.

Define Drop unnecessary columns Unnecessary columns were dropped, making the tables easier to read.

1.2.10 Issue #10: Img_num column is unnecessary

The field of Img_num in the images file was unnecessary.

Drop img_num column The img_num column was dropped.

1.3 Analysis

After cleaning operations were complete, I used the information to create insights. Based on the information provided I discovered that the most common dog breed associated with the tweets was Golden Retriever. In regard to the WeRateDogs dog typing system, the most common one was Pupper, however some dogs did have more than one kind of type. Also there is a positive correlation between retweet count and favorite count. This is to be expected. There was also a large number of retweets in which it was not also favorited.