Wrangle Report

November 2, 2020

1 Wrangle Report:

This report summarizes the steps and efforts made to wrangle the data into a cleaner version appropriate for visualization and analysis.

2 Quality Issues:

3 Some of the observations/entries are retweets. We need to delete these because we only want original data and don't want to double count anything.

We fix this issue by looping through the data and deleting any of our tweets that have a non-NaN value in the "Retweeted Status ID" column because this column will have a tweet ID listed in the case of a retweet. (I.e. a NaN value indicates that the tweet is NOT a retweet which is what we want to keep)

4 Some of the entries are reply tweets from the WeRateDogs account on other Twitter accounts' tweets. We don't want to include these reply tweets in our analysis because they skew the data due to receiving far less exposure than the regular tweets that come directly from WeRateDogs themselves. The breed prediction model also doesn't work on these since the picture is not a direct part of the reply tweet.

We fix this issue similarly to how we fixed issue #1: loop through the data and delete any of our tweets that have a non-NaN value in the "In_reply_to_status_id" column.

5 Some of the remaining entries after cleaning the first two quality issues are not actually dog ratings and are just tweets about unrelated matters like bands or hotdogs or the date.

We loop through the data and delete any of our tweets that do not have a value in the "Expanded URLs" column because all proper submissions to WeRateDogs should have an embedded picture in the tweet whose url link should show up in this column.

6 There are a few tweets that seem to have been deleted at some time between its original posting and this project's commencement (October 2020). This means that Tweepy was unable to gather tweet favorite/retweet data since the tweet no longer exists.

We loop through the data and delete any of our tweets that Tweepy failed to extract a favorite or retweet count number on. Sampling some of these tweets and checking them live on Twitter.com verifies that many of these tweets have been deleted or no longer exist.

7 There are a few tweets in which the image prediction algorithm was unable to produce a result.

p1 is the prediction that the model is most confident in when assessing each tweet's photo. If there is no value in p1, that means the tweet has some issue in which the picture is either copyrighted/not available, is a video, or contains multiple dogs.

8 There are some scores which have a denominator that is not equal to 10. This should not be the case since WeRateDogs is rigorously consistent in using a x/10 rating system. (Many of these appear to be because there are multiple dogs in the photo - we want to make sure each entry that we perform our data analysis/visualization on are all standardized to be one dog per rating and a rating out of 10).

Loop through our data and drop any tweets which have a denominator that is not out of 10.

9 Timestamp values do not need to have the "+0000" at the end. It should be in the "datetime" data type, not a string/object.

Self-explanatory fix - We remove the +0000 and then convert the resulting string into a datetime format.

10 Retweet and Favorite Count should be integer datatype, not a float.

You cannot have a decimal amount of a retweet or favorite count, only whole positive integers.

11 There are a few entries which have incorrect names for the dogs. This is either because there isn't a provided name in the tweet (which usually results in the name saying "an" or "a"), or the name was faultily extracted during the data gathering process.

Although some of the entries have an incorrect name listed due to missing data (i.e. there was no name provided at all in the tweet), we can fix the ones that DO have a name but were incorrectly extracted which we do so in our code inside the jupyter notebook.

12 Tidiness Issues:

13 There are a lot of extra columns, some of which don't provide any data or variables (e.g. Retweet ID and Reply To ID) particularly after deleting the retweeted and reply tweet entries (part of our Quality Issues cleaning process)

Simple fix - clean up the dataframe by deleting our extra columns we no longer need.

14 The stage of the dog's life should be one column instead of four.

First we create a new column called "other" which assigns an "other" value to any dog that is not classified as one of the existing four stages. Then we use panda's melt function to consolidate this new column with the other four for the stages (doggo, pupper, puppo, and floofer) to ensure that we follow proper data tidiness rules.