Wrangle report

-By imen bakir

In this report I will be explaining my wrangling methods of the data that I've gathered in different sources and in different formats.

Firstly, I have gathered 3 datasets : the tweet archive of twitter user @dog_rates, also known as WeRateDogs; It is a twitter account that rates people's dogs with a humourous comment about the dog. The archive contains basic tweet data(tweet ID, timestamp, text, etc.) for all 5000+ of their tweets till August 1, 2017.The second dataset is the tweet image prediction data downloaded from Udacity's server. It contains additional tweet data(tweet ID, jpg_url or the dog's image, confidence interval of prediction, predictions of the dogs...).The last dataset is gathered through the twitter API using python's Tweepy package to get retweet count and favorite count of dogs.

After gathering all the data needed for the wrangling, the first step to be considered is to assess the data to improve its quality and avoid costly mistakes that can affect later, on the machine learning model.

I have assessed each dataset and gathered some quality and tidiness issues. In the twitter archive data, I noticed that there are many rows in the column 'name' that contains None as value, I have tried to change it to 'unknown' to make it more understandable for reader. There are many columns that are unnecessary and not useful to us such as the retweet columns because we only need the original posts, so we had to delete them.

Also, I have noticed that the column 'timestamp' is not in the correct data type which is date time and is in type 'object', therefore I changed its type.

There was repetitiveness in the 'expanded_url' column so I had to split the urls by commas using str.split function. Another thing I have noticed is that the url is combined with a hyperlink tag (<a href='...>) so I had to split it and remove the hyperlink.

As for the tidiness issues, I had to combine dog stages in one column since we don't need multiple columns ('doggo', 'floofer', 'pupper', 'puppo'). And for the 'text' column I have changed it to 'tweet' so it's more readable using the rename function.

In the second dataset, we had multiple predictions columns (p1_dog, p2_dog, p3_dog), so I had to melt those columns in one column, so it is cleaner.

Also, there was predictions that the image doesn't contain a dog, so I had to delete those observations since we don't need them in our analysis.

I've encountered some tidiness issues such as some of the dog names in p1, p2 and p3 columns contained an underscore "_", so I edited it to make it tidier.

Another thing is that some column names had to be changed so the dataset would be more understandable.

In the last dataset, there were much fewer quality issues. I've noticed that the date column wasn't in the correct format and as a tidiness issue, it was not written in a tidy structure.

As a last step before visualization, I had to store all the datasets gathered and wrangled in one master dataset that is tidy and clean using the 'merge' function and saving it to a csv file using the function 'to_csv'.