wrangle_report

August 28, 2023

The data was first gathered into three separate data frames. The first data frame containing the scraped archive data was provided and simply read into the program. The second consisted of downloading the image data programmatically with a URL and reading it into the program. The third tweet data frame was originally meant to have the tweepy API utilized but the data needed required a paid subscription. Instead, I opted to manually upload the provided text file, read it line by line into a list, and create a data frame.

Each data frame was then assessed visually and programmatically to identify quality and tidiness issues. I found some of the data in the archive data frame to be retweets that needed to be dropped since we were only focusing on original tweets for this analysis. Several columns in the archive data frame were either unnecessary or were tied to the retweet data. The timestamp column in the archive data frame was an object data type when timedate would have been a more accurate choice. The dog adjective rating system employed by the account in the archive data frame was unique but difficult to understand so those values needed to be replaced with more concise wording.

In the image data frame, it was found that 66 duplicate images existed. These need to be dropped. Each data frame had an ID column with an integer data type, but these ID's were more unique addresses than numeric values so the data type needed to be changed to object for accuracy. Every dog that was numerically rated was meant to be rated out of 10 as the denominator but it seemed that was not the case for the rows in the archive data frame. It would need to be changed so that every denominator was 10.

There were several probability-based columns in the image data frame that addressed each attempt the neural network made in identifying dog breeds. Each attempt besides the first had no accuracy percentage higher than 70% so they were best dropped from the set all together. According to the rules of tidiness dog stage is one variable and hence should form single column. Finally, each data frame would be combined and any duplicate or unnecessary columns in this final set would be dropped.

Each issue addressed was corrected. All the cleaning efforts were documented and tested. The final result was saved to a single master csv file.