Cooper Hepworth

12/14/2023

Wrangle Report

Observing the three given datasets I got to work looking at what each one was holding. The archive dataset was compiled to hold data on the content of tweets. I noticed that whoever compiled it used the text to pull the name of the dog, the numerator and denominator for the user rating, and the type of dog listed. The name of the dog was found from grabbing the word that followed “This is <name>” in the text. This worked most of the time but when the text read something else like “This is a” or “This is the” it would grab “a” and “the”, which obviously are not names. To separate these “non-names” from the rest I searched for words that were all lowercase. I was able to review the list visually quickly to see that there were no actual names that were just lowercase. There was also a problem with the name O’Malley being stored as “O”.

The Images dataset dealt with images and the algorithm’s predictions as to what dogs were in the images. I made sure all the dogs predicted were uniform in format with no caps and spaces instead of underscores. There wasn’t much else I would change with this dataset.

The tweets dataset contained descriptive information about tweets such as retweets and favorites. The id column needed to be renamed to “tweet_id” to be able to be joined with the other two datasets.

After the datasets were joined, I wanted to look at the predictions of images. Since the image predictions were in three separate values, I needed to combine them and “stack” them into a new dataset to observe them. This allowed me to see the correlation of the confidence

percentages and the counts of predictions. The other analysis insights were achieved by querying the original “master” dataset.