wrangle_report

July 14, 2020

0.1 Data Wrangling Report

In this project, tweet data from weratedogs was retrieved from Twitter's API and stored in a json txt file, which was then read into a dataframe in the pandas module in python. Given for this project was also a csv file with some extracted data from the twitter archive and a neural network that analyzes the tweet images from twitter and attempts to predict the breed of the dog for each individual tweet.

These three dataframes were merged into one throughout the analysis.

While perusing the data, several issues were discovered:

1) First off, due to the fact that the data extract picked up retweets, there was duplicate data present that needed to be dropped.
2) Secondly, the data on the dog stages was spread across four different columns in the dataframe when it really should be represented by one column
3) Next, the timestamp need to be converted to a datetime object
4) Also, some of the rating systems were not accurate
5) Due to the fact that we are dropping the retweet data, we can discard the retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns
6) There are large outlier dog ratings that made it difficult to analyze and visualize the data and should be removed for analysis
7) The dog stages should be expressed as categorical variables
8) Dog ratings should be expressed as a single column
9) In the dog stage column, the "None" string should be replaced with a null value
10) All non-dog image predictions needed to be removed from the image dataframe

These issues were all fixed throughout the process of the data cleaning process.