wrangle_report

April 24, 2018

1 Wrangling Report:

Twitter_archive_enhanced.csv has a fair amount of issues that needed to be addressed in the wrangling process. The data had issues with its tidiness in terms of having columns that could be combined or eliminating columns that weren't necessarily useful since we wanted only original tweets with images. There were a large amount of quality issues found in the data too, where dog rankings weren't necessarily accurate or simply that some of the columns used suboptimal data types, such as floats instead of strings or timestamps not being datetime objects. Despite the cleaning efforts that were made, there remains a fair amount of issues that weren't touched upon for the sake of time and can be revisited in the future.

The tweet image predictions weren't extensively cleaned since the importance of each column isn't necessarily defined. However, it can be argued that there are some tidiness issues that can be addressed in the future, where maybe prediction accuracy scores can be consolidated. Regardless, because we wanted tweets that had images associated with them, we used this dataframe to filter out only the ones we need from the twitter archive and the most recent data we scraped from the Twitter API.

From the twitter API, we used a couple of columns of interest, although there are plenty more we could've used in the final cleaned master. We renamed the dataframe as twitter_json in the jupyter notebook and used the id, retweet_count, favorite_count, and full_text data from the JSON file and merged that into our final cleaned master dataframe. Although only retweet count and favorite count were necessary for the final dataframe, we also pulled the extended text data from each individual tweet since the archive only stored a limited amount of text. This can be useful in the future since perhaps some of the data quality issues from the archive stem from incomplete text information.

From these three dataframes, we cleaned and merged them together based on the tweet ids that had an image and were currently on twitter into the final master that can be used for more in depth analysis. This can be useful in the future in order to better train a machine learning algorithm to recognize certain breeds of dogs. Also you could gather insight into what breeds of dogs are the most popular, or perhaps gain understanding on how users rate dogs and see whether or not that's based on factors such as breed or simply because they're more prevalent on twitter.