Alexander Rodriguez

Wrangle Report

There were three files that needed to be gathered for this project. The first file was the twitter-archive-enhanced.csv which was a file on hand. The file contained a lot of important and useful information such as ‘tweet_id’, ‘timestamp’,’text’, ‘rating_numerator’, ‘rating_denominator’, and a dog stage. This file had both quality and tidiness issues. One glaring tidiness issue was that the dog stages were recorded in multiple columns. To fix this tidiness issue, the columns doggo, floofer, pupper, and puppo were melted into one column called ‘stage’. Most dogs did not have a stage classification and ‘None’ was used as their stage.

Another quality issue with twitter-archive-enhanced.csv was the fact that there were misclassified rating_numerator values. This error arose from the regular expression that was used to extract the rating_numerator, which failed to notice floating point values. Below are the issues that were discovered during the wrangling process:

- Index 1689 has a dog that is rated as a 5. The dog's rating is actually 9.5.
- Index 313 is 980 when the rating should actually be 13.
- Index 340 rating numerator is 75 when it should be 9.75.
- Index 695 rating numerator is 75 when it should be 9.75
- Index 763 rating numrator is incorrect.
- Index 1202 rating numerator is 50 and should be 11.
- Index 1712 rating numerator is 26 and should be 11.26.

The next step was to download the ‘image_predictions.tsv’ file programmatically. Since there was only one file to gather the process was straightforward and no real issues were encountered. The file did not have too many quality/tidiness issues that needed to be fixed. The file contained inconsistent naming of dog breed with some letters capitalized and some not. Furthermore, the neural net algorithm could not correctly classify the breed of the dog and instead honed in on other objects such as computer screens in the background. This issue could not be fixed because although I had access to the dog’s picture I could not properly identify every dog’s breed. The ‘images_clean’ data frame contains the ‘jpg_url’, ‘tweet_id’, and all of the dog breed classifications.

The final gathering step was the most challenging since the Twitter API was queried to obtain tweet data in JSON format. Learning tweepy was challenging and I subscribed to ‘Datacamp’ to learn more about how to import data. However, a lot of important details were not taught in the course, which did not make the process of learning tweetpy any easier. I utilized the API documentation and online tutorials and I learned that gathering data is extremely important step of the data science process. A process that I need to do much more of to get comfortable with it. The content of interest in this dataset was the ‘favorite_count’ and the ‘retweet_count’. In the final ‘tweets_clean’ dataframe every column was removed except ‘favorite_count’, ‘tweet_id’, ‘retweet_count’, and ‘text’.

Lastly, all three data frames were merged together by making sure that the ‘tweet_id’ was converted to a string. The final data frame only had 1300 observations and was used to create visualizations of the data. The interesting insights will be displayed in the act_report.