Mid Term Project

Sam Loyd

Bellevue University

January 2020

I selected a data set on adoption outcomes from Austin Animal Shelter available from Kaggle. The dataset had 37 variables and 29,421 rows. Preliminary exploratory data analysis showed 75,619 missing values. Each category was evaluated for type from a list of numerical, boolean, empty, time related, text, alphanumeric and unknown with a count provided for each. I added alphanumeric which deviated from the examples in our text. I found it useful for evaluating the “Animal id” variable.

Analysis was performed to determine a valid unique key. The “Animal Id” variable was evaluated and proved to be inadequate as the same animal was returned and adopted again on multiple occasions. The “Outcome Age Timestamp” variable was added which provided a valid unique key. However, seven rows had to be removed as they were complete copies of each other.

There were formatting issues with the “Outcome Month-Year” variable as most of its values had a “T” placed in between the data and time portion of the data. This was removed. For the human readable portion of the assignment “Outcome Age in Years” was rounded to two decimal places for practicality and readability. In this stage, the termcolor library was used to highlight the answers in red.

Fuzzy logic using the fuzzywuzzy library was performed to match duplicate answers for color. In this data set, problems were commonly created due to the position of an extra “/” in the data. Values that could readily be merged were returned in red. This would potentially reduce the count from 153 to 122. The library was also used to attempt to reduce the number of color possibilities. A new list was returned proposing which colors could potentially be combined based on similarity using a score of 75 or higher to match likeness. If transformed using this method, the number of unique colors would be reduced to 20 from the original count of 153.

In conclusion, there were several challenges that were readily addressed using commonly available Python libraries. Seven lines that were duplicated were removed for ongoing analysis. Two examples of fuzzy logic were provided. One example was created to address data entry concerns around redundant colors. The other provides a rational attempt at reducing color values.