Allison Forte

DSC 530 Final Project

June 4, 2022

Which animals are most likely to be adopted in Texas?

This project explored the relationship between age and time spent in a shelter in Texas. The data used from 2 sources recorded the intake and outcome for 166,000 animals from Oct 2013–Jan 2022. Data from the 2 datasets was merged before analysis.

The exploratory data analysis showed an interesting relationship between age at intake and time spent in shelter. My initial hypothesis was that older animals would stay longer but that was not the case. It should be noted though; the data did have some inaccuracies which could be indicative of the reliability of the whole data set. (For example, some animals were listed at negative 3 years old when arriving at the shelter.)

A limitation of this EDA was the data used. Much of the data was initially coded as categorical variables and so the entries were strings. In order to work with the data several variables needed to be coded into numbers for analysis. Age was broken from a string (2 years) into 2 columns, and sex was changed from a description to a number for each of 4 options. This could be continued to further breakdown outcome types and subtypes.

An interesting addition to the data set would be an evaluation of the animal’s health on intake. There was a designation of ‘sick’ that could be selected but if each animal was given a standardized score of say 1­–5, we could look closer at whether intake condition played a large role in length of time in the shelter or outcome type.

After completing this analysis, I think I could have been more thorough with how the data frames were merged. While I was able to match by animal ID, I think filtering out incorrect entries early on would have helped the analysis. The assumption that records are accurate is also always slightly problematic. Knowing more about how and why these records were kept could help us determine the validity of them and therefore, the validity of any results.

A challenge of this project was working with many non-numeric variables. The data set was certainly easier to read and interpret because it used words rather than codes, but this made the analysis more challenging. Thinking through how to convert variables into numeric entries in a more thorough manner early on could have helped improve this analysis.