An Analysis of Factors Influencing Pet Adoption

DSCI 415 Final Report

Matthew Barclay

12/16/2020

Introduction

Animal shelters across the United States receive nearly 6.5 million animals per year and are tasked with housing, feeding, and returning animals to owners or adopting these animals out. Invariably, with cost and size constraints many shelters must resort to euthanization to avoid overcrowding. According to the ASPCA, 1.5 million animal euthanizations take place in animal shelters. Reducing the time spent in a shelter by an animal and conversely increasing the adoptions of animals spending a large amount of time in the shelter may be a solution to reduce the volume of euthanizations.

Using shelter data to analyze the factors that influence the time spent in a shelter before adoption may offer insight to assist shelter staff. By understanding the factors which increase time spent in the shelter, staff may be able to identify animals at risk of long stays and focus more attention to them to reduce their time until adoption. The goal of this analysis is to report the most influential factors affecting time spent in a shelter.

All data were retrieved from the public database data.AustinTexas.gov which reports all animal intakes and outcomes for the Austin Animal Center since October, 2013. This shelter is a "No Kill" shelter and therefore animals leave the shelter by adoption, transfer to rescues, or returned to owners. Because the shelter does not engage in euthanizations, the data are not influenced by overcrowding and the time elapsed represents the true time until adoption for the animals. The data includes information on the intake type, intake condition, animal type, animal sex, outcome type, foster status, animal age, and the time spent in the shelter. The majority of this data are categorical, with animal age and time elapsed being the only numeric data.

The data will be analyzed using both multiple linear regression (MLR) and the Cox proportional hazards models. These models will be used to give insight on the factors influencing pet adoption.

Methods

Data Cleaning

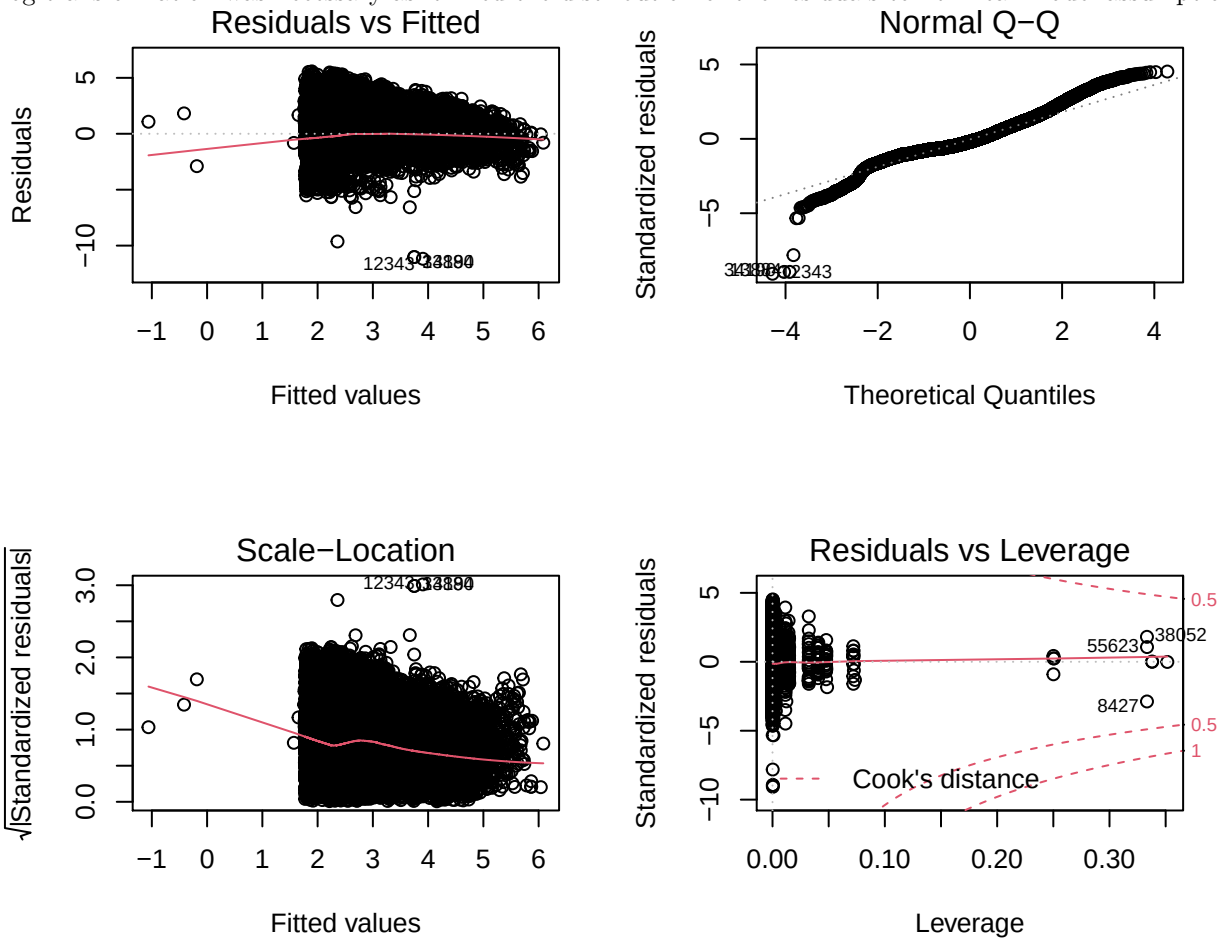
The raw data from the City of Austin is separated into two different datasets: Intakes and Outcomes. Therefore, the data was left joined via a unique Animal ID that is common between both datasets. The combined dataset included 122,000 entries, with the majority being animals returned to owners and some NA entries. These were filtered out to leave only adoptions in the dataset, shrinking it to 54,000 entries. The datasets did not include information on the time elapsed to adoption, so this was created by subtracting the date of adoption from the intake date to find the time in weeks spent in the shelter. Animal age was determined by subtracting the intake date from the animal's date of birth. The categorical variables were converted to factors to allow for use with R functions. Finally, birds, livestock, and "other" animals were removed from the dataset, as they accounted for less than 200 entries and skewed the regression values, leaving animal types of cats and dogs only. Breed and color data were also available, but there would be too many levels to the factors for it to make sense to use. The following table shows the structure of the cleaned dataset:

Variable	Type	Factor Levels/Info
Animal.ID	Character	Unique animal ID
Name	Character	Animal name if known
DateTime.x	Date	Intake date
Intake.Type	Factor	"Abandoned" "Euthanasia Request" "Owner Surrender" Public Assist" "Stray" "Wildlife"
Intake.Condition	Factor	"Aged" "Behavior" "Feral" "Injured" "Medical" "Normal" "Nursing" "Other" "Pregnant" "Sick"
Animal.Type	Factor	"Cat" "Dog"
Sex.upon.Intake	Factor	"Intact Female" "Intact Male" "Neutered Male" "Spayed Female" "Unknown"
Age.upon.Intake	Numeric	Animal age in weeks
Outcome.Subtype	Factor	"Normal" "Barn" "Foster" "Offsite"
DateTime.y	Date	Date of adoption
Time.Elapsed	Numeric	Time spent in the shelter in weeks (DateTime.y - DateTime.x)

Models

MLR

A multiple linear regression model was fit to explain the time spent in the shelter from the data collected on the animal. The response variable Time.Elapsed was highly right skewed and required a log transformation. The log transformation was necessary as it fixed the distribution of the residuals to fit linear model assumptions.



After log the transformations, the model much more closely follows the assumptions for a linear model.

Forward stepwise search was used to select the optimal linear model using both AIC and BIC, with both selecting the following model:

log(Time.Elapsed) ~ Outcome.Subtype + Intake.Type + Age.upon.Intake + Animal.Type.x + Intake.Condition + Sex.upon.Intake + Age.upon.Intake:Animal.Type.x

To test for multicollinearity, the variable inflation factor (VIF) was tested and all factors were below the reasonable level.

Proportional Hazards

In addition to the MLR model, a Cox proportional hazards model was fit, as it is a strong method for determining the influence a factor, or "risk", has on an event happening. As the dataset contains only adopted animals, it was not necessary to censor the data because all entries have the "event" of adoption. The proportional hazards model was selected via stepwise forward AIC. The stepwise algorithm selected the same factors as MLR:
Intake.Type + Intake.Condition + Animal.Type.x + Sex.upon.Intake + Age.upon.Intake + Outcome.Subtype + Animal.Type.x *Age.upon.Intake

Similar to MLR, the response variable is log transformed. While this model works well to highlight the effect of each factor, it may be inaccurate as the proportionality assumption, indicating that significant conclusions cannot be drawn using this model. As this report is focused only on suggesting possible effects, the significance of the model is not as important as its effects, and therefore should be used but with a grain of salt.

Mixed Effects Modelling

A mixed effects model was considered for use, but the dataset did not contain any usable information on sub-groupings and therefore was not possible to be used. If data on groupings such as zip code of adopters, adopter age group, or found location were available, it would be interesting to use a mixed effects model to explain more of the variance.

Results

MLR

Because the response variable only is log transformed, the coefficients must be exponentiated before interpretation. The formula used for exponentiation was $(exp(coefficient) - 1) * 100$. The interpretation for exponentiated coefficients is the percent increase or decrease in the response for a one unit increase in the independent variable. For categorical variables, the coefficient is simply the percent change in the response variable compared to the dropped (baseline) category if that category is true and has zero effect if false. The coefficients for the multiple linear regression are as follows:

Coefficient	Estimate	Std. Error	Pr(>|t|)
(Intercept)	423.4	19.96	9.76E-20
Outcome.SubtypeBarn	-26.64	104.7	0.6655
Outcome.SubtypeFoster	348.8	1.417	0
Outcome.SubtypeOffsite	27.06	6.735	0.0002389
Intake.TypeEuthanasia Request	47.42	42.31	0.2713
Intake.TypeOwner Surrender	-16.88	13.26	0.1376
Intake.TypePublic Assist	185.1	13.79	5.25E-16
Intake.TypeStray	29.33	13.21	0.03822
Age.upon.Intake	0.1599	0.007317	3.36E-105
Animal.Type.xDog	-36.5	1.339	7.54E-253
Intake.ConditionBehavior	199.5	87.86	0.08195
Intake.ConditionFeral	144	32.51	0.001529
Intake.ConditionInjured	253	14.54	1.61E-20
Intake.ConditionMedical	87.73	34.94	0.03557
Intake.ConditionNormal	109.3	14.18	2.54E-08
Intake.ConditionNursing	281.9	14.89	4.98E-22
Intake.ConditionOther	237.9	21.8	6.72E-10
Intake.ConditionPregnant	345.3	29.43	7.10E-09
Intake.ConditionSick	178.3	14.96	2.14E-13
Sex.upon.IntakeIntact Male	3.302	1.251	0.00897
Sex.upon.IntakeNeutered Male	4.844	1.924	0.01306
Sex.upon.IntakeSpayed Female	1.781	1.952	0.3612
Sex.upon.IntakeUnknown	-94.18	103.8	6.45E-05
Age.upon.Intake:Animal.Type.xDog	0.05241	0.008667	1.49E-09

Observations	Residual Std. Error	R^2	Adjusted R^2
54102	1.233	0.2714	0.2711

For the significant coefficients, the interpretation of the coefficients can be used to determine the expected change in stay length for an animal given its characteristics.

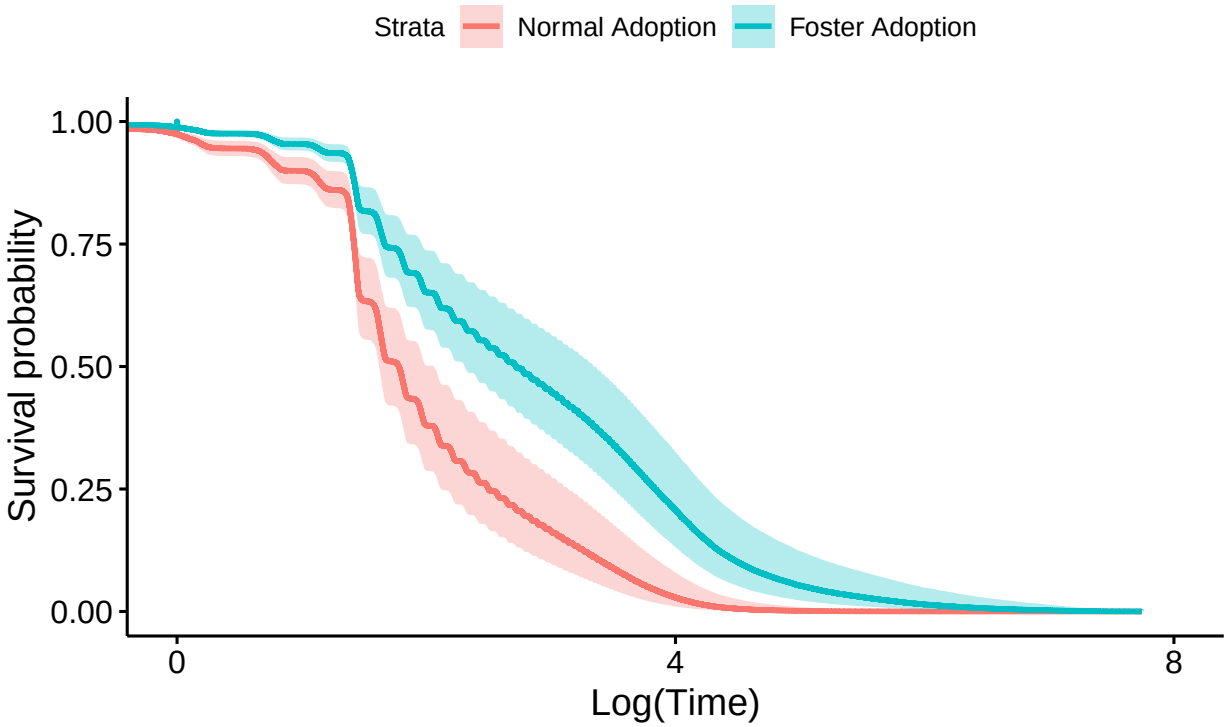
Proportional Hazards

The proportional hazards coefficients can be interpreted as having no effect if the coefficient is 1, reducing the hazard if the coefficient is less than 1, and increasing the hazard if the coefficient is greater than 1. In this context, the event is adoption and therefore a shorter time to event is preferable. Therefore a coefficient greater than 1 decreases the time in the shelter and a coefficient less than 1 increases the time in the shelter. The coefficients are as follows:

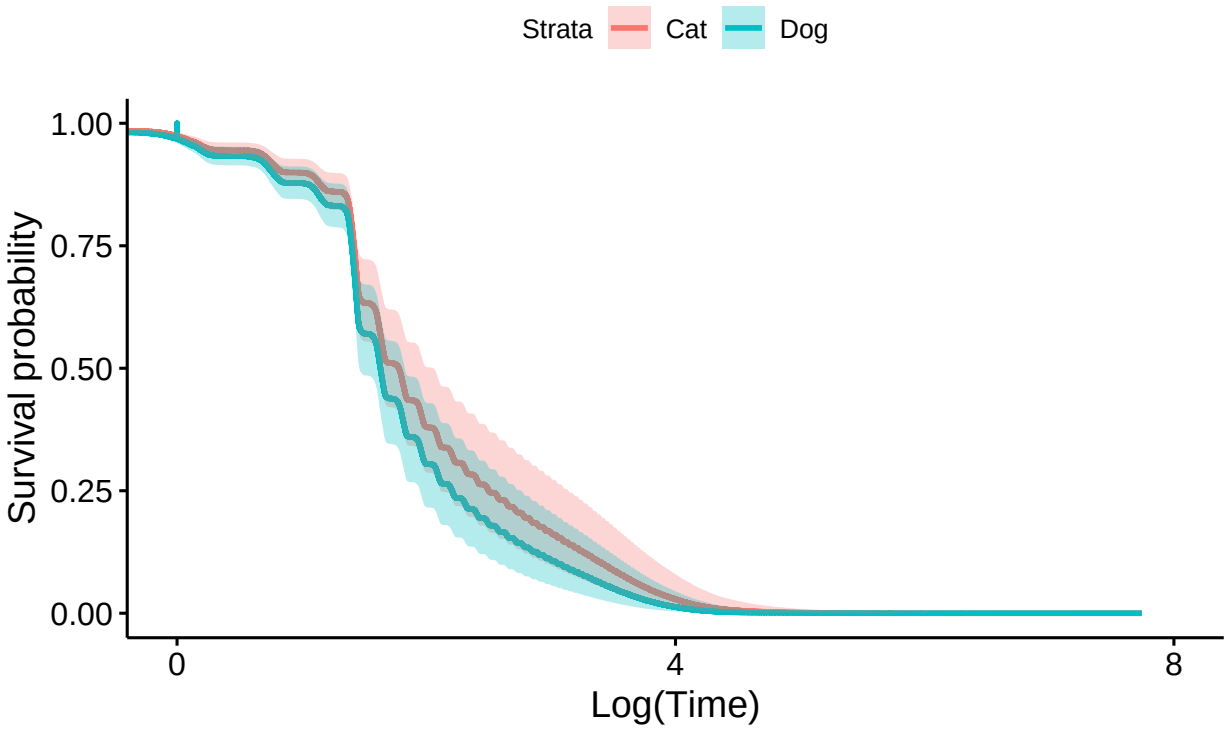
	coef	se(coef)	Pr(>|z|)
Outcome.SubtypeBarn	0.5131	0.5824	0.3783
Outcome.SubtypeFoster	-0.8145	0.01133	0
Outcome.SubtypeOffsite	-0.1252	0.05288	0.01789
Intake.TypeEuthanasia Request	-0.2318	0.2863	0.4182
Intake.TypeOwner Surrender	0.003312	0.101	0.9738
Intake.TypePublic Assist	-0.754	0.1049	6.542e-13
Intake.TypeStray	-0.2236	0.1007	0.02638
Age.upon.Intake	-0.0007424	6.05e-05	1.292e-34
Animal.Type.xDog	0.2795	0.01112	2.607e-139
Intake.ConditionBehavior	-0.3291	0.5115	0.52

	coef	se(coef)	Pr(>|z|)
Intake.ConditionFeral	-0.5697	0.2286	0.0127
Intake.ConditionInjured	-0.7589	0.11	5.332e-12
Intake.ConditionMedical	-0.2934	0.2431	0.2274
Intake.ConditionNormal	-0.4443	0.1074	3.519e-05
Intake.ConditionNursing	-0.8549	0.1124	2.909e-14
Intake.ConditionOther	-0.7489	0.1599	2.822e-06
Intake.ConditionPregnant	-1.039	0.2092	6.8e-07
Intake.ConditionSick	-0.6004	0.113	1.077e-07
Sex.upon.IntakeIntact Male	-0.02865	0.01009	0.004531
Sex.upon.IntakeNeutered Male	-0.1588	0.01531	3.158e-25
Sex.upon.IntakeSpayed Female	-0.152	0.0156	1.899e-22
Sex.upon.IntakeUnknown	2.86	0.5775	7.369e-07
Age.upon.Intake:Animal.Type.xDog	-0.0008301	7.495e-05	1.662e-28

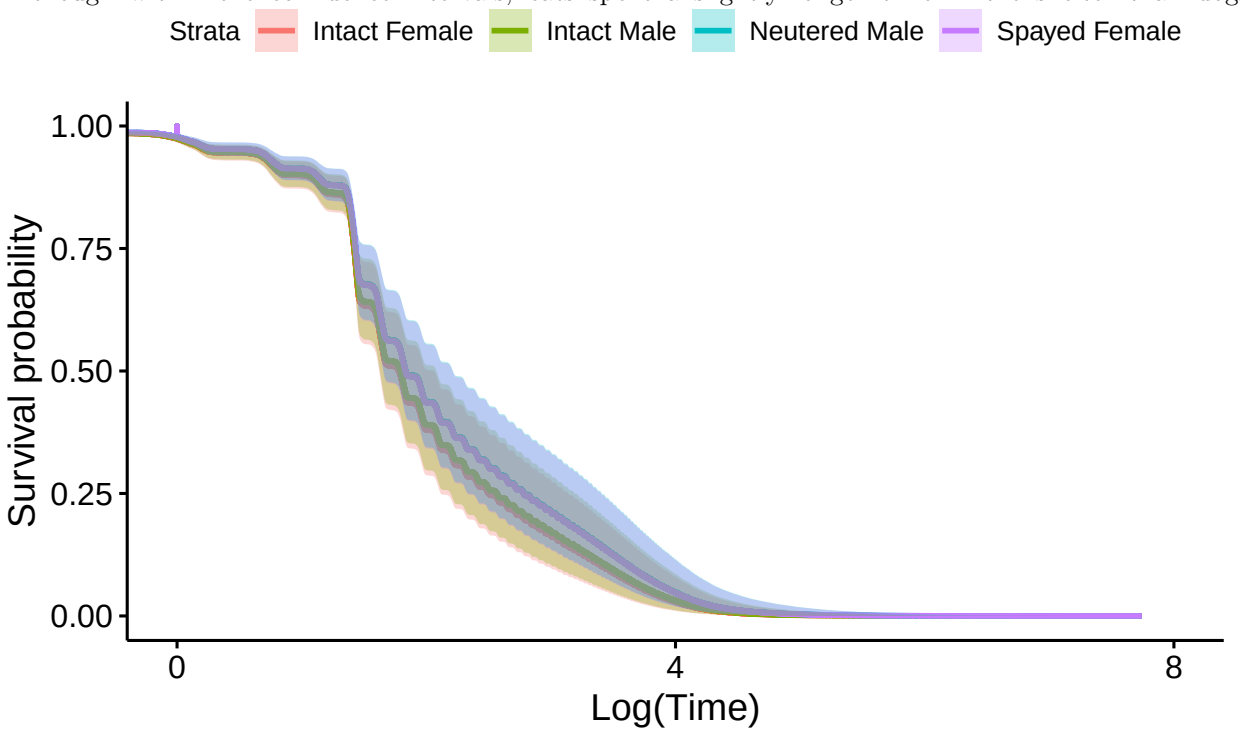
It is possible to analyze the differences within a categorical variable in a plot, where reaching a survival of zero means every animal has been adopted.



This plot shows that fostered animals are adopted out at a slower rate than ordinary adoptions and therefore spend more time in the shelter.



Although within the confidence intervals, cats spent a slightly longer time in the shelter than dogs.



This plot shows that no sex has a significant impact on time to adoption. All three of these plots are confirmed by the MLR coefficients; foster animals spend significantly longer, dogs spend slightly less time, and sex is not a significant indicator.

Discussion

The R^2_{adj} value of 0.2711 for the MLR model indicates that the factors account for only 27% of the variance in the time elapsed. While this may be considered low, the available data and highly subjective nature of choosing a pet for adoption mean that this number is relatively high given these circumstances. Both models can be used to identify the effects that influence time to adoption. As verified by both models, as animal age increases, so does time in the shelter. While this is typically accepted by shelters, proving the relationship between animal age and time elapsed further encourages shelters to adopt out older animals, as younger ones are adopted much faster. Additionally, both models confirm that the sex of the animal has almost no effect on time spent in the shelter.

By analyzing the coefficients from both models, it is possible to identify similarities which indicate a factor is significant in practice. Therefore, shelter staff should use this to focus adoptions on animals which have one or more factors that indicate long stay times.

This data is from only one shelter and therefore may not be representative for other shelters. However, the differing preferences of adopters is a possible future point of research that may explain more about the time spent in the shelter.

Finally, it would be helpful to have more data on each animal, such as weight, size, personality, and ability to get along with other animals. Although these data were not collected by the shelter, it may allow the models to explain more of the variance in the time spent in the shelter until adoption.