Exploring the Factors Influencing Pet Adoption

Matthew Barclay

11/20/2020

Contents

Problem Statement

According to the ASPCA, nearly 6.5 million dogs and cats enter animal shelters each year, and of these about 1.5 million must be euthanized, mostly due to shelter overcrowding[1]. Therefore, increasing adoptions or decreasing the time an animal spends in the shelter will free up space and decrease the amount of euthanizations. In order to assist shelters on increasing adoptions and decreasing time in shelter, it is useful to determine the factors that influence the time in shelters. Therefore, a model can be fit to predict the time spent in a shelter by an animal using its characteristics. The goal of this report is to find a model that explains the time an animal spends in a shelter until it is adopted.

Introduction

Background Knowledge

According to many sources, older animals tend to be adopted less than young animals[2]. This can be due to many different preferences of future pet owners which cause them to prefer the youngest available adoptions. Therefore, the age variable is particularly of interest to determine if age really does increase the stay of an animal in a shelter. Additionally, as cats and dogs are adopted at a nearly equal rate[1], the animal type will be of interest to determine if there is a difference in expected shelter time for cats and dogs. Finally, both male and female animals are adopted at similar rates, but it will be interesting to see if the sex of an animal plays a part in its expected time to adoption.

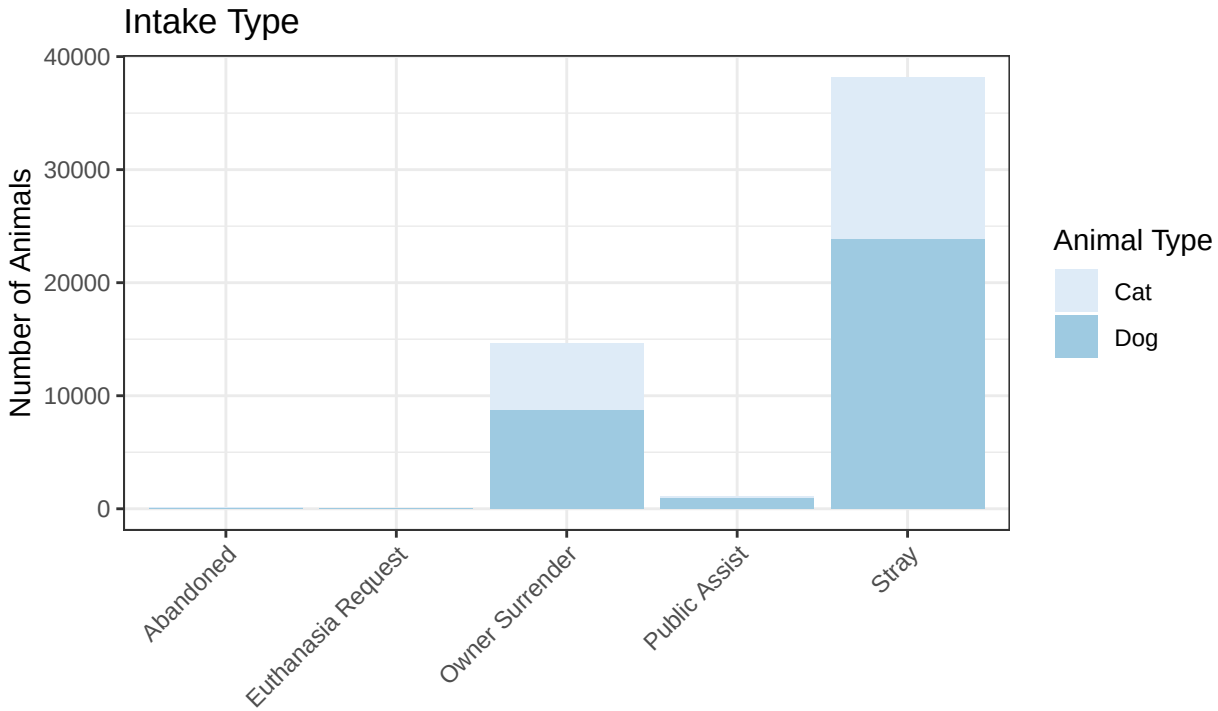
Analysis plan

The data will be cleaned and transformed to be ready for regression. Only adopted animals will be analyzed from the dataset. A linear model will be constructed useng stepwise fit by AIC and BIC values. The coefficients will be analyzed through p-value and confidence intervals. Diagnostic plots will confirm regression assumptions. R^2_{adj} will be used to indicate model fit.

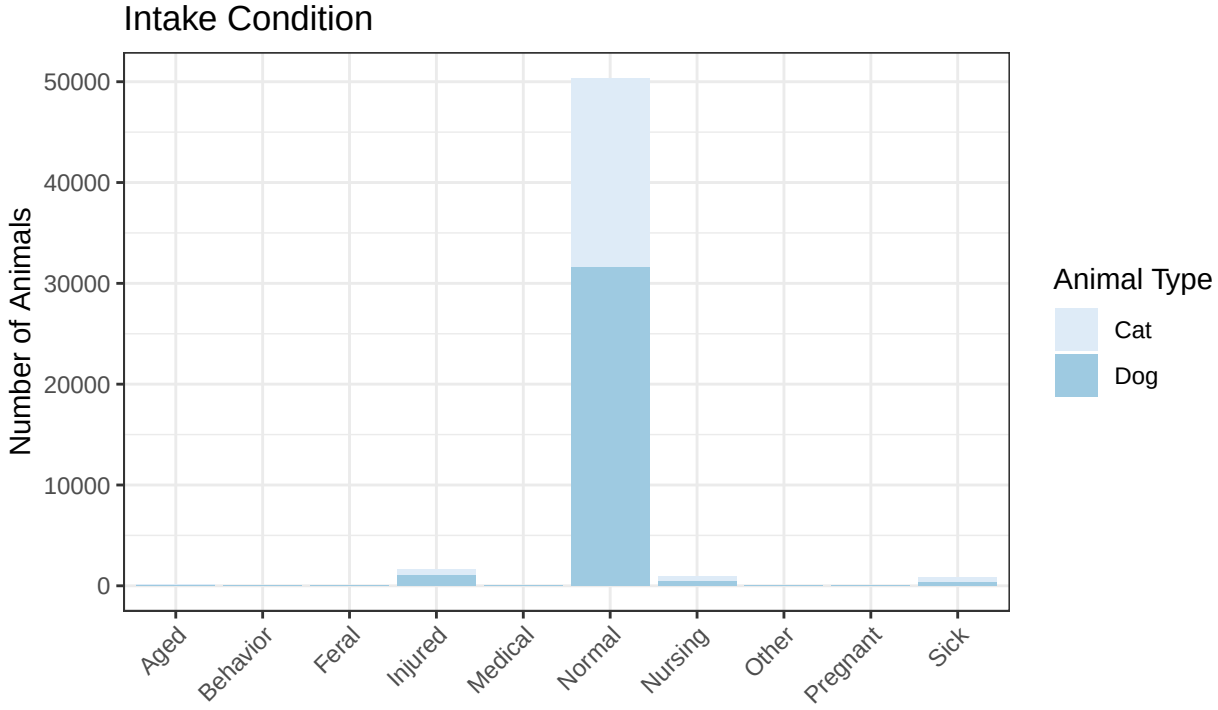
Data Collection

The data comes from two datasets: Austin Animal Center Intakes[3] and Austin Animal Center Outcomes[4]. These datasets contain all intakes and outcomes from October 2013 to the present day at the Austin Animal Center, an animal shelter in Austin Texas. The raw data contains information on over 120,000 animal intakes and outcomes. The datasets were left joined by Animal ID, an ID number assigned at intake for all animals. The data was then cleaned to filter out non-adopted animals and format categorical variables to factors which is usable in regression.

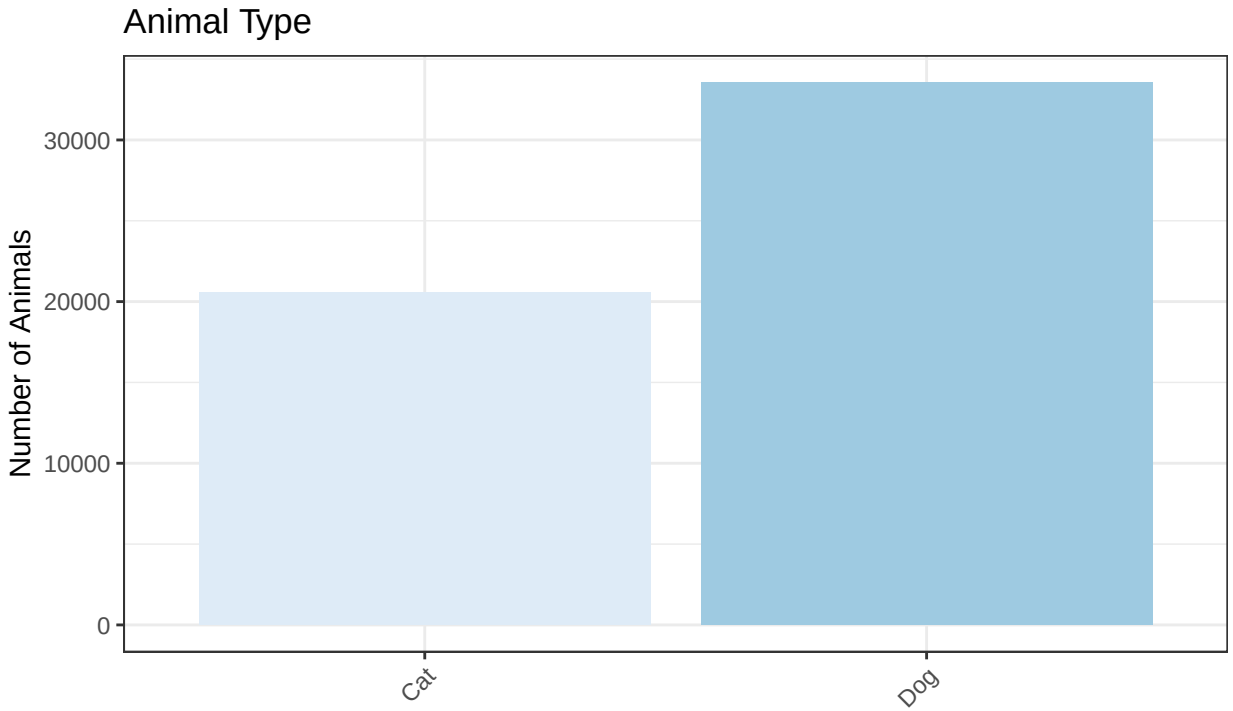
Variable	Type	Factor Levels/Info
Animal.ID	Character	Unique animal ID
Name	Character	Animal name if known
DateTime.x	Date	Intake date
Intake.Type	Factor	"Abandoned" "Euthanasia Request" "Owner Surrender" Public Assist" "Stray" "Wildlife"
Intake.Condition	Factor	"Aged" "Behavior" "Feral" "Injured" "Medical" "Normal" "Nursing" "Other" "Pregnant" "Sick"
Animal.Type	Factor	"Bird" "Cat" "Dog" "Livestock" "Other"
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)



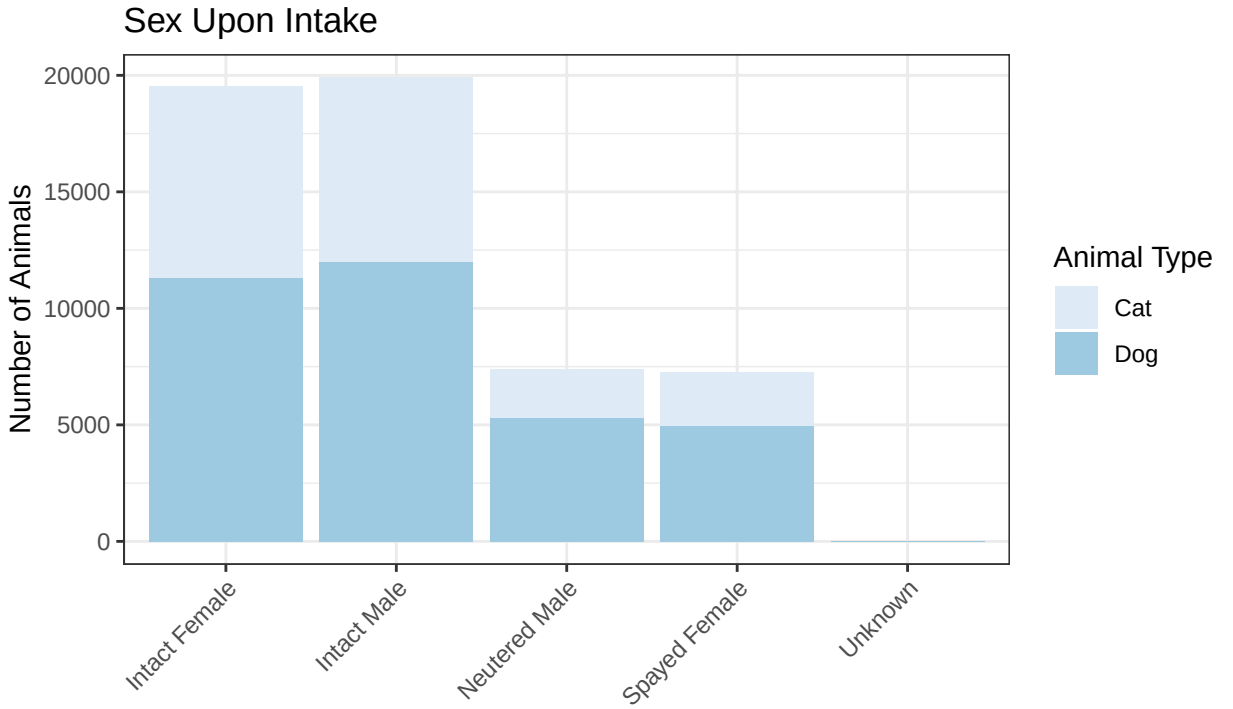
The majority of intakes are strays or owner surrenders, with the distribution remaining similar among animal types.



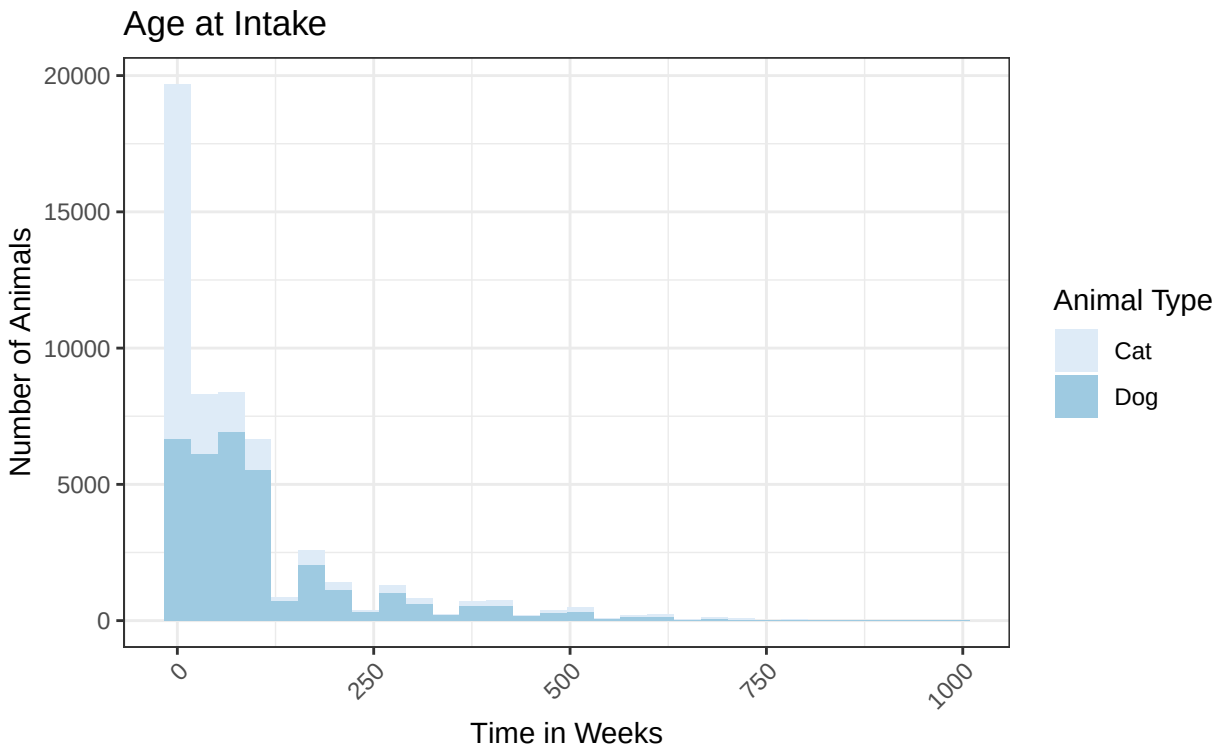
The vast majority of intakes are normal, with the distribution remaining similar among animal types.



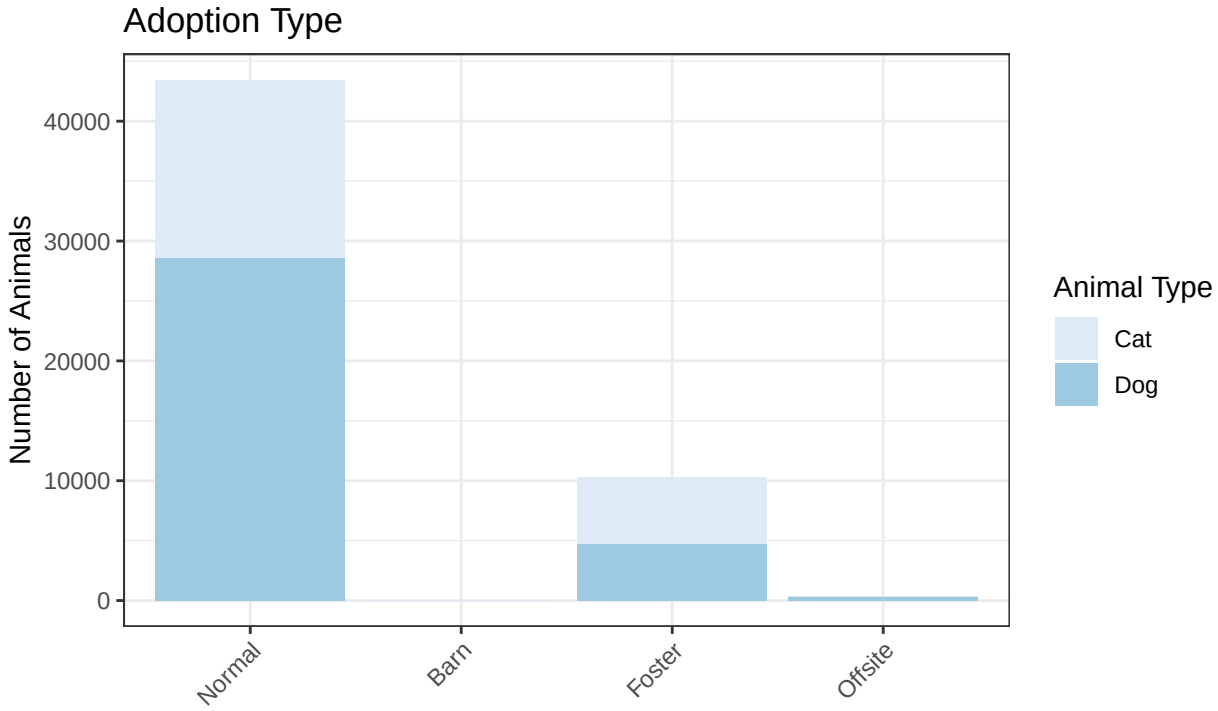
There were 33,576 dog intakes and 20,601 cat intakes. 197 bird intakes, 8 Livestock intakes, and 500 "other" intakes were not taken into account by this model for simplicity.



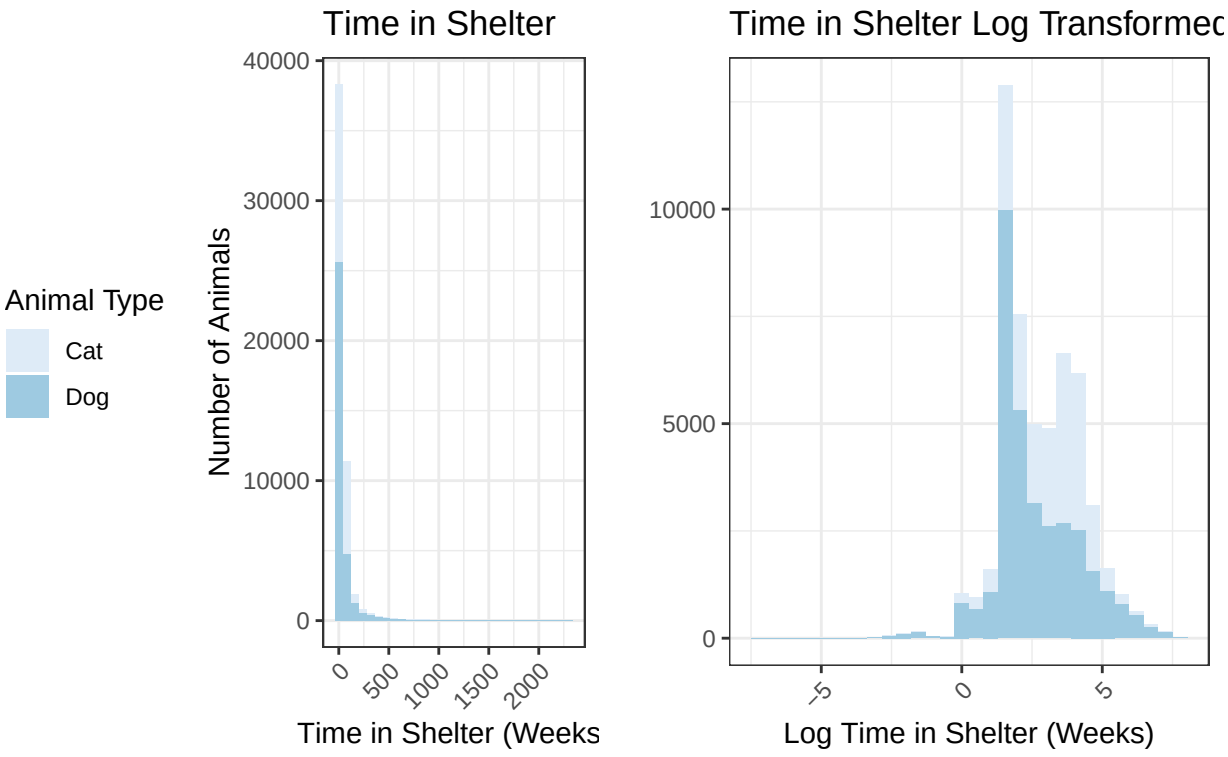
The majority of both cats and dogs are not neutered. The female to male ratio is nearly evenly split for both neutered and intact animals, as well as across the different species.



The majority of cats enter the shelter at a very young age, while dogs appear to have a larger age range when entering the shelter.



The majority of adoptions are standard adoptions. Cats appear to be the majority of animals being fostered despite being a smaller population than dogs.



The response variable time spent in the shelter is highly skewed with the vast majority of animals spending less than 100 weeks and very few spending over 2,000 weeks. Therefore, a log transformation was applied to the response variable to reduce the skewness and center the values around the mean.

Data Analysis and Results

Model Selection

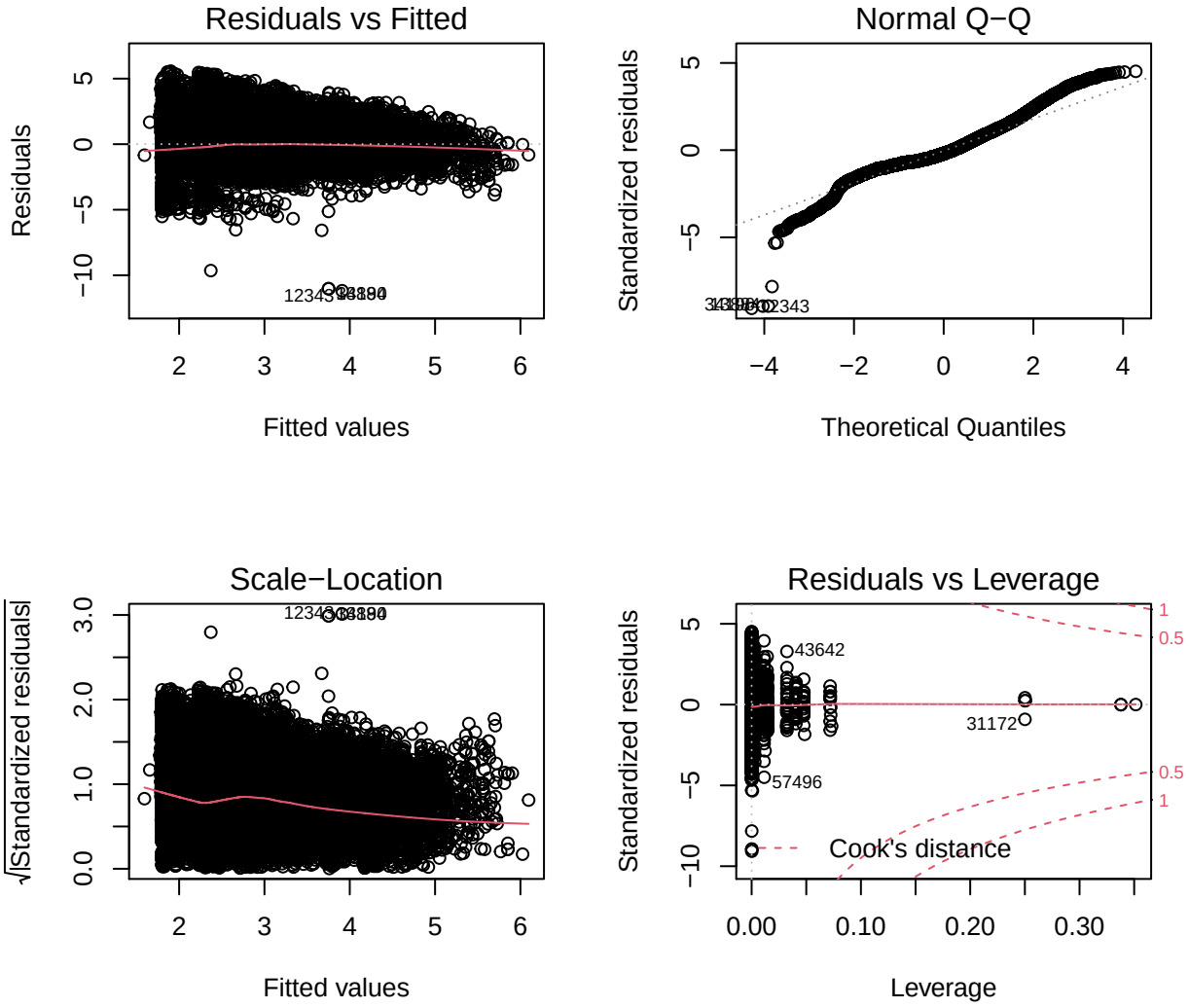
Stepwise forward selection was used to build the model using the step() function in R. The possible effects were Intake.Type, Intake.Condition, Animal.Type.x, Sex.upon.Intake, Age.upon.Intake, Outcome.Subtype, and an interaction term between Animal.Type.x and Age.upon.Intake. The stepwise function determined the optimal model by AIC. The process was repeated using BIC and resulted in the same model. The final model included all possible effects except Sex.upon.Intake with an initial AIC of 39749 and final AIC of 22854. The final equation is:

$$Y = \beta_0 + \beta_{1-5} Intake.Type + \beta_{6-14} Intake.Condition + \beta_{15} Animal.Type.x +$$

$$\beta_{16} Age.upon.Intake + \beta_{17-19} Outcome.Subtype + \beta_{20} Animal.Type.x * Age.upon.Intake + \epsilon$$

Where: $\epsilon \sim N(0, \sigma^2)$

Model Diagnostics



The model appears to fit all necessary assumptions for linear regression:

1. There appears to be a linear relationship between the regressors and response variables
2. The variance appears to be mostly evenly distributed for any x, as observed in the Residuals vs Fitted Values plot
3. The data are independent, as each row is a unique adoption
4. The response variable is mostly normally distributed, as observed by the Normal Q-Q plot, with deviations occurring only at the extreme tails

Interpreting Coefficients

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[5]. 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 exponentiated coefficients are as follows:

	Estimate	Std. Error	p-value
(Intercept)	428.9	19.94	5.414e-20
Intake.Type Baseline:Abandoned	0	0	0
Intake.TypeEuthanasia Request	47.86	42.31	0.2677
Intake.TypeOwner Surrender	-16.46	13.25	0.1485
Intake.TypePublic Assist	186.1	13.79	4.196e-16
Intake.TypeStray	29.41	13.22	0.0378
Intake.Condition Baseline: Aged	0	0	0
Intake.ConditionBehavior	205.2	87.88	0.07682
Intake.ConditionFeral	145.9	32.51	0.001396
Intake.ConditionInjured	255.6	14.54	9.358e-21
Intake.ConditionMedical	89.16	34.94	0.03343
Intake.ConditionNormal	110.5	14.18	1.978e-08
Intake.ConditionNursing	283.4	14.89	3.788e-22
Intake.ConditionOther	240.9	21.8	5.017e-10
Intake.ConditionPregnant	339	29.43	9.829e-09
Intake.ConditionSick	180	14.96	1.519e-13
Animal.Type.x Baseline: Cat	0	0	0
Animal.Type.xDog	-36.43	1.337	2.127e-252
Age.upon.Intake	0.1622	0.006887	6.702e-122
Outcome.Subtype Baseline: Normal	0	0	0
Outcome.SubtypeBarn	-27.06	104.8	0.6598
Outcome.SubtypeFoster	348.4	1.415	0
Outcome.SubtypeOffsite	27.02	6.736	0.0002435
Animal.Type.xDog:Age.upon.Intake	0.05261	0.008654	1.229e-09

Observations	Residual Std. Error	R^2	Adjusted R^2
54102	1.233	0.271	0.2708

The Confidence Intervals for each exponentiated coefficient for the variables of interest are as follows:

	2.5 %	97.5 %
Animal.Type.xDog	-38.06	-34.75
Age.upon.Intake	0.1487	0.1758
Animal.Type.xDog:Age.upon.Intake	0.03564	0.06958

The confidence intervals for the coefficients of interest are relatively tight and therefore offer strong insight. The interval for the coefficient for dogs does not include zero, so dogs appear to spend a significantly shorter time in the shelter than cats do. The confidence interval for age is relatively tight and indicates that increased age does increase time spent in the shelter. For the interaction term, the small, positive slope indicates that for dogs, the age slope increases slightly, indicating that older dogs will spend a longer time in the shelter

than older cats.

Summary and Discussion

The constructed model can be interpreted to determine the factors which increase or decrease an animal's time spent in the shelter. Animal sex was not selected in model selection, indicating that the sex of the animal does not influence its time spent in the shelter. The model shows that on average, as a cat's age increases by one week, its expecter stay in the shelter increases by 0.16% with all other factors held constant, while for dogs the expected stay increases by 0.21% for each week of age.. Compared to cats, dogs spend on average 36% shorter time in the shelter. Compared to normal adoptions, animals in foster care will spend 348% longer there on average. Each coefficient can be interpreted in this way. All of these interpretations can be used by shelter employees to identify animals most at risk of long term stays. With this information, an animal with a very high expected stay can be given more staff attention and greater priority for adoption to attempt to decrease its time in the shelter, while animals that are expected to spend very little time in the shelter require less attention. In theory, by decreasing time spent in the shelter, a shelter will be able to decrease euthanizations and therefore both cut costs and save animal lives- both very positive outcomes. The R^2_{adj} value of 0.2708 indicates that the model accounts for 27.08% of the variance in the response. While this may seem low, with the subjectivity of adoptions and preferences of adopters, this makes sense to observe a low R^2_{adj}.

Appendix

Executable R code: ‘markdown{r}

intake <- read.csv(“Austin_Animal_Center_Intakes.csv”) outcome <- read.csv(“Austin_Animal_Center_Outcomes.csv”)

adopted <- left_join(intake, outcome, by = “Animal.ID”) %>% filter(Outcome.Type == “Adoption”, !is.na(Animal.ID)) %>% mutate(DateTime.x = as.POSIXct(DateTime.x, format = “%m/%d/%Y %I:%M:%S %p”), DateTime.y = as.POSIXct(DateTime.y, format = “%m/%d/%Y %I:%M:%S %p”)) %>% mutate(Time.Elapsed = as.numeric(difftime(DateTime.y, DateTime.x, units = “days”))) %>% filter(Time.Elapsed >= 0) %>% mutate(Date.of.Birth = as.POSIXct(Date.of.Birth, format = “%m/%d/%Y”)) %>% mutate(Age.upon.Intake = as.numeric(difftime(DateTime.x, Date.of.Birth, units = “weeks”))) %>% filter(Age.upon.Intake >=0) %>% filter(Animal.Type.x == “Dog” | Animal.Type.x == “Cat”) #Long and complex, but essentially joins the two data frames, converts dates to date objects, removes negative times, and filters out to see only adopted animals

adopted <- adopted[!duplicated(adopted[,1:4]),] #Removes duplicated rows as a result of left join for animals that have had multiple adoptions, removes the incorrect rows

adopted[,c(6,7, 8, 9, 11, 12, 17, 18, 19, 20, 23)] <- lapply(adopted[,c(6,7, 8, 9, 11, 12, 17, 18, 19, 20, 23)], factor)

#Convert categorical rows to factors to be used in regression

#Data: #Animal.ID- Assigned ID #Name.x- Animal name #DateTime.x- Date and time of arrival to shelter #MonthYear.x- Unformatted DateTime.x #Found.Location- Address of origin of animal #Intake.Type- Type of intake (categorical) – “Abandoned” “Euthanasia Request” “Owner Surrender” “Public Assist” “Stray” “Wildlife” #Intake.Condition- Animal condition at arrival (categorical) – “Aged” “Behavior” “Feral” “Injured” “Medical” “Normal” “Nursing” “Other” “Pregnant” “Sick”
#Animal.Type.x- Animal type (categorical) – “Bird” “Cat” “Dog” “Livestock” “Other”
#Sex.upon.Intake- Animal sex (categorical) – “Intact Female” “Intact Male” “Neutered Male” “Spayed Female” “Unknown”
#Age.upon.Intake- Animal age at arrival in weeks (numerical (weeks old)) #Breed.x- Animal Breed #Color.x- Animal color #Name.y- Animal name #DateTime.y- Date and time of adoption #MonthYear.y- Unformatted DateTime.y #Date.of.Birth- Animal DOB #Outcome.Type- If an animal was adopted, uniform Adopted for this dataset #Outcome.Subtype- If the animal was foster adopted #AnimalType.y- Animal type (categorical) #Sex.upon.Outcome- Animal sex (categorical) #Age.upon.Outcome- Animal age at adoption #Breed.y- Animal breed (categorical) #Color.y- Animal color #Time.Elapsed- Time spent in the shelter in weeks (numerical (weeks elapsed)) attach(adopted) #model_all <- lm(Time.Elapsed ~ Intake.Type + Intake.Condition + Animal.Type.x + Sex.upon.Intake + Age.upon.Intake, data = adopted) #summary(model_all) #plot(model_all)

#fwd selection fwd <- lm(log(Time.Elapsed)~1, data = adopted) #AIC step(fwd, scope = list(lower = ~1, upper = ~Intake.Type + Intake.Condition + Animal.Type.x + Sex.upon.Intake + Age.upon.Intake + Outcome.Subtype + Animal.Type.x *Age.upon.Intake), direction = “forward”, data = adopted) #BIC n <- length(fwd$residuals) step(fwd, scope = list(lower = ~1, upper = ~Intake.Type + Intake.Condition + Animal.Type.x + Sex.upon.Intake + Age.upon.Intake + Outcome.Subtype + Animal.Type.x* Age.upon.Intake), direction = “forward”, data = adopted, k = log(n))

Full model with response variable log transformed

model_log <- lm(log(Time.Elapsed) ~ Intake.Type + Intake.Condition + Animal.Type.x + Age.upon.Intake + Outcome.Subtype + Animal.Type.x *Age.upon.Intake, data = adopted) #summary(model_log) #extractAIC(model_log) #plot(model_log)

#Transform Sex variable to Male/Female only, drops neutered dimension # Adj R^2 before: 0.2717 , AIC: 23197.35 #adopted$Sex.upon.Intake <- ifelse(adopted$Sex.upon.Intake == “Intact Male” | adopted$Sex.upon.Intake == “Neutered Male”, “Male”, “Female”) # Adj. R^2 after: 0.2697, AIC:

#Backward Selection #summary(step(model_log, direction = “backward”, data = adopted))

#Exponentiate coefficients #(exp(model_log$coefficients) - 1) * 100

#Plots: ggplot(adopted) + aes(Intake.Type, fill = Animal.Type.x) + geom_bar() + theme_bw() + labs(title = “Intake Type”, y = “Number of Animals”, x = "“) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + guides(fill=guide_legend(title=”Animal Type“)) + scale_fill_brewer() ggplot(adopted) + aes(Intake.Condition, fill = Animal.Type.x) + geom_bar() + theme_bw() + labs(title =”Intake Condition“, y =”Number of Animals“, x =”“) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + guides(fill=guide_legend(title=”Animal Type")) + scale_fill_brewer()

ggplot(adopted) + aes(Animal.Type.x, fill = Animal.Type.x) + geom_bar() + theme_bw() + labs(title = “Animal Type”, y = “Number of Animals”, x = "“) + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position =”none“) + guides(fill=guide_legend(title=”Animal Type“)) + scale_fill_brewer() ggplot(adopted) + aes(Sex.upon.Intake, fill = Animal.Type.x) + geom_bar() + theme_bw() + labs(title =”Sex Upon Intake“, y =”Number of Animals“, x =”“) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + guides(fill=guide_legend(title=”Animal Type“)) + scale_fill_brewer() ggplot(adopted) + aes(Age.upon.Intake, fill = Animal.Type.x) + geom_histogram() + theme_bw() + labs(title =”Age at Intake“, y =”Number of Animals“, x =”Time in Weeks“) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + guides(fill=guide_legend(title=”Animal Type“)) + scale_fill_brewer() ggplot(adopted) + aes(Outcome.Subtype, fill = Animal.Type.x) + geom_bar() + theme_bw() + labs(title =”Adoption Type“, y =”Number of Animals“, x =”“) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + guides(fill=guide_legend(title=”Animal Type“)) + scale_fill_brewer() + scale_x_discrete(labels = c(”Normal“,”Barn“,”Foster“,”Offsite"))

grid.arrange(ggplot(adopted) + aes(Time.Elapsed, fill = Animal.Type.x) + geom_histogram() + theme_bw() + labs(title = “Time in Shelter”, y = “Number of Animals”, x = “Time in Shelter (Weeks)”) + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = “left”) + guides(fill=guide_legend(title=“Animal Type”)) + scale_fill_brewer(), ggplot(adopted) + aes(log(Time.Elapsed), fill = Animal.Type.x) + geom_histogram() + theme_bw() + labs(title = “Time in Shelter Log Transformed”, y = "“, x =”Log Time in Shelter (Weeks)“) + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position =”none“) + guides(fill=guide_legend(title=”Animal Type")) + scale_fill_brewer(), ncol = 2)

par(mfrow= c(2,2)) plot(model_log)

“‘

References

1. Pet Statistics. *ASPCA*.

2. What Kinds of Pets Get Adopted? *Priceonomics*.

3. Services, A. Austin Animal Center Intakes: Austin Animal Center Intakes. 3KB (2015) doi:10.26000/025.000002.

4. Services, A. Austin Animal Center Outcomes: Austin Animal Center Outcomes. 3KB (2016) doi:10.26000/025.000001.

5. Interpreting Log Transformations in a Linear Model University of Virginia Library Research Data Services + Sciences.

Self-Reflection

I spent probably 18 hours total on this project, including data selection, data cleaning, model fitting, and troubleshooting. I learned a little more on using the lm() function for large MLR, as well as applying the step() model for model selection. I had issues in trying to log transform the age variable and was never able to do it, as I always received an error, no matter what I tried. I think this dataset was difficult, as having

only one numeric variable limited the possibility of interaction terms or regressor transformations. I am happy with my project and I think it turned out very good.

Thank you to Dr. Li for an interesting and informative semester.