Animal Shelter

April Yang, Haoyue Yang, Xin Jin, Yaxi Lou, Yu Zhang

Spring 2017

Contents

1 Abstract

Our research mainly focused on finding the factors that can best predict the adoption rate of intaken dogs in animals'shelters. Besides, we also examined whether it is true that Pit bulls are more likely get euthanized due to its aggressive characteristics. To do the analysis, we built several models including logistic regressions and naive bayes classifier. From our models, we concluded that variables "Intake Type", "Time Spent in Animal Shelter", "Age Range"and "Size"can predict the outcome type of a dog better than other variables. Also, from our model, we concluded that Pit bulls are more likely being euthanized than non Pit bulls.

2 Introduction

According to American Society for the Prevention of Cruelty to Animals (ASPCA), approximately 6.5 million animals enter animal shelters nationwide each year, and there are approximately 1.5 million shelter animals are euthanized. In 2015, over 6,000 animals were killed in Los Angeles city shelters simply because they could not find safe places to call home. Meanwhile, Austin celebrated its five-year anniversary of being Americas largest no-kill city at the end of fiscal year 2015, saving more than 90 percent of its homeless animals since 2011.
As a group of animal lovers, we wanted to explore the conditions of sheltered animals in Austin, Texas, hoping to gain some helpful insight from Austin's success in order to improve the living condition of LA's homeless animals in the future. To get more precise results, we decided to focus on dogs, the most popular choice for domestic pets.
We downloaded two datasets from Austin governments open data portal, which contain information of all animal intakes and outcomes for fiscal year 2015 (October 1st 2014 to September 30th 2015).
This research focuses on finding the variables that best predict the opportunity of a dog being adopted, and exploring the exact effect of related predictors on the outcome. Besides, we also tested the stigma of Pit bulls that they are dangerous and less likely to be adopted.

2.1 Research Questions

Main research question
We first built a model to predict if a dog would be adopted or die with given conditions. We subset the original datasets and created a response variable, "if_died", for the first model which has two levels, 1, which means a dog is predicted to die, and 0, which means a dog is predicted to be adopted. In order to test the accuracy of the model, we split the subset data into a training dataset and a testing dataset. By randomly selecting observations, 70% of the data was used to train the model and 30% was used to test it. We constructed our model using the training data, and calculated the misclassification rate of the model using the testing data.

Does the stigma of Pit bulls that they are dangerous make them less likely to be adopted?
First, we grouped the "dog breed" variable into 2, "Pit bulls" and "non Pit bulls". We ran logistic regression to find significant variables. The results are: age, outcome type, sterilization, time spent in the shelter. Then we made multiple plots of outcome types to see the difference between Pit bulls dogs and non Pit bulls dogs.
Finally we ran the logistic model between Pit bulls and non Pit bulls with response variable being outcome type. Since outcome type has three levels, and one of them, "died" is a very small portion in the outcome, we dropped died level. Therefore, our logistic model have only two level: 0, other or non-adopted and 1, adopted.

2.2 Data Preparaton

Table 1: Raw Datasets

Data Name	Discription	Dimension
Austin_Animal_Center_FY15_ Intakes_Updated_Hourly_	all animal intakes for Fiscal Year 2015 (October 1st 2014 to September 30th 2015) provided by Austin Animal Center	18627 observations, 12 variables
Austin_Animal_Center_FY15_ Outcomes_Updated_Hourly_	all animal outcomes for Fiscal Year 2015 (October 1st 2014 to September 30th 2015) provided by Austin Animal Center	18447 observations, 11 variable

Table 2: Final Cleaned Up Dataset

Data Name	Discription	Dimension
dogs	A dataset that takes variables from both of the original datasets. After merging, cleaning and recoding, the final dataset contains 12 variables: "Breed", "Intake_Type", "Intake_Condition", "Age_upon_Intake", "Outcome_Subtype", "Age_upon_Outcome", "Time_Diff", "Size", "Sterilization", "Outcome", "Age Range", "Time_Spent"	7794 observations, 12 variables

Table 3: Variables Selected

Name	Type	Definition	Level	Description
Breed	Categorical	The breed information of intaken dogs	"Airedale Terrier Mix ", "Airedale Terrier/Irish Terrier", , "Yorkshire Terrier/Toy Poodle"(685 levels in total)	A factor with 685 levels indicating the breeds of intaken dogs
Intake Type	Categorical	The reasons why the dogs were taken in	"Euthanasia Request" "Owner Surrender" "Public Assist" "Stray"	

Table 4: Variables Created/ Recoded

Name	Type	Definition	Level	Description
Time Spent*	Categorical	Length of time in the shelter before adoption.	1 (< 5 days), 2 (6-11 days),3 (12 + days)	This variable in the original dataset ranges from hours to years which is too wide. In order to lower the number of levels, we divided them into 3 groups, 1 (< 5 days), 2(6-11 days), and 3 (12+ days)
Sterilization**	Categorical	whether a dog has been neutered or spayed	Y (Spayed or neutered dogs), N (Otherwise)	In the original dataset, variable "gender"has 5 levels indicating whether a dog is male or female, spayed/neutered or not, and NA's. We grouped by their sterilization condition; "Y"means it's spayed or neutered and "N"means it's intacted. We assigned "Y"or "N"to the NA's randomly by proportion of Y or N in other observations
Age Range***	Categorical	the age range of dogs	Baby(0-1 year), Teen(1-3 year), Adult(3 + year)	The range of this variable in the original dataset is too wide, from days to years. In order to lower the number of levels, we divided them into 3 groups, Baby (< 1 year), Teen (1-3 years), and Adult (3+ years)
Outcome Type****	Categorical	The outcomes of the intaken dogs	"Adopted", "Died", "Other"	The "Outcome Type" in the original dataset has 8 levels. To avoid an over-complicated model, we grouped outcomes we were not really interested in together as "Other"

Size	Categorical	the size of dogs (determined by average weight)	S (Terrier, Chihuahua, Dachshund, Poodle),M (Corgi, Schnauzer, Border Collie, Pit Bull, Husky), L (Shepherd, Retriever, Catahoula, Boxer mix)	In the raw data, there isn't a variable indicating the sizes of intaken dogs, which could be a very important factor that affects rate of adoption. In order to determine their sizes, we used "average weight" as the criterion. A breed with average weight less than 25 was classified as "Small Dogs", between 25 and 50 as "Medium Dogs", and larger than 50 as "Large Dogs". For the mixed bred dogs, we took both of the breeds into account. If both breeds belong to the same size group, we put them into the corresponded size group; otherwise we classified them as "medium"
Intake Condition	Categorical	The condition of intaken dogs	"Injured", "Normal ", "Other "	A categorical variable with three levels. In the original datasets, there are 9 different types of intake conditions. In order to reduce the levels of this variable, we grouped some categories with only a few observations to "Other"

***Note I**

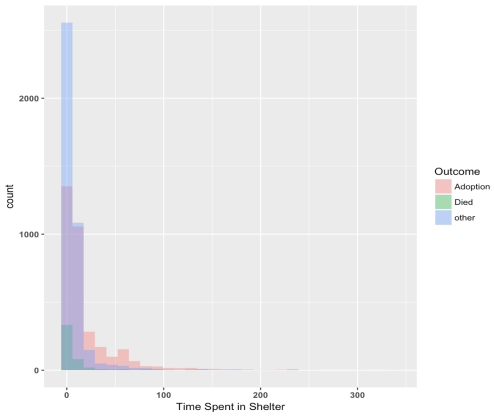

Figure 1: time spent in shelter

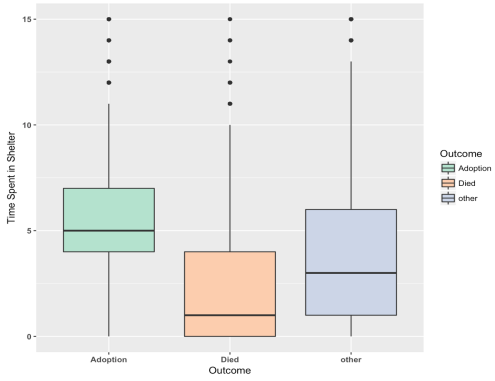

Figure 2: timespent vs outcome

This plot shows that regardless of the outcome type, time spent in the animal shelter is very right skewed. Knowing that using this variable directly would cause errors for our further analysis, we recoded this variable as categorical variable. We divided the time as

following levels: 1 (stay in the shelter less than 5 days), 2 (stay in the shelter from 5 to 11 days), 3 (otherwise)

****Note II**

Table 5: Sterilization Frequency

	Sterilization	Frequency
1	Yes	5426
2	No	2368

The original data set has 4 levels of genders which are Neutered Male, Intact Female, Spayed Female, and Intact Male. For simplicity, we decided to group them differently. In our new data set, Y means sterilized and N means not sterilized. For NA's, we assigned "Y" or "N" to the N's randomly by proportion of "Y" or "N" in other observations.

*****Note III**

Table 6: AgeRange Frequency

	AgeRange	Frequency
less than 1 year	Baby	2425
1-3 year	Teen	3889
3+ year	Adults	1480

In our original data set, there are too many units of age including days, weeks, months, and years. Knowing that too many levels of a categorical variable will lead to large error rate of our model for further analysis, so we recoded the age variable as following: If the dog is less than 1 year old, we classified it as baby. If the dog is in the range of 1 to 3 years, then it is be recode as teen, otherwise adults.

******Note IV**

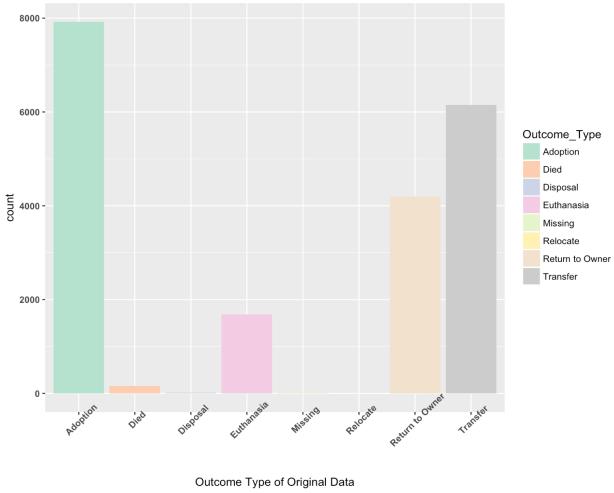

Figure 3: outcome type of original data

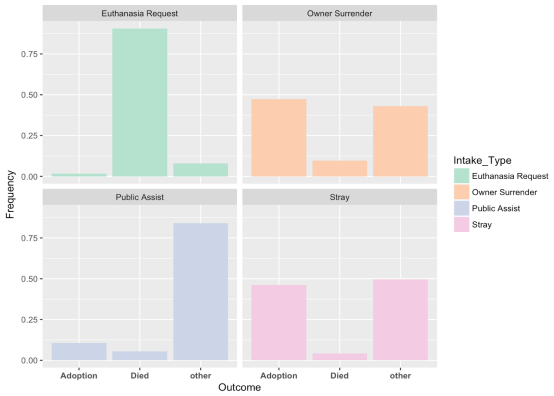

Figure 4: outcome type of new data

As this table shown, there are few issues in the response variable, Outcome Type, in the original data set. The first issue is that there are way many levels for one predictor than we actually need. Having too many levels can cause the complexity of the model, which can lead to higher error rate of the model. The other issue is that the proportion of each level is not even, particularly "Missing", "Relocation", and "Disposal". In order to make our model more efficient, we merged "Euthanasia" in to "Died". Since we were more interested in why certain dogs are being adopted or killed, we combined "Transfer" and "Return to owner" together as our new level "Other".

3 Results

3.1 Main research question

3.1.1 Logistic Regression

Adoption versus Died

As described above, we ran the logistic model using the training data. The R output shows that most of the parameters chosen are statistically significant at 99% confidence level. Specifically, Intake Type, Intake Condition, time spent in the shelter, age range (baby) of dogs, and size of dogs are significant at 99% confidence level. Age range (teen) is significant at 90% confidence level. Sterilization condition is not statistically significant at 90% confidence level.

We then tested if the model was accurate. We calculated the misclassification rate based on our testing data, and the R output shows that the misclassification rate is 8.9%, which indicates that the model predicts the response variable well.

Finally, we ran the logistic model with the full subset again. The statistical significance of predictors are the same as those in our training model.

Table 7: Logistic Model: Adoption versus Died

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.5331	1.1896	-2.97	0.0030
Intake_TypeOwner Surrender	1.3310	1.1837	1.12	0.2608
Intake_TypePublic Assist	-0.9031	1.1894	-0.76	0.4477
Intake_TypeStray	1.2445	1.1826	1.05	0.2926
Intake_ConditionNormal	0.9678	0.1177	8.22	0.0000
Intake_ConditionOther	0.0229	0.1883	0.12	0.9032
time_spent2	0.9320	0.0687	13.58	0.0000
time_spent3	1.6869	0.0659	25.60	0.0000
ragebaby	1.2221	0.0805	15.19	0.0000
rageteen	0.4306	0.0738	5.83	0.0000
SizeM	-0.3153	0.0694	-4.55	0.0000
SizeS	0.3415	0.0614	5.56	0.0000
sterilizeY	0.0567	0.0564	1.01	0.3147

Interpretations

Based on the coefficients of the predictors, we have the following conclusions:

1. Keeping all others constant, the odds of a dog with intake type owner surrender to be died is 99% less than the odds of a dog with intake type Euthanasia request.

2. Keeping all others constant, the odds of a public assistant dog to be died is 95% less than the odds of a dog with intake type Euthanasia request.

3. Keeping all others constant, the odds of a stray dog to be died is 99.5% less than the odds of a dog with intake type Euthanasia request.

4. Keeping all others constant, the odds of a dog with normal intake condition to be died is 97% less than the odds of an injured dog.

5. Keeping all others constant, the odds of a dog with other intake condition to be died is 71% less than the odds of an injured dog.

6. Keeping all others constant, the odds of a small dog to be died is 62.52% less than the odds of a large dog.

7. Keeping all others constant, the odds of a medium dog to be died is 158% greater than the odds of a large dog.

8. Keeping all others constant, the odds of a baby dog (age less than a year) to be died is 88% less than the odds of an adult dog (age greater than three years).

9. Keeping all others constant, the odds of a dog which stays in the shelter for 6 to 11 days to be died is 72.9% less than the odds of a dog which stays in the shelter within 5 days.

10. Keeping all others constant, the odds of a dog which stays in the shelter for more than 11 days to be died is 88% less than the odds of a dog which stays in the shelter within 5 days.

Adoption versus Others

We again ran the logistic model using R. The R output shows that only several predictors are statistically significant at 99.9% confidence level. Specifically, Intake

Condition(normal), time spent in the shelter, the age range of dogs, and the size of dogs are statistically significant at 99.9% confidence level. However, intake type, intake condition(other), sterilization condition are not statistically significant at 95% confidence level.
The misclassification rate of this model is around 37.5%, which indicates this model can be improved. To further improve the model, we considered to take more predictors from external sources.
Finally, we ran the same model on the full subset. The statistical significance of predictors are the same as our training model.

Table 8: Logistic Model: Adoption versus Others

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.5436	1.1232	6.72	0.0000
Intake_TypeOwner Surrender	-4.5657	1.0966	-4.16	0.0000
Intake_TypePublic Assist	-3.0104	1.1208	-2.69	0.0072
Intake_TypeStray	-5.3307	1.0939	-4.87	0.0000
Intake_ConditionNormal	-3.5815	0.1774	-20.19	0.0000
Intake_ConditionOther	-1.2497	0.3887	-3.22	0.0013
time_spent2	-1.3068	0.1949	-6.70	0.0000
time_spent3	-2.1478	0.1716	-12.51	0.0000
ragebaby	-2.1437	0.2165	-9.90	0.0000
rageteen	-0.3034	0.1678	-1.81	0.0706
SizeM	0.9383	0.1637	5.73	0.0000
SizeS	-0.9891	0.1768	-5.59	0.0000
sterilizeY	-0.1637	0.1420	-1.15	0.2492

Interpretations
Based on the coefficients of the predictors, we have the following conclusions:
1. Keeping all else constant, the odds of a dog with normal intake condition to be adopted is 164% higher than the odds of a dog with injured intake condition to be adopted.
2. Keeping all others constant, the odds of a dog which stays in the shelter for 6 to 11 days to be adopted is 154% greater than the odds of a dog which stays in the shelter for within 5 days.
3. Keeping all others constant, the odds of a dog which stays in the shelter for more than 11 days to be adopted is 440% greater than the odds of a dog which stays in the shelter within 5 days.
4. Keeping all else constant, the odds of a baby dog (age less than a year) to be adopted is 240% higher than the odds of an adult dog (age greater than three years) to be adopted.
5. Keeping all else constant, the odds of a teenager dog (age between a year and three years) to be adopted is 54.3% higher than the odds of an adult dog (age greater than three years) to be adopted.

6. Keeping all else constant, the odds of a small dog to be adopted is 40.6% higher than the odds of a large dog.

3.2 Bayes Classifier

After grouping the outcomes of dogs into three categories, Adoption, Died and Other, we were able to build a prediction model of outcome based on Bayesian Classification. We used six predictors including "Intake Type", "Intake Condition", "Size", "Sterilization", "Age Range"and "Time Spent", then we obtain the table:

Table 9: naive Bayes Classifier

	Adoption	**Died**	**Other**
A-priori probabilities:	0.4287	0.0642	0.5070
Conndditional Probabilities			
Intake_Type			
Enthanasia Request	0.000	0.1200	0.0013
Owner Surrender	0.1948	0.3175	0.1528
Public Assist	0.0217	0.0750	0.1452
Stray	0.7831	0.4875	0.7007
Intake_Condition			
Injured	0.0441	0.3900	0.0664
Normal	0.9256	0.5625	0.8947
Other	0.0303	0.0475	0.0389
Size			
L	0.3366	0.2600	0.3214
M	0.2274	0.4325	0.3018
S	0.4361	0.3075	0.3768
sterilize			
N	0.2812	0.3050	0.3189
Y	0.7188	0.6950	0.6811
AgeRange			
adults	0.1298	0.2375	0.2290
baby	0.3983	0.1425	0.2534
teen	0.4720	0.6200	0.5176
time_spent			
1	0.4159	0.7325	0.6197
2	0.2143	0.0775	0.1955
3	0.3699	0.1900	0.1847

The a-priori probabilities are prior probabilities in Bayes' theorem which tells us how frequently each level of class occurs in the training dataset. The rationale underlying the

prior probability is that if a level is rare, then it is unlikely appear in the test dataset. In other words, the prediction of an outcome is not only influenced by the predictors, but also by the prevalence of the outcome. Conditional probabilities are calculated for each variable. It is actually the likelihood table as shown in the table above. For example, the likelihood of adoption given "Intake Type" being Euthanasia Request P(Adoption—Euthanasia Request) equals to 0.0003739716. This model can be applied for outcome prediction with these six predictors.

Table 10: Confusion Matrix

Prediction	Adoption	Died	Other
Adoption	952	28	513
Died	4	39	15
Other	523	1480	

The training error rate is calculated by the fraction of the sum of off-diagonal entries and the total number, which is (513+39+523)/(952+28+513+4+39+15+523+116+1276)=31.18%.

Table 11: Model Accuracy

ACC	65.41%		
	Adoption	Died	Other
Precision	63.76%	67.24%	66.63%
TPR	64.37%	21.31%	70.73%
F1-score	64.06%	32.37%	68.62%

We are also able to get the confusion matrix and coefficients of ACC, precision and F1-score, which are measurements of the accuracy of the model. From the results, we get a training error rate of 31.18% and ACC of 65.41%. The precision rate of outcome types "adoption" and "other" are between 63% to 70%. The precision of the dog will die is quite low, which is only about 21.31%. The reason behind the low accuracy is that "Died" only takes up a small proportion among all the outcomes in our dataset. For further study, we can increase the accuracy by collecting more data.

3.3 Pit bulls vs Non Pit bulls

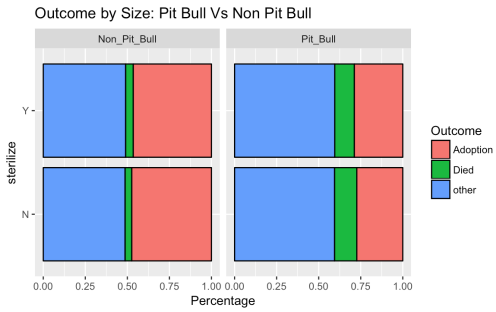

Figure 5:

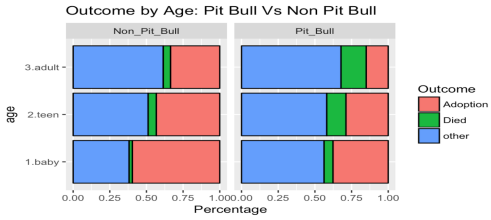

Figure 6:

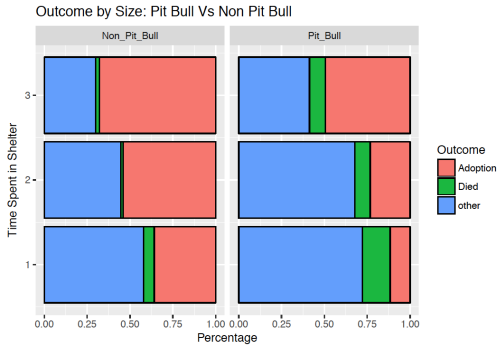

Figure 7:

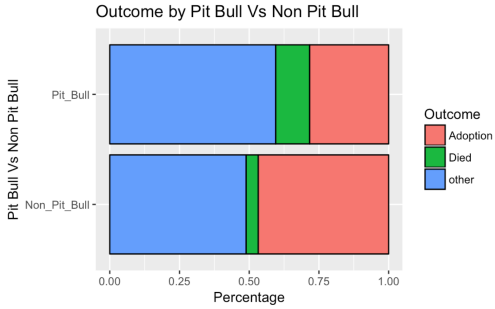

Figure 8:

Table 12: Ratio

Ratio of Outcome Type by Breed			
	Adopt	Died	Other
Non Pit Bull	47%	4%	49%
Pit Bull	28%	12%	60%

Table 13: Outcome SubType

Outcome Sub Type of 198 Died Pit Bull				
	Aggressive	Foster	Rabies	Suffering
Pit Bull	120(61%)	0	12(6%)	65(33%)

In the plots, we can see that different sterilization conditions do not have significant effect on the outcome in terms of dog breed. While dogs' ages differ, their time spent in the shelter have different ratio of outcome type.
In general, Pit bulls' adoption rate is notably lower than that of other dogs and their death rate is much higher than that of other dogs. While Pit pulls'adoption rate is almost twice lower than that of other dogs, their death rate is three times higher than that of other dogs.

Moreover, among the 198 died Pit bulls, 120 of them, which is about 61% were labeled as aggressive.

Table 14: logistic model Pit Bul vs Non Pit Bull

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1730	0.0690	-2.51	0.0122
BreedPit_Bull	-1.0883	0.0898	-12.12	0.0000
SizeM	0.2755	0.0851	3.25	0.0012
SizeS	0.3349	0.0598	5.60	0.0000
sterilizeY	0.0273	0.0551	0.050	0.6199
rage2.teen	-0.7808	0.0584	-13.37	0.0000
rage3.adult	-1.3183	0.0776	-16.99	0.0000
time_spent2	0.7722	0.0656	11.77	0.0000
time_spent3	1.6354	0.0648	25.25	0.0000

Table 15: coefficients

Exponentiate the coefficients from model	
(Intercept)	0.84
BreedPit_Bull	0.34
SizeM	1.32
SizeS	1.40
sterilizeY	1.03
rage2.teen	0.46
rage3.adult	0.27
time_spent2	2.16
time_spent3	5.13

To examine more closely on the difference between Pit bulls and other dogs, we ran a logistic regression. From the table, "Exponentiate the coefficient from model", we can see that the odds of Pit bulls compare to non Pit bulls to be adopted is 66% less.

4 Conclusion and Reflection

4.1 Conclusion

We constructed two logistic models. The first one is to predict if a dog would be adopted or died, and the significant predictors are "Intake Type", "Intake Condition", "Time Spent in Shelter", "Age Range (baby) of dogs", and "Size of Dogs". The second model is to predict if a dog would be adopted or have other outcomes, and the significant predictors

are “Intake Type (normal)”, “Time Spent in Shelter”, “Age Range of Dogs”, and “Size of Dogs”.
While the death rate of Pit bulls is three times higher than that of other dogs, its adoption rate is almost half of the other dogs. In addition, in our logistic model, the odds of Pit bulls being adopted is 66% less than other dogs.

4.2 Reflection

We created the size variable by ourselves based on the dog breeds. We used average weight as our classification criterion. It is not the most efficient way of classifying the size because the size of dog varies even in the same breed. For the mixed breeds, the variation is even less predictable. Yet with limited recourse, that is the best we can do.
With limited time, we are not able to compare the model in different years to see if the trend or pattern is similar. In the future, we will take datasets from other years into account to improve our models and results.

References

[1] *"Austin government website."*
https://data.austintexas.gov/Government/Austin-Animal-Center-FY15-Outcomes-Updated-Hourly-/fb53-k8de
https://data.austintexas.gov/Government/Austin-Animal-Center-FY15-Intakes-Updated-Hourly-/hjeh-idye

[2] *"Dog Groups by Size"*
http://onlydogbreeds.com/groups/size.html

[3] https://www.austinchronicle.com/news/2016-12-23/five-years-of-no-kill-in-austin/