Finding Shelter Dogs a Home

Kit Hui

**Table of Contents**

1. **Abstract. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**
2. **Introduction. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3**
3. **Data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**
4. **Data Wrangling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4**

4a. Merging the Dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

4b. Cleaning the Dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1. **Exploratory Analysis and Statistical Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6**

5a. Breed. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

5b. Gender. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

5c. Age. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

5d. Color. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

5e. Intake Type. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

5f. Health. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

5g. Time Dependency. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

5h. Summary of Findings. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17

1. **Machine Learning. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17**

6a. Machine Learning Models. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .18

6b. Choosing the Right Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

1. **Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19**
2. **Statistical Methods Used. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19**

7a. Bootstrap Proportion Test. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19

7b. Bootstrap Permutation Test. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

7b. z-Proportion Test. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .20

1. **Supporting Figures/Tables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .20**
2. **Abstract**

1 in 5 animals at shelters are euthanized according to American Society for Prevention of Cruelty to Animals. A method to mitigate this inhumane crisis is to identify root cause by use of visual analysis and machine learning to characterize attributes leading to mass euthanization and predict at risk animals. The purpose of this project is to create a universal pipeline for all shelters and all animal types to minimize euthanization and instead transfer at risk animals to location where they can be safely adopted to loving family. To limit the scope and dive deeper into understanding the behavior of outcome, only dogs would be analyzed. Four datasets from Louisville, KN, Austin, TX, Dallas, TX, and Sonoma, CA were analyzed. The machine learning was done on the four datasets separately owing to a distinct data collecting behavior and uncommon features. At risk animals could be identified with up to 87% f1-score in Sonoma.

1. **Introduction**

According to The American Society for the Prevention of Cruelty to Animals, US animal shelters takes in approximately 6.5 million companion animals nationwide every year and 1.5 million shelter animals are euthanized every year. This means 1 in 5 animals entering shelters never leave! The public has always had compassion for companion animals. Animals have integrated into many as a part of their daily lives. They act as companions during emotionally difficult times and provide a sense of joy in day to day life for regular folks. Furthermore, dogs have been assimilated into many industries trained as support dogs to provide services from medical alerts for diabetic patients to drug detection on suspicious cargo crossing the border. With the long history between human and animals, the public has felt the need provide moral and sensible support for helpless animals. This project would provide awareness to animal loving individuals and inspire some to support animal shelters and other animal rights organization.

Therefore, I propose to investigate the use machine learning to decrease the rate of euthanization by creating a smart transferring system between pet adoption/foster agencies to improve outcomes for shelter animals. To achieve this the following questions would be explored. What constitutes the decision to such high euthanize animals? Are there certain key features that predispose the shelter animals to euthanization? What are the key characteristics of entering animals that lead to easy adoption? Are these characteristics different by region of the United States? This project would attempt to shed light to these questions and gaining understanding of the facing moral problem. The project would also limit its scope to dogs as to provide a more in-depth analysis of one type of animal and screen for patterns that would hopefully provide a general insight into other animals.

The discovery can help animal shelters identify high risk sheltered animals and provide extra aids to improve outcomes. This could remedy long sheltered times, freeing up sheltered spaces for in-needed animals and reallocating resources to other areas of the shelter. Financially speaking this would provide monetary remediation for animal shelters by reducing operating expenses in this area and provide funding to improve conditions in the shelter or other parts of the organization. Ultimately this would help to provide a closer step to their mission of providing care and rehabilitation to animals.

1. **Data**

According to the American Humane, animal shelters and other care agencies are not required to keep data on animals, hence the data available on the web are difficult to acquire and must be scoured to find. Currently six datasets from different regions of the US were found dating as far back as 2003 with feature vectors such as ‘type of animal’, ‘breed’, ‘color’, ‘outcome’, ‘entering date’, etc. However, the datasets were mostly incomplete with many important features missing. Only Dallas, Austin, Sonoma, and Louisville datasets have enough important features that can make insightful project.

<https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238>

<https://www.dallasopendata.com/City-Services/Dallas-Animal-Shelter-Data/7h2m-3um5>

<https://data.louisvilleky.gov/dataset/animal-service-intake-and-outcome>

<https://data.sonomacounty.ca.gov/Government/Animal-Shelter-Intake-and-Outcome/924a-vesw>

However, the datasets still have some important features that are still missing. For example, Austin did not have intake date and do not have animal gender. These are fundamental identifying characteristics of a shelter animal that would reduce effectivity of the machine learning. Also, even if the features are the same, since no universal data collecting method is made on national level, same feature may contain different information. For example, Austin labels their gender as intact male, neutered male, spayed female and intact female whereas for other locations, male and female are the only distinct categories.

1. **Data Wrangling**

The biggest challenge to the data wrangling portion of this project was effectively merging the datasets. Since there is no universal standard in organizing and collecting shelter data, these datasets were likely collected for the purpose of record keeping instead of data science purposes. Merging the datasets resulted in abundance of missing data and important features. The data wrangling process took take a 2-step approach.

1. Merging the datasets

a. dropped unarguably unimportant features

b. universalized different names of common features

c. feature engineered dataset

2. Cleaning the merged dataset

a. cleaned dataset by removing uncommon labeling

b. feature engineered dataset

**4a. Data Wrangling: Merging the Dataset**

For Austin, ‘animal ID’, ‘Name’, ‘MonthYear’, ‘Age upon Outcome’, and ‘Outcome Subtype’ were removed. ‘Animal ID’, ‘Name’, and ‘MonthYear’ were random features that do not correlate to outcome by any means. ‘Age’ upon Outcome was removed because the feature contains the ‘Date of Birth’ column already and the column rounded to whole years. The exact age could be extracted again from the ‘Data of Birth’ feature. ‘Outcome Subtype’ is a subset of the ‘Outcome Type’ column and although ‘Outcome Subtype’ can provide direct information on ‘Outcome Type’ (the target variable), it would be illogical in practical settings.

'AnimalID','IntakeReason','IntakeAsilomarStatus', 'ReproductiveStatusAtIntake','OutcomeSubtype', 'OutcomeReason', 'OutcomeInternalStatus', and 'ReproductiveStatusAtOutcome' were dropped from the Louisville dataset. ‘AnimalID’, ‘OutcomeSubtype, 'ReproductiveStatusAtIntake', and 'ReproductiveStatusAtOutcome' were removed by the same reasons were mentioned in prior instances. 'OutcomeReason', 'OutcomeInternalStatus', 'IntakeReason' were removed for incomplete data, containing less than half of the total population. 'IntakeAsilomarStatus' (health condition at intake) was removed because the evolving information was captured by ‘OutcomeAsilomarStatus’. The ‘Color’ and ‘Breed’ features were also created by merging the ‘PrimaryBreed’ and ‘SecondaryBreed’ columns and ‘PrimaryColor’ and ‘SecondaryBreed’ columns respectively.

In the Dallas dataset, 'Animal ID', 'Kennel Number', 'Tag Type', 'Kennel Status', 'Activity Number', 'Activity Sequence', 'Source ID', 'Census Tract', 'Council District’, ‘Reason', 'Staff ID', 'Intake Time', 'Outcome Time', 'Due Out', 'Intake Condition', 'Hold Request', 'Outcome Time', 'Receipt Number', 'Impound Number', 'Service Request Number', 'Animal Origin', 'Additional Information', 'Month', and 'Year' were removed. ‘Intake Condition’, ‘Month’, and ‘Year’ were removed because the information could be captured by ‘Outcome Condition’ and ‘Outcome Date’. All other information were numbers generated to identify the shelter dog or had incomplete data.

For Sonoma Dataset, 'Name', 'Size', 'Impound Number', 'Kennel Number', 'Animal ID', 'Days in Shelter', 'Intake Condition', 'Outcome Subtype', 'Intake Jurisdiction', 'Outcome Jurisdiction', 'Outcome Zip Code', 'Location', 'Count' were dropped from the dataset. Reasons for elimination were associated with reasons mentioned prior.

All the datasets had names of their feature changed to be better understood and normalized for merging. The data set was combined by concatenating common features.

**4b. Data Wrangling: Cleaning the Dataset**

‘Outcome Type’ was the target variable and the cleaning process was dealt carefully. Any confusion with the meaning was categorized in the ‘other’ category. Adoption and Foster were merged to ‘Adoption/Foster’ because the dogs are being taken care of. All other types were easily identified with corresponding outcome types.

The ‘Health’ column was cleaned by grouping the categorization into healthy, untreatable, treatable and contagious treatable. Because shelters may manage animals with contagious disease differently, the distinction was kept offsetting possible differences to outcome type.

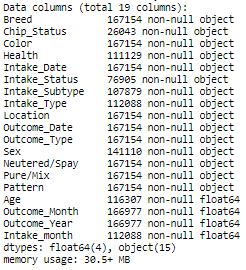
‘Intake Type’, ‘Sex’, and ‘Intake Subtype’ were straightforward renaming processes to make the language used consistent.

‘Chip Status’ was modified to a binary category with ‘no chip’ and ‘unable to scan’ as ‘no chip’. Since no chip does not necessarily indicate the dog isn’t preowned, I assumed both would be treated equally.

Nominal change was made to ‘Intake\_Status’. Only ‘Heartworm’ and ‘Ringworm’ were merged with ‘Sick’. Other were maintained the same since they may be useful categories.

Neutered/Spay column was formulated to separate out some dataset without this classification to dataset with this information. Since many owners like their pets neutered or spay, this could be a useful feature to determine outcome.

Breed was categorized using the American kennel club classification because the feature encompasses 4635 distinct naming/type of dogs. The feature resulted in ‘Herding’, ‘Hound’, ‘Toy’, ‘Non-Sporting’, ‘Sporting’, ‘Terrier’, ‘Working’, and ‘Miscellaneous’ categories. Mix breeds were categorized depending on whether the dog was a mix of distinct groups of dogs or of one class. For example, if the dog was a mix of ‘non-sporting’ and ‘non-sporting’ type of dog, it would be classified as ‘non-sporting’. On the other hand, if the dog was a blend of ‘non-sporting’ and ‘terrier’, it’s classified as mix. Pure/mix column was created to make distinction between pure dog types or a mix of same category of dog. Dogs not grouped by the American kennel club were classified as miscellaneous. This method of grouping the dogs is a huge assumption and one part to do future works on.

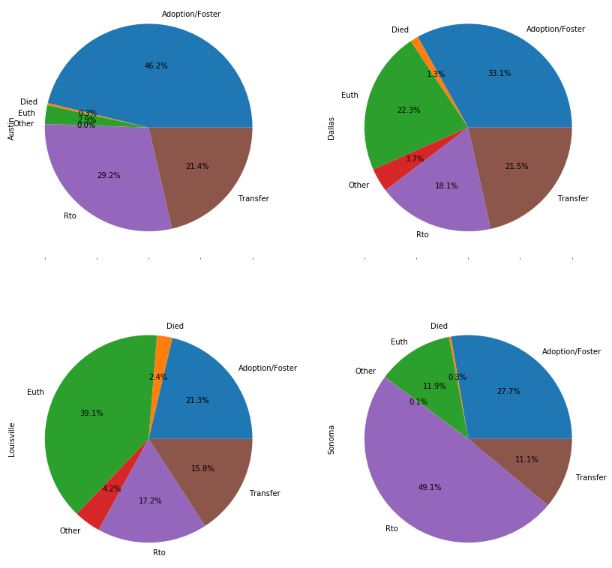
‘Color’ was classified to colors of the color of the rainbow with white and black, simplifying the numbers of categories.

From the ‘outcome date’, year and month were extracted as separate features.

According to Support Table 1, some ages were mislabeled with age below at 0 days old and have max age at 14075 days old (38.6 years old), older than any dogs ever existed. ‘Age’ under 0 years old and over 20 years old (dogs according to google do not live to 20 years old) were considered mislabeled and set to average age. This again is a huge assumption considering the importance of age to adoptability. It could that recorders documented unknown dogs to unrealistic age to denote their old age. This part could be further investigated for better predictions.

Even though, the data isn’t fully clean with null values in the dataset, it’s now useful for data analysis. (fig 1)

Figure 1: Information of the resulting merged data

1. **Exploratory Data and Statistical Analysis**

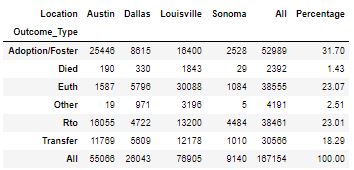
In Austin, a mere 2.9% of dogs are euthanized with over 46% of their dogs adopted/fostered. Dogs transferred and returned to owner accounts for the remaining half of their shelter dogs. Dallas has the second to best adoption/foster rate at 33.1%, however with euthanization rate of 22.3%. Louisville leads euthanization rates with almost 40% of dogs euthanized at the shelter whereas for other regions less than 25% of dogs are euthanized. The shelter also has the lowest adoption/foster rate out of the four location at 21.3% and has the second to lowest transfer rate at 15.8% only to be trumped by Sonoma with a transfer rate of 11.1%. Largest portion of dogs at the shelter appears to have owners with almost 50% of the dogs returned to their owners. Considering this, the euthanization rate is still second to lowest. The data collected for all four location has adoption rate at 31% and euthanization at 23%, on par with the expected euthanization rate on the national level. Notably, Austin beats other location with the greatest number of adopted dogs at 25k with only 7 years of data whereas in Louisville, only 16k was adopted/fostered with over 15 years of data (Fig. 2).

Figure 2: Pie Chart and Table of Dog Outcome by Location

**5a. Breed**

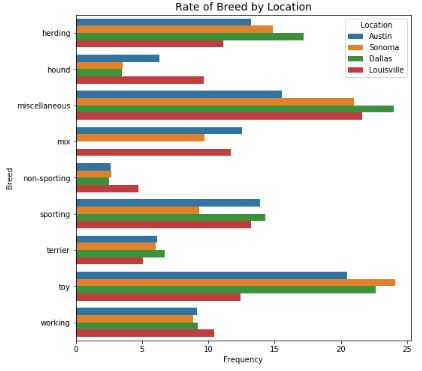
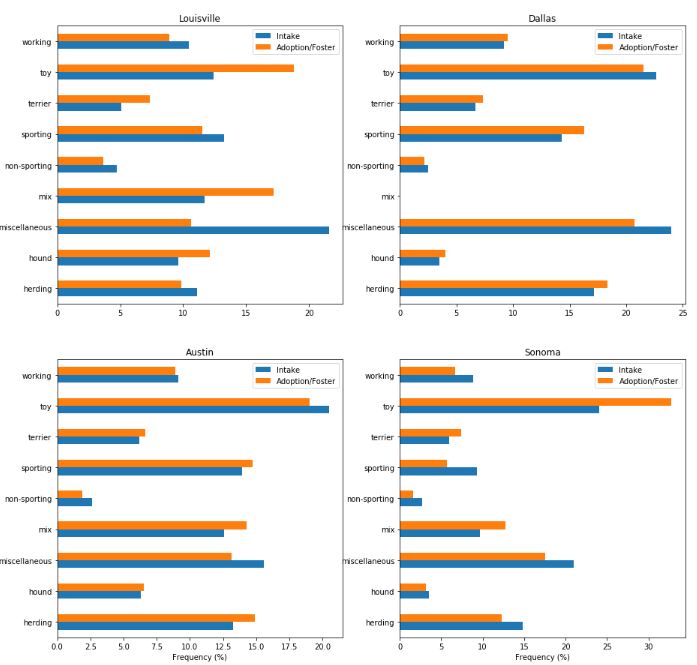
Generally, all the facilities tend to have similar distribution of the types of dogs taken in at these facilities with few exceptions. Toy and miscellaneous groups lead with the highest intake rates compared the other groups of dogs. A hypothesis for high miscellaneous breed at shelters could be because most dogs are mix of breeds and sometimes intakers can find it difficult to identify what breed mix was the dog. As a result, it’s far easier to classify them as miscellaneous. More surprisingly are the high rates of toy dogs among all regions. However, toy dogs in Louisville are more than 36% lower compared to Austin, Sonoma and Dallas. This indicates toy groups are uncommon in Louisville shelter. Also, Louisville tends to have higher intake of hound and non-sporting dogs than the other locations. In fact, Louisville, visually, has the most uniform distribution of breeds compared to other locations. Sonoma has 33% lower in sporting intake than the next lowest. Austin and Dallas have similar dog diversity. (Fig. 3)

Figure 3: Breed by Location. (Dallas do have mix breeds)

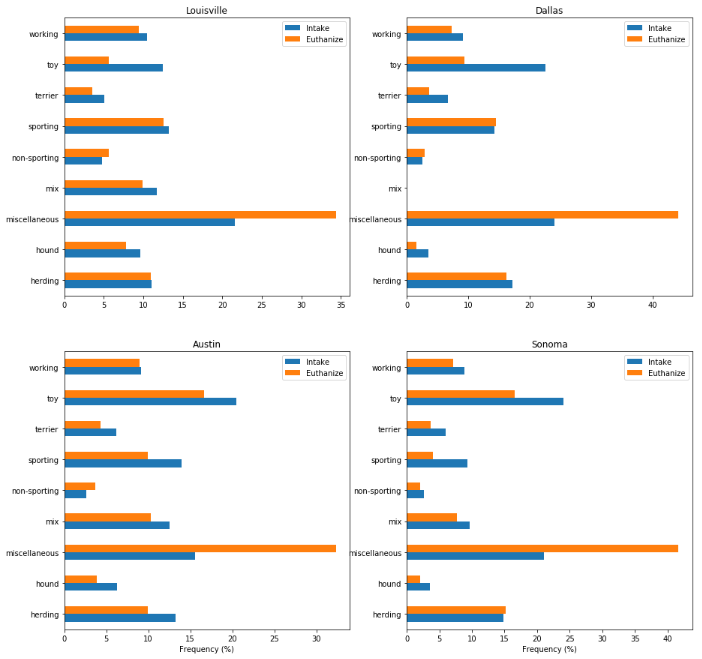
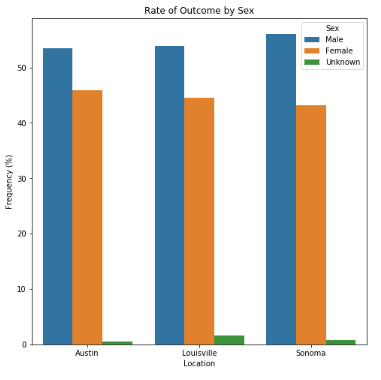
Adoption/foster frequency tends to align well with intake dog diversity at shelters. It’s most notably clear by the bar graphs of Dallas and Austin. However, miscellaneous dogs have lower adoption/foster rates than the intake rates and it’s most extreme in Louisville with percent difference by more than 50%. Toy and mix breeds are favored in Louisville and that of Sonoma overtakes intake frequency by 30%. However, do note these percentage is the profile of the adoption/foster and do not imply the demand is not met at these shelters. What could be said is in Louisville, toy and mix breed are most in demand, whereas toy, sporting, miscellaneous, and herding are most favored in Dallas. For other locations toy breeds are most in demand. (Fig 4)

Figure 4: Adoption/Foster Behavior by Breed

Miscellaneous breeds are most vulnerable for euthanization in all locations. They are over 50% more at risk than other breeds of dogs whereas toys seem to be least at risk for Louisville and Dallas with euthanization rate below 50% compared to the intake rate. In Austin and Sonoma, without accounting for miscellaneous breeds, the diversity for euthanization generally match up with the intake frequency of each group of dogs. Herding and non-sporting is most at risk in Austin and hound is least at risk proportional wise. Sporting is disproportionally least at risk in Sonoma with more than 50% slip from the intake distribution. (Fig. 5)

Figure 5: Euthanization Profile of Dogs by Location

**5b. Gender**

Across all the shelters, male dogs end up at the shelter most where the difference is more than 10% in Sonoma. Counting up the genders, male accounts for 53.8% and female accounts for 45% of the gender diveristy in shelters. To test the statistical significance of the difference, one-sided bootstrap proportion testing and one-sided z proportion test were conducted with the conditions stated below.

Null hypothesis: p(male intake) = p(female intake)

Alternative hypothesis: p(male intake) > p(female intake)

alpha = 0.05

Both statistical testings yield p-values of 0.0 and reject the null hypothesis and accept the alternative hypthesis that male dogs are taken to shelters more frequently than female dogs.

Figure 6: Gender Profile by Location

In terms of outcome, 16.3% and 15.2% were adoption rates of male and female dogs respectively. Since the proportion is close bootstrap proportion testing was conducted to determine statistical significance of the difference.

Null hypothesis: p(male adoption/foster) = p(female adoption/foster)

Alternative hypothesis: p(male adoption) > p(female adoption)

Alpha = 0.05

With a p-value of 0.0, the null hypothesis was rejected and the alternative hypothesis was accepted that male adoption is greater than female adoption.

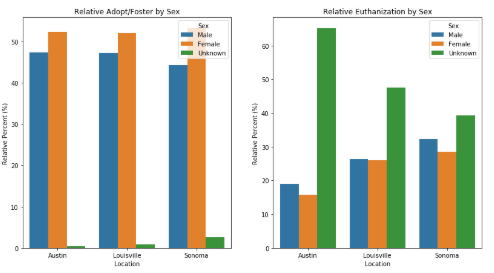
Even though percentage of female dogs picked for adoption is lower than male, the relative proportion suggest that adopters/fosterers tend to favor female dogs at 16.9% and 15.1% for female and male adoption rates respectively (Supporting Table 1). Fig 7. reveals that across all location, female is more favorable to adopt than male dogs using relative percentage. Sonoma has the largest difference with 6%. Unknown gender is the least likely to be adopted. Instead unknown gender dogs are disproportionally euthanized across all location with the most extreme case in Austin. Unknown gender is almost 2 times more likely to be euthanized in Austin. Generally male dogs are more likely to be euthanized than female dogs.

Figure 7: Relative Frequency of Adoption/Foster and Euthanization by Gender

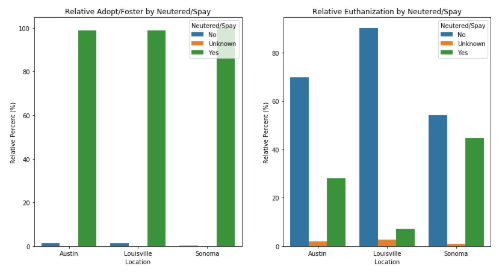
According to Fig 8., overwhelming number of dogs adopted or fostered are nuetered/spayed. Negligible relative percentage of adoption/foster dogs are not sterile or have unknown gender. On the other hand, euthanization tends to target nonsterile dogs where in Louisville over 80 relative percentage of dogs euthanized are not sterile. Although for other locations the relative percentages differ, sterilization is undoubtly one of the key contributor to the adoption and euthaniztion of dogs.

Figure 8: Relative Frequency of Adoption and Euthanization by Neutered/Spay

**5c. Age**

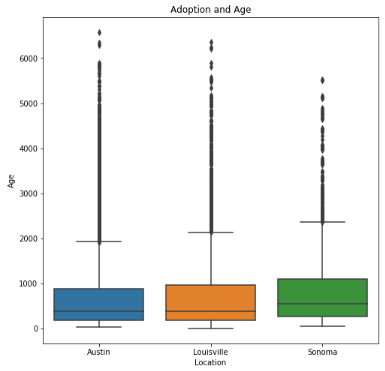
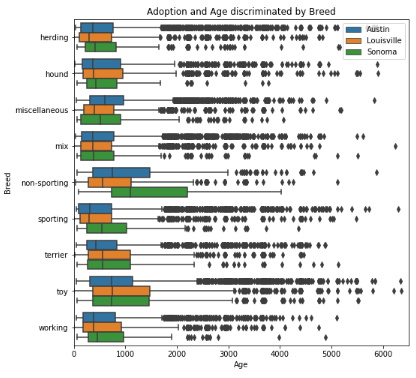
Median adoption ages are 1.1, 1.1 and 1.5 years old for Austin, Louisville, and Sonoma respectively. 25% of adopted/fostered dogs are above 6.5 years old. As noted in Fig. 9, range for adoption/foster spread from less than a couple days old to over 6000 days (16.4 years old). In Austin, there is a large gap between dogs under 6000 days old and those over and these could be mislabeled. As mentioned before Age is an area that may require further feature engineering. Prior to this, all dogs over 20 years old were considered mislabeled and set to the average age. Similar gapping phenomenon could be seen in Louisville and Sonoma. Sonoma has the lowest spread for adoption/foster with max adoption/foster age just around 5500 days old. The mean adoption/foster ages were calculated as 2.1, 2.0, and 2.4 years old for Austin, Louisville and Sonoma respectively. Austin and Louisville share similar mean age for adoption/foster, but Sonoma differs by 0.3 years old (14%) compared to the next highest. Because Sonoma has a highest mean age of adoption, Sonoma adopters/fosterers may be less age discriminating than Austin and Louisville. Bootstrap Permuatation testing was conducted to determine the significance.

Figure 9: Age of Adopted/Foster Dogs

Null hypothesis: μ(adoption age of Austin) = μ(adoption age of Sonoma)

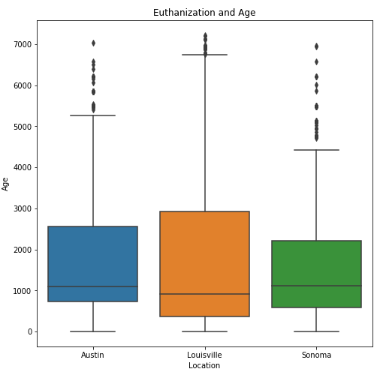
Alternative hypothesis: μ(adoption age of Austin) < μ(adoption age of Sonoma)

alpha = .05

With a p-value of 0.013, the null hypothesis was rejected, and the alternative hypothesis was accepted that the average age of adoption/foster in Sonoma is statistically higher than that of Austin.

To further investigate the age effects, breed was used to further categorized the adoptability. For the most part, median adoption/foster age among all the breed aligns well by location. Nonsporting stood out in Sonoma having median age far surpassed the other location by almost a year old. Sonoma also stood out in sporting breeds with median adoption/foster age around 700 days compared to 500 days for Louisville and Austin. In fact, the lower 25 percentile for adoption/foster at Sonoma is the median for Louisville and Austin in the sporting breed. Toy breed has the highest median adoption/foster age across all location at around 750 days old. (Fig. 10)

Figure 10: Age of Adoption/Foster by Breed and Location

The median ages of euthanized dogs at the three shelters are 3.0, 2.5, and 3.1 years old at Austin, Louisville and Sonoma respectively. Louisville has the largest spread and highest 75 percentile at almost 7000 days old but with lowest median. However, the average euthanization ages are similar with 4.7, 4.8, and 4.6 years old for Austin, Louisville and Sonoma respectively. Bootstrap Permuatation testing was conducted to determine the significance.

Null hypothesis: μ(euthanization age of Louisville) = μ(euthanization age of Sonoma)

Alternative hypothesis: μ(euthanization age of Louisville) > μ(euthanization age of Sonoma)

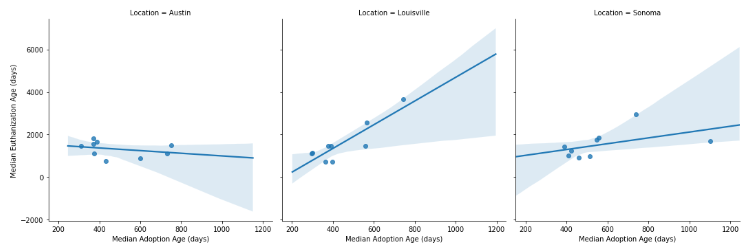
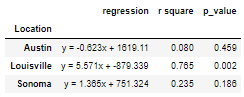
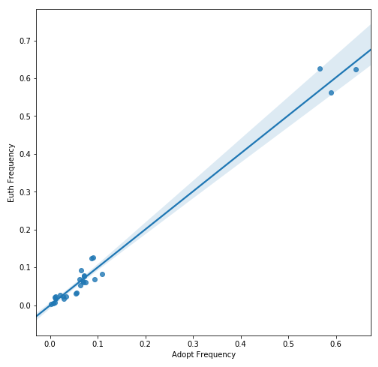
With a p-value of 0.19, we fail to reject the null hypothesis and that mean euthanization of Louisville and Sonoma is the same. This in respect suggests that all three locations have similar mean euthanization age since age for euthanization at Austin is between both those of Louisville and Sonoma.

Figure 12: Adoption/Foster and Euthanization Relationship by Median Age of Dog Breeds

Figure 11: Age of Euthanized Dogs

With euthanization median and the adoption/foster median of the dog groups, regression lines of the relationship between adoption/foster and euthanization were constructed. (Fig. 12) Louisville shows a strong linear relationship between adoption and euthanization ages with a p-value of 0.002 and a r2 value of 0.765. For the other locations, the regression lines are weaker and do not show statistical significance with p-value over 0.15. However, for Sonoma, there appears to be a slight positive correlation and it seems the median age for adopting/fostering non-sporting offset the entire regression curve. If that data point is taken out, the correlation will be stronger than the current value. For Austin the regression is essentially flat with weak correlation value. Austin almost seems as if there is a cut off median age for euthanization at around 2000 days old regardless of the median age of adoption.

Table 1: Statistical Information of Adoption/Foster and Euthanization Relationship by Median Age of Dog Breeds

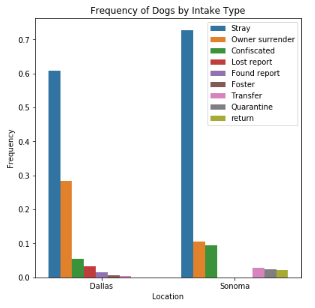
**5d. Color**

The regression line of adoption/foster and euthanization frequency by the color of the animal is linear with r-value of 0.987. However, this does not indicate that color plays a huge role in determining the survival of the dog. The slope and intercept of the regression line shows 1:1 proportional relationship. This essentially mean that the adoption/foster and euthanization profiles by color is the same. In other words, there is no bias in selecting dogs by one color over the other for adoption/foster and euthanization. Thus, there is no favoritism of color. (Fig 13)

y = 01.004x -0.0

r-value: 0.987

Figure 13: Adoption/Foster and Euthanization Relationship by Color of Dogs



**5e. Intake Type**

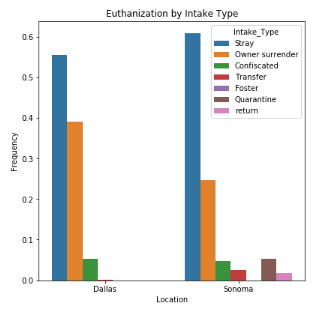
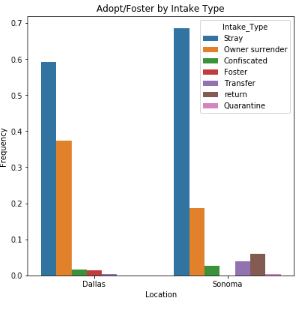
Over 60% of shelter dogs are stray dogs and previously owned dogs take the next highest proportion of dogs at the shelter. Dallas shelter has almost 30% of their dogs coming from owners who gave their dogs to shelter where only 10% gave up their dogs from Sonoma. Confiscated dogs take 10% of the shelter dogs Sonoma. Compared this to euthanized dogs, euthanized dogs are disproportionately owner surrendered dog at 38% and 25% of total dogs euthanized in Dallas and Sonoma respectively. Proportion-wise owner surrendered dogs are 33% and more than two times more likely to be euthanized in Dallas and Sonoma shelters respectively. However, they do tend to be more likely to be adopted/fostered as well. Stray have equal proportion of adoption/foster rate as the intake rate by intake type, however, they are less likely to be euthanized. Quarantine dogs are disproportionately more likely to be euthanized at 5% compared to intake rate of about 2%. Returned dogs are also more likely to be adopted/fostered as well.

Figure 14: Dog Profile of Intake Type. Upper: General profile, Lower Left: Euthanization Profile, Lower Right: Adoption/Foster Profile (Only Dallas and Sonoma have this information)

**5e. Health**

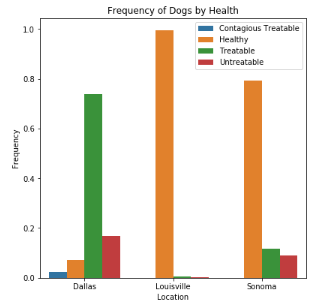
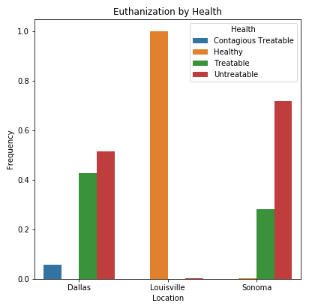
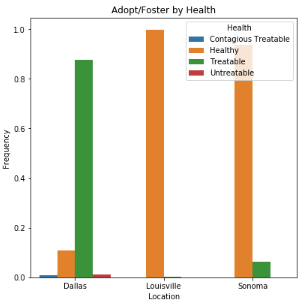
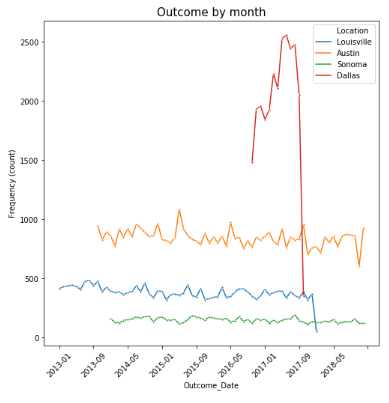
Over 95% of all dogs taken in by Louisville is healthy whereas in Sonoma shelter, it drops down to 80% and even less in Dallas at 10%. In Dallas, 75% of the dogs taken in is sick/injured but is treatable. In Sonoma only about 10% is sick or injured. Even though most shelters take in mostly stray dogs and their health is presumed to be similar, there seems to be a real bias in how health is determined. Unsurprisingly adoption/foster and euthanization in Louisville are primarily healthy dogs.

Figure 15: Dog Profile of Health. Upper: General profile, Lower Left: Euthanization Profile, Lower Right: Adoption/Foster Profile (Austin do not have health information)

On the other hand, untreatable dogs are disproportionately euthanized in all locations. In Dallas, a higher relative percentage of dogs adopted/fostered is healthy at 10% compared to about 7% for general profile. Healthy dogs are not euthanized in Dallas or Sonoma. Contagious but treatable disease dogs are disproportionately euthanized. Treatable but not contagious dogs are less likely to be euthanized in Dallas but more likely in Sonoma. This may be due to the number of healthy dogs taken in at the shelter and do not speak to the actual euthanization tendency at the shelters.

**5e. Time Dependency**

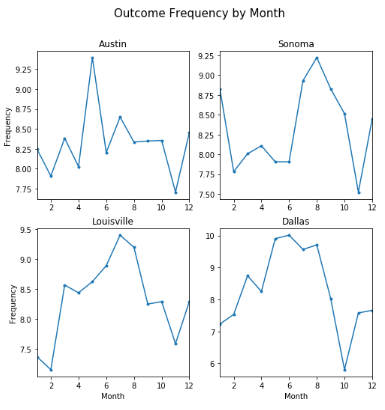
Outcome rates at the various shelters are stable hovering around 800, 400 and 200 per month at Austin, Sonoma, and Dallas respectively. Only one year of data was collected for Dallas and as a result stability cannot be immediately determined. One of the first phenonomenon that is observed in the time series graph is the steep decay at the last month of the dataset as seen for Dallas and Louisville. These steep deviations suggest that the data was not collected entirely at the end and this section should be excluding from frequency analysis to aviod bias. Also there is noticable periodicity in the time series suggesting possible seasonal effect on outcome. It’s most noticable in the Austin data where sharp peaks follow immediately by a deep tough in an approximately equal distance. Austin has high outcome rates around May every year and Louisville experiences similar pattern around August with double peak. (Fig 16)

Figure 16: Time Series of Outcome Count by Month

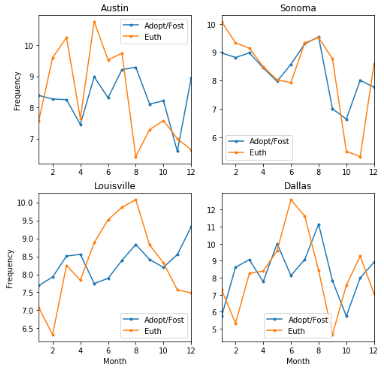
All the frequency-time graphs generated henceforth were constructed with equal number of years for each months to aviod biasing the data. The last and first month of the each dataset were left out to avoid incomplete data. Fig 17 shows a clear pattern for seasonal outcome rates. For Austin, peak outcome rates occurs in May at 9.25% whereas for the remainder of the year, outcome frequency is consistent until November trough of 7.7%, differing by 1.55% or percent change of 17% from the peak. In Sonoma, outcome peaks during the summer reaching to maximum outcome rate of 9.2% in August and slumping in November to about 7.6%, differing by 1.6% and percent change of 17% from the peak. Lousiville peaks in the summer as well to around 9.5% in July and falls in the winter to 7.2% in February, differing by 2.3% and a change of 24% from the peak. Dallas has the most extreme change from 10% in June to 6% in october, differing by 4% and a change of 40% from the peak. The observation from Fig 17. Is consistent with the pattern observed from Fig 16.

Figure 17: Outcome Frequency by Month

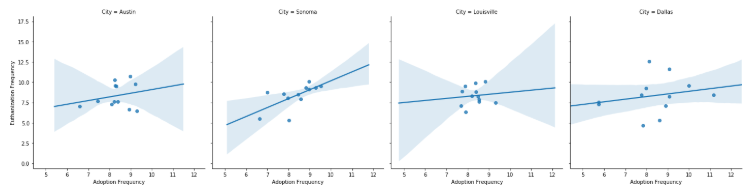
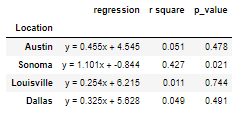
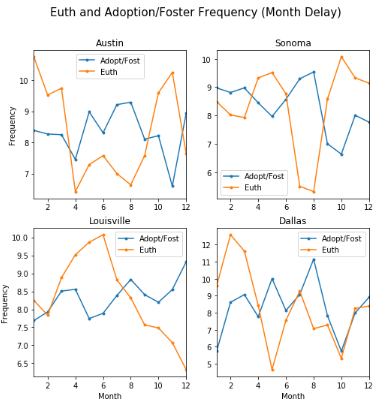
Fig 18 shows seasonal effect on adoption/foster and euthanization. In Austin euthanzation mostly occurs in the first half of the year peaking in May at about 11% and dropping to almost 6% in August. In Sonoma euthanization peaks in January and in August and in Louisville euthanization peaks generally in the summer and mellows in the winter. Dallas also peaks in the summer and bottoms out in November on euthanizations. For adoption/foster, it doesn’t appear to have a clear peak and trough, instead adoption/foster seems to come in waves and is especially apparent for Louisville and Dallas. This is whether surprising because there isn’t a true match up between euthanization and adoption/foster rate. One would postulate that as adoption/foster improves, euthanization rates would decrease. However, Fig. 18 shows an almost independence from one another, suggesting other factors may contribute more to euthanization rate aside from adoption rates. Sonoma is the only outlier in that euthanization frequency almost match up entirely with adoption/foster frequency. To test this, a regression is constructed.

Figure 18: Adoption/Foster Frequency Compared to Euthanization Frequency by Month

Figure 19: Adoption/Foster and Euthanization Relationship by Monthly Frequency

Fig 19 and Table 2 shows the relationship between adoption and euthanization by monthly frequency. Almost all locations show weak correlation value and high p-value, depicting weak or no correlation between euthanization and adoption in seasonal perspective. However, Sonoma exposes direct postive relationship with r2 of 0.427 where as adoption/foster rate increases, euthanization rate increases. This is counter intuitive even if the p-value suggest statistical significance. I postulate that there could be time delay between adoption/foster and euthaniztion.

Table 2: Statistical Information of Adoption/Foster and Euthanization Relationship by Monthly Frequency

The time delayed seasonal effect of adoption/foster and euthaniztion frequency is shown in Fig 20. The time delay effect have two restrictions:

1. The delay time does not exceed 4 months
2. The euthanization and adoption/foster frequency have an inverse relationship (negative slope)

The resulting time delay is as follows:

Austin euthanization delay time: 4 months

Sonoma euthanization delay time: 3 months

Louisville euthanization delay time: 2 months

Dallas euthanization delay time: 4 months

Figure 20: Adoption/Foster Frequency Compared to Euthanization Frequency by Month with Time Delay

Visually, troughs match with peaks nicely where Austin and Sonoma have the largest improvement. Time delay for Dalllas and Louisville did not significantly improve the relationship.

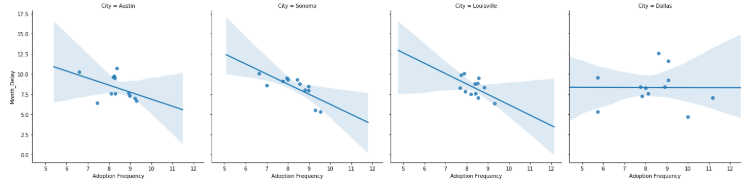
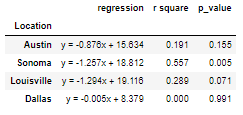
Significant improvement on r2 and p-value except for Dallas where p-value actually increased and r2 value fell to 0.0. This is because of the negative slope criteria established above. r2 value of Sonoma improved from 0.427 to 0.557. The r2 value of Louisville increased from .011 to 0.289 and that of Austin increased from 0.051 to 0.191. Better correlation values do not suggest definite relationship, especially the p-values suggests there is no effect. Even for sonoma case, where p-value is low enough to statistically suggest the model could be meaningful, the periodicity of adoption frequency may just coincidentially overlap with that of euthanization frequency if the two outcomes have the same time period.

Table 3: Statistical Information of Adoption/Foster and Euthanization Relationship by Monthly Frequency with Time Delay

Figure 21: Adoption/Foster and Euthanization Relationship by Monthly Frequency with Time Delay

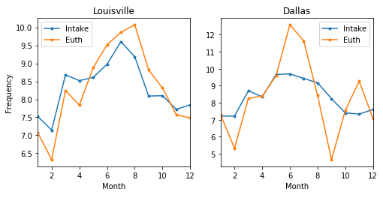
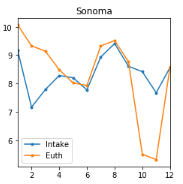


Figure 23: Intake Frequency Compared to Euthanization Frequency by Month without Time Delay

Comparing intake frequency with euthanization frequency, strong direct correlation emerges. Both Louisville and Sonoma almost have exactly the same distribution. Louisville in particular have peaks at the same month and toughs at the same month. Even though visually intake frequency and euthanization frequency do no line up directly for Dallas shelter, it’s apparent that the peak for intake frequency matches up to the more rounded peak of intake frequency. This is true for the toughs as well. Since Intake frequency correlates with the space avaliable at shelters, this sheds light the effect of space capacity on euthanization. (Fig 23)

Louisville showed exceptional performance with r2 of 0.842 and a p-value of 0.0. Dallas also has remarkable r2 of 0.461 with p-value of 0.015. Both p-values suggest statistical significant in the model for predicting euthanization rates. However, regression did not fit Sonoma well and is not a good predictor for euthanization. But a closer inspection Fig 24 reveals two particular outliers that skewed the model. The remaining data points appears to have a strong linear relationship. (Fig 24 and Table 4)

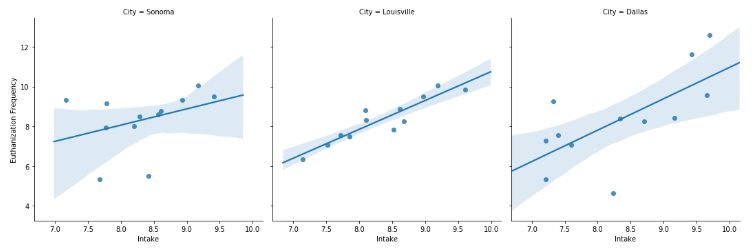
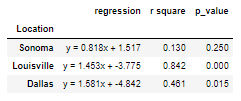
**5h. Summary of Findings**

Figure 24: Intake and Euthanization Relationship by Monthly Frequency without Time Delay

The breed of dogs plays some role in predicting adoption/foster and euthanization. Miscellaneous breed has higher euthanization rate and lower adoption/foster rate compared to intake rate whereas toy breed has higher adoption/foster rate and lower euthanization rate compared to intake rates.

Table 4: Statistical Information of Intake and Euthanization Relationship by Monthly Frequency without Time Delay

Male dogs are more commonly taken in by shelter and have a higher rate of adoption, however, female dogs have higher relative adoption rate. Unknown gender dogs are disproporationately euthanized where male and female dogs are roughly the same in terms of euthanization. All adopted/fosters dogs at all the locations are neutered/spay whereas it is not a predictor for euthanization.

Age of dogs plays a huge role in euthanization and adoption/foster rate. Sonoma tends to be less age discriminating for adoption with mean age of 2.4 years old whereas for other location the mean age is 2 years old. Also mean euthanization age is the same at Austin, Louisville and Sonoma shelters at 4.6 years old.

The color of the dogs plays no role in determining adoption/foster and euthanization of a dog.

Intake Type has some significance in evaluating for euthanization of a dog. Owner surrender and quarentine dogs tend to be euthanized. Intake type have moderate affect on adoption/foster rate.

All adopted/fostered dogs are disproportionally healthy but health is not a good indicator for Louisville since over 95% of shelter dogs are healthy. Dogs with untreatable or contagious disease are disproporationately euthanized at all locations.

Outcome Month may play a slight role in determining adoption/foster and euthanization of dogs but intake Month plays a significant role in euthanization of dogs. This is believed to be caused by limitation to capacity at the shelter.

1. **Machine learning**

Machine learning algorithms were conducted on the four datasets separately. Features not native to the original dataset was removed and all null values were removed. Color and pattern features were drop because they fail to predict the adoption/foster and euthanization rates. Age column was scaled to speed up learning. All the categorical features were converted to binary representation for learning with OneHotEncoding.

**6a. Machine Learning Algorithms**

*Naïve Bayes Bernoulli:* From the Naïve Bayes algorithms Bernoulli was chosen because it natively deals with categorical dataset. Since almost all features in the dataset was categorical, therefore this was an obvious choice. Age was converted to young, old and middle aged to fit the categorical setting of Bernoulli. Randomized search 3-fold cross validation was used to reduce variance and alpha was hypertuned to improve accuracy. The scoring metric was set for accuracy.

*Logistic Regression:* Logistic regression was performed with solver set to lbfgs and multi class set to one vs rest. Randomized search cross validation 3-fold was used to reduce variance and regularizer (C) was hypertuned to improve accuracy. The scoring metric was set for accuracy

*Support Vector Machine:* kernel was set to Linear for faster learning and randomized search cross 3-fold validation was used to reduce variance. C was hypertuned to improve accuracy and the scoring metric was set for accuracy

*K Nearest Neighbors:* The weights were set to uniform for faster learning. Randomized search cross 3-fold validation was used to reduce overfitting and n neighbors was hypertuned to improve accuracy. The scoring metric was set for accuracy.

*Random Forest:* Grid search 3-fold cross validation was used to improve overfitting and the number of trees was hypertuned. The scoring metrics was set to accuracy.

*Deep learning:* 6 hidden layers with a final layer of softmax with 5 nodes for classification. Dropout was used every other layer to reduce variance and batch normalization was used on layers without Dropout to avoid conflict and speed up learning and remove the risk of vanishing gradient or exploding gradient. Relu was used for all the activation layers.

All models presented weren’t hypertuned entirely with all available parameters and it could be done for future work and study.

**6b. Choosing the Right Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Louisville | Dallas | Austin | Sonoma |
| Best Accuracy | 68.1% Deep Learning | 55.4% Deep Learning | 60% Deep Learning | 72% Deep Learning |
| Best Euthanization Precision | 0.79 Deep Learning | 0.78 SVM | 1.0 k Nearest Neighbor | 0.89 Logistic Regression |
| Best Euthanization Recall | 0.90 Logistic Regression | 0.65 Deep Learning | 0.08 Random Forest | 0.88 Naïve Bayes |
| Best Adoption Precision | 0.64 Deep Learning | 0.52 Deep Learning and Random Forest | 0.65 Deep Learning | 0.72 Deep Learning |

Table 5: Table of Best Accuracy, Euthanization Precision, Euthanization Recall and Adoption Precision from the four shelters

Even though for most models, the accuracies are subpar, the models do provide some value. (Table 5) The purpose of this project was to identify dogs at risk for euthanization and transfer them to locations with better outcome. The first part could be achieved with high euthanization precision and recall for some location. For example, Sonoma, the euthanization precision and recall could be achieve close to 90% with two separate models providing great confidence of which dogs are at risk. However, this is not to say we are happy with the models. Adoption/foster rate is an area that needs more work since no models were able to achieve a precision score of 80%. This is essential to determining the location to send at risk dogs. Also, RTO and Transfer dogs are creating a real problem for predicting adoption and euthanization. According to the confusion matrixes generated in the models, most mistakes made are from adoption and euthanization predicted as RTO and Transfer or RTO and transfer predicted as euthanization and adoption. This is the largest area of potential improvement to rectify the model from error and significantly improve the model.

Also choosing the right model for the shelters depends on the economic and business model of the shelter. For shelters that want to be more humane towards dogs, they can run the model prioritizing euthanization precision to identify the predicted dogs at risk. While more economical shelter can use the model with highest recall to be confident the dog is at risk. Once identified the same features could be run against all locations with models to find the best adoption/foster precision for transfer.

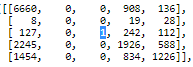
However, once the model is chosen, the confusion matrix should be inspected. In this project, one of the most surprising finding was having a 1.0 recall. However, when the confusion matrix was inspected, only one euthanization was predicted for euthanization and was a true positive. Therefore, the recall, although 1.0, is extremely bias and the model shouldn’t be used for predicting euthanization.

Figure 25: Confusion Matrix of KNN for Dallas. Highlighted is the number of true positive for euthanization

1. **Future works**

Although the resulting models were less than ideal, it is encouraging to yield some strong precision and recall for the prediction of euthanization and adoption/foster. This could be improved further and require changes in several sections of the project. Following are some areas that could improve the model:

1. Improve the featuring engineering and assumptions made for Age, Breed and Color
2. Study the effects of features on RTO and Transfer
3. Study the interdependencies of RTO, Transfer, Adoption and Euthanization rates
4. Simplify outcome types with binary euthanized column to improve accuracy and better target for euthanized dogs.
5. Cross validate with other hyperparameters to improve accuracy
6. Weight average all the models to create an ensemble model
7. Engineer model to improve accuracy of euthanization rate and adoption rate
8. Use transfer learning for Deep Neural Network
9. **Experimental methods**

**7a. Bootstrap Proportion Test**

The null hypothesis and alternative hypothesis were generated with an alpha of 0.05. 10000 bootstrap replicates were sampled. Each replicate was the mean of randomly generated binary number with population equal to the testing data. The percentage of samples surpassing the cut off proportion was measured and tested against alpha.

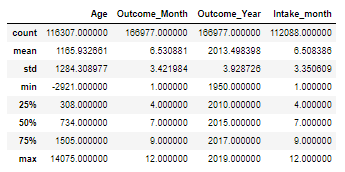
**7b. Bootstrap Permutation Test**

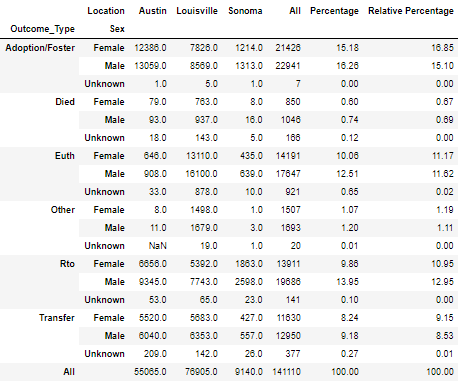
The null hypothesis and alternative hypothesis were generated with an alpha of 0.05. Mean difference between the two testing populations was calculated for testing and the two populations were merged. 10000 permutations of the mean difference were generated by randomly drawing samples with size equivalent to the two testing populations. Then the p-value is calculated for null hypothesis testing.

**7c. z-Proportion Test**

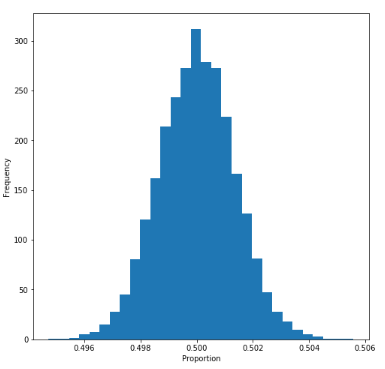
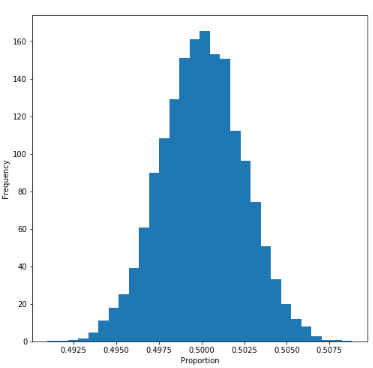
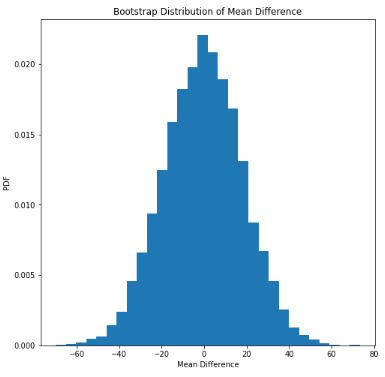
The null hypothesis and alternative hypothesis are generated with an alpha of 0.05. Standard error and proportion were measured. The p-value was calculated and compared and tested again alpha.

1. **Supporting Figure**



Supporting Table 1: Data Description before Modifying Age

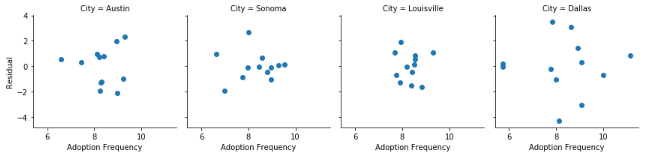
Supporting Table 2: Gender and Outcome



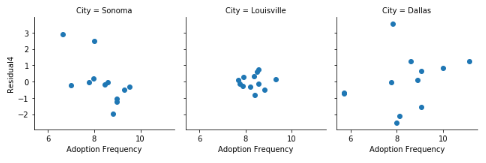
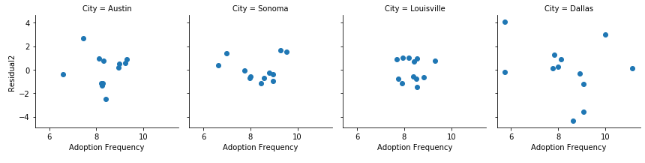
Supporting Figure 1: Bootstrap distribution of Male and Female in Shelter

Supporting Figure 2: Bootstrap distribution of Male and Female Adoption rates

Supporting Figure 3: Permutation Distribution for the Adoption Mean Difference between Sonoma and Louisville



Supporting Figure 5: Residual Plot of Regression comparing Euthanization and Adoption/Foster Frequency in Seasonal Perspective

 **

Supporting Figure 7: Residual Plot of Regression comparing Euthanization and Intake Frequency in Seasonal Perspective without Time Delay

Supporting Figure 6: Residual Plot of Regression comparing Euthanization and Adoption/Foster Frequency in Seasonal Perspective with Time Delay