edX-Harvard Data Science Capstone Project: Shelter Animal Outcomes

Eric Tucker

March 3rd, 2022

Introduction

Every year, approximately 1.5 million animals are euthanized in shelters across the United States. Some of these deaths may be preventable by identifying particularly at risk animals and allowing shelters to focus their budgets towards additional medical procedures, marketing, or play time with potential owners. Austin Animal Center recently released 3 years worth of intake and outcome data on its animals in hopes of doing just this. Here we analyzed these data in an attempt to model the likelihood of euthanization for any incoming animal so that funds can be allocated appropriately and unneccesary deaths can be avoided.

Data Cleaning, Exploration, and Feature Transformation

The data was first downloaded and imported from its raw CSV format, the first 10 rows of which are shown below. On each animal, we were provided with several predictors including age, sex, spayed/neutered status, breed, color, outcome, outcome subtype, and the date and time of the outcome.

```
##    AnimalID    Name            DateTime     OutcomeType OutcomeSubtype
## 1   A671945 Hambone 2014-02-12 18:22:00 Return_to_owner
## 2   A656520   Emily 2013-10-13 12:44:00      Euthanasia      Suffering
## 3   A686464  Pearce 2015-01-31 12:28:00        Adoption         Foster
## 4   A683430         2014-07-11 19:09:00        Transfer        Partner
## 5   A667013         2013-11-15 12:52:00        Transfer        Partner
## 6   A677334    Elsa 2014-04-25 13:04:00        Transfer        Partner
## 7   A699218   Jimmy 2015-03-28 13:11:00        Transfer        Partner
## 8   A701489         2015-04-30 17:02:00        Transfer        Partner
## 9   A671784    Lucy 2014-02-04 17:17:00        Adoption
## 10  A677747         2014-05-03 07:48:00        Adoption        Offsite
##    AnimalType SexuponOutcome AgeuponOutcome                            Breed
## 1         Dog  Neutered Male         1 year            Shetland Sheepdog Mix
## 2         Cat  Spayed Female         1 year           Domestic Shorthair Mix
## 3         Dog  Neutered Male        2 years                     Pit Bull Mix
## 4         Cat    Intact Male        3 weeks           Domestic Shorthair Mix
## 5         Dog  Neutered Male        2 years      Lhasa Apso/Miniature Poodle
## 6         Dog  Intact Female        1 month Cairn Terrier/Chihuahua Shorthair
## 7         Cat    Intact Male        3 weeks           Domestic Shorthair Mix
## 8         Cat        Unknown        3 weeks           Domestic Shorthair Mix
## 9         Dog  Spayed Female       5 months     American Pit Bull Terrier Mix
## 10        Dog  Spayed Female         1 year                    Cairn Terrier
##          Color
## 1  Brown/White
## 2  Cream Tabby
## 3   Blue/White
```

```
## 4   Blue Cream
## 5          Tan
## 6    Black/Tan
## 7   Blue Tabby
## 8  Brown Tabby
## 9    Red/White
## 10       White
```

The outcome subtype was ignored, as this information is not provided until the outcome is determined, so it could not be used as a predictor. Prior to splitting the data set into a training and validation set, the original outcomes, shown in the table below, were condensed into a binary outcome (Euthanized or Not Euthanized), also shown below.

```
##
## Original Outcomes:

## .
##        Adoption            Died      Euthanasia Return_to_owner        Transfer
##           10769             197            1555            4786            9422

##
## Binary Outcomes:

## .
## FALSE  TRUE
## 25174  1555
```

This transformation simplied the classification problem and ensured that similar ratios of euthanized vs. non-euthanized pets would be split into the validation and training sets. Once transformed, the data was then split into training and validation sets at an 80:20 ratio.

We treated the validation data as if it were future observations on which predictions need to be made, and so all data exploration was carried out on only the training data set. However, any cleaning and transformations carried out as a result of the exploratory findings were applied to both data sets in parallel.

Age Effects

First, we looked at age effects. Before being able to assess these, however, we had to standardize the entries, as they were reported as character strings such as:

```
## [1] "1 year"  "2 years" "3 weeks" "2 years" "1 month" "1 year"
```

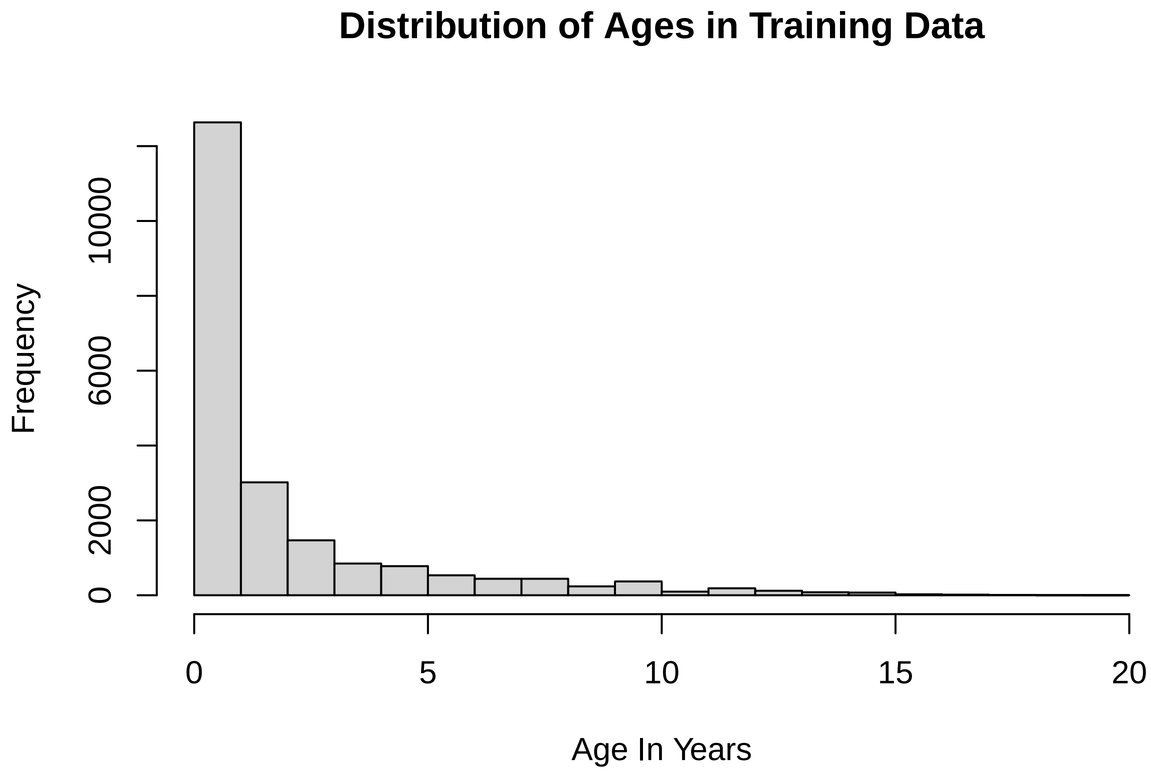
and in a variety of units, including:

```
## [1] "day"    "days"   "month"  "months" "week"   "weeks"  "year"   "years"
```

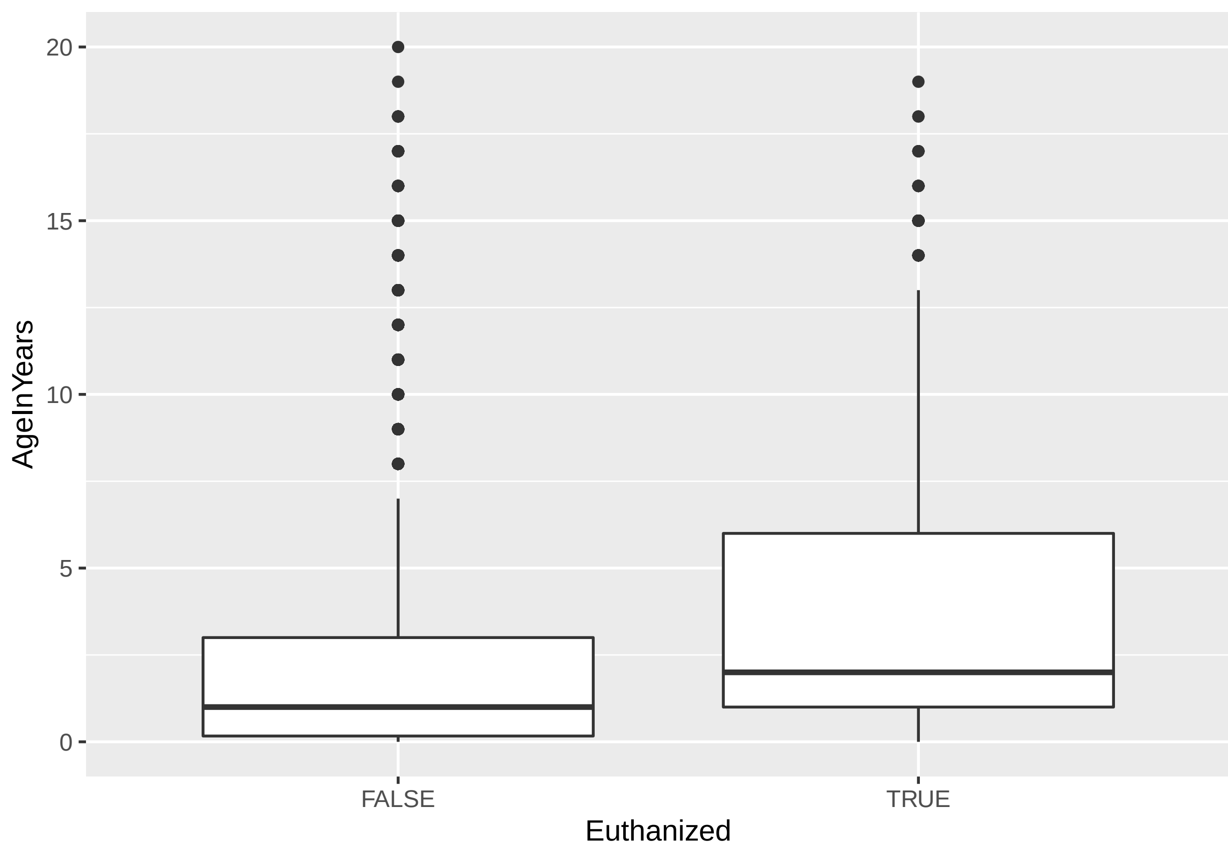
The numbers were extracted from the strings, and the “s” was removed from all plural entries to convert everything to singular units. The extracted numbers were then divided by an appropriate conversion factor to convert everything into years. (Month values were divided by 12, weeks divided by 52, and days divided by 365.) A new feature, AgeInYears, was then created to store these values, the distribution of which is shown below.

```
##
##    day   days  month months   week  weeks   year  years
##      6     61    262   1686     33    356    816   2123

##
##   day month  week  year
##    67  1948   389  2939
```



We then looked to see if there were age differences between euthanized and non-euthanized animals, and we indeed found some, with statistical significance confirmed by t-test as shown below.



```
##
##  Welch Two Sample t-test
##
## data:  pull(filter(data, Euthanized == TRUE), AgeInYears) and pull(filter(data, Euthanized == FALSE)
## t = 15.771, df = 1311.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.679930 2.157242
## sample estimates:
## mean of x mean of y
##  4.000169  2.081583
```

Given these findings, we kept AgeInYears as a feature to carry forward into modeling attempts.

Animal Type Effects

There were only two types of animals in this data set: Dogs and Cats. We checked to see if there was any difference in euthanization rates between them, and indeed found there was a small but significant difference, as confirmed by a Chi Squared test (shown below), where cats were euthanized slightly more frequently than dogs (6.4% vs 5.4%). AnimalType was therefore carried forward into the models.

```
## # A tibble: 2 x 3
##   AnimalType     n euth_prop
##   <chr>      <int>     <dbl>
## 1 Cat         8854    0.0645
## 2 Dog        12530    0.0537
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data$AnimalType and data$Euthanized
## X-squared = 10.807, df = 1, p-value = 0.001011
```

Sex and Fixed Status Effects

The sex information was provided as string, combined with the spayed/neutered status of each animal. Each combination is shown below along with the euthanization proportions observed in the training set, ordered by most frequently to least frequently euthanized.

```
## # A tibble: 6 x 3
##   SexuponOutcome      n euth_prop
##   <chr>           <int>     <dbl>
## 1 "Intact Male"    2807    0.139
## 2 "Intact Female"  2807    0.111
## 3 "Unknown"         863    0.0962
## 4 "Neutered Male"  7786    0.0354
## 5 "Spayed Female"  7120    0.0257
## 6 ""                  1    0
```

As shown, males are generally more frequently euthanized than females, and “intact” animals are euthanized more than spayed or neutered animals. We separated out these variables with regex extraction, labeling the separate variables as Sex and, though not an flattering term, “Fixed” Status - representing either neutered (if male) or spayed (if female). Chi Squared tests were run to confirm statistical significance of euthanization rates between each variable, the results of which, along with the corresponding rates, are shown below.

```
## # A tibble: 3 x 3
##   Sex         n euth_prop
##   <fct>   <int>     <dbl>
```

```
## 1 Unknown   864    0.0961
## 2 Male    10593    0.0630
## 3 Female   9927    0.0498
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Sex and data$Euthanized
## X-squared = 39.897, df = 2, p-value = 2.17e-09
```

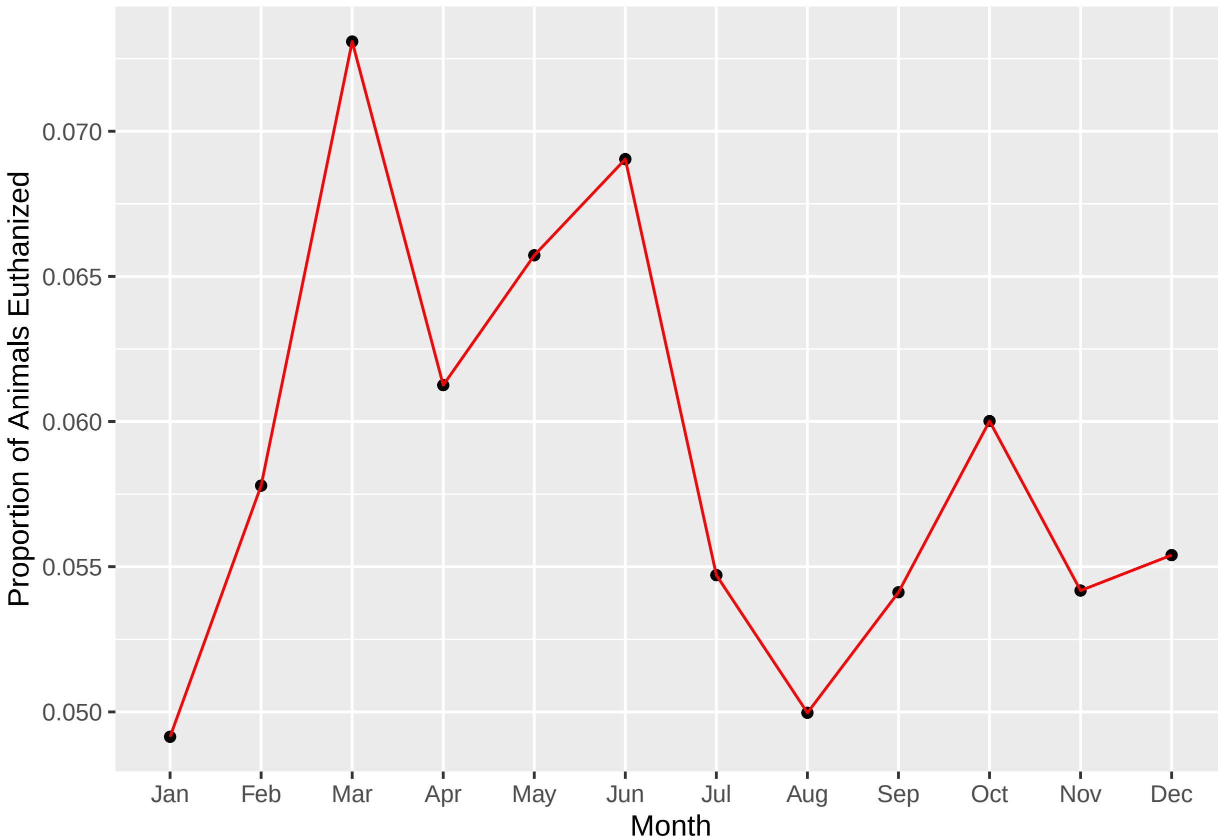
```
## # A tibble: 3 x 3
##   FixedStatus         n euth_prop
##   <fct>           <int>     <dbl>
## 1 Intact           5614    0.125
## 2 Unknown           864    0.0961
## 3 Spayed-Neutered 14906    0.0308
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$FixedStatus and data$Euthanized
## X-squared = 684.79, df = 2, p-value < 2.2e-16
```

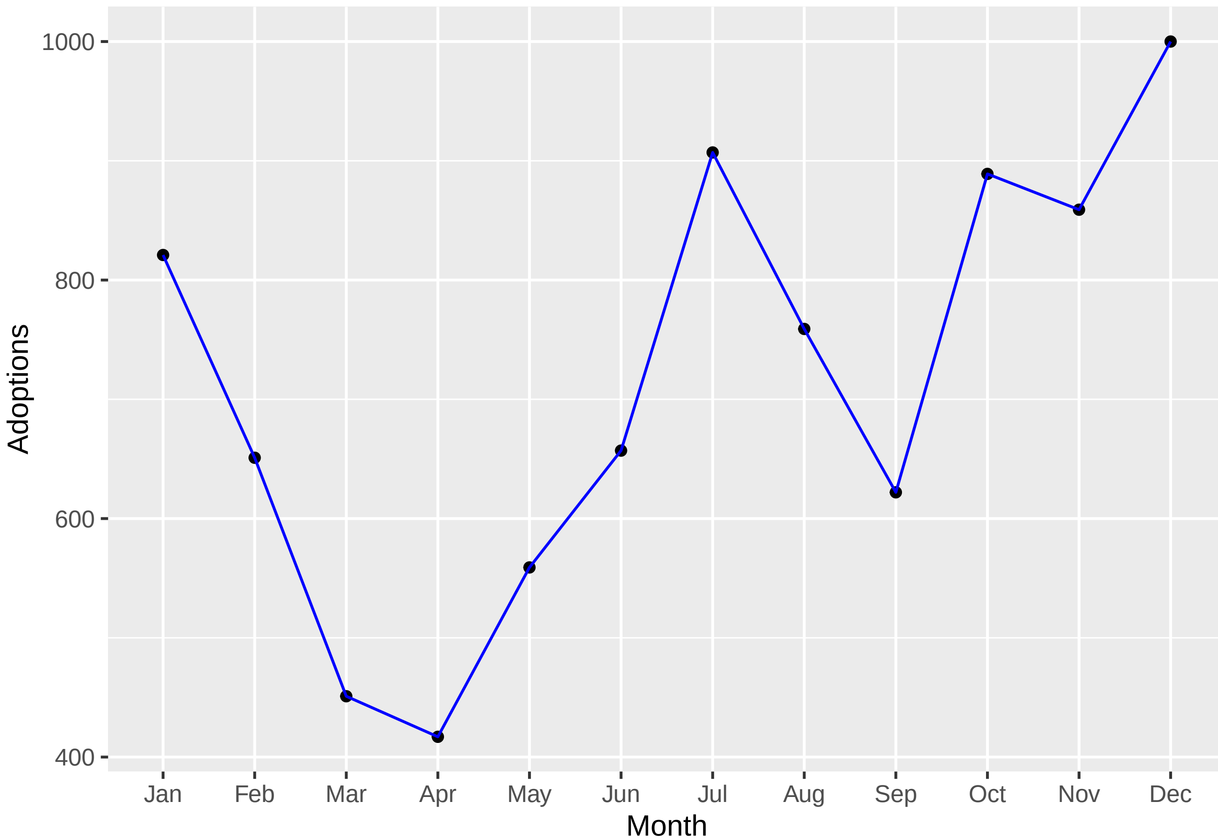
Through this separation, we saw that unknown sex animals were most likely to be euthanized, followed by males, followed by females (a slightly different order than when viewing the combined variables). The FixedStatus variable followed the same trends as previously observed. Both of these variables were carried forward into the models.

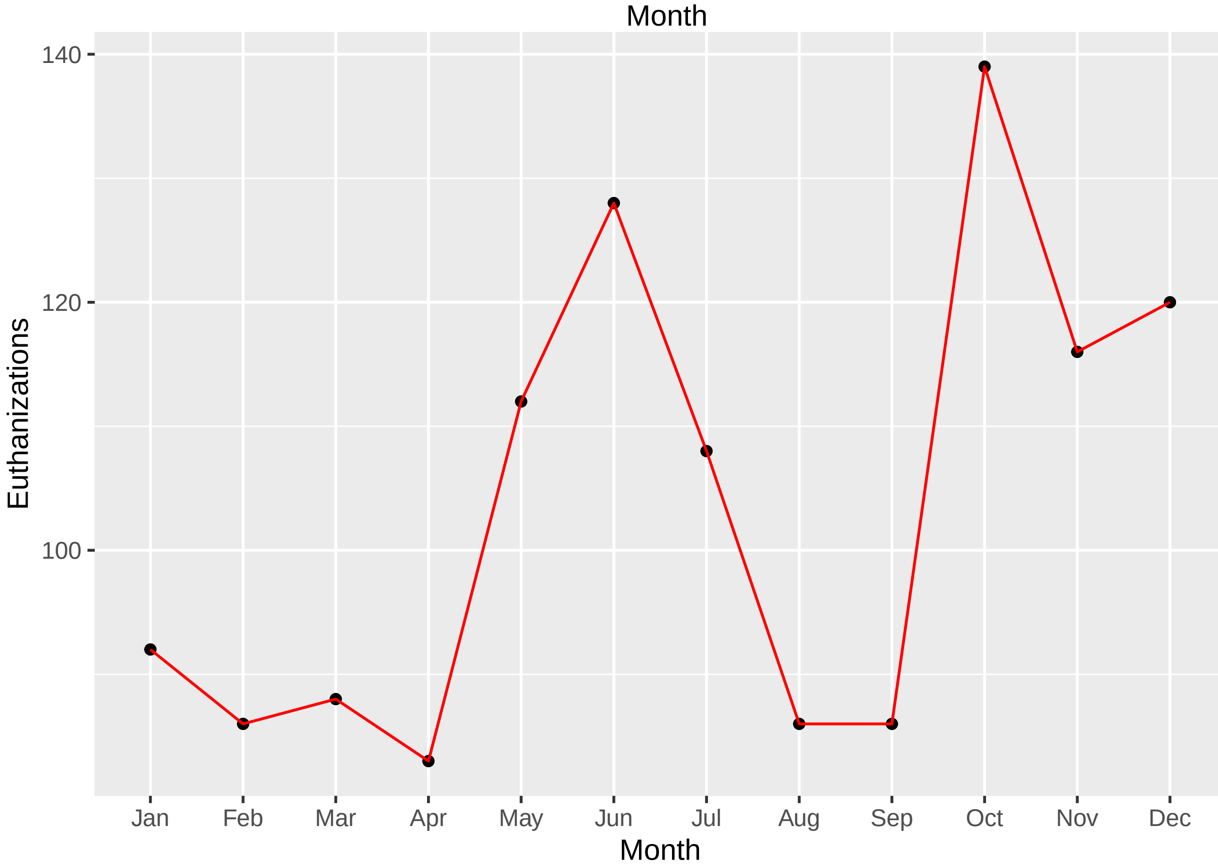
Seasonal Effects

We reasoned that shelters would be likely to experience variability in both intake and adoption rates throughout the year, so we looked for any seasonal effects on euthanization rates. The exact meaning of the DateTime variable was not made clear by the publishers, so it could either correspond to the intake time or the outcome time. In either case, it was interpreted as an approximate activity time during which an animal was present in the shelter and would arrive at some outcome. We extracted the month from each timestamp and viewed euthanization rates for each, shown below.

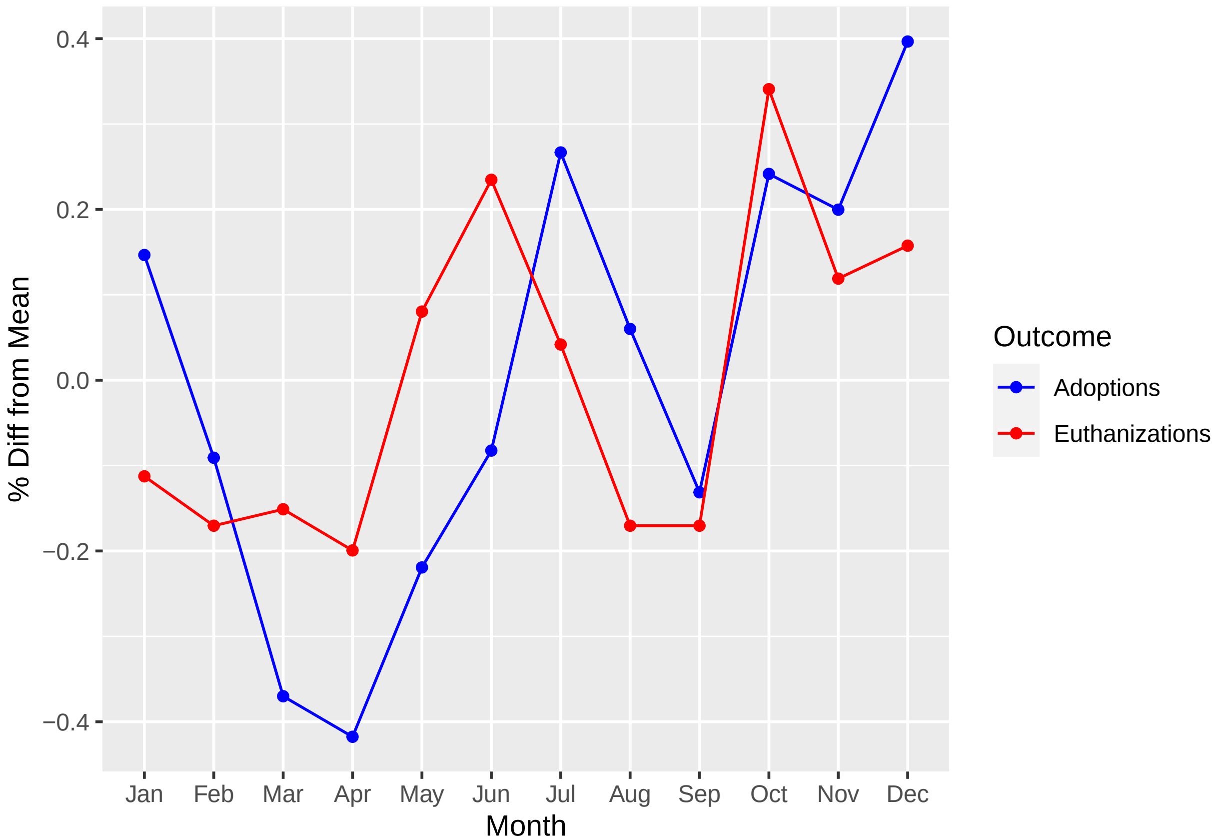


It appeared that euthanization rates spiked in the Spring (March through June). To get a clearer picture of what was happening, we looked at both euthanization and adoption counts by month, and saw that they both followed similar but not identical trends (shown below), generally dropping in Winter and Spring but rising in the Fall.





Given the differences in scale for adoptions vs. euthanizations by month, we instead compared them by plotting as percentage deviations from their respective averages (shown below), and we were able to show that although euthanization counts dropped below average in the months of March and April, they did not drop nearly as much as the adoptions, and so the chances of an animal getting euthanized during those months was indeed higher. In April and May, the euthanization counts actually increased above average, while the adoption counts fell below. Collectively, these months were labeled as "Danger Months", and we coded a matching binary variable set to a new feature called "Month Status", the two statuses being "Danger" or "Neutral". In an effort to reduce dimensionality, the MonthStatus variable, not the Month itself, was carried forward into the modeling process.



Breed Effects

Determining breed effects was more challenging, as there were 1233 unique breeds listed with much of the variety coming from mixed breed combinations. As an example, 20 randomly selected breeds and the number of animals belonging to them are shown below.

```
## # A tibble: 20 x 2
##    Breed                                      n
##    <chr>                                  <int>
##  1 Miniature Pinscher/Chihuahua Shorthair    13
##  2 Border Collie/Labrador Retriever          25
##  3 Lhasa Apso Mix                            37
##  4 Labrador Retriever/Border Collie          33
##  5 Toy Poodle Mix                            30
##  6 Miniature Pinscher                        14
##  7 Standard Poodle Mix                       11
##  8 Maltese Mix                               39
##  9 Domestic Medium Hair Mix                 670
## 10 Cocker Spaniel Mix                        35
```

```
## 11 Border Collie Mix                          185
## 12 American Pit Bull Terrier Mix               49
## 13 Collie Smooth Mix                           24
## 14 Chihuahua Longhair Mix                     107
## 15 Miniature Poodle                            16
## 16 Parson Russell Terrier Mix                  12
## 17 Rottweiler Mix                              92
## 18 Soft Coated Wheaten Terrier Mix             18
## 19 Chihuahua Shorthair/Pug                     15
## 20 Dachshund Wirehair Mix                      20
```

It turned out there was significant variation in euthanization rates among breeds as well. The 30 most frequently and 30 least frequently euthanized breeds with at least 10 animals per breed are shown below. As shown, some breeds are euthanized far above the average rate, while some have no record of ever being euthanized.

```
##
## Most Frequently Euthanized Breeds
```

```
## # A tibble: 163 x 3
##    Breed                                  n euth_prop
##    <chr>                              <int>     <dbl>
##  1 Pit Bull/Chinese Sharpei              10    0.5
##  2 Standard Poodle Mix                   11    0.273
##  3 Boxer/Labrador Retriever              13    0.231
##  4 Chow Chow Mix                         48    0.229
##  5 Himalayan Mix                         14    0.214
##  6 Pit Bull/Boxer                        14    0.214
##  7 Pit Bull/Labrador Retriever           25    0.16
##  8 Cocker Spaniel                        13    0.154
##  9 Pembroke Welsh Corgi Mix              14    0.143
## 10 German Shepherd/Labrador Retriever    51    0.137
## 11 Pit Bull Mix                        1539    0.133
## 12 Whippet Mix                           16    0.125
## 13 American Bulldog Mix                  89    0.124
## 14 Pit Bull                              49    0.122
## 15 Rottweiler Mix                        92    0.120
## 16 Carolina Dog Mix                      34    0.118
## 17 Rottweiler                            34    0.118
## 18 Shetland Sheepdog Mix                 17    0.118
## 19 Labrador Retriever/Pit Bull           61    0.115
## 20 Domestic Longhair Mix                404    0.111
## 21 Domestic Longhair                     18    0.111
## 22 Labrador Retriever/Chow Chow          18    0.111
## 23 Beauceron Mix                         10    0.1
## 24 Dachshund Wirehair Mix                20    0.1
## 25 Great Pyrenees                        10    0.1
## 26 Australian Cattle Dog                 21    0.0952
## 27 Queensland Heeler Mix                 42    0.0952
## 28 American Staffordshire Terrier Mix    76    0.0921
## 29 Siamese Mix                          307    0.0912
## 30 American Staffordshire Terrier        11    0.0909
## # ... with 133 more rows
```

```
##
## Least Frequently Euthanized Breeds
```

```
## # A tibble: 163 x 3
##    Breed                                         n euth_prop
##    <chr>                                     <int>     <dbl>
##  1 Anatol Shepherd                              12         0
##  2 Australian Shepherd                          11         0
##  3 Australian Shepherd/Labrador Retriever       10         0
##  4 Basset Hound                                 13         0
##  5 Beagle                                       16         0
##  6 Beagle/Chihuahua Shorthair                   12         0
##  7 Belgian Malinois Mix                         13         0
##  8 Bichon Frise Mix                             10         0
##  9 Black Mouth Cur Mix                          53         0
## 10 Black/Tan Hound Mix                          15         0
## 11 Border Collie                                14         0
## 12 Border Terrier Mix                           22         0
## 13 Boston Terrier                               12         0
## 14 Boston Terrier Mix                           35         0
## 15 Boxer                                        25         0
## 16 Bruss Griffon Mix                            17         0
## 17 Cairn Terrier/Chihuahua Shorthair            11         0
## 18 Catahoula/Labrador Retriever                 11         0
## 19 Chihuahua Shorthair/Cardigan Welsh Corgi     15         0
## 20 Chihuahua Shorthair/Jack Russell Terrier     17         0
## 21 Chihuahua Shorthair/Rat Terrier              16         0
## 22 Collie Smooth Mix                            24         0
## 23 English Bulldog Mix                          22         0
## 24 English Pointer Mix                          12         0
## 25 German Shorthair Pointer Mix                 17         0
## 26 Golden Retriever                             11         0
## 27 Harrier Mix                                  12         0
## 28 Jack Russell Terrier                         13         0
## 29 Jack Russell Terrier/Chihuahua Shorthair     10         0
## 30 Labrador Retriever                           60         0
## # ... with 133 more rows
```

Again in an attempt to reduce dimensionality, we set out to divide the breeds into "Danger Breeds" (those at high risk of euthanization), "Safe Breeds" (those unlikely to get euthanized), and "Neutral Breeds" (the rest in between). First we removed any "Mix" labels and separated top 30 breed strings into distinct breeds as shown in the table below. 30 breeds were selected as potential "Danger Breeds" because, as the data show, these generally represent breeds that are euthanized at least twice as frequently as average breeds. Any fraction could be chosen though, and this would be a good opportunity for future tuning.

```
## # A tibble: 30 x 5
##    `1`                            `2`                `3`       n euth_prop
##    <chr>                          <chr>              <chr> <int>     <dbl>
##  1 Pit Bull                       Chinese Sharpei    <NA>     10     0.5
##  2 St. Bernard Smooth Coat        <NA>               <NA>     10     0.3
##  3 Boxer                          Labrador Retriever <NA>     13     0.231
##  4 Chow Chow                      <NA>               <NA>     53     0.226
##  5 Pit Bull                       Boxer              <NA>     14     0.214
##  6 Standard Poodle                <NA>               <NA>     14     0.214
##  7 American Eskimo                <NA>               <NA>     10     0.2
##  8 Himalayan                      <NA>               <NA>     16     0.188
##  9 Persian                        <NA>               <NA>     11     0.182
## 10 Pit Bull                       Labrador Retriever <NA>     25     0.16
```

```
## 11 German Shepherd                 Labrador Retriever <NA>     51    0.137
## 12 Pit Bull                        <NA>               <NA>   1588    0.132
## 13 Whippet                         <NA>               <NA>     16    0.125
## 14 Rottweiler                      <NA>               <NA>    126    0.119
## 15 Pembroke Welsh Corgi            <NA>               <NA>     17    0.118
## 16 American Bulldog                <NA>               <NA>     94    0.117
## 17 Labrador Retriever              Pit Bull           <NA>     61    0.115
## 18 Carolina Dog                    <NA>               <NA>     35    0.114
## 19 Domestic Longhair               <NA>               <NA>    422    0.111
## 20 Labrador Retriever              Chow Chow          <NA>     18    0.111
## 21 Shetland Sheepdog               <NA>               <NA>     19    0.105
## 22 Chesa Bay Retr                  <NA>               <NA>     10    0.1
## 23 Ragdoll                         <NA>               <NA>     10    0.1
## 24 Dachshund Wirehair              <NA>               <NA>     21    0.0952
## 25 Queensland Heeler               <NA>               <NA>     43    0.0930
## 26 American Staffordshire Terrier <NA>               <NA>     87    0.0920
## 27 Staffordshire                   <NA>               <NA>     87    0.0920
## 28 Akita                           <NA>               <NA>     11    0.0909
## 29 Beauceron                       <NA>               <NA>     11    0.0909
## 30 Labrador Retriever              Beagle             <NA>     11    0.0909
```

We then identified the unique individual breeds that showed up anywhere in this matrix. The results are shown below in order of frequency.

```
##                                 . Freq
## 1              Labrador Retriever    6
## 2                        Pit Bull    5
## 3                           Boxer    2
## 4                       Chow Chow    2
## 5                           Akita    1
## 6                American Bulldog    1
## 7                 American Eskimo    1
## 8  American Staffordshire Terrier    1
## 9                          Beagle    1
## 10                      Beauceron    1
## 11                   Carolina Dog    1
## 12                 Chesa Bay Retr    1
## 13                Chinese Sharpei    1
## 14             Dachshund Wirehair    1
## 15              Domestic Longhair    1
## 16                German Shepherd    1
## 17                      Himalayan    1
## 18           Pembroke Welsh Corgi    1
## 19                        Persian    1
## 20              Queensland Heeler    1
## 21                        Ragdoll    1
## 22                     Rottweiler    1
## 23              Shetland Sheepdog    1
## 24        St. Bernard Smooth Coat    1
## 25                  Staffordshire    1
## 26               Standard Poodle    1
## 27                        Whippet    1
```

Next, we analyzed the 60 least frequently euthanized breeds. 60 was chosen because more breeds are in the safe zone than in the danger zone, and the majority in the lower 60 are never euthanized (48 to be exact).

The remainder are euthanized at a rate less than half of the average animal rate, with the highest group member facing a euthanization rate of 2.7%. Again, this threshold could be set anywhere and should be the subject of optimization in future studies.

The unique breed names extracted from this "Safe Breeds" list are shown below in order of frequency counted.

```
##                              . Freq
## 1          Chihuahua Shorthair   10
## 2           Labrador Retriever    6
## 3         Jack Russell Terrier    3
## 4             Miniature Poodle    3
## 5          Miniature Schnauzer    3
## 6          Australian Shepherd    2
## 7                       Beagle    2
## 8                    Dachshund    2
## 9               Great Pyrenees    2
## 10          Miniature Pinscher    2
## 11                 Rat Terrier    2
## 12               Alaskan Husky    1
## 13              Anatol Shepherd    1
## 14          Australian Kelpie    1
## 15           Belgian Malinois    1
## 16                Bichon Frise    1
## 17                       Black    1
## 18             Black Mouth Cur    1
## 19               Border Collie    1
## 20              Border Terrier    1
## 21              Boston Terrier    1
## 22               Bruss Griffon    1
## 23                 Bullmastiff    1
## 24               Cairn Terrier    1
## 25        Cardigan Welsh Corgi    1
## 26                   Catahoula    1
## 27               Collie Smooth    1
## 28                   Dalmatian    1
## 29             English Bulldog    1
## 30             English Pointer    1
## 31             German Shepherd    1
## 32    German Shorthair Pointer    1
## 33            Golden Retriever    1
## 34                   Greyhound    1
## 35                     Harrier    1
## 36                    Havanese    1
## 37                  Lhasa Apso    1
## 38                  Maine Coon    1
## 39                     Maltese    1
## 40                     Mastiff    1
## 41             Norfolk Terrier    1
## 42             Norwich Terrier    1
## 43                   Pekingese    1
## 44                 Plott Hound    1
## 45                     Pointer    1
## 46                  Pomeranian    1
## 47                         Pug    1
## 48              Siberian Husky    1
```

```
## 49                       Snowshoe    1
## 50 Soft Coated Wheaten Terrier    1
## 51          Standard Schnauzer    1
## 52                      Tan Hound    1
## 53                     Weimaraner    1
## 54                  West Highland    1
## 55        Wire Hair Fox Terrier    1
```

It was apparent that some breeds, such as Labrador Retrievers, show up frequently on both lists, indicating that they are not necessarily any more or less likely to be euthanized, but rather that they are mixed with other breeds quite frequently. To account for this, we extracted only the breeds that were unique between the two lists, and used these as our final lists of "Safe" vs. "Danger" breeds. Each observation in the training and validation data sets were then analyzed to see if the Breed string contained any element of either list. The results were stored in a new feature called "Breed Status", the summary of which is shown for the training data below.

```
##
## Counts of Breed Statuses in Training Set
```

```
##
## Neutral    Safe  Danger
##   11826    5855    3703
```

The euthanization rates among these groups were then analyzed and tested with a Chi Squared test to confirm statistically different proportions (results shown below). Significance was observed, and so these features were carried into the model development.

```
## # A tibble: 3 x 2
##   BreedStatus euth_prop
##   <fct>           <dbl>
## 1 Danger         0.111
## 2 Neutral        0.0561
## 3 Safe           0.0289
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$BreedStatus and data$Euthanized
## X-squared = 283.24, df = 2, p-value < 2.2e-16
```

Color Effects

Lastly, we tried to perform a similar analysis on the animals' colors, which also came in various combinations, making up a total of 342 unique colors reported. The 30 most frequently and 30 least frequently euthanized colors are shown below, separated into unique strings and organized by frequency of occurrence within each group.

```
##
## Most Frequently Euthanized Unique Colors
```

```
##            . Freq
## 1      White   13
## 2      Brown    6
## 3       Blue    5
## 4      Point    5
## 5       Gray    4
## 6      Tabby    4
## 7      Black    3
```

```
## 8    Brindle    3
## 9        Red    3
## 10       Tan    3
## 11      Gold    2
## 12     Merle    2
## 13    Calico    1
## 14 Chocolate    1
## 15     Cream    1
## 16      Fawn    1
## 17      Lynx    1
## 18    Orange    1
## 19      Seal    1
## 20     Smoke    1
## 21    Tortie    1
## 22  Tricolor    1
## 23    Yellow    1
```

```
##
## Least Frequently Euthanized Unique Colors
```

```
##            . Freq
## 1      White   14
## 2        Tan    5
## 3      Black    3
## 4       Buff    3
## 5      Merle    3
## 6      Sable    3
## 7   Tricolor    3
## 8       Blue    2
## 9  Chocolate    2
## 10     Cream    2
## 11       Red    2
## 12    Silver    2
## 13      Tick    2
## 14    Torbie    2
## 15    Yellow    2
## 16   Apricot    1
## 17   Brindle    1
## 18     Brown    1
## 19      Fawn    1
## 20     Liver    1
## 21    Orange    1
## 22     Point    1
## 23      Seal    1
## 24     Smoke    1
## 25     Tabby    1
```

There was clearly a lot of overlap between these groups (white, black, blue, tan, etc), so we did not use color as a predictor in the models. However, there are a few colors that do appear to be unique to the groups such as "tricolor", so perhaps additional analysis of these would be helpful in future studies.

Final Data

The data columns that were chosen as likely predictors were then separated out and saved as final data sets for model training and evaluation. The first 10 rows of the final training set are shown below.

```
##    AnimalType AgeInYears     FixedStatus    Sex MonthStatus BreedStatus
## 2         Cat 1.00000000 Spayed-Neutered Female     Neutral     Neutral
## 3         Dog 2.00000000 Spayed-Neutered   Male     Neutral      Danger
## 4         Cat 0.05769231          Intact   Male     Neutral     Neutral
## 5         Dog 2.00000000 Spayed-Neutered   Male     Neutral        Safe
## 6         Dog 0.08333333          Intact Female      Danger        Safe
## 10        Dog 1.00000000 Spayed-Neutered Female      Danger        Safe
## 12        Dog 2.00000000 Spayed-Neutered Female     Neutral        Safe
## 13        Dog 4.00000000 Spayed-Neutered   Male     Neutral      Danger
## 14        Dog 2.00000000 Spayed-Neutered   Male      Danger     Neutral
## 15        Dog 1.00000000 Spayed-Neutered   Male     Neutral        Safe
##    Euthanized
## 2        TRUE
## 3       FALSE
## 4       FALSE
## 5       FALSE
## 6       FALSE
## 10      FALSE
## 12      FALSE
## 13      FALSE
## 14      FALSE
## 15      FALSE
```

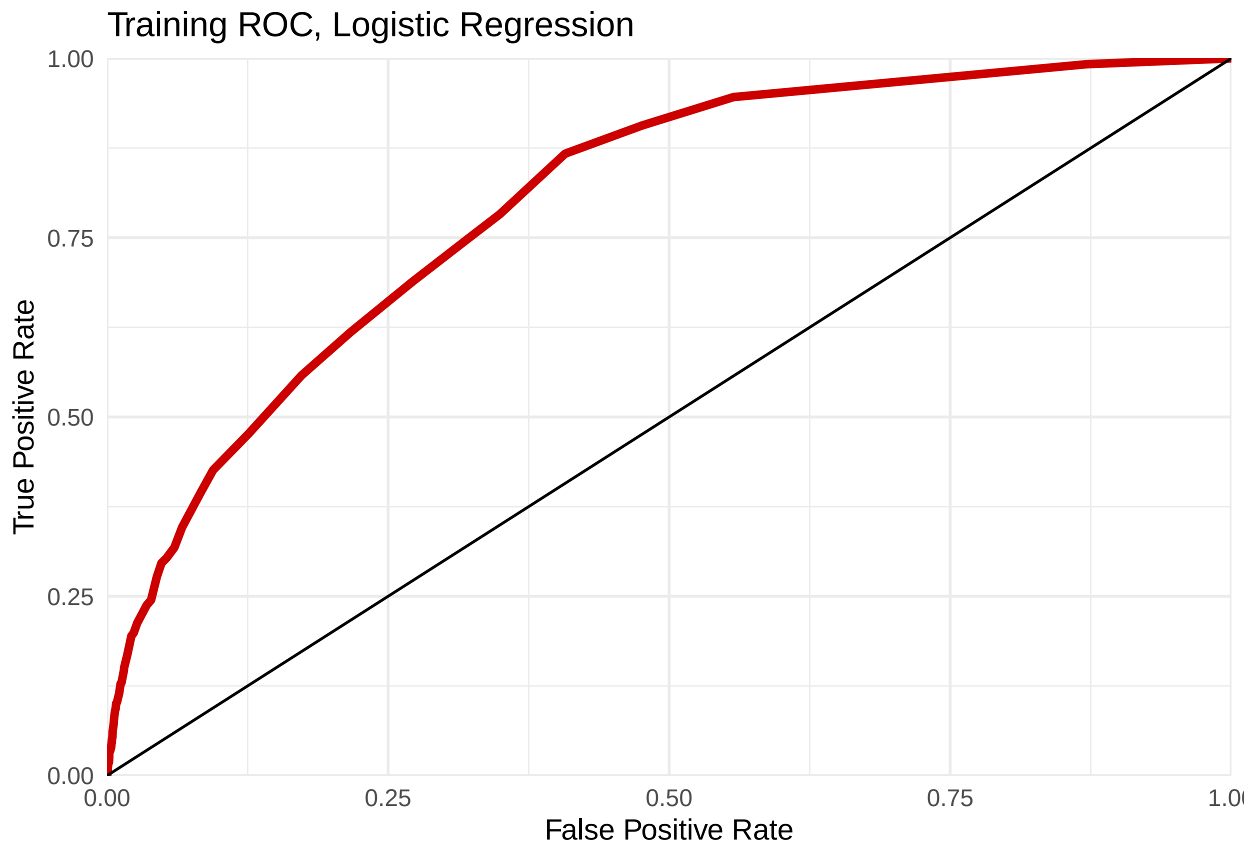
Predictive Models

Logistic Regression

We first fit a logistic regression model to the training data, the coefficients of which are shown below. With this model, Breed Status and Fixed Status demonstrated the strongest predictive power, though all predictors, with the exception of Month Status, were significant.

```
##                        Estimate  Std. Error    z value      Pr(>|z|)
## (Intercept)         -4.43221463 0.086009057 -51.531952  0.000000e+00
## AnimalTypeDog       -0.24529674 0.083498294  -2.937745  3.306083e-03
## AgeInYears           0.21021030 0.008018265  26.216432 1.726448e-151
## FixedStatusIntact    1.82641984 0.068905092  26.506312 8.197395e-155
## FixedStatusUnknown   1.92102418 0.141614509  13.565165  6.443449e-42
## SexMale              0.33698154 0.064935573   5.189475  2.108877e-07
## MonthStatusDanger    0.09931582 0.066244230   1.499237  1.338121e-01
## BreedStatusSafe     -0.53156546 0.109625615  -4.848917  1.241376e-06
## BreedStatusDanger    0.97196166 0.084598433  11.489121  1.496262e-30
```

Because euthanization is a rare event, overall accuracy was likely to be a poor measure of model performance. Instead we generated an ROC curve for each model and analyzed the area under the curve (AUC). An AUC of 1 would indicate a perfect model, that is, a model that can identify all the true positives while generating zero false positives. We did not expect to achieve this result, but our goal was to get as close as possible. The ROC curve and resulting AUC on the training data are shown for the logistic regression model below.




```
##
## Training AUC, Logistic Regression:
##  0.799136
```

0.799 is not a terrible AUC, but it was certain to go down when we ran the model on the validation data, so we tried a few other models first to see if we could improve upon this value.

K-Nearest Neighbors

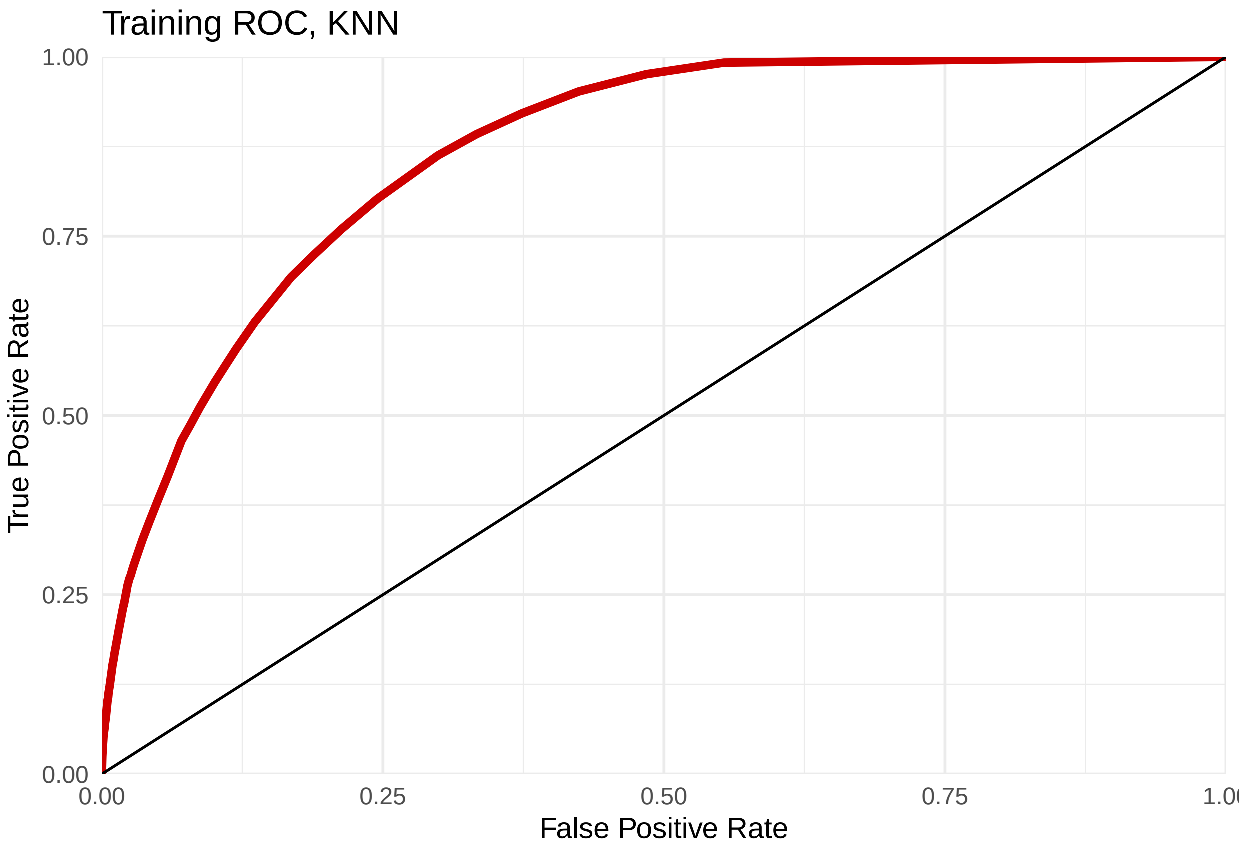
Then then tried to fit a KNN model to the training data. Because euclidean distance was to be used to calculate the proximity of each observation, we first converted all predictors to numerical values. We also standardized them to all be within a range from 0 to 1. This way no predictor would be weighted more heavily than another. In cases where more than two factors existed in a feature, such as in Fixed Status and Breed Status, the distance between 0 and 1 of the middle factor was assigned based on the observed euthanization frequency in the training data, relative to the frequencies of the outer factors.

The results of the KNN model on the training data are shown below.

```
## k-Nearest Neighbors
##
## 21384 samples
##     6 predictor
##     2 classes: 'FALSE', 'TRUE'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 21384, 21384, 21384, 21384, 21384, 21384, ...
## Resampling results across tuning parameters:
##
```

```
##    k  Accuracy   Kappa
##    5  0.9393932  0.10150219
##    7  0.9402072  0.09667184
##    9  0.9407518  0.08937729
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

The tuning parameters within the KNN function resulted in an optimal k size of 9, but this was trained based on accuracy, so we again checked the ROC and AUC (shown below), and we shown a marked improvement over the logistic regression model.




```
##
## Training AUC, K Nearest Neighbors:
##  0.8657912
```

Random Forest

The last model attempted as a random forest (decision tree) model. This model has a lot of tuning options, so we tested a few of the key tuning parameters in order to optimize our performance.

First, we tuned mtry, the number of features that may be randomly selected from at any split point. The resulting AUCs for each input between 1 and 6 are shown below. Mtry = 2 was carried forward.

```
##   mtry       AUC
## 1    1 0.8032319
## 2    2 0.8187758
## 3    3 0.8167635
## 4    4 0.8138891
```

```
## 5    5 0.8107167
## 6    6 0.8088201
```

We then tuned the ntree parameter, which is the number of trees generated in the model and averaged together to produce final predictions. The resulting AUCs from tree counts ranging from 1 to 850 are shown below.

```
##    ntree       AUC
## 1      1 0.5915135
## 2      3 0.6980635
## 3      5 0.7442994
## 4     10 0.7898309
## 5     15 0.7989039
## 6     25 0.8076264
## 7     50 0.8128319
## 8     75 0.8152438
## 9    125 0.8178124
## 10   200 0.8163121
## 11   325 0.8165972
## 12   525 0.8170482
## 13   850 0.8186056
```

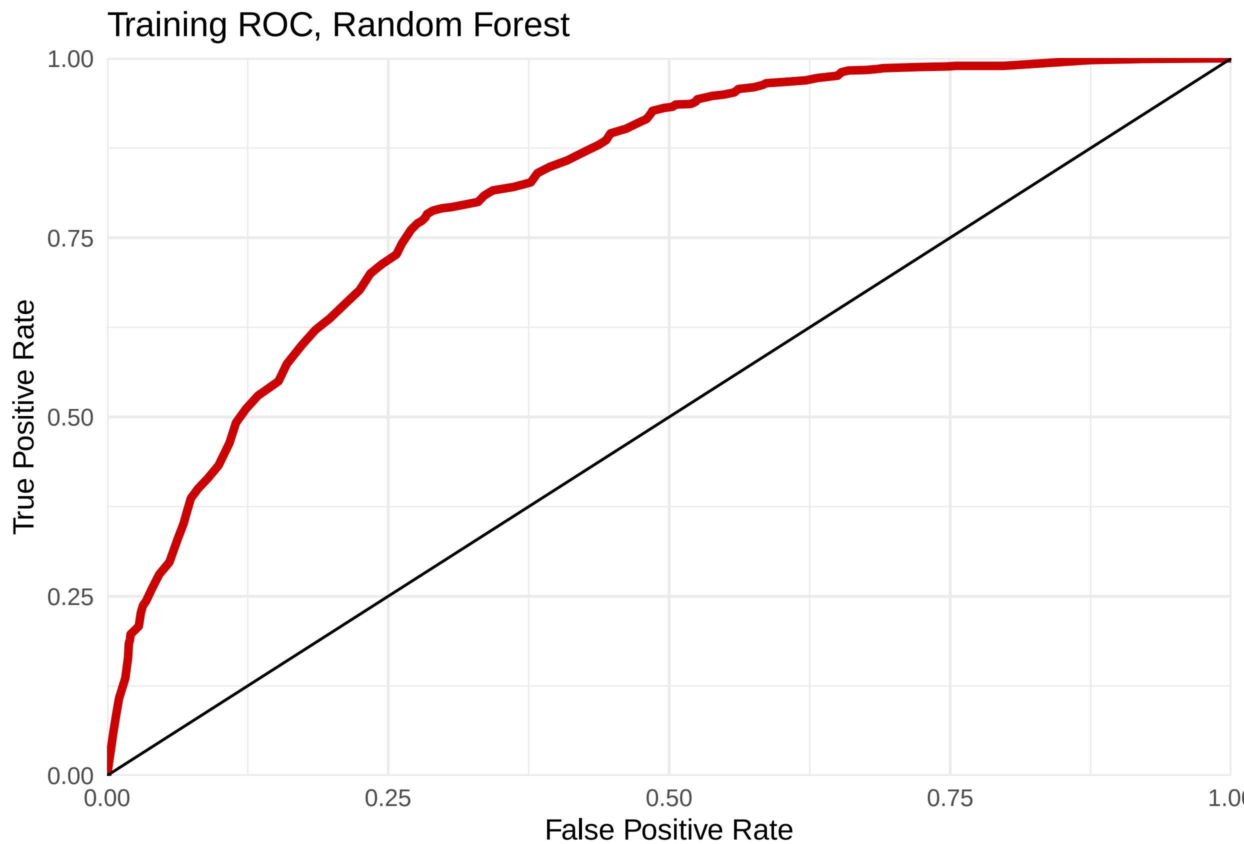
It appeared that the AUC started to stabilize after 200 trees or so, so an additional tuning test was run in this range, the results of which are shown below.

```
##   ntree       AUC
## 1   200 0.8179755
## 2   225 0.8164414
## 3   250 0.8160695
## 4   275 0.8168484
## 5   300 0.8187427
## 6   325 0.8162768
## 7   350 0.8166236
## 8   375 0.8176376
## 9   400 0.8165765
```

Ntree = 200 was carried forward into the final tuning test, which tested the sample size. Note that in each test, the samples drawn from both euthanized and non-euthanized observation groups were equal. Deviations from equality were tested as well but yielded poor results (not shown). The results of sample size tuning on AUC values are shown below.

```
##   sampsize       AUC
## 1       25 0.8128496
## 2       50 0.8157768
## 3       75 0.8188324
## 4      100 0.8192373
## 5      200 0.8171617
## 6      350 0.8180369
## 7      500 0.8173343
## 8      750 0.8176858
## 9     1000 0.8146774
```

A sample size of 100 was found to perform best, and so the final model was fit on the training data with these parameters, and the resulting ROC curve and AUC (shown below) were recorded.




```
##
## Training AUC, Random Forest:
##  0.8171473
```

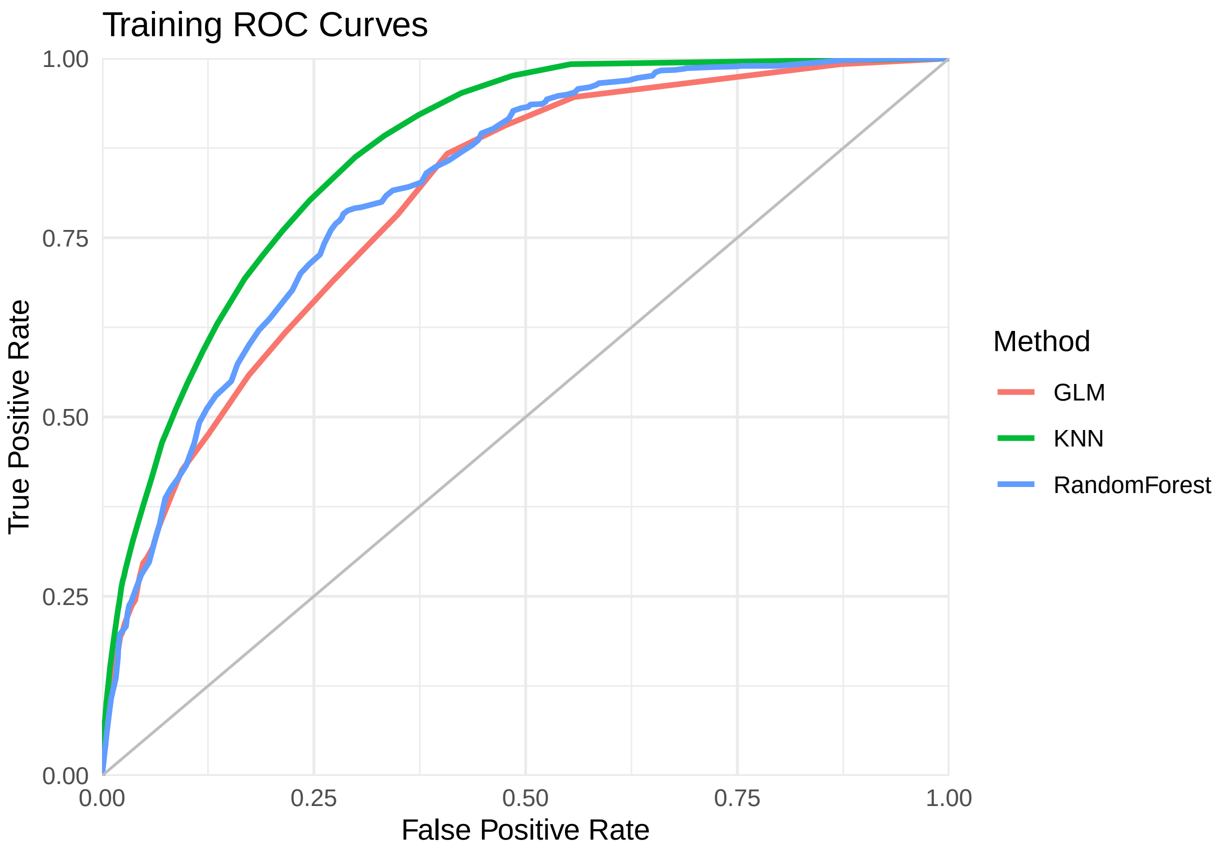
Variable importance data from the final random forest model was captured and is presented below. The most impactful predictors in this model seemed to be Fixed Status and Age, followed by Breed Status. It is notable that Age was not as impactful in the logistic regression model, indicating that perhaps this feature exhibits a highly non-linear relationship with outcome.

```
##
## Importance of Variables in Random Forest Model:
```

```
##                     FALSE         TRUE MeanDecreaseAccuracy MeanDecreaseGini
## AnimalType   0.0004118305 0.0209109126         0.0015223567         2.469614
## AgeInYears   0.0175250115 0.0980966750         0.0218901786        23.391805
## FixedStatus  0.0151390249 0.1564000816         0.0227922241        13.893959
## Sex          0.0017853910 0.0052549745         0.0019731535         3.588316
## MonthStatus  0.0009850260 0.0005663265         0.0009623312         2.819940
## BreedStatus  0.0007142699 0.0581884743         0.0038278255         7.345997
```

Model Comparison

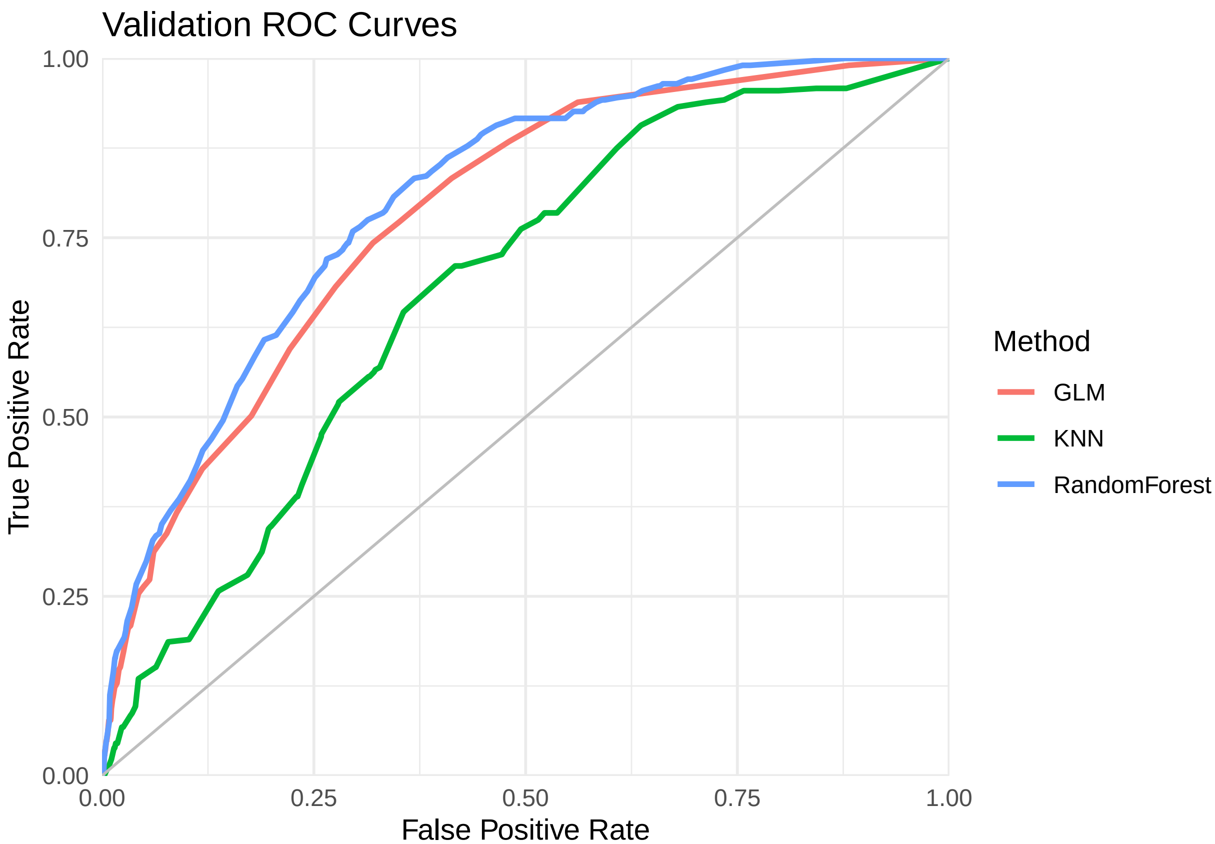
The three models were then compared, as shown below. The best performing model, at least on the training data, was KNN, followed by random forest, followed by logistic regression.




```
##                 Model Training_AUC
## 1 K Nearest Neighbors    0.8657912
## 2       Random Forest    0.8171473
## 3 Logistic Regression    0.7991360
```

Model Validation

We then tested our three models on the validation data set. As expected, the resulting AUCs were lower than on the training data, but still not terrible. In terms of comparison, the random forest model actually performed best on the validation data, retaining a similar AUC from training into validation (0.817 to 0.803). The KNN performance was reduced the most, making it the worst performing model on the validation set.




```
##                    Model Validation_AUC
## 1        Random Forest      0.8038490
## 2 Logistic Regression      0.7818297
## 3 K Nearest Neighbors      0.6779632
```

Conclusions

In the end, we were able to use a random forest model to identify animals that are at highest risk of euthanization with reasonable accuracy. The real-world implications of the final ROC curve would be how much false positive tolerance the clinic has, and this would most likely come down to budget. For example, if they are able to tolerate 25% false positives, meaning they'd be devoting that much budget to supplying extra attention to animals who were likely to be adopted anyway, then they would be able to correctly identify approximately 65% of the animals that are accurately at risk, and they'll ideally be able to prevent some of their deaths.

Assuming 25% false positive rate is too high (after all, they see thousands of animals per year), there are several improvements that could be made here. First, there were clear interaction effects that were not explored. We eliminated some of them by separating out the sex and fixed statuses. It's possible that particular color and breed combinations may be predictive as well, or breeds and age. Secondly, the categories and thresholds around which the months and breeds were aggregated have a lot of flexibility, and no model tuning was done while varying these. Thirdly, better validation techniques could be performed on the training data such as k-fold cross-validation or leave-one-out cross-validation. Finally, other classification models could be explored, such as support vector machines or neural networks. Overall, the models presented here serve as a great starting point to help the clinic start identifying at risk animals, and through some of the methods described here, along with additional data collection, they can only improve.