Effects of age and sex on Dog Outcomes at the Austin Animal Center

By: Garrett Johnson
SDS 322E
Fall Semester 2024

Introduction

Background

- Every year, millions of dogs enter shelters or animal centers, but only a fraction of them get the opportunity to live in a permanent home. Understanding what makes dogs more likely to be adopted is important for improving adoption rates and reducing overcrowding. According to data by the American Veterinary Medical Association, since 1996 the population of dogs has increased from 52.9 million to 89.7 million. With a growing population an answer must be found.

Motivation

- The two predictors we will be using to find trends in dog outcomes are age of the dog and the sex of the dog, along with if the animal has been fixed or not. All of this data is being taken from the Austin Animal Center and is being leveraged to find patterns between those predictors and the outcome variable “outcome_type.”

Research Questions

- How large of an impact does dog age have on the certain outcomes they receive at the animal center?
- How does the sex of a dog influence its chances of being adopted and does being fixed matter as a male or female dog?

Methods

To Start

Began with only filtered data for the animal type “dogs” between the years 2021 and early 2024 for both dog intakes and outcomes.

Combining

```
# Load in the datasets
dog_intakes <- read.csv("aacidogs.csv")
dog_outcome_data <- read.csv("aacodogs.csv")

# Join the intakes and outcomes data
combined_dog_data <- full_join(dog_intakes, dog_outcome_data, by = "Animal.ID")
```

Combined intakes and Outcomes

Same Unit of Measurement

```
# Covnert all units to months
mutate(age_in_months = case_when(
  # Convert Years to Months
  Age.Unit == "years" | Age.Unit == "year" ~ Age.Number * 12,
  # Keep Months as is
  Age.Unit == "months" | Age.Unit == "month" ~ Age.Number,
  # Convert Weeks to months
  Age.Unit == "weeks" | Age.Unit == "week" ~ Age.Number * (12 / 52),
  # Convert Days to Months
  Age.Unit == "days" | Age.Unit == "day" ~ Age.Number * (12 / 30.44))) |>
```

Converting all ages to Months

Popular Breeds

```
# Filter the dog breeds by if they are mentioned over 100 times
combined_dog_data <- combined_dog_data |>
  group_by(breed) |>
  filter(n() > 100)
```

Filtering for Breed Count above 100

Methods

```
# Select the variables we want to keep in our dataset
combined_dog_data |>
  select(
    animal_id,
    animal_type,
    breed,
    sex,
    outcome_type,
    outcome_subtype,
    age_in_months,
    new_sex)
```

Selecting only the variables I want to include in the study

Statistical Modeling and Data Improvements

Using Logistic Regression I used two outcomes variables, “Transfer” and “Adoption” and created binary dummy variables where Transfer equals 0 and Adoption equals 1. This Logistic Regression will help make predictions on the outcome_type for dogs.

```
new_dog <- combined_dog_data |>
  filter(sex != "Unknown",
         outcome_type == "Adoption" |
         outcome_type == "Transfer")
```

```
# Change the outcomes into numerical variables
new_dog <- new_dog |>
  mutate(outcome_type = case_when(
    outcome_type == "Transfer" ~ 0,
    outcome_type == "Adoption" ~ 1))
```

Exploratory Data Analysis on Univariate Visuals

Figure 1

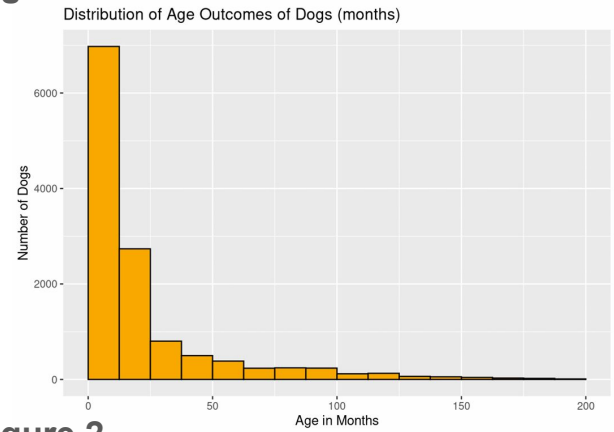
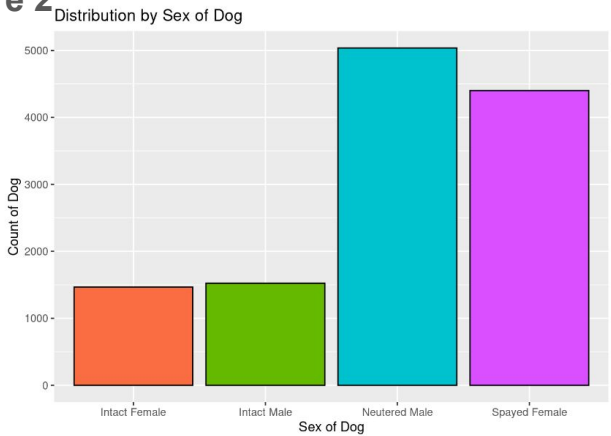


Figure 2



Distribution Statistics

- Median Age: 12 months
- Mean Age: 24.54 months
- Max Age: 240 months

Dog Count

- Neutered Male: 5037
- Spayed Female: 4401
- Intact Dogs: Less than 3000

Exploratory Analysis on Multivariate Visuals

Figure 3

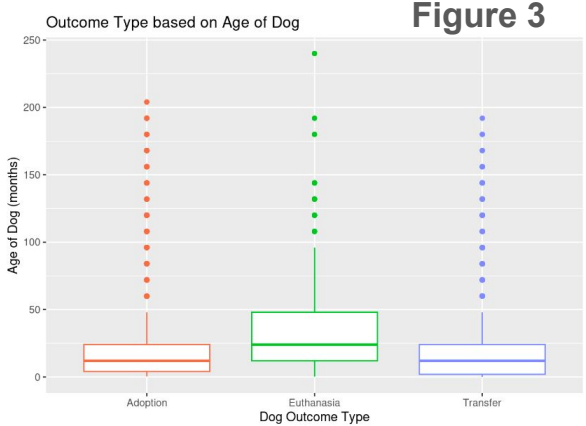
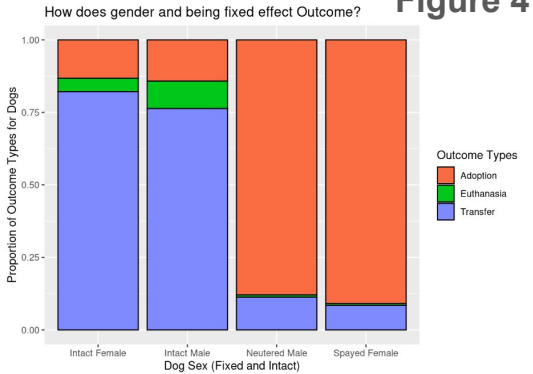


Figure 4



Adoption

- Mean: 21.12
- Median: 12
- IQR: 20

Euthanasia

- Mean: 39.22
- Median: 24
- IQR: 36

Transfer

- Mean: 21.12
- Median: 12
- IQR: 22

Adoption %

- Spayed Female: 80.21 76.39
- Neutered Male: 76.39
- Intact Male: 8.14
- Intact Female: 8.94

Improvements

Factored to create proportions on Figure 4 instead of just calculating adoption rate. Now, we can see proportions for all 3 outcomes. Also added IQR in summary statistics for Figure 3 so we could see the difference in 1Q and 3Q

Logistic Modeling

```
dog_log <- glm(outcome_type ~ age_in_months + sex, data = new_dog, family = "binomial")

summary(dog_log)
```

```
##
## Call:
## glm(formula = outcome_type ~ age_in_months + sex, family = "binomial",
##     data = new_dog)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.686275   0.095245 -17.705   <2e-16 ***
## age_in_months    -0.009826   0.001043  -9.425   <2e-16 ***
## sexIntact Male    0.140801   0.136117   1.034    0.301
## sexNeutered Male  3.999789   0.107438  37.229   <2e-16 ***
## sexSpayed Female  4.297785   0.111872  38.417   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10759  on 9935  degrees of freedom
## Residual deviance:  6696  on 9931  degrees of freedom
## AIC: 6706
##
## Number of Fisher Scoring iterations: 5
```

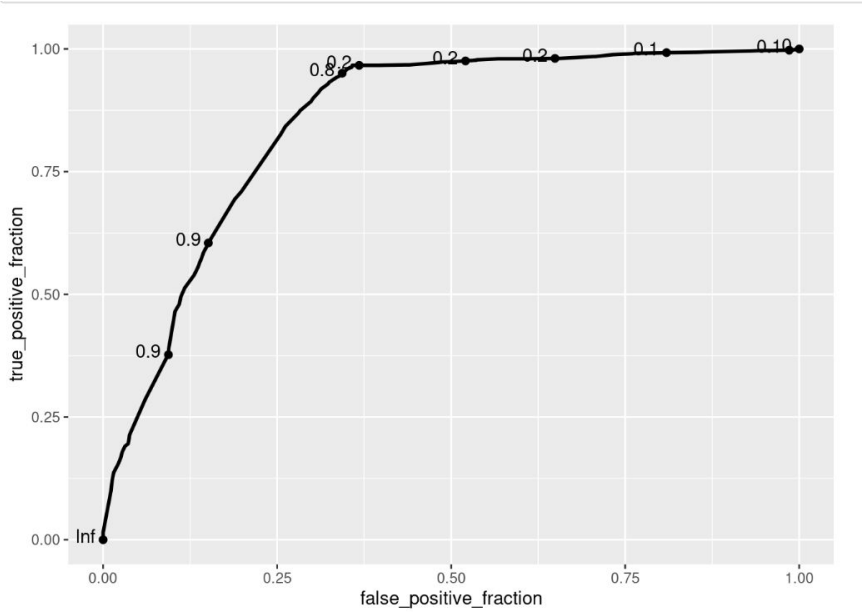
```
new_dog %>%
  ungroup() %>%
  mutate(dog_predictions = predict(dog_log, newdata = new_dog, type = "response"))
```

```
# Calculate performance with AUC
calc_auc(
  # Make a ROC curve
  ggplot(new_dog) +
    geom_roc(aes(
      # Outcome is Survived
      d = outcome_type,
      # Probability of outcome based on the logistic model
      m = predict(dog_log, type = "response")))
  )$AUC
```

AUC = .85

```
## [1] 0.8504174
```

```
# Recalculate predictions and plot ROC curve with grouping
ROC <- new_dog |>
  group_by(breed) |>
  mutate(dog_predictions = predict(dog_log, newdata = cur_data(), type = "response")) |>
  ggplot(aes(d = outcome_type, m = dog_predictions)) +
  geom_roc(n.cuts = 10)
```



Takeaways:

- All variables were statistically significant
- Being a neutered male or spayed female greatly increases the odds of the adoption occurring shown from the summary of the logistic model

Cross Validation

Average Performance over all k folds:

- .8495563

Standard Deviation over all k folds:

- .009752791
- Hardly differing from the average

```
# Choose number of folds
k = 5

# To have the same random sample, use set.seed
set.seed(322)

# Randomly order rows in the dataset
data <- new_dog[sample(nrow(new_dog)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

# Initialize a vector to keep track of the performance for each k-fold
perf_k <- NULL

# Use a for-loop to get performance for each k-fold
for(i in 1:k){
  # Split data into train and test data
  train_not_i <- data[folds != i, ] # train data = all observations except in fold i
  test_i <- data[folds == i, ]  # test data = observations in fold i

  # Train model on train data (all but fold i)
  # CHANGE: what model/predictors should be included
  train_model <- glm(outcome_type ~ age_in_months + sex,
                     data = train_not_i,
                     family = "binomial")

  # Performance listed for each test data = fold i
  # CHANGE: how the performance is calculated
  perf_k[i] <- calc_auc(
  # Make a ROC curve
  ggplot(test_i) +
    geom_roc(aes(
      # Outcome is outcome_type
      d = outcome_type,
      # Probability of outcome_type based on the logistic model
      m = predict(train_model, newdata = test_i, type = "response")))
  )$AUC
}
```

```
## [1] 0.8323028 0.8552639 0.8522660 0.8525919 0.8553571
```

```
# Average performance over all k folds
mean(perf_k)
```

```
## [1] 0.8495563
```

```
# Standard Deviation
sd(perf_k)
```

```
## [1] 0.009752791
```

Results

Univariate Visual Takeaways:

- Most dogs at the Center are on the younger side. From Figure 1 we saw a median age of 12 months and a mean age around 24 months meaning our histogram is skewed right with a majority of dogs on the younger side.
- From Figure 2 we saw that the majority of the dogs in the Center were fixed at 5037 for neutered male dogs and 4401 for spayed female dogs. We noticed intact dogs held the least amount of space in the Center with each less than 2000.

Multivariate Visual Takeaways:

- From Figure 3 we can see that the euthanized outcome has the largest gap in ages with an IQR of 36. Adopted and Transferred dogs are pretty similar in their box plots both with the same medians and means, transferred just has a slightly larger IQR meaning the spread of ages in the middle of the data is bigger.
- From Figure 4 we mainly looked at adoption rates and were able to see that fixed dogs get adopted at much higher percentages. Neutered males at 76. 39% while spayed females were at 80.21%. For intact dogs, both get adopted at a rate of under 10%, but female dogs do get adopted at higher rates than males for both.

Logistic Regression:

- From the summary of the logistic model, the predictors of age and sex are significant. The AUC of our model is at .85 which means the model is relatively strong, but we'd probably want something with a higher performance. The model is also consistent within our all k-folds during cross validation at an average of .84955 which means it generalizes well to unseen data. The average compares almost exactly to the AUC from the logistic model.

Discussion

R1 Response:

- Euthanized dogs tend to be on the older side in the shelter
- Transferred and Adopted Dogs are very similar in age

R2 Response:

- Female dogs get adopted at higher rates, which matched my expectations.
- Fixed dogs get adopted at much higher rates than intact dogs
- When creating a model with both predictors it was able to rank the adopted cases(1) higher than transfer cases(0) in 85% of comparisons.

Major Findings:

- Fixed dogs get adopted at higher rates
- Should consider resource allocation to the Animal Center to potentially get more dogs fixed.

Ethical Issues:

- Working on a dataset with animals comes at a great cost because the most important thing is to prioritize the well-being of the animals. This is why if we were to make decisions like this that affect dog owners and the dogs themselves we would need to take a look at a lot more data.

Future Exploration:

- Considering Austin is a pet friendly city, I would like to see how the animal center data for dogs compares to other cities across Texas. Maybe we could see similar or unexpected different patterns across cities.

State of the Dataset

- Some inconsistencies in the data and missing values

Reflection, Acknowledgement, and References

What did I learn?

- Best to have robust data that tells the whole picture and allows you to see the whole picture. Furthermore, statistical modeling really helps with drawing meaningful insights besides the actual data itself.

Most Challenging Aspect?

- Dealing with inconsistent data which creates formatting issues and also deciding how to work with missing values.

Acknowledgements

- Would like to thank Dr. Guyot and the TAs for their help throughout the data science process, and I would like to thank the City of Austin for providing me with their datasets.

References

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

Intakes: https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm/about_data

Population Facts: https://www.avma.org/news/pet-population-continues-increase-while-pet-spending-declines, by Melinda Larkin

Peer Reviewed Source: The Shelter Charade: The Dilemmas of Urban Animal Control | JSTOR