Animal Shelter Adoption Analysis

Vinay Agarwal
Amber Huang
Elaine Jennings
Joseph Jiang
Aysouda Vatanparast

BANA 212: Data and Programming for Analytics
Dec 7, 2022

Animal shelters play an important role in our community as they constantly strive to reunite pets with their owners. It's often a thankless job for the shelters as they balance housing the animals, encouraging and promoting pet adoptions and motivating people to responsibly spay and neuter their pets with limited resources. An average of 6.3 million companion animals enter U.S shelters each year, out of which 2.7 million are humanely killed due to overcrowding. Our goal is to analyze the animal shelter data and create awareness to more effectively help respond to and identify the animals that need more support to avoid unwanted outcomes.

In this project, we analyzed the population animal adoption behavior, particularly, characteristics of sheltered animals that lead to "positive outcome", i.e. adoption. The data involve various information and characters of animals that are sheltered at The Austin Animal Center, the largest no-kill animal shelter in the United States that "provides care and shelter to over 18,000 animals each year".

First, we reviewed and cleaned the data set. This involved getting familiar with the attributes and values available as well as addressing any null values. Then, we applied various data visualization methods to assist in identifying interesting trends and patterns. Finally, we applied the predictive analysis method Decision Tree to further identify trends and relationships in the data and draw conclusions.

The data set we're using started with 41 columns and almost 80,000 rows. In order to get it ready for the data analysis we cleaned the data by removing a redundant column, searching for blank cells and null values, and renaming the column names to be more user-friendly. Upon further data cleaning, there were 10 rows that were missing an Outcome value. Since there were so few rows missing this variable, these were removed from the dataset and ended with 79,661 rows. The format of the remaining columns was already what we wanted for the analysis, so we were able to leave them as is. The cleaned dataset includes three different types of information: information on the animal, information about the intake, and information about the outcome. Here is a breakdown of the values given for the three different categories:

Information on the Animal		
Attribute	**Values (Either a range, or the largest percents)**	**Unique Values**

Animal ID	Unique ID for each individual animal Ranges from A006100 - A769067	71955
Date of Birth	Date of birth of the animal, some are estimates Ranges from 12/11/1991 - 9/9/2017	5923
Animal Type	Dog (57%), Cat(37%), Other(6%), Bird(<1%)	4
Breed	Domestic Shorthair Mix(29%), Pit Bull Mix (8%), Chihuahua Shorthair Mix (6%)	2155
Color	Black/White (10%), Black(8%), Brown Tabby(6%),	529
Found Location	Either a specific address or just the city name Austin (18%), Outside Jurisdiction(1%), Travis(1%)	36570
Time in Shelter	Time between intake and outcome, includes days and hours. Ranges from 0 days 00:00:00 to 1606 days 04:40:00	29315

In addition to the values outlined above, there are additional columns with the breakdowns of certain values for ease of data analysis. For example, the Date of Birth value is also displayed by only year, only month, and month-year. Time in Shelter is also displayed in fractions of days.

Information on the Intake		
Attribute	**Values (Either a range, or the largest percents)**	**Unique Values**
Age Upon Intake	Age of the animal at the time of the intake Ranges from 1 day to 25 years	46
Intake Condition	Normal (88%), Injured(5%), Sick(4%)	8
Intake Type	Stray (70%), Owner Surrender (19%), Public Assist (6%)	5
Sex Upon Intake	Intact Male (32%), Intact Female (30%), Neutered Male (16%),	5

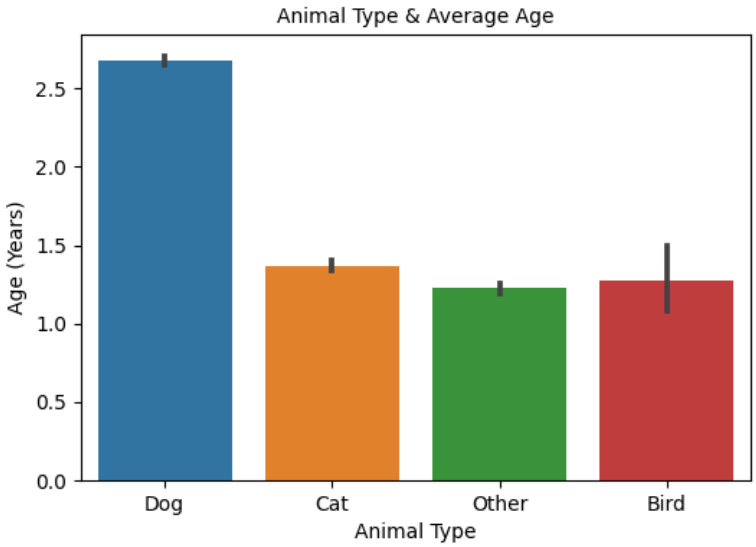
Age Upon Intake Age Group	Grouped bins of the animal age by 2.5 years (0,2.5] - 75%, (2.5,5.0] - 14%, (5.0,7.5] - 4%	10
Intake Datetime	Date and time when the intake occurred: mm/dd/yyyy hh:mm	56738
Intake Weekday	Day of the week of the intake. Saturday (15%), Monday(15%), Wednesday(15%)	7
Intake Number	The number of times the animal has been brought into the shelter. 1(90%), 2(7%), 3(1%)	13

There are additional columns that break down the animals' age upon intake into total days as well as total fraction of a year. The animals' time of intake is also split into more specific categories like month, year, year-month, and hour.

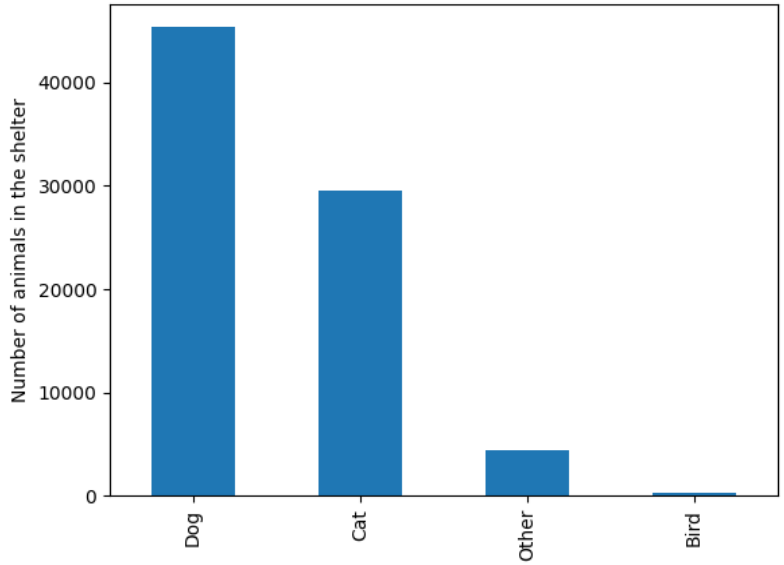
Information on the Outcome		
Attribute	**Values (Either a range, or the largest percents)**	**Unique Values**
Age Upon Outcome	Age of the animal at the time of the outcome Ranges from 1 day to 25 years	46
Outcome Type	Adoption(42%), Transfer(30%), Return to Owner(18%), Euthanasia(8%)	9
Sex Upon Outcome	Neutered Male(35%), Spayed Female(32%), Intact Male(12%)	5
Age Upon Outcome Age Group	Grouped bins of the animal age by 2.5 years (0,2.5] - 75%, (2.5,5.0] - 14%, (5.0,7.5] - 4%	10
Outcome Datetime	Date and time when the outcome occurred: mm/dd/yyyy hh:mm	65677
Outcome Weekday	Day of the week of the outcome Saturday (16%), Sunday(15%), Tuesday(14%)	7
Outcome Number	How many times that animal has had an outcome through the shelter; 1(90%), 2(8%), 3(1%)	13

Again, there are additional columns that break down the animals' age upon outcome into total days as well as total fraction of a year. The animals' time of outcome is also split into more specific categories like month, year, year-month, and hour.

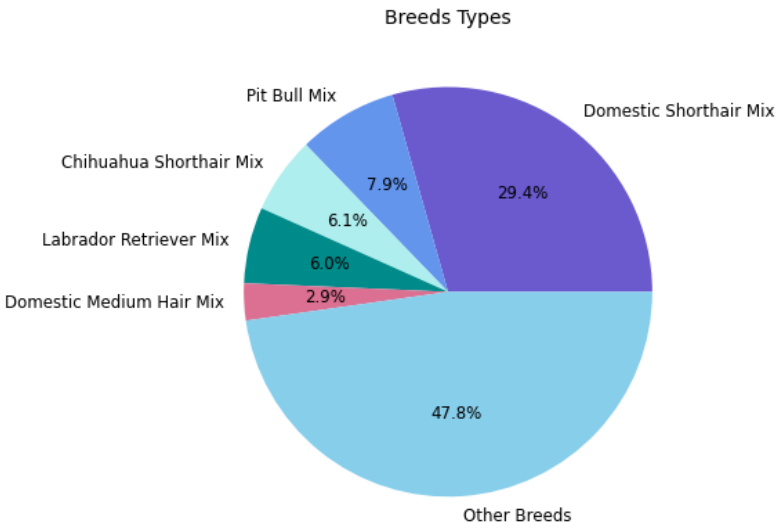
With so much data on so many values we created a number of visualizations to help us get a better understanding of the data. The below plot shows animal types and their average ages upon intake in the shelter:



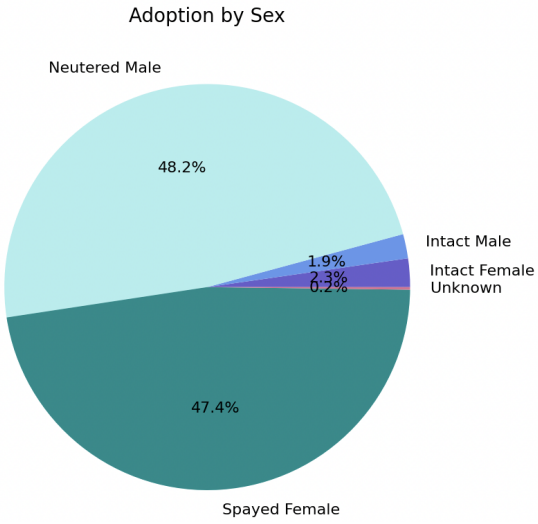
We see dogs to be the type of animals that have the highest age in the shelter, with an average age of ~2.6 upon intake. Cats have the second highest age of animals in the shelter with an average age less than 1.5 upon intake. The below plot shows the numbers of animals types in the shelter that were adopted:



The plot suggests that dogs are the most popular type of animals in the shelter with a total number of more than 4000. Second popular type of animal to be cats with a number of almost 3000. The below pie chart shows the top 5 breeds of animals in the shelter:



According to the chart, we see Domestic Shorthair Max (Cats) to be the most popular breed in the shelter. This is interesting because dogs take the most number of seats in the shelter, however, the most popular breed among all animals in the shelter is the breed of cats. The below pie charts shows a breakdown of the sexes of animals in the shelter:



The most popular two types of sexes among all animals in the shelter are Neutered Male and Spayed Female.

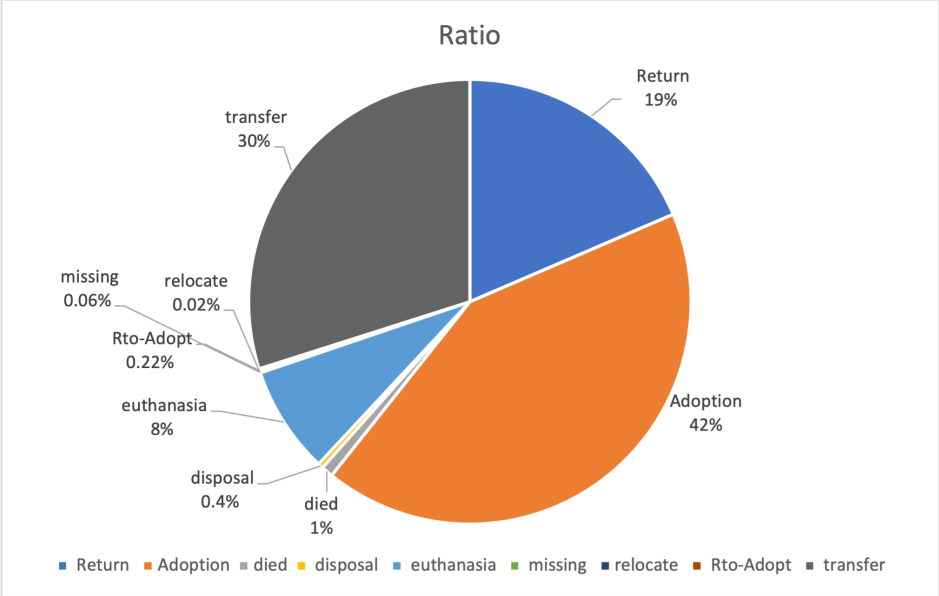
The illustration indicates that the most popular colors are black, brown, and white, resulting in more adoptions of animals in these colors. People usually adopt dogs and cats as pets, and the majority of them are these three colors.

Considering the proportion of adopted animals in each color gives a different conclusion. Rare colors like flame point and agouti have better adoption rates than black and brown. Despite their small number, they're popular.

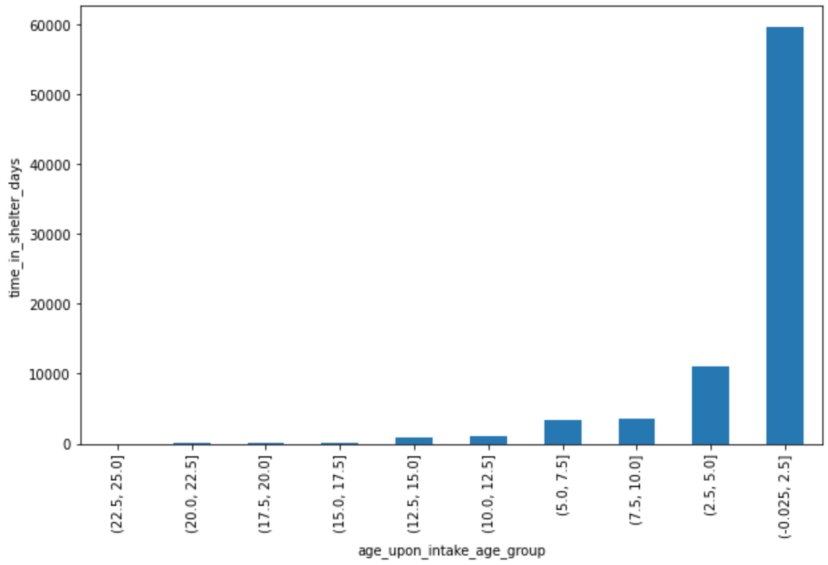
The chart below compares demographics and adoptions by color. The outcome_type_x column shows the overall number of animals for each color, whereas the outcome_type_ y column shows the number adopted.

	color	outcome_type_x	outcome_type_y	Percentage
0	Agouti	16	8	50.000000
1	Apricot	70	32	45.714286
2	Black	19984	8629	43.179544
3	Blue	5975	2667	44.635983
4	Brown	17165	6656	38.776580
5	Buff	562	248	44.128114
6	Calico	1499	696	46.430954
7	Chocolate	1299	594	45.727483
8	Cream	1462	654	44.733242
9	Fawn	592	234	39.527027
10	Flame Point	207	107	51.690821

The ratio of animals returned to owners is 0.185, which indicates 18% were returned; also, adoption has the greatest ratio, at 42%, and transfers are second, at 30%.



In the evaluation of time in the shelter by age, as the age of animals upon intake grows, their duration in the shelter reduces, and this figure for animals older than 15 years is relatively small. Older animals in shelters probably die sooner than younger ones, so they spend less time there. This bar has been sorted based on time spent in the shelter (y-axis).



In order to gain insights about the predictions of the outcome of animals in the shelter, we constructed a decision tree to see what attributes can help predict the outcomes and even help all sheltered animals reach a positive outcome.

Transforming the variables of each attribute into dummy variables helps construct the decision tree using Python. The dataset becomes as follows, with 21 columns and 79661 rows.

```
Data columns (total 21 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   age_upon_outcome_age_group   79661 non-null  object
 1   time_in_shelter_days         79661 non-null  float64
 2   outcome_type                 79661 non-null  object
 3   Aged                         79661 non-null  uint8
 4   Feral                        79661 non-null  uint8
 5   Injured                      79661 non-null  uint8
 6   Normal                       79661 non-null  uint8
 7   Nursing                      79661 non-null  uint8
 8   Other                        79661 non-null  uint8
 9   Pregnant                     79661 non-null  uint8
 10  Sick                         79661 non-null  uint8
 11  Euthanasia Request           79661 non-null  uint8
 12  Owner Surrender              79661 non-null  uint8
 13  Public Assist                79661 non-null  uint8
 14  Stray                        79661 non-null  uint8
 15  Wildlife                     79661 non-null  uint8
 16  Intact Female                79661 non-null  uint8
 17  Intact Male                  79661 non-null  uint8
 18  Neutered Male                79661 non-null  uint8
 19  Spayed Female                79661 non-null  uint8
 20  Unknown                      79661 non-null  uint8
```

To understand what attributes can lead to a higher chance of adoption, and further apply adjustments according to the insights gained from data, characteristics that could be changed after the animals entered the shelter were chosen and those that are already decided when the animal was born were left out. “Intake condition”, “intake type”, “age upon outcome”, “sex upon outcome”, and “time in shelter” were used to construct the decision tree.

Entropy is used as the criteria to measure the split of each node, while the minimum number of samples for splitting is set as 700, which is about 1% of the instances in the dataset. The decision tree has an accuracy of 75.02%, which is not a bad model. The confusion matrix for this model is as follows:

```
[[6309,    0,    0,    0,    0,    0,   93,    0,  358],
 [  10,    0,    0,   29,    0,    0,   12,    0,   83],
 [   1,    0,    0,   38,    0,    0,    1,    0,   17],
 [ 112,    0,    0,  778,    0,    0,   91,    0,  296],
 [   6,    0,    0,    0,    0,    0,    1,    0,    5],
 [   0,    0,    0,    1,    0,    0,    0,    0,    1],
 [ 430,    0,    0,   41,    0,    0, 2014,    0,  426],
 [  21,    0,    0,    0,    0,    0,    3,    0,    4],
 [1600,    0,    0,   34,    0,    0,  266,    0, 2852]]
```

It seems like a great model for classifying those animals whose outcome is adoption. However, overfitting problems might exist, and also the tree is too deep and wide to visualize and be understood by people, further methods are applied to build a new decision tree. A for loop

was used to see what depth and the maximum number of leaves can lead to the highest accuracy of the decision tree. The confusion matrix and accuracy haven't changed much. Following is the visualized decision tree:

Looking into the decision tree, it is easily seen that there are two branches that mainly led to the adoption outcome. The following rules can be generated according to the decision tree:

- Time in shelter < 3.867 & intake type owner surrender & intake condition normal & sex upon outcome not intact female and also not intact male & age group 0-2.5 years old

- Time in shelter > 3.867 & intake type not public assist & sex upon outcome not intact male and also not intact female

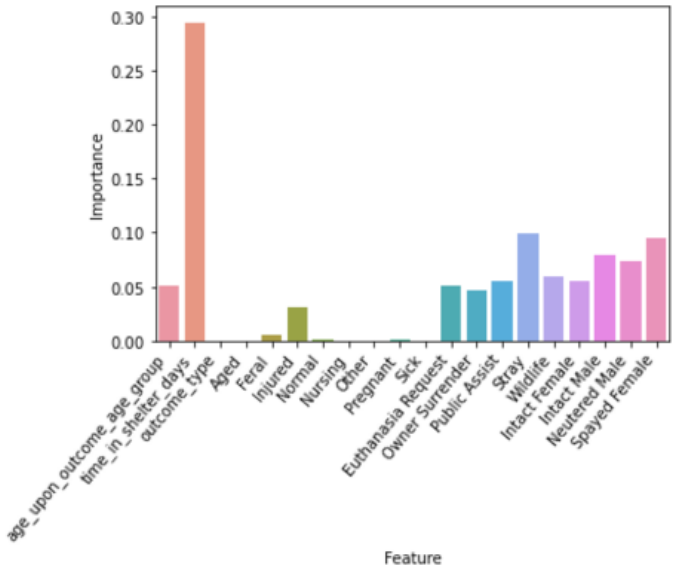
It can be concluded that time in shelter is important when it comes to deciding whether an animal will be adopted or not. However, people are more likely to adopt animals that were surrendered by their owners with normal condition. When it comes to sex and age, people prefer to adopt animals that are intact, no matter male or female, and especially those in a lower age range.

The Decision Tree might end up with overfitting problems, as a result random forest is used to see the results. The accuracy of predicting using the random forest built is 74.27%, which is a bit lower but not much of a big difference compared with the decision tree constructed. The following is the confusion matrix of the model:

```
[[6369,    0,    0,    0,    0,    0,  102,    0,  289],
 [  13,    0,    0,   24,    0,    0,    9,    0,   88],
 [   0,    0,    0,   46,    0,    0,    1,    0,   10],
 [ 128,    0,    0,  639,    0,    0,   95,    0,  415],
 [   6,    0,    0,    0,    0,    0,    1,    0,    5],
 [   0,    0,    0,    2,    0,    0,    0,    0,    0],
 [ 468,    0,    0,    1,    0,    0, 1941,    0,  501],
 [  24,    0,    0,    0,    0,    0,    3,    0,    1],
 [1641,    0,    0,    9,    0,    0,  217,    0, 2885]]
```

This is quite similar to the results we get from the decision tree we constructed. The random forest consists of the feature importance attribute, which gives us a peek into which

features are most predictive of the target variable. The feature importance of the random forest model we construct is as follows:



From the graph above, we can conclude that “time in shelter days” is an important factor in predicting the outcome of the animals in the shelter. While “intake type” and “sex upon outcome” are also both important features of our model compared to other attributes.

In this project, we have analyzed the population’s animal adoption behavior based on the different characteristics of sheltered animals like the type of animal, breed, sex, and color, that leads to a “positive outcome”, i.e. adoption. We were successfully able to visualize our findings using different packages in python.

To take it a step further, we performed predictive analysis using a decision tree and random forest. We conclude that “Time in shelter” is one of the most important factors attributing to the adoption of an animal. The other trailing important factors include the “Intake_type” and “sex upon outcome” of the animal at the time of adoption The efficacy of the decision tree was at 75.02% and that of random forest is at 74.27%, which is a bit lower but not much of a big difference compared with the decision tree constructed.

After our rooted analysis of this dataset by Austin Animal Center, the largest no-kill animal shelter in the United States we can conclude that we have identified the most crucial factors that lead to the adoption of an animal.

References

Jones, Oliver. “14 Animal Shelter Statistics and Facts to Know in 2022.” Pet Keen, 1 Aug. 2022, https://petkeen.com/animal-shelter-statistics/

Nscott. “The Importance of Animal Shelters: Blog: Richell USA.” Richell USA Inc., 25 Nov. 2019, https://www.richellusa.com/the-importance-of-animal-shelters/

SCHLEGEL, AARON. “Austin Animal Center Shelter Outcomes.” Kaggle, 16 Feb. 2018, https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes?select=aac_intakes_outcomes.csv