Analyzing and Visualizing Austin Animal Center Status

Authors: Darwin Lopez, Ryan Chu, Casey Connolly

Summary of research questions:

1. At what age are most animals adopted?

   The average age of animals at adoption is about 23 months, or about 2 years. The age group with the most adoptions is between 0 and 36 months, or 0 and 3 years.
2. What is the most common outcome for an animal?

   The most common outcome for an animal at the Austin Animal Center is adoption, followed by transfer, return to owner, euthanasia, death, and return-to-owner-adoption.
3. What animals are most common at the shelter?

   Dogs and cats are the most housed animals at the Austin Animal Center.
4. Are fixed animals more popular for adoption?

   Fixed animals are overwhelmingly more adopted than pets left intact.
5. Can we create a classifier that will tell us the most likely outcome for an animal based on certain characteristics?

   Our decision tree classifier has very high accuracy on the training set, often between 0.990 and 1, while its testing set accuracy ranged between 0.667 and 0.788 in different trials.

Motivation:

Knowing the answers to the research questions above could allow adoption centers to have a better idea of what to expect when taking in a new animal. If they can predict how long an animal will be with them based on its characteristics, they could take steps in increasing the chance of adoption, as well as take into consideration the capacity of the adoption center when receiving new adoptees. Also, knowing how long an animal will be with them could lead to better long-term medical care and be able to prioritize certain treatments.

Dataset:

We're using two datasets from Austin Animal Center, the first being data regarding animal intakes and the second being the outcomes for those animals. Data collection is still technically ongoing for the center, but the dataset we'll be using looks at intakes and outcomes from June 2021. We had originally planned to look at multiple years of data, but due to the size of the dataset, investigating a single month was the most manageable. In the dataset, each animal is represented by a row. Important columns for the intake dataset include: Intake Type, Animal Type, Age Upon Intake, and Breed. Important columns for the outcome dataset include: Outcome Type, Animal Type, Sex Upon Outcome, Age Upon Outcome, Breed, and Color.

Austin Animal Center Outcomes:

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

Austin Animal Center Intakes:

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

Method:

Question 1: At what age are most animals adopted?

Developing code

1. We will import the pandas, scikits-learn, and matplotlib library.
2. Download and import the dataset for cleaning and merging in python
3. We can use the age at outcome column to group animal outcomes together by age and calculate the average age at adoption.
4. Using a small custom testing data set from the larger dataset to examine the dataset with the intake datetime as the feature and the outcome datetime as the label.
5. Using the dataset we will create a bar graph representing the number of animals in different age groups at time of outcome.
6. Repeat the steps with the actual datasets.

Test: Test another dataset with given and predicted variables

Question 2: What is the most common outcome for an animal?

Developing code

1. We will import the pandas, scikits-learn, and matplotlib library.
2. Download and import the dataset for cleaning and merging in python.
3. Using a small costum testing data set from the larger dataset we can find the outcome types and how many each was and analyze with k-means.
4. We can create a chart or graph that best represents the most to least outcome type for the animals.
5. Repeat the steps with the larger dataset.

Test: Test another dataset with given and predicted amounts.

Question 3: What animals are most common at the shelter?

Developing code

1. We will import the pandas, scikits-learn, and matplotlib library.
2. Download and import the dataset for cleaning and merging in python.
3. Using a small costum testing data set from the larger dataset we can use the animal type for outcome and we can find whether there is a difference between cat and dog outcomes.
4. We can graph our findings by count or ratio of findings.
5. Repeat the steps with the larger dataset.

Test: Test another dataset with given and predicted variables. Using the new dataset and PCA as a testing method.

Question 4: Are fixed animals more popular for adoption?

1. We will import the pandas, scikits-learn, and matplotlib library.
2. Download and import the dataset for cleaning and merging in python.
3. Using a small costum testing data set from the larger dataset we will use the sex upoon outcome column to find a correlation in outcome type.
4. Graph the results.
5. Repeat the steps with the larger dataset.

Test: Test another dataset with given and predicted variables. Using the new dataset.

Question 5: Can we create a classifier that will tell us the most likely outcome for an animal based on certain characteristics?

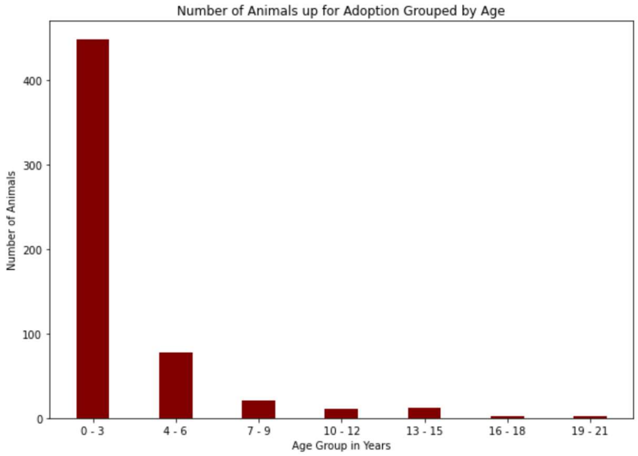
1. We will import the pandas, scikits-learn, and matplotlib library.
2. Download and import the dataset for cleaning and merging in python.
3. Using a small testing data set from the larger dataset to find correlations between characteristics of cats or dogs and find highest correlations with rate of adoption.
4. Graph the results to visualize.
5. Repeat the steps with the larger dataset.

Testing: Test another dataset with given and predicted variables. Using the custom dataset and PCA as a testing method.

Results:

Question 1: What is the average age of animals in the shelter?

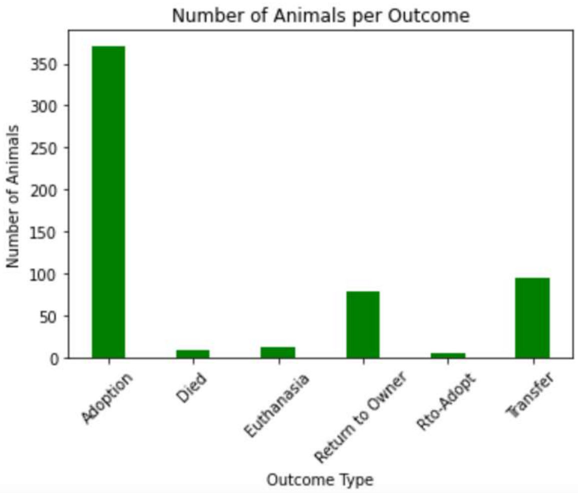
We discovered that the average age of animals up for adoption was 23 months, or about 2 years old. Because this was a rather surface level question, we decided to look deeper into the distribution of animals across age groups of 3 years. This led us to find that the largest age group up for adoption is animals between 0 and 36 months, or 0 and 3 years.



We generally expected the average age and age group of animals at the shelter to be fairly young, especially because adopting kittens and puppies is so popular. This undoubtedly causes our data to be skewed more towards younger animals. In order to keep shelter capacity as open as possible, our data would suggest increasing the adoption rate of younger animals.

Question 2: What is the most common outcome for an animal?

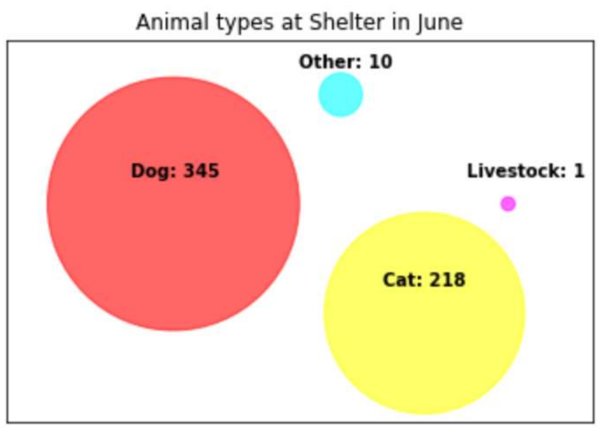
Our data showed that adoption is the most common outcome for animals, followed by transfer and return to owner. Paired with the knowledge that the Austin Animal Center is a no-kill facility, this conclusion is not a surprise.



One limit of this visualization is that we can’t tell which animals are being adopted, transferred, or euthanized. This would give us further insight as to how certain groups of animals could be assisted better.

Question 3: What animals are most common at the shelter?

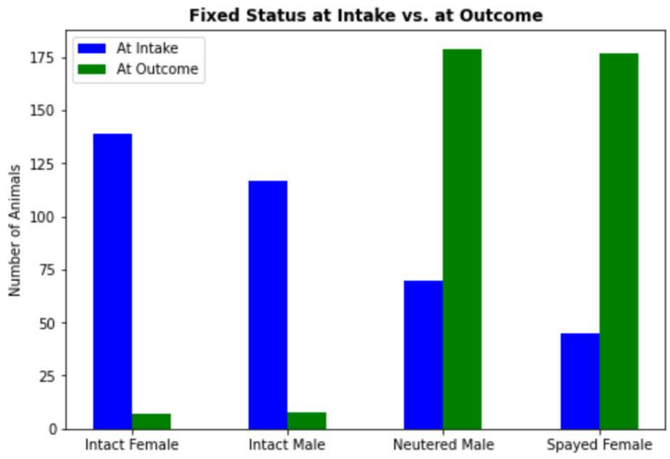
People typically think of animal shelters as places that solely house cats and/or dogs, but making this generalization may impact our conclusions in a biased manner. By comparing the number of different kinds of animals at the shelter during a single month we may be able to interpret what type of care the shelter is best suited to provide.



Data from June of 2021 shows that the majority of shelter space was taken up by dogs, followed by cats and then other animals such as raccoons, guinea pigs, and livestock. While this data may seem surface level, it is important in determining what kind of resources should be allocated to different types of animals. Providing more green space and walking areas for the dogs may be more beneficial than another cat shelter or room.

Question 4: Are fixed animals more popular for adoption?

An overwhelming majority of adopted animals are spayed or neutered. An aspect that we added to this question is what animals' original fixed status is at intake. This showed that nearly all animals who came into the shelter in June, 2021 were not fixed when they arrived at the shelter, but the majority of animals who left the shelter during that same month were.



This shows that the shelter puts in considerable amounts of time ensuring that animals are spayed and neutered. By increasing the number of fixed animals, the shelter is benefiting itself by reducing the likelihood that stray animals can breed. In turn, this reduces the number of stray intakes, especially puppies and kittens, which creates more space for other intakes, such as those with medical needs.

Question 5: Can we create a classifier that will tell us the most likely outcome for an animal based on certain characteristics?

> We were impressed to find that a decision tree classifier did so well in finding what animals would have a certain outcome. While we recognize that having testing accuracy between 0.990 and 1 is high, the depth of these tests resulted in good numbers for our testing set accuracy, which ranged between 0.667 and 0.788. A classifier like this may be helpful for the shelter in determining whether or not they should take in a certain type of animal or number of animals, depending on whether or not the animal can be given a favorable outcome.

Impact and Limitations:

This research helps the shelter and public further investigate the raw numbers they have already collected and how to properly handle the animals they have, as well as what they can expect for future months of intakes and outcomes. The graphs created are important because they can be used to further assist the shelter with materials needed. For example, there was a noticeable amount of cats and dogs being spayed and neutered. Having this data the shelter can apply for further funds towards this. On the other hand, we also found that the shelter could collect further information on what animals are being most euthanized, adopted and transferred (as mentioned in question 2). While many times the shelter has this information, the study helps interpret so that more assistance becomes available for the animals and so that the staff is well prepared to receive and send them off.

A limitation of our study is that we were only able to use one month of data from the animal shelter. The reason why we did this is because when merging data from Intakes and Outcomes, when applied to more than one month, most of the data is removed or unusable due to most animals being adopted within one months, thus not existing on more than a month's dataset. The way to mitigate this limitation is to apply the study month by month and aggregate the data afterwards. That way, much more animals can be included in the results of the study. However, our machine learning model could bypass this limitation

Challenge goals:

- **Multiple datasets**: Our challenge will definitely meet this goal since it is unlikely that one dataset will contain the necessary information to solve our research question. Thus, we will need to join and compare multiple datasets in order to answer our question. For each dataset, given that the features are categorical, we will use one-hot encoding to be able to run analyses on the data.
- **Machine Learning**: Our research question will require some type of correlation between the datasets. We can utilize the machine learning models that we learned in class to predict adoption outcomes such as the time it takes for an animal to be adopted. In addition to the machine learning techniques that we learned in class, we are looking to use Principal Component Analysis (PCA) as well as KMeans clustering to better understand the most important features, observe trends in categorical data, and build more efficient and accurate machine learning models.
- **Messy Data:** The original dataset includes ages for animals in both years and months, and are represented with numbers and strings. This makes any calculations regarding time we plan on asking more difficult and requires us to transform multiple columns so that we are working with only month values.

The two goals that we are more passionate about are data cleaning and machine learning. This is because the products of those outcomes would actually benefit adoption centers. This is because it would allow the adoption centers to predict outcomes and also make the data more understandable to a larger audience.

Work Plan Evaluation:

Proposed:

1. Import datasets and python libraries and clean the datasets by turning categorical data into numeric data by using one-hot encoding. Estimated time: 1 hour.
2. Data visualization: This step involves taking our data and finding useful ways to visualize the data using matplotlib and seaborn. Doing this first will give us better insight into answering the research questions and having a better approach to building machine learning models. Estimated time: 4 hours.
3. Create machine learning models to answer the research questions. We will not only use models that were used in class like Decision Tree Classification, but also PCA analysis and KMeans clustering to compare the models and figure out which model is the most accurate. Estimated time: 5 hours.
4. Testing machine learning models and testing the code itself to ensure the data that is returned is accurate. Estimated time: 3 hours.

Reflection:

The proposed work plan was quite similar to what we expected. The time spent on the project was a fair estimate although we did spread it out into several blocks of times to complete it. Due to schedule changes and finals we had to move around our meeting times but we did end up meeting every week. Due to having messy data the cleanup took longer than expected in the proposal.

Testing:

Because our data is visualization heavy, we decided it would be best to compare output with another dataset. In our testing file we load in and merge two other datasets from the large dataset, specifically data from July 2021. We compared our output to the original dataset to see if our findings were similar, both with calculations and visualizations.

It is important to note that there was no testing done on the ML since the value was in a range that is considered strong. Also, the decision tree's classifications are predicted and therefore cannot be tested since the outcome can change.

Collaboration:

Only team members and course staff were involved in this project.