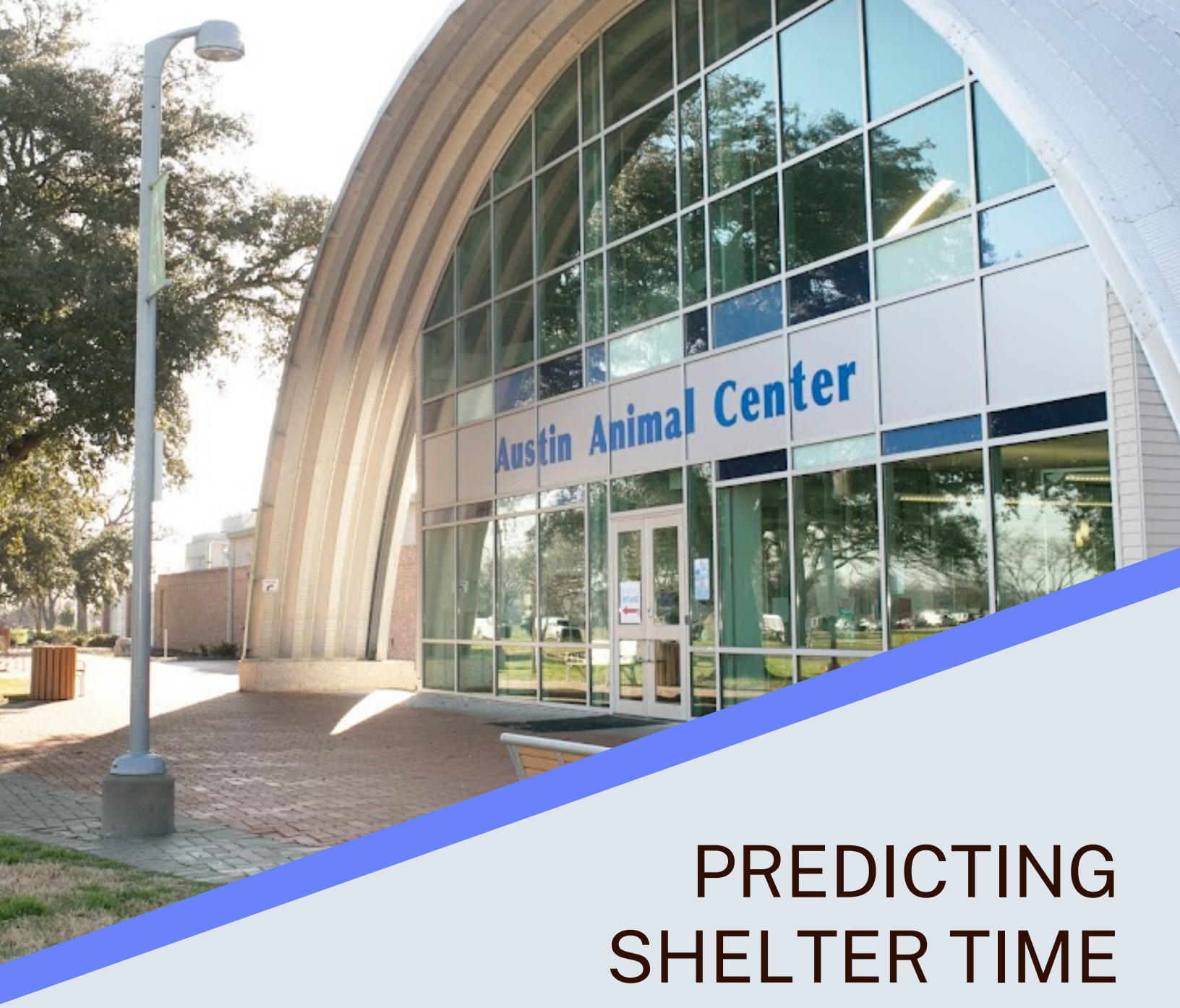



# PREDICTING SHELTER TIME FOR DOGS AT THE AUSTIN ANIMAL CENTER

Tai Tony Tran

Using the intake and outcome data provided by the Austin Animal Center to create a predictive model that predicts the time dogs would spend potentially spend at the shelter

# Introduction

The Austin Animal Center is the municipal shelter for City of Austin. The Austin Animal Center accepts both stray and owned animals regardless of status or species. The shelter prides itself in providing a live outcome such as adoption, foster care, or transfer to another rescue or facility,  to over 90% of the approximately 20,000 animals that come into the shelter each year. The Austin Animal Center hosts adoption events, shelter transfer programs, owner surrender consultations, and an out of state transfer program to help increase their odds of having a live outcome and decreasing the amount of animals left in the shelter.

*Patches*, a dog that waited more than 5 years at the Austin Animal Center to be adopted, and her owner

Due to the large influx of animals accepted into the Austin Animal Center, overcapacity is a major issue that the shelter often faces and makes it difficult for the shelter to continue providing  high rates of live outcomes to the animals it receives. This project aims to help  mitigate that issue for the Austin Animal Center by using features such as name, age, breed, gender, intake condition, and potential outcome to predict the estimated time a dog will spend in the Austin Animal Shelter before its outcome. With an accurate estimated time that each dog will spend in the shelter, the Austin Animal Center can more easily prepare for vacancies in their shelter for future animals.

https://www.austintexas.gov/austin-animal-center

# Data

## About the Data

The Austin Animal Center uploads its intake and outcome data publicly and the data is updated on a daily basis. The data used was split into two different databases, one for intake and one for outcome. Both sets include a unique animal ID that is used to label each individual animal and it is equivalent in both data sets for the same animal. The two datasets were merged upon their unique animal ID to create a large data set containing the following features:

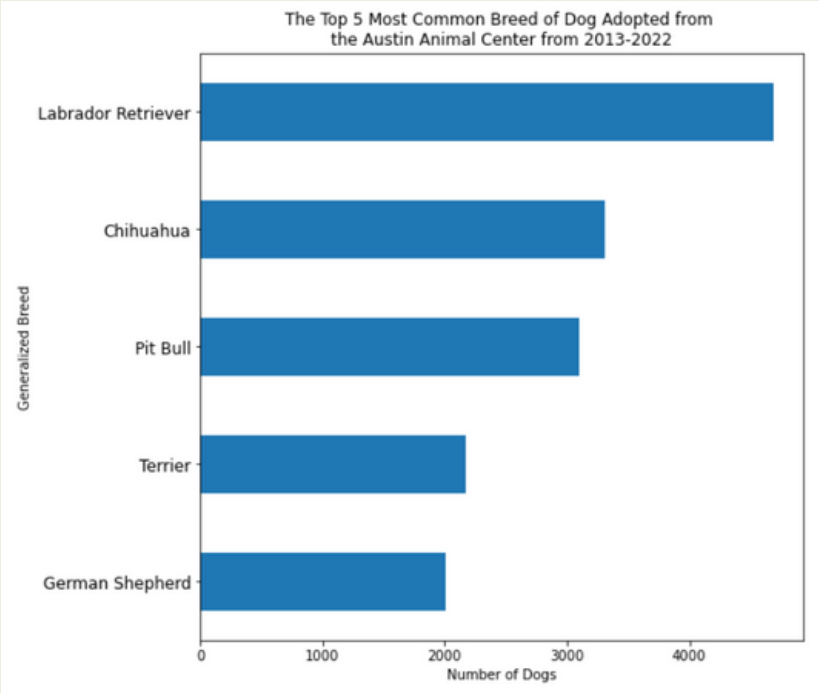
- Animal ID
- Name
- Intake Date/Time
- Intake Month/Year
- Intake Condition
- Animal Type
- Sex Upon Intake
- Age Upon Intake
- Outcome Date/Time
- Outcome Month/Year
- Outcome Type
- Outcome Subtype
- Sex Upon Outcome
- Age Upon Outcome
- Breed
- Color

Each row of the merged data set contains one encounter per animal ID containing both their income and outcome data. Both intake and outcome data sets were extracted as CSV files from their site and the dates of data collection range between October 2013 to April 2022.

Both data sets can be found at https://data.austintexas.gov/

# Data Cleaning

Data is often incomplete and can host a variety of inconsistencies or errors due to incorrect data entry. Upon exploring the Austin Animal Center intake and outcome merged data set, the messy data were presented many issues that would hinder the machine learning process for our predictions.

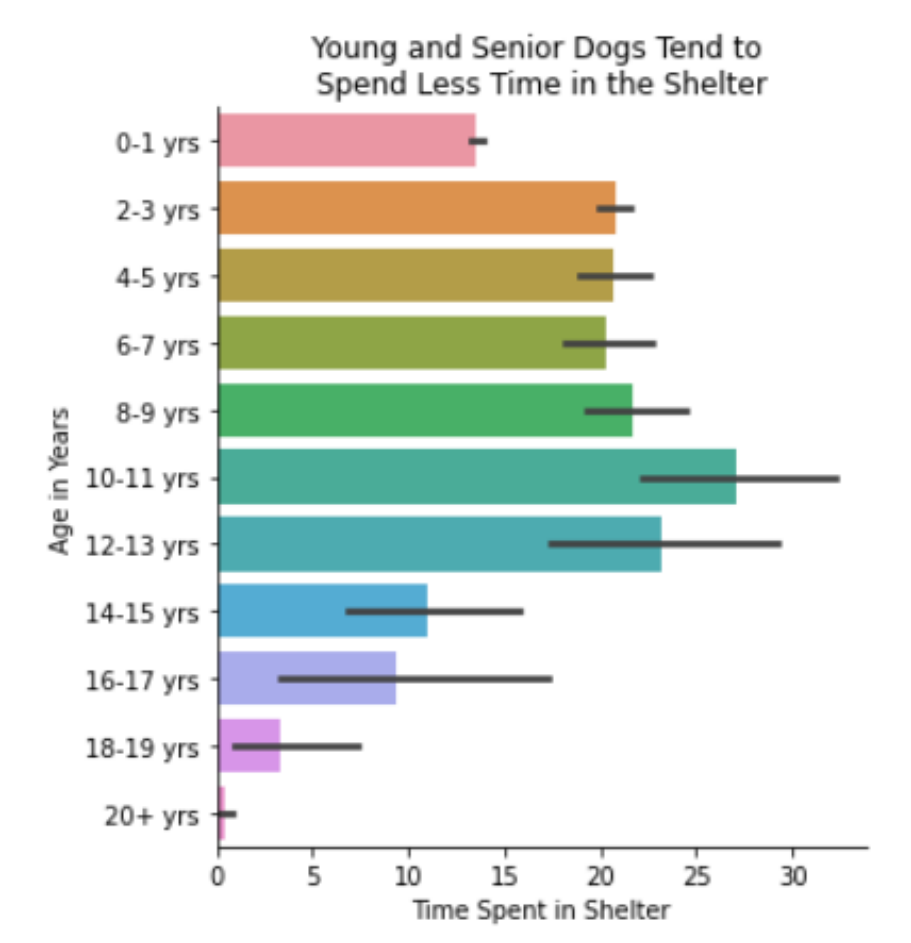



One major problem faced was that the Breed category was not consistent between entries. Breeds, in general, are vague to identify and breed mixes make identification even more difficult. The data set contained over 2000 different breed combinations throughout the entire set. This issue was remedied by binning the breeds into a list of more generalized breeds based off the 50 most popular breeds of 2021 created by the American Kennel Club. For example, if the dog breed was a "Pit Bull / Dachshund Mix", it would categorize into the Pit Bull bin as it is higher up on the list of most popular breeds. Any entries that did not fit any of the breed categories were given a breed category of "Other".

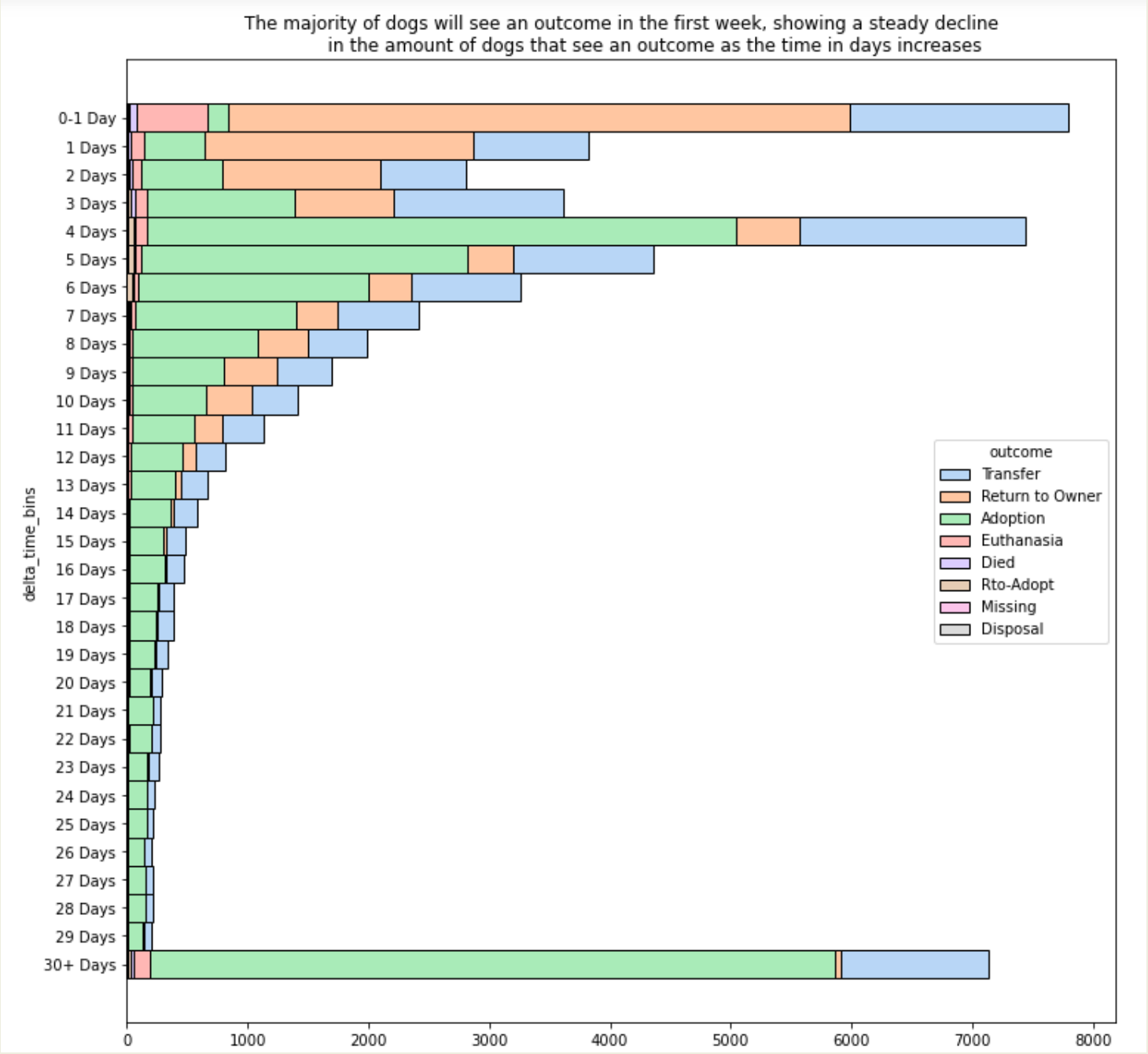
# EDA

I calculated a new column in the table for the time that each dog spent in the shelter by subtracting the outcome date/time column minus the intake date/time column. I begin my exploratory data analysis by investigating this new column and see that dogs at the Austin Animal Center typically see their outcome in an average of around 16 days, with a maximum time over 1500 days, which is over 5 years. This quick exploration shows that although most over 75% of the dogs at the shelter see their outcome in under 2 weeks, there are many outliers that skew the average.

| | Time in Shelter |
|---|---|
| **count** | 49186.000000 |
| **mean** | 16.375851 |
| **std** | 45.472987 |
| **min** | 0.000694 |
| **25%** | 3.023611 |
| **50%** | 5.217361 |
| **75%** | 11.092882 |
| **max** | 1521.979861 |




Age also affects the time that each dog spent in the shelter. It was shown that puppies under 1 year old and senior dogs typically spent less time in the shelter, with the dogs within the ages of 10-13 years spending the most time in the shelter.




Upon exploring the different outcomes, we discovered that on average, dogs were returned to their owners in an average time of 3 days, which was a much smaller time frame than the other outcome types. This can be seen by the decreasing size of the orange bars in the graph above as the days spent in the shelter increases. Dogs that were adopted spent on average 27 days in the shelter. As shown in the chart above by the green sections, most dogs were adopted within the first week and the amount adopted steadily declined after 4 days.

# Algorithm and ML

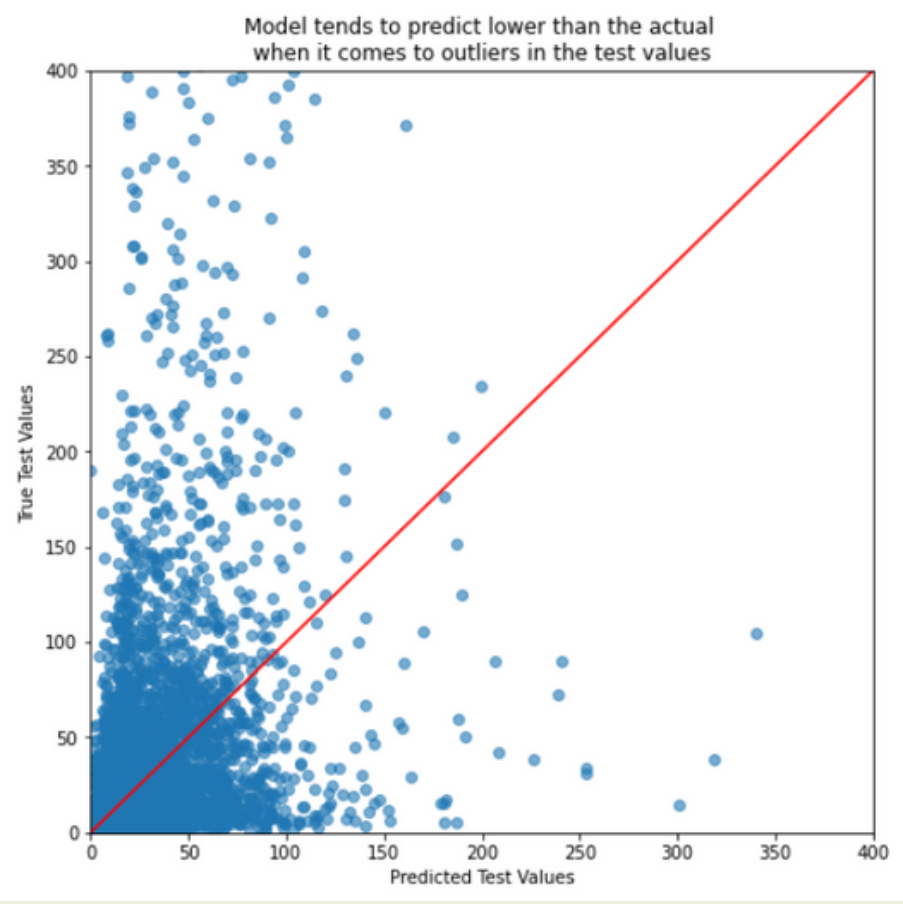
Since the target value of this model was to predict the amount of days spent in the shelter, a continuous value or a number, I chose to go with a regression model. Each model that was tried was evaluated on its mean absolute error (MAE), which calculates the difference between the actual value and the predicted value from the model. I calculated the MAE of the test data with the mean of the training data to form a baseline to compare my models to.

| Model | MAE |
|---|---|
| **Baseline** | 19.685 |
| **Linear Regressor** | 18.562 |
| **Simple Random Forest Regressor** | 16.875 |
| **Dirty Cat Random Forest Regressor** | 16.395 |
| **Dirty Cat histGradientBoostingRegressor** | 15.660 |

I began to use the Dirty Cat library to transform the data before training due to its ability to encode high cardinality categorical data such as the breed categories and names. The use of dirty cat decreased the MAE of the random forest regressor by around 0.5. I decided on the HistGradientBoosting regressor model since it gave the lowest MAE and because it is more easily able to generalize and map missing data values.

# Accuracy

Upon settlings on the HistGradientBoosting Regressor algorithm, the model was used to predict values for the time each dog spent in the shelter. The plot below shows the true values of each dog versus the values predicted by the model. The model seems to have difficulties with predictions for dogs with shelter times that were outliers, more than 1.5 times the IQR. More predictions fell short showing predictions lower than the actual value, than predictions that were greater than the actual, as shown by the vertical rise of points above the red line.




| index | True Value | Predicted Value | Error |
|---|---|---|---|
| 23166 | 1.878472 | 8.245602 | 6.367130 |
| 44047 | 10.035417 | 12.968438 | 2.933021 |
| 50096 | 4.274306 | 13.906139 | 9.631834 |
| 45498 | 3.167361 | 4.023633 | 0.856272 |
| 54750 | 7.081944 | 9.039869 | 1.957925 |
| 29839 | 2.172917 | 8.499728 | 6.326811 |
| 41836 | 4.179167 | 18.607531 | 14.428364 |
| 19928 | 5.835417 | 4.196842 | -1.638575 |
| 8610 | 0.142361 | 2.049705 | 1.907344 |
| 34689 | 0.951389 | 3.770437 | 2.819048 |
| 9877 | 25.178472 | 19.098496 | -6.079976 |
| 49271 | 6.312500 | 32.348648 | 26.036148 |
| 9722 | 7.136806 | 7.804273 | 0.667467 |
| 46578 | 4.002778 | 3.501329 | -0.501449 |
| 43366 | 15.809722 | 38.587128 | 22.777405 |
| 54017 | 5.582639 | 56.585996 | 51.003358 |
| 48499 | 4.863194 | 17.514751 | 12.651556 |
| 20338 | 3.898611 | 9.239796 | 5.341185 |
| 9421 | 4.052083 | 8.210339 | 4.158255 |

# Conclusion

The model created using a histGradientBoosting Regressor model can predict the estimated time a dog spends in the Austin Animal Shelter to a degree of certainty with a mean absolute error of approximately 15 days. Using our model, we predicted that Patches, an injured spayed female pit bull of around 9-10 years old would have a shelter time of over 2 years. Even though the actual amount of time for Patches was over 5 years before she was adopted, it would alert the Austin Animal Center to her situation and allow then to improve her live outcome odds via social media promotions or through shelter exchange programs.


*Cali*, an American Pit Bull Terrier currently at the shelter and waiting for adoption!

## Further Improvements and Research

- Comparison of Austin Animal Center to shelters in other areas
- Research regarding animals at the Austin Animal Center other than dogs
- Include more features to the model to help improve performance regarding outliers (dogs that for some reason stayed in the shelter for long periods of time)
- Allow for more data collection to increase the size of the data set
- A/B testing for their online adoption page to see if certain factors increase adoption rates, thus lowering shelter retention rates