Project Proposal

Amanda, Tony, Harry, Geena

Data Sets

- Austin Animal Center Intakes (https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm/about_data): This dataset logs all of the Austin Animal Shelter Intakes (the animals that enters the shelter), and keeps track of the time the animal was taken in, where it was found, the intake type (stray, owner surrenders, etc), intake conditions of the animal, what type/breed the animal is, age upon intake, and a color description of the animal. The dataset begins from Oct 1st, 2013, and is updated regularly.
- Austin Animal Center Outcomes (https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238/about_data): This dataset mirrors the Intakes dataset, except it records data about the animals that leave the shelter, whether it is because of an adoption, transfer (of shelter/facilities), or euthanasia. This dataset has the same variables as the intakes dataset, with the addition of the outcome type (in place of intake type/condition), and has date of birth as well.

For this project, because of the immense size of the data set, we will be using only data from 2024 (Jan 1-Dec 31, 2024). Our final dataset will join the intakes and outcomes dataset into one that only has information about animals that were taken in and left the shelter in 2024, with the key variables: Intake Type, Intake Condition, Animal Type, Sex (including if they were spayed/neutered), Age, Outcome Type, Outcome Date, Date of Birth, and Length of Stay (which was calculated from the outcome date - intake date).

Here is a sample of our final data set:

```
## # A tibble: 6 × 16
##   `Animal ID` Name         IntakeType     IntakeCondition AnimalType Sex     Age
##   <chr>       <chr>        <chr>          <fct>           <fct>      <fct> <dbl>
## 1 A495162     Mr Manly Man Public Assist  Medical         Cat        Neut…    16
## 2 A510858     Shiva        Owner Surrend… Normal          Cat        Spay…    16
## 3 A557091     Bartina      Owner Surrend… Normal          Cat        Spay…    16
## 4 A557091     Bartina      Owner Surrend… Normal          Cat        Spay…    16
## 5 A566659     Buddy        Stray          Medical         Dog        Neut…    16
## 6 A566837     Chica        Stray          Normal          Cat        Spay…    15
## # ℹ 9 more variables: OutcomeType <chr>, OutcomeDate <date>,
## #   LengthofStay <dbl>, DOB <date>, log_LOS <dbl>, log_Age <dbl>,
## #   sqrt_Age <dbl>, Adoption <dbl>, age2 <dbl>
```

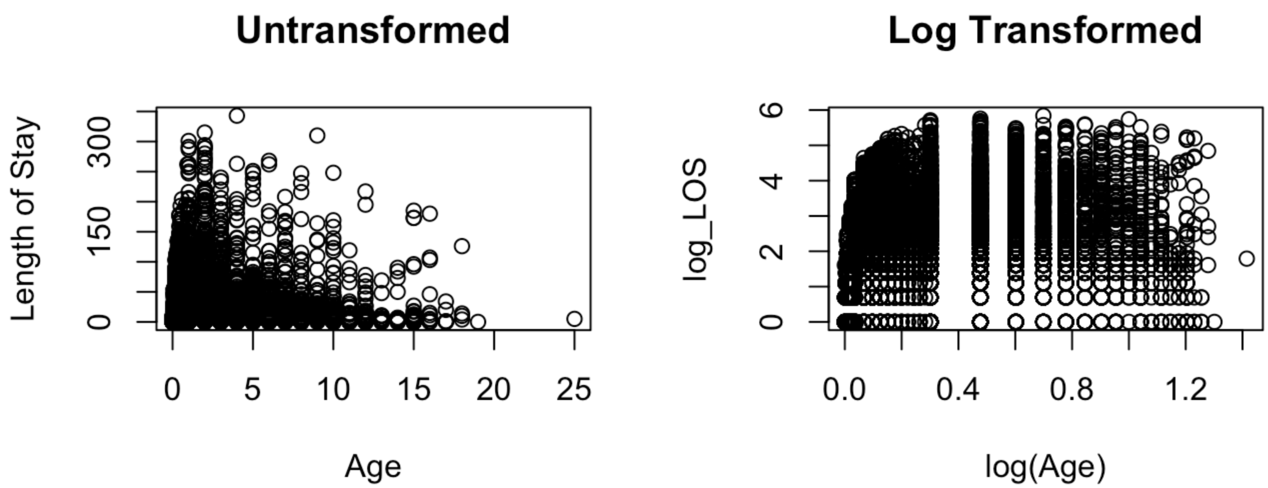
Questions

For each question, you need to include initial results, such as figures/ plots and some simple analyses (e.g., linear and non-linear models, tree-based methods, etc.).

Q1 - Using the available predictors, is it possible to predict the length of stay of an animal that comes into the Austin animal shelter?

Looking at the variables in our dataset, it seems like most of the predictors would have a relationship with the length of stay - especially the animal type, gender, and age, but maybe also how the animal was introduced to the shelter.

There seems to be a pretty extreme skew in both of our numeric variables, so we look at the relationship between the log transforms of both:



We can run an initial linear regression (with the log of both numeric variables) to see if there are any linear relationships.

```
## [1] "Adjusted R Squared 0.385994438482036"
```

As there are several categorical variables with many levels, the summary output is left out of this document. However, our adjusted R-squared value is 0.396, which means our linear model explains less than half of the variability in our length of stay variable. Our RMSE is also fairly low, at 1.063. If we use the AIC step criterion to select our best subset, we are still given the full model with all of the predictors (log_Age + AnimalType + IntakeCondition + Sex + OutcomeType).

```
## [1] "Mean RMSE from k=10 CV:  1.06424876386094"
```

The Cross-Validation to our linear model also produces a mean RMSE of 1.06, which means our model does not overfit too much, and does fairly well at predicting new data compared to the RMSE of the training data, which is good.

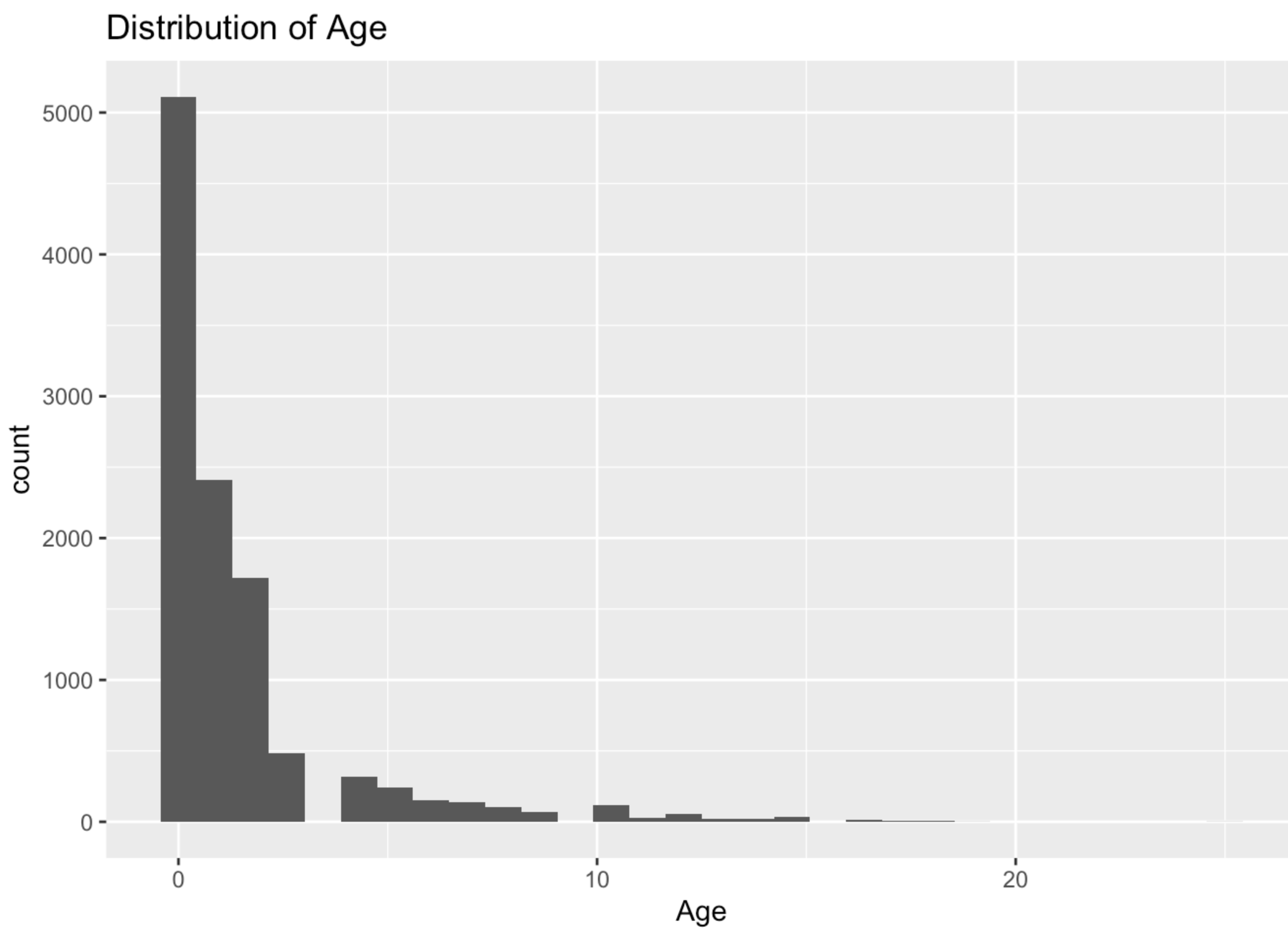
For future considerations, since our coefficients are all fairly small and pretty equal in magnitude, we could implement ridge or lasso regression in another attempt to perform feature selection so that the more important predictors are weighted more in our model, which can hopefully tell us more information. Additionally, since there are so many categorical predictors in our data set, perhaps we could use decision trees or a step function to better work with these categorical variables.

Q2 - Can we predict adoptability based on the age upon intake?

```
mean(shelter_clean$Adoption)
```

```
## [1] 0.6103414
```

About 61% animals were adopted in the dataset.



We can see from the graph that the data is skewed heavily left. Let's fit a logistic model to the data.

```
##
## Call:
## glm(formula = Adoption ~ Age, family = "binomial", data = shelter_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.62524    0.02354   26.55   <2e-16 ***
## Age         -0.11009    0.00811  -13.57   <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: 14767  on 11042  degrees of freedom
## Residual deviance: 14571  on 11041  degrees of freedom
## AIC: 14575
##
## Number of Fisher Scoring iterations: 4
```

The output gives the logit-form of the model which is: ln(p hat / 1-p hat) = 0.62524 - 0.11009 * Age, where p hat is the probability of the animal being adopted (1 = adopted).

Here is a visualization of how our logistic regression is categorizing our variables. As you can see, our model is making a lot of errors. Maybe in the future we could consider more predictors, or try another classification method such as a classification tree.