For this problem, we created a graph with linear regression lines for each breed. This was done because we selected age as the most significant factor to consider due to the domain knowledge we gathered while analyzing the data. Sine this one factor was what we wanted to analyze, and the outcome was a binary outcome, we used linear regression to find the probability of a positive outcome (adoption) based on age. The graph is limited to dogs, since monitoring multiple species proved to be too complex.

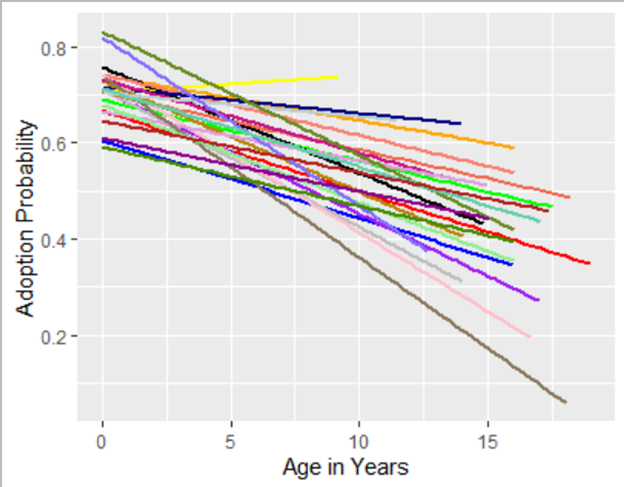
To create this linear regression graph, the OutcomeType (whether an animal was adopted, not adopted, euthanized, or transferred) was converting to binary in an R function. Thus, 1 stood for adopted and 0 for any other outcome. Repeated animal ID’s were then removed (these represented dogs that ran away and were returned afterwards) as well as a few animals that had their birth date and arrival date reversed. These animals were recorded as arriving at the animal shelter before being born, and no explanation was provided by the Austin Animal Center. Thus, it was left to us to derive a conclusion.

The most likely case is the animal shelter made an error in inputting the data, perhaps flipping the arrival date and birth date on accident. Another possible conclusion would be the mother arrived at the shelter, then gave birth to the pup. Finally, an input error switching the day and month could have also been to blame. Regardless, since there was no way for us to know the reason, we removed these observations from our model.

Another issue was recording the age at adoption. The age at adoption was recorded as %y years, %m months and %d days. This changed animal to animal, with some only have years, or animals less than a year old only had months.

To convert the age to numeric Joshua made a function in R that would take the String and depending on its value assign an age to the animal. ((if year || years) #assign number of years)

Unfortunately, this function came with a lot of glitches, and we instead made a simple excel function in the csv that would give age as a numeric value based on year. (1.25 = 1 year and 3 months). This numeric style made it much easier to make calculations and graph the age.

Below is the resulting graph. 

Legend (would not fit on graph):

Pit Bull Mix = blue

Chihuahua Shorthair Mix = red

Labrador Retriever Mix= green

Anatol Shepherd Mix = yellow

Australian Cattle Dog Mix = orange

Australian Shepherd Mix = purple

Border Collie Mix = black

Boxer Mix = gray

Cairn Terrier Mix = salmon

Catahoula Mix = navy

Chihuahua Longhair Mix = coral

Dachshund Mix = pink

German Shepherd = darkgoldenrod

Jack Russell Terrier Mix = medium violet red

Miniature Poodle Mix = plum

Miniature Schnauzer Mix = navajo white

Pointer Mix = azure

Rat Terrier Mix = light green

Siberian Husky Mix = light blue

Staffordshire Mix = olive drab

Yorkshire Terrier Mix = dark magenta

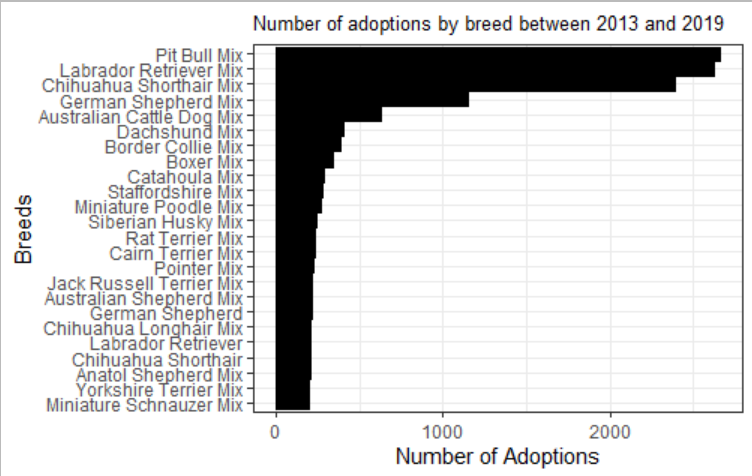
German Shepherd Mix = aquamarine

Labrador Retriever = chartreuse

Chihuahua Shorthair = firebrick

Each color represents a breed, with x = age in years, and y = adoption probability.

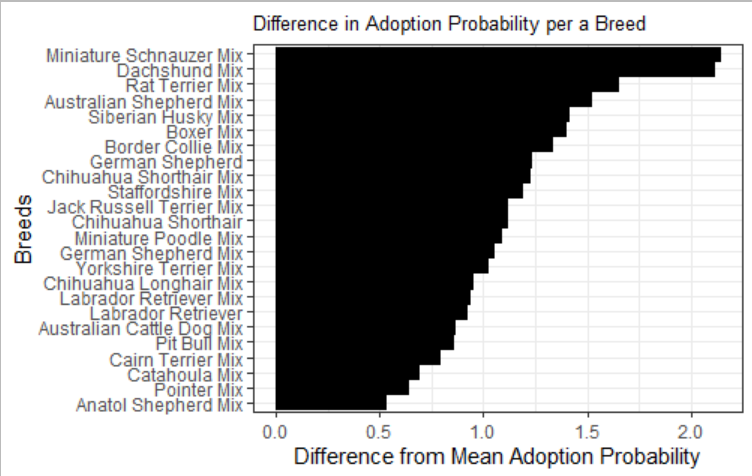
We chose to limit our graph to only the top 24 breeds, since according to our frequency chart: (there are 1473 breeds so this is just a snippet)



The most frequent breed was the Pit Bull Mix with a frequency of 2,663. Thus, the Miniature Schnauzer Mix is the 24th most popular breed at the shelter. After the Schnauzer, the breeds dramatically begin to decrease in count, as the median frequency of all the breeds was only 16.34.

We decided to limit our data to just the top 24 breeds, or breeds with a frequency over 200, as these breeds had enough quantity to make more accurate predictions and made up roughly 50% of the overall adoptions. Also,we felt that showing more breeds would overcomplicate the graph. However, we still found 24 breeds to be very difficult to graph. As the multiple linear regression shows, it is almost unreadable.

For our next step we calculated the mean adoption age regression line, then took the difference in the slope of the mean regression line and each breed’s regression line and plotted it in a histogram.



On this graph, the bars near 1.0 represent almost no difference from the overall relationship between age and adoption probability from the average. Notice that the Miniature Schnauzer and Dachshund Mix are much less likely to be adopted in their later years than the other breeds. This could be explained as the puppy phenomenon combined with their cult following. Both of these breeds are small dogs that possibly attract people with a more serious desire for the specific breed, which may lead to them desiring a puppy they can raise from a young age rather than an older dog.

In addition, the Anatol Shepherd Mix is the only breed that has a positive slope (yellow line on regression curve). This could be because the Anatol Shepherd is commonly used as a ranch dog, meaning there is an inherent demand for the breed in an area like Austin, Texas where huge ranches are plentiful. Older dogs could potentially already be trained, and since the breed does not suffer much in performance as they age, it would make sense that this breed would be very popular in this part of the country.

Also, remember only the Anatol Shepherd Mix is in the top 24 dog breeds. Thus, the pure breed Anatol Shepherd is not available in enough quantity to view, but would assumedly have a similar positive slope.

Finally, we would like to point out that most of the top 24 breeds are mixes. This makes sense as pure breeds are generally far less common due to the inherent difficulty of keeping a pure breed lineage, and most of these animals being acquired from professional breeders..

Finally, we calculated the error rate for our model. We approached this calculation based on the k-fold method. For our folds we used the subset of breeds and calculated the difference between the prediction of the whole set and compared it to the outcome of the subsets, and took the average of the squared errors. This gave us an error rate of .1478, or about 15%. This told us that age is the source of 85% or the variance in adoption rate, and the other 15% is from other factors, most likely breed. However, with the extreme variance in breed type, and some breeds having such low numbers represented, our attempts at testing the error of both of these predictors was outside of our models ability. This error rate tells us that our model is fairly accurate at determining whether or not a dog will be adopted given its age, but could be more accurate if we were able to include more factors.

Finally, to give a few use cases for this model, we can look at the difference between two dogs of different breeds but of the same age, the difference between a dog of the same breed but different ages, and the probability that a random dog rescued from the street will be adopted given its age and breed.

Case 1: If we take a Chihuahua Shorthair Mix and a Border Collie Mix both five years of age, our model can tell us that they have a 58% and 62% chance of being adopted respectively. This tells us that the Border Collie has a slightly higher chance of being adopted and can give the shelter owners the knowledge that they should probably try to push the chihuahua a little harder in order to get it adopted.

Case 2: In this case, let’s consider a Border Collie Mix puppy and a ten-year-old dog of the same breed. From this model we can see that the puppy has a 76% chance of being adopted whereas the ten-year-old has a 55% chance. So the shelter can make sure to put aside more resources for the adult dog since it’s likely to stay in the shelter longer.

Case 3: Using this model a shelter can also determine how likely any new dog they take in is to be adopted. This can provide much needed data for budgeting and planning expansions of the shelter if necessary. In this case, let’s consider a seven-year-old Pit Bull Mix. This dog has a 50% chance of being adopted. Being so low, the shelter can then prepare to keep this dog on a long-term basis.