STATS 15 - Austin Animal Center Dataset Report

Thomas Peeler, James Cobb, Ciara Smith, Mackenzie Smith

2022-12-08

Section 1: Background and Explanation of Data

Introduction and Motivating Question

In the United States alone, 6.5 million dogs, cats, and other former pets are abandoned or lost every year and end up in shelters. In reality, only 3.2 million of these animals are adopted, while others who are not adopted within a certain time frame are euthanized. The Austin Animal Center, however, objects to this quota by being the largest no-kill animal shelter in the United States and providing care and shelter to over 18,000 animals each year. In contrast to other shelters, the animals that are not adopted within a certain time period are either transferred, returned to their original owner, and in very rare cases, euthanized. In this project, we will seek to answer the motivating question: *what attributes affect how quickly a dog is adopted?*

Background

The Austin Animal Center

The Austin Animal Center is the largest no-kill animal shelter in the United States that provides care and shelter to over 18,000 animals annually. They accept strays and owned animals regardless of age or health, and they accept all species and breeds of animals. However, for our analysis, we will be solely looking at the intake and outcomes of dogs at the center. As part of the AAC's efforts to help and care for animals in need, the organization makes available its accumulated data and statistics as part of the city of Austin's Open Data Initiative. The AAC datasets that we will be analysing outline the intakes and outcomes of animals at the center, of which many happen each day.

Intakes and Outcomes

An *intake* refers to an instance of an animal being admitted to the shelter, whether it be stray or owned. An *outcome* refers to an animal leaving the shelter for any reason, whether they are adopted, transferred to another shelter, or pass away while staying in the shelter. Our data consists of two separate tables, one detailing outcomes and one detailing intakes. Each animal has a unique ID that is shared between the two tables.

How Animals Come Into the Shelter

Some are brought in by owners who can no longer keep the animals. Others are found roaming the streets and brought in by animal control officers. In an ideal situation, an animal will only stay in a shelter until its owner returns or it is adopted. However, if that is not the case and they don't have enough room to indefinitely house all of the animals they receive, they will transfer them to a partner location.

Data Structure

The intakes table has one entry for each instance of an animal coming into the shelter, and the outcomes table has one entry for each instance of an animal leaving the shelter. Each entry has details about the intake and outcome situations for the animals, which are specified in the "Variables" section. In our dataset, we will only be looking at dogs. Before any cleaning, the intakes and outcomes tables had a total of 145,851 entries and 145,914 entries, respectively, with 12 variables each.

Importance of the Data

The importance of this data lies in the fact that it not only identifies characteristics people are looking for in dogs but also helps shelters that house dogs of all kinds decide whether to keep animals that exhibit certain characteristics at their current location or move them to a shelter that specializes in such characteristics. Suppose an elderly dog was staying at a shelter that couldn't get adopted due to a trend of people not adopting elderly dogs. If this was the case, the shelter might be better off transferring the dog to a shelter that is specifically designed for elderly dogs, where people looking to adopt elderly dogs are more likely to adopt the dog.

Variables

- Note: the variables started off as being in two separate tables, but the tables will be binded together during cleaning

Explanatory Variables

Age.upon.Intake and ***Age.upon.Outcome***: (char) the dog's age at the time of the intake/outcome; an estimate if the dog is a rescue

Breed: (char) the dog's primary and secondary breed (if it has one)

Color: (char) the dog's primary and secondary color (if it has one); the one or two main colors that make up its coat

Response Variable

Outcome.Type: (char) the specifics of the intake/outcome situation for the dog; we will be looking only at adoptions, and specifically, how long it takes before dogs are adopted from the shelter.

Section 2: Data Loading and Cleaning

```
## Warning: package 'rsample' was built under R version 4.2.2
```

```
## Warning: package 'scales' was built under R version 4.2.2
```

First, we load our data in.

```r
austin_intakes <- read.csv("aac_intakes.csv")
austin_outcomes <- read.csv("aac_outcomes.csv")
```

Then, we of course need to limit our dataset to just dogs.

```r
outcomes_filter <- austin_outcomes %>%
  filter(Animal.Type=="Dog")
intakes_filter <- austin_intakes %>%
  filter(Animal.Type=="Dog")

outcomes_filter %>%
  select(Age.upon.Outcome, Animal.Type, Breed) %>%
  head()
```

```
##   Age.upon.Outcome Animal.Type                                Breed
## 1           1 year         Dog              Chihuahua Shorthair Mix
## 2         4 months         Dog   Anatol Shepherd/Labrador Retriever
## 3          7 years         Dog              Chihuahua Shorthair Mix
## 4          2 years         Dog American Foxhound/Labrador Retriever
## 5          2 years         Dog   Border Collie/Cardigan Welsh Corgi
## 6          2 years         Dog                             Pit Bull
```

Then, we need to eliminate all of the unused columns. The `Animal.Type` column will be unused since we've already filtered to just dogs.

```r
outcomes_s <- outcomes_filter %>%
  select(Animal.ID, Age.upon.Outcome, Breed, Color, DateTime, Outcome.Type)
intakes_s <- intakes_filter %>%
  select(Animal.ID, Age.upon.Intake, Breed, Color, DateTime, Intake.Type)

outcomes_s %>%
  select(Age.upon.Outcome, Breed, Color) %>%
  head()
```

```
##   Age.upon.Outcome                                Breed       Color
## 1           1 year              Chihuahua Shorthair Mix White/Brown
## 2         4 months   Anatol Shepherd/Labrador Retriever        Buff
## 3          7 years              Chihuahua Shorthair Mix       Brown
## 4          2 years American Foxhound/Labrador Retriever White/Brown
## 5          2 years   Border Collie/Cardigan Welsh Corgi Black/White
## 6          2 years                             Pit Bull  White/Blue
```

Then, we'll be binding the two tables together for the sake of analysis. An extra column will be added that will specify if each entry is an outcome/intake.

```r
o1 <- outcomes_s %>%
  rename("Age"="Age.upon.Outcome",
         "Type"="Outcome.Type")
o1$Action <- "outcome"

i1 <- intakes_s %>%
  rename("Age"="Age.upon.Intake",
         "Type"="Intake.Type")
i1$Action <- "intake"

aac <- rbind(o1, i1) %>% arrange(Animal.ID)
aac %>%
  select(Age, Breed, Color) %>%
  head()
```

```
##        Age                Breed        Color
## 1  7 years Spinone Italiano Mix Yellow/White
## 2 10 years Spinone Italiano Mix Yellow/White
## 3  6 years Spinone Italiano Mix Yellow/White
## 4  7 years Spinone Italiano Mix Yellow/White
## 5 10 years Spinone Italiano Mix Yellow/White
## 6  6 years Spinone Italiano Mix Yellow/White
```

Now, there is some cleaning to do in the `Type` column: some outcome/intake types are very similar to each other, and it will make more sense for our analysis if we lump them all together. Specifically,

- `Rto-Adopt` is the same as `Return to Owner`
- `Wildlife` and `Abandoned` are the same as `Stray`
- `Disposal` is the same as `Died`
- `Stolen` is the same as `Lost`.

```
aac <-
  aac %>%
  mutate(Type=case_when(
    Type=="Rto-Adopt" ~ "Return to Owner",
    Type=="Stolen" ~ "Lost",
    Type=="Disposal" ~ "Died",
    Type %in% c("Wildlife", "Abandoned") ~ "Stray",
    Type==Type ~ Type
  ))
```

Then, we'll use `lubridate` to add a column for all of the `DateTime` values converted to `date` instead of `char` objects, which will make it easier to compare them.

```
aac <- aac
aac$DateTime <- mdy_hms(aac$DateTime)

aac$DateTime %>%
  head(3)
```

```
## [1] "2014-12-20 16:35:00 UTC" "2017-12-07 00:00:00 UTC"
## [3] "2014-03-08 17:10:00 UTC"
```

Next, note that the `Age` column is made up of character strings at the moment.

```
aac$Age %>%
  head()
```

```
## [1] "7 years"  "10 years" "6 years"  "7 years"  "10 years" "6 years"
```

Such a variable will be much easier to analyse as an integer, so we will convert these to doubles representing the dog's age in years.

```
conv_to_num <- function(string)
{
  conv <- data.frame(
    c("week", "month", "year"),
    c(7.0, 30.0, 365.0)
  )
  tmp <- 1
  for (n in (1:3))
  {
     if(str_detect(string, conv[n, 1]))
    {
      tmp <- conv[n, 2]
    }
  }
  ret_str <- str_extract(string, "^[0-9]+")
  ret_dub <- as.double(ret_str)
  return(round((ret_dub*tmp)/365.0, 3))
}

aac$Age.New <- map_dbl(aac$Age, conv_to_num)

aac <- aac %>%
  select(-c(Age)) %>%
  rename("Age"="Age.New") %>%
  relocate(Age, .before=Breed)

aac$Age %>%
  head()
```

```
## [1]  7 10  6  7 10  6
```

Then, we'll simply remove all duplicate entries in our dataset.

```
aac <- aac[!(duplicated(aac)), ]
```

Lastly, we'll be implementing a new column to denote the time between the last action and the one in the entry; this will be our response variable, specifically for adoptions. The first entry for each dog will have an NA in this column.

```
len <- function(d1, d2)
{
  retval <- int_length(interval(d1, d2))
  return(retval)
}
aac <-
aac %>%
  arrange(Animal.ID, DateTime) %>%
  group_by(Animal.ID) %>%
  mutate(Num.Action=row_number()) %>%
  ungroup() %>%
  mutate(Days.From.Last.Action=case_when(
    (Num.Action > 1) ~ len(lag(DateTime), DateTime)/86400
  )) %>%
  select(-c(Num.Action))
aac %>%
  select(Animal.ID, Action, Days.From.Last.Action) %>%
  head()
```

```
## # A tibble: 6 x 3
##   Animal.ID Action  Days.From.Last.Action
##   <chr>     <chr>                   <dbl>
## 1 A006100   intake                 NA
## 2 A006100   outcome                 1.11
## 3 A006100   intake                286.
## 4 A006100   outcome                 1.26
## 5 A006100   outcome              1082.
## 6 A006100   intake                  0.588
```

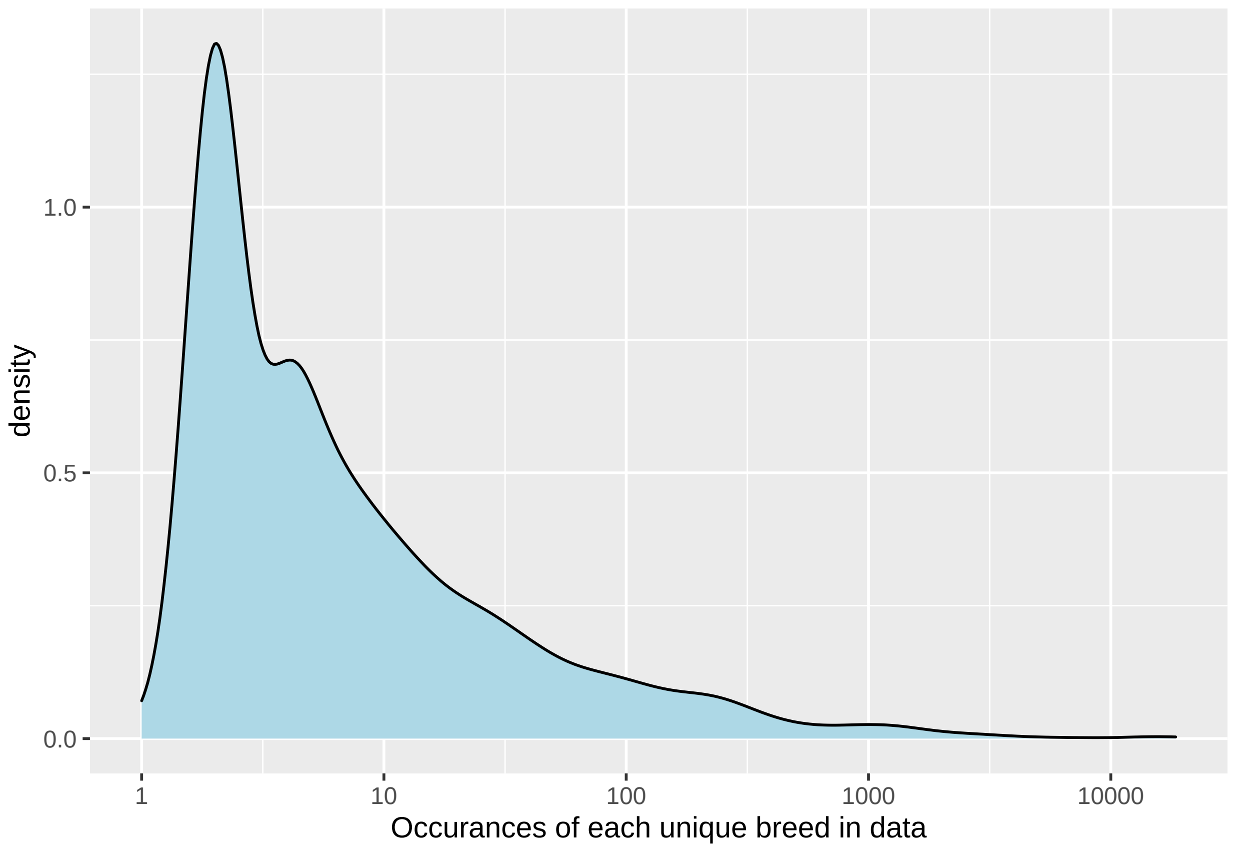
Section 2.1: Univariate Exploratory Analysis

Breed

To start, we'll be looking at the `Breed` column.

It currently has the problem of being a bit too specific, where some breeds are unmixed, some are just labeled "Mix" while others have the breed mix specified, among other naming specifics; this leads to a large majority of the breed categories having <10 entries.

```
aac %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=n)) +
  geom_density(fill="light blue") +
  scale_x_log10() +
  xlab("Occurances of each unique breed in data")
```



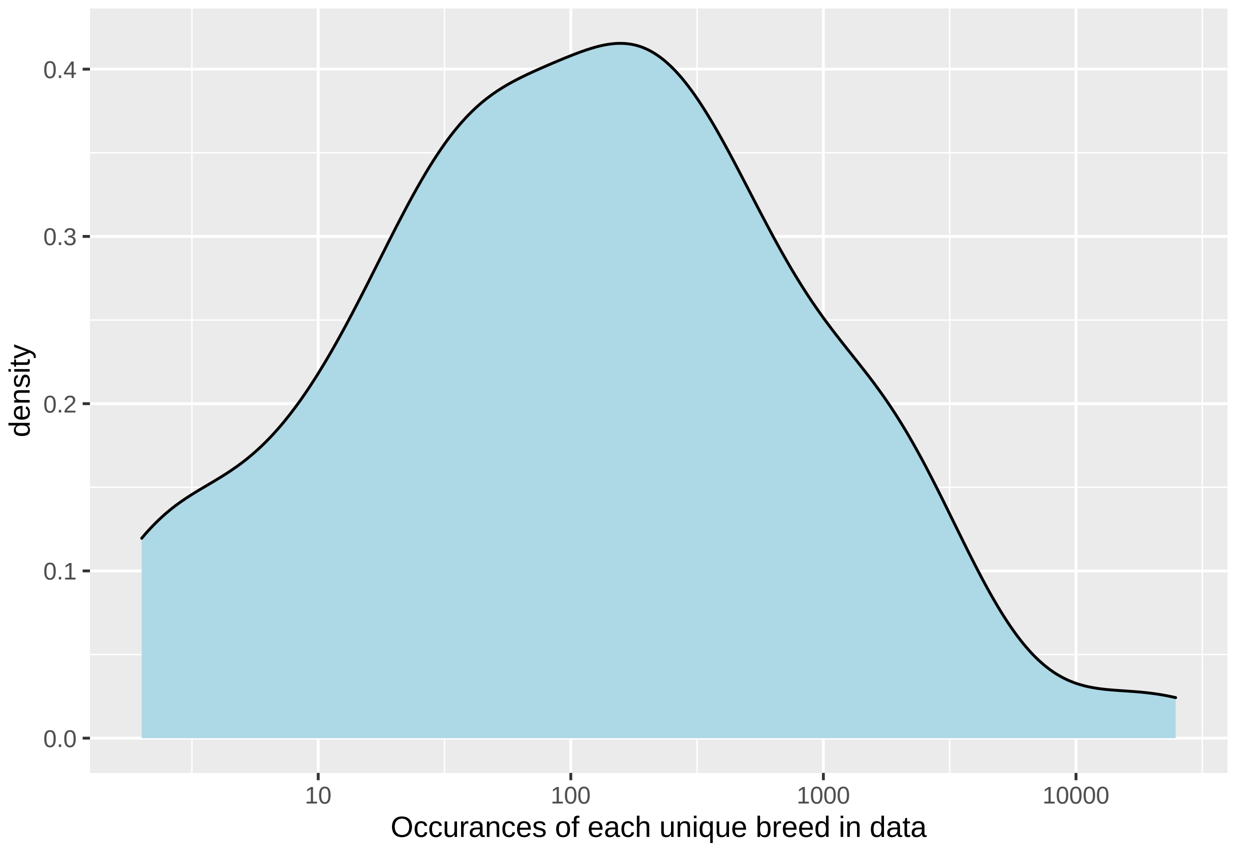
```
aac %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  summarise(breeds=n())
```

```
## # A tibble: 1 x 1
##   breeds
##    <int>
## 1   2484
```

In order for our analysis to be a bit more useful, it will helpful to create less breed categories; we will be removing any "mix" labels and specifying all dogs by their primary breed.

```
aac <- aac %>%
  mutate(Breed=str_extract(Breed, "^[A-Za-z ]+")) %>%
  mutate(Breed=str_replace(Breed, " Mix", ""))

aac %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=n)) +
  geom_density(fill="light blue") +
  scale_x_log10() +
  xlab("Occurances of each unique breed in data")
```



```
aac %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  summarise(breeds=n())
```

```
## # A tibble: 1 x 1
##   breeds
##    <int>
## 1    207
```

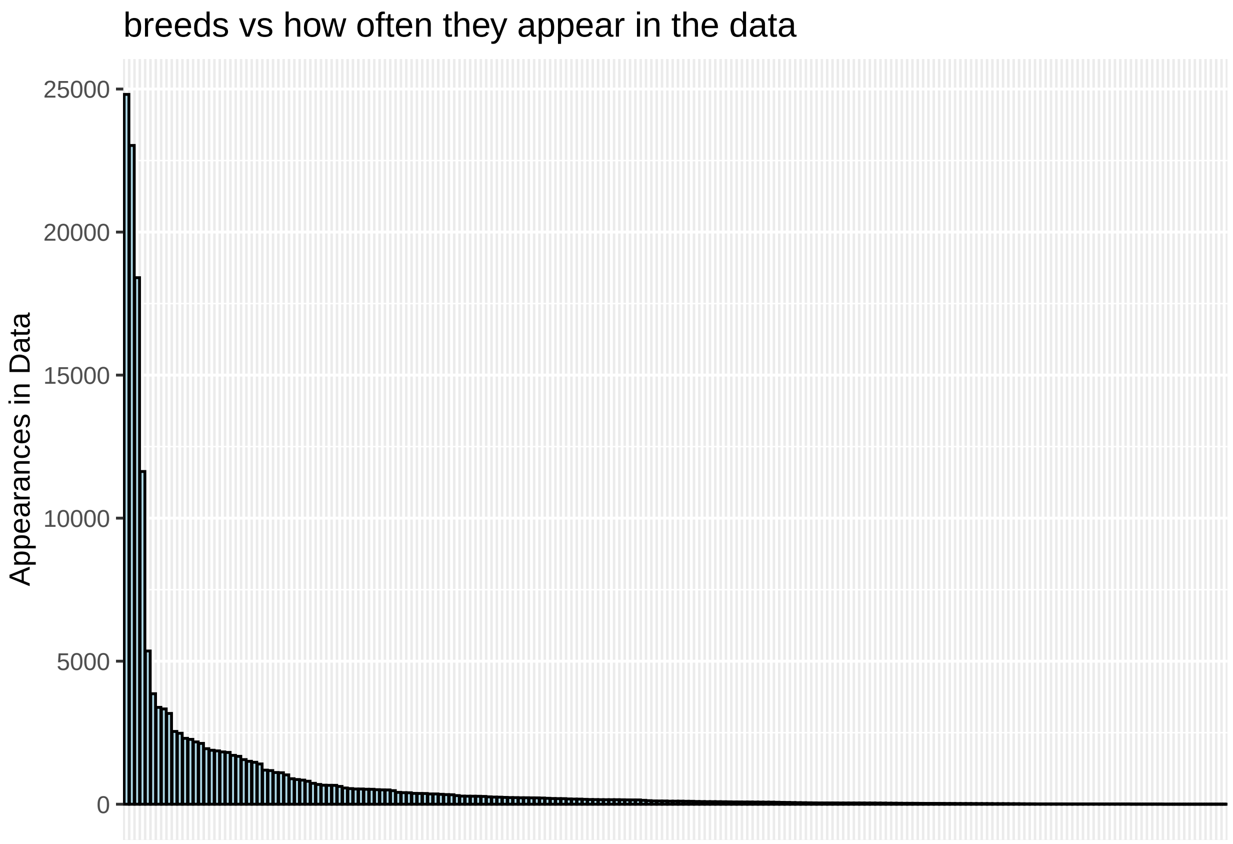
```
aac %>%
  select(Age, Breed, Color) %>%
  head()
```

```
## # A tibble: 6 x 3
##     Age Breed            Color
##   <dbl> <chr>            <chr>
## 1     6 Spinone Italiano Yellow/White
## 2     6 Spinone Italiano Yellow/White
## 3     7 Spinone Italiano Yellow/White
## 4     7 Spinone Italiano Yellow/White
## 5    10 Spinone Italiano Yellow/White
## 6    10 Spinone Italiano Yellow/White
```

While there are still some breeds that show up <10 times, they now make up a minority of the data, and our analysis will be far more useful.

Let's look at the distribution of breeds in the data after modification.

```
aac %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=fct_rev(fct_reorder(Breed, n)), y=n)) + geom_col(fill="light blue", color="black")  +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  ylab("Appearances in Data") +
  ggtitle("breeds vs how often they appear in the data")
```



The distribution of breeds throughout the data appears *very* right skewed, so there are a small number of breeds that show up far more than any others. Let's look at those at the top:

```
aac %>%
  group_by(Breed) %>%
  summarise(appearances=n()) %>%
  arrange(desc(appearances)) %>%
  head(10)
```

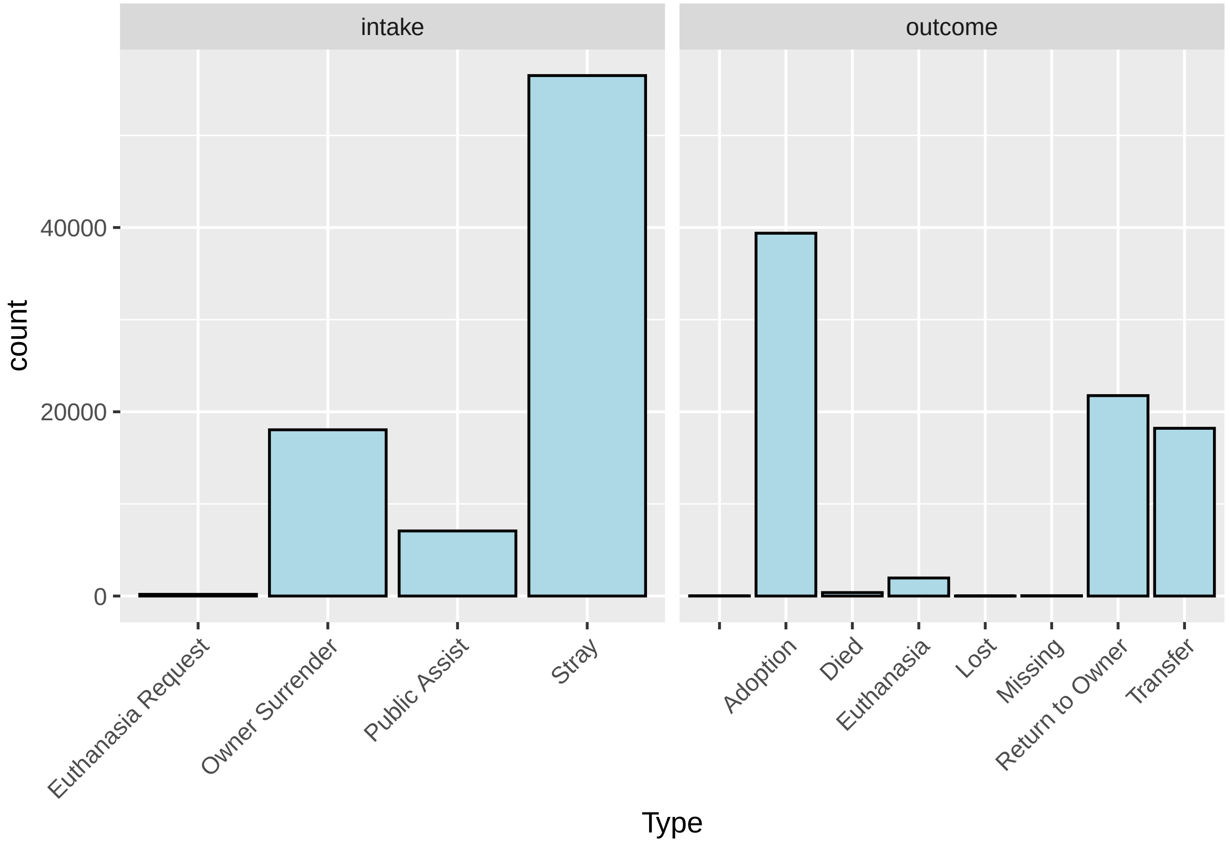
```
## # A tibble: 10 x 2
##    Breed                    appearances
##    <chr>                          <int>
##  1 Pit Bull                       24812
##  2 Labrador Retriever             23028
##  3 Chihuahua Shorthair            18406
##  4 German Shepherd                11631
##  5 Australian Cattle Dog        5356
##  6 Dachshund                    3863
##  7 Boxer                        3386
##  8 Siberian Husky               3332
##  9 Border Collie                3174
## 10 Miniature Poodle             2544
```

Unsurprisingly, Pit Bulls appear most often in the data, appearing in 24,799 entries out of 163,450 total entries (i.e. nearly a sixth). Pit Bulls are very common in animal shelters, since they tend to have a hard time finding owners due to the aggressive and dangerous reputation of the breed, and they are often abandoned by their owners for the same reason. Retrievers are not far behind Pit Bulls, however dogs of the breed are far more numerous than Pit Bulls, which can account for their many occurrences in the data.

Action Type

Let's look at the distribution of action types in the data.

```
aac %>%
  ggplot(aes(x=Type)) +
  geom_bar(fill="light blue", color="black") +
  facet_wrap(~ Action, scale="free_x") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



```
n_o <- (filter(aac, Action=="outcome") %>% summarise(n=n()))[1,1]
n_i <- (filter(aac, Action=="intake") %>% summarise(n=n()))[1,1]

aac %>%
  filter(Action=="outcome") %>%
  group_by(Type) %>%
  summarise(percent_of_outcomes=round((n()/n_o)*100, 2))
```

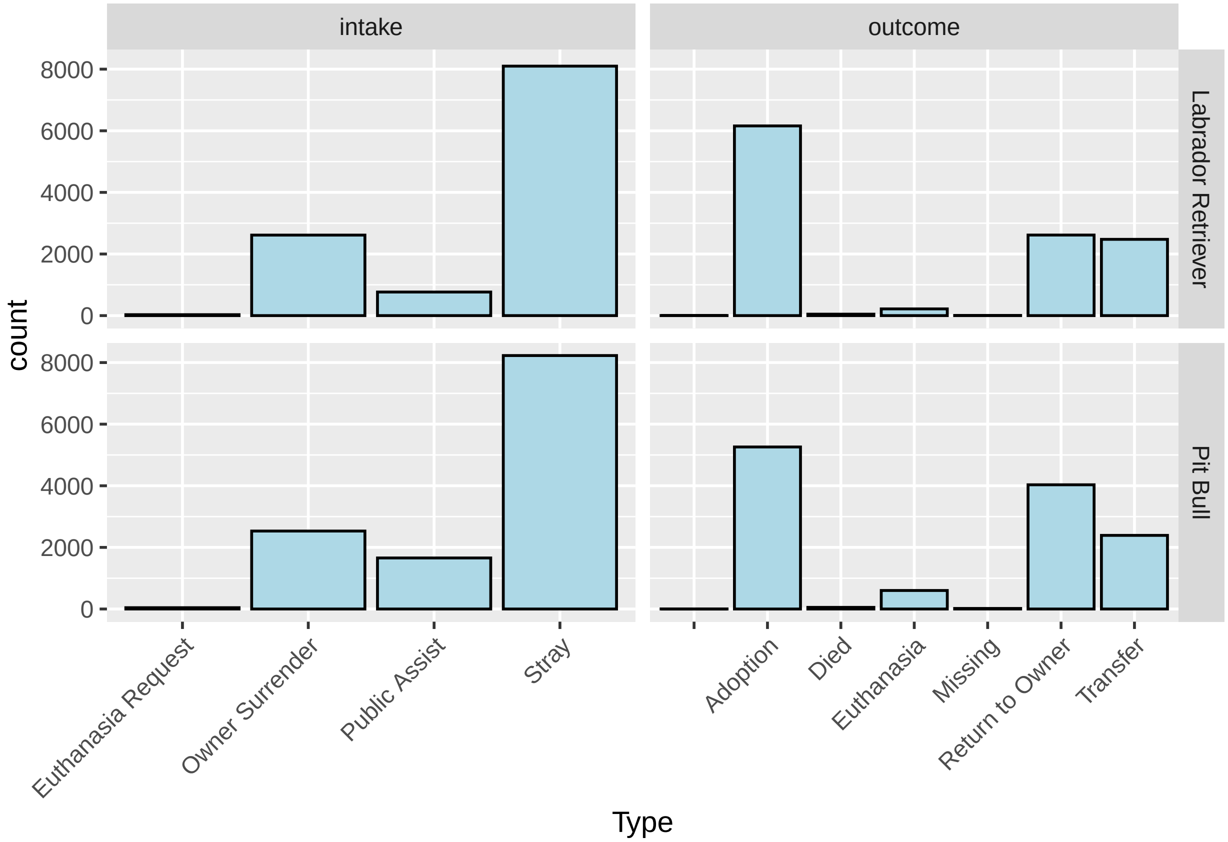
```
## # A tibble: 8 x 2
##   Type              percent_of_outcomes$n
##   <chr>                             <dbl>
## 1 ""                                 0.02
## 2 "Adoption"                        48.2
## 3 "Died"                             0.46
## 4 "Euthanasia"                       2.4
## 5 "Lost"                             0
## 6 "Missing"                          0.04
## 7 "Return to Owner"                 26.6
## 8 "Transfer"                        22.3
```

```
aac %>%
  filter(Action=="intake") %>%
  group_by(Type) %>%
  summarise(percent_of_intakes=round((n()/n_i)*100, 2))
```

```
## # A tibble: 4 x 2
##   Type               percent_of_intakes$n
##   <chr>                             <dbl>
## 1 Euthanasia Request                 0.22
## 2 Owner Surrender                   22.1
## 3 Public Assist                      8.64
## 4 Stray                             69.1
```

We can see that almost 50% of the dogs that leave the shelter are adopted, and well over 50% that enter are strays. Considering the findings from our analysis of the Breed column, it may be interesting to facet further by Breed comparing Pit Bulls and Retrievers, since they both appear very frequently in the data but likely for different reasons.

```
aac %>%
  filter(Breed %in% c("Pit Bull", "Labrador Retriever")) %>%
  ggplot(aes(x=Type)) +
  geom_bar(fill="light blue", color="black") +
  facet_grid(Breed ~ Action, scale="free_x") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



```
n_o_r <- (filter(aac, Action=="outcome", Breed=="Labrador Retriever") %>% summarise(n=n()))[1,1]
n_i_r <- (filter(aac, Action=="intake", Breed=="Labrador Retriever") %>% summarise(n=n()))[1,1]
n_o_p <- (filter(aac, Action=="outcome", Breed=="Pit Bull") %>% summarise(n=n()))[1,1]
n_i_p <- (filter(aac, Action=="intake", Breed=="Pit Bull") %>% summarise(n=n()))[1,1]

aac %>%
  filter(Action=="outcome", Breed %in% c("Labrador Retriever", "Pit Bull")) %>%
  group_by(Type) %>%
  summarise(percent_of_outcomes_LR=round((sum(Breed=="Labrador Retriever")/n_o_r)*100, 2),
            percent_of_outcomes_PB=round((sum(Breed=="Pit Bull")/n_o_p)*100, 2))
```

```
## # A tibble: 7 x 3
##   Type              percent_of_outcomes_LR$n percent_of_outcomes_PB$n
##   <chr>                                <dbl>                    <dbl>
## 1 ""                                    0.03                     0.02
## 2 "Adoption"                           53.4                     42.6
## 3 "Died"                                0.41                     0.43
## 4 "Euthanasia"                          1.89                     4.86
## 5 "Missing"                             0.04                     0.14
## 6 "Return to Owner"                    22.7                     32.6
## 7 "Transfer"                           21.5                     19.3
```

```
aac %>%
  filter(Action=="intake", Breed %in% c("Labrador Retriever", "Pit Bull")) %>%
  group_by(Type) %>%
  summarise(percent_of_intakes_LR=round((sum(Breed=="Labrador Retriever")/n_i_r)*100, 2),
            percent_of_intakes_PB=round((sum(Breed=="Pit Bull")/n_i_p)*100, 2))
```

```
## # A tibble: 4 x 3
##   Type               percent_of_intakes_LR$n percent_of_intakes_PB$n
##   <chr>                                <dbl>                   <dbl>
## 1 Euthanasia Request                    0.26                    0.34
## 2 Owner Surrender                      22.7                    20.3
## 3 Public Assist                         6.65                   13.3
## 4 Stray                                70.4                    66.0
```

Indeed, while about 53% of Retrievers' outcomes are adoptions, only about 43% of Pit Bulls' outcomes are adoptions: a difference of 10%! Conversely, "return to owner" for Retrievers was 10% greater than that for Pit Bulls.

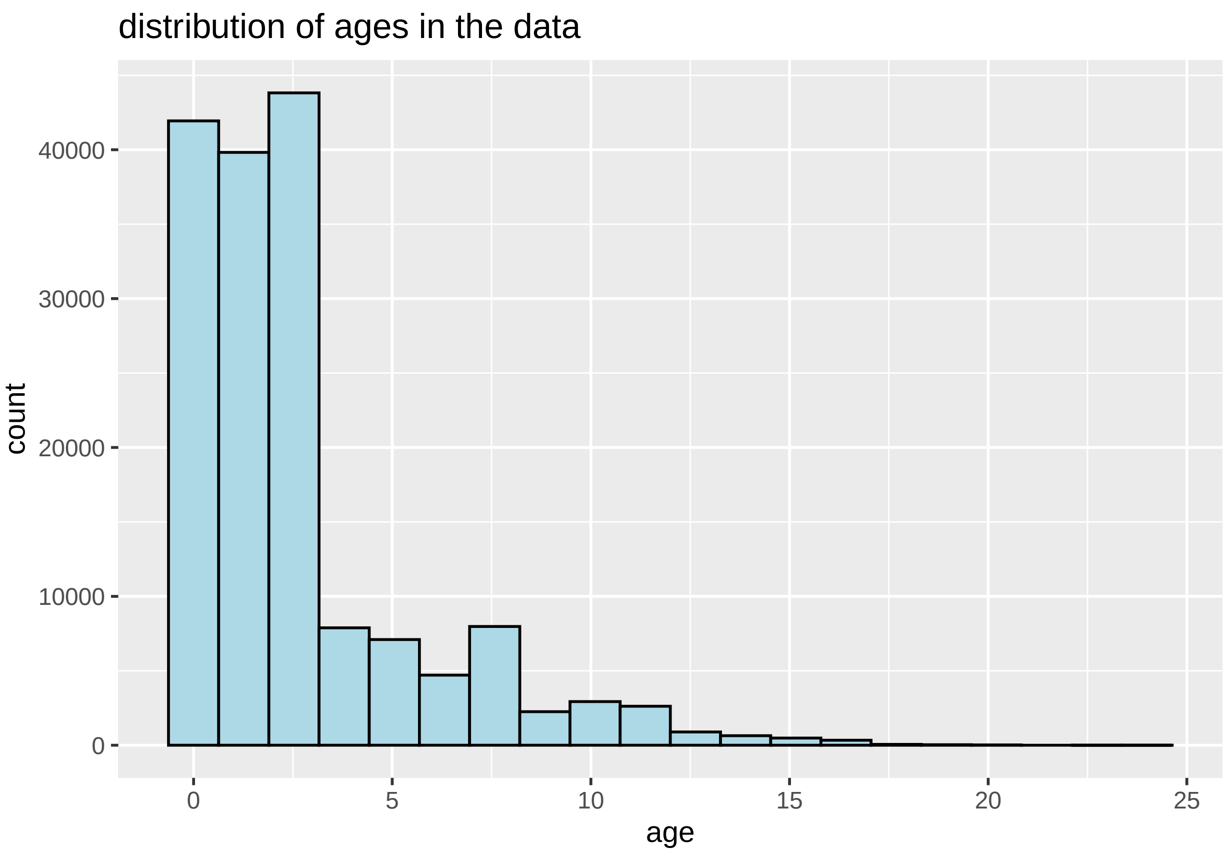
Additionally, while the actual difference is only about 3%, the percentage of outcomes for Pit Bulls that were euthanasia is around 2.5 times greater than that for Retrievers. All of these differences can be attributed to the aggressive reputation that Pit Bulls carry (or, in some cases such as euthanasia, the aggressive temperament they have).

Age

Let's look at the distribution of ages throughout the data.

```
aac %>%
  ggplot(aes(x=Age)) +
  geom_histogram(fill="light blue", color="black", bins=20) +
  xlab("age") +
  ggtitle("distribution of ages in the data")
```

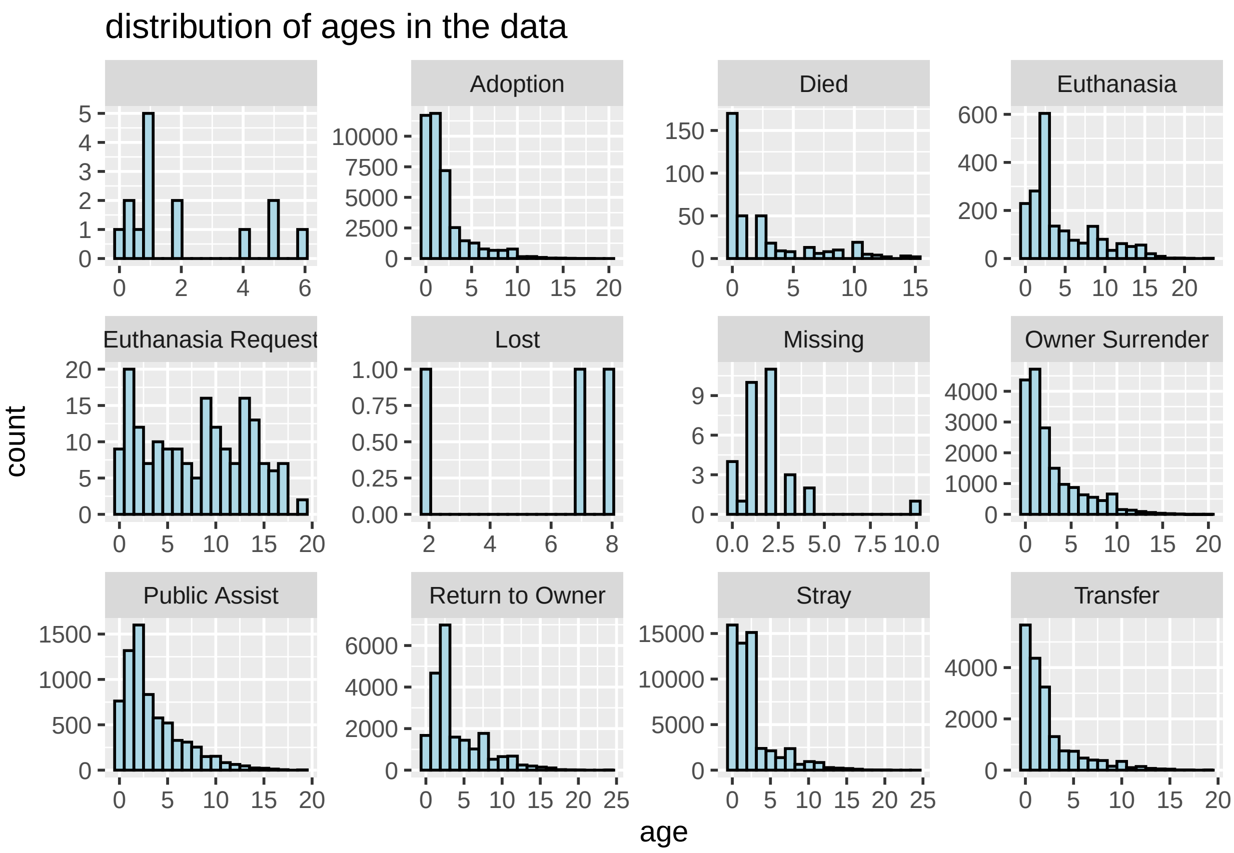
```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```



The ages are certainly right-skewed, but it may be useful to dig deeper. The distribution of ages may change depending on what is happening to the dog, so we'll facet by `Type` and see if it's any different.

```
aac %>%
  ggplot(aes(x=Age)) +
  geom_histogram(fill="light blue", color="black", bins=20) +
  xlab("age") +
  ggtitle("distribution of ages in the data") +
  facet_wrap(~ Type, scales="free")
```

```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```



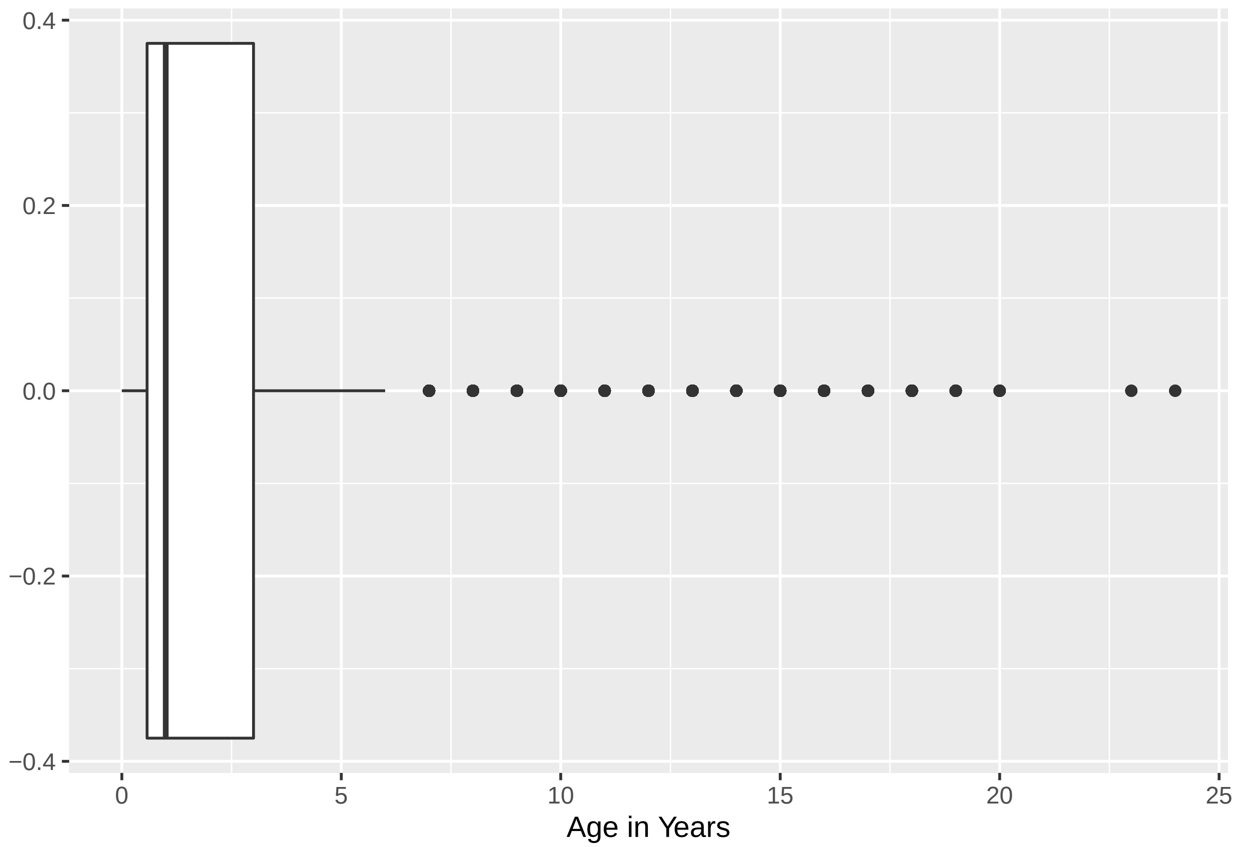
It seems as though almost all of the facets are also right skewed, except for `Euthanasia Request`. this could reflect how every other intake/outcome is far more likely to happen to younger dogs; younger dogs are more likely to be abandoned or run away, and are also more likely to be adopted. However, dogs can become sick or unsuitable for care at any point in their life, and thus need to be euthanized at any age. Interestingly, the `Euthanasia` outcome is right skewed unlike the `Euthanasia Request` intake, however the plot also shows that there are far more dogs that are euthanised than are requested to be euthanized by the owner. Thus, many euthanasia outcomes are likely from strays that were picked up early in their life, but not early enough before they got rabies or some other disease.

It may be useful to bin the dogs' ages in categories for later. Some numerical and visual distribution may better help us understand how to bin them.

```
quantile(aac$Age, na.rm=TRUE)
```

```
##     0%    25%    50%    75%   100%
##  0.000  0.575  1.000  3.000 24.000
```

```
ggplot(aac, aes(x=Age)) +
  geom_boxplot(na.rm=TRUE) +
  xlab("Age in Years")
```

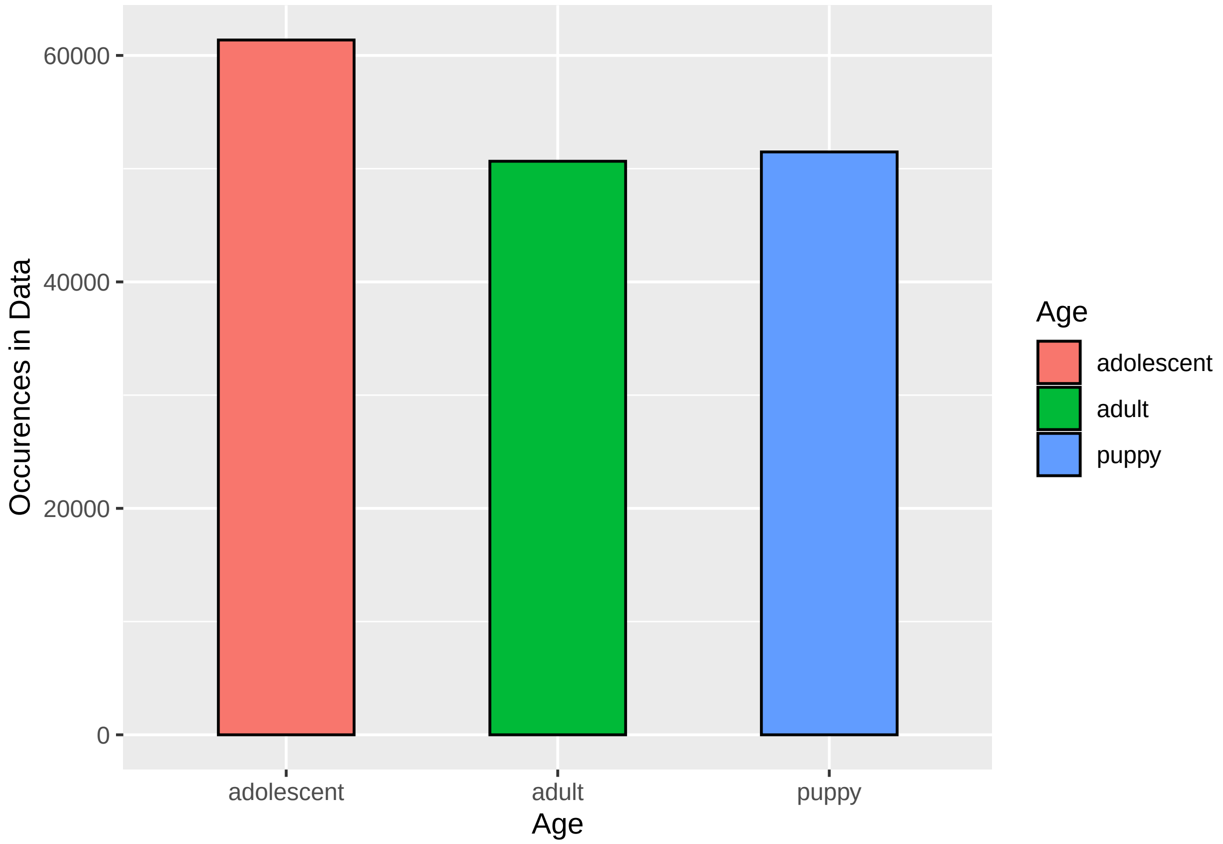


The American Kennel Club considers dogs to be adults once they reach one year old, and seniors once they reach 10 years of age. However, in our data, there are a small number of dogs that are 10 years or older, so binning them by puppy-adult-elderly would not be terribly helpful to out analysis; in fact, any dog older than 6 years old is considered an outlier. Instead, we can group them by puppy-adolescent-adult, reflecting the changes that younger dogs go through while also allowing each bin to have a good amount of dogs. So, we'll say that puppies are 0-1 years old, adolescents are 1-3, adults are 3 or older.

```
aac <-
  aac %>%
  mutate(Age.Bin=case_when(
    Age<1 ~ "puppy",
    Age<3 ~ "adolescent",
    Age>=3 ~ "adult"
  ))
head(aac$Age.Bin)
```

```
## [1] "adult" "adult" "adult" "adult" "adult" "adult"
```

```
aac %>%
  drop_na(Age.Bin) %>%
  group_by(Age.Bin) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=Age.Bin, y=n, fill=Age.Bin)) +
  geom_col(color="black", width=0.5) +
  xlab("Age") +
  ylab("Occurences in Data") +
  scale_fill_discrete(name="Age")
```



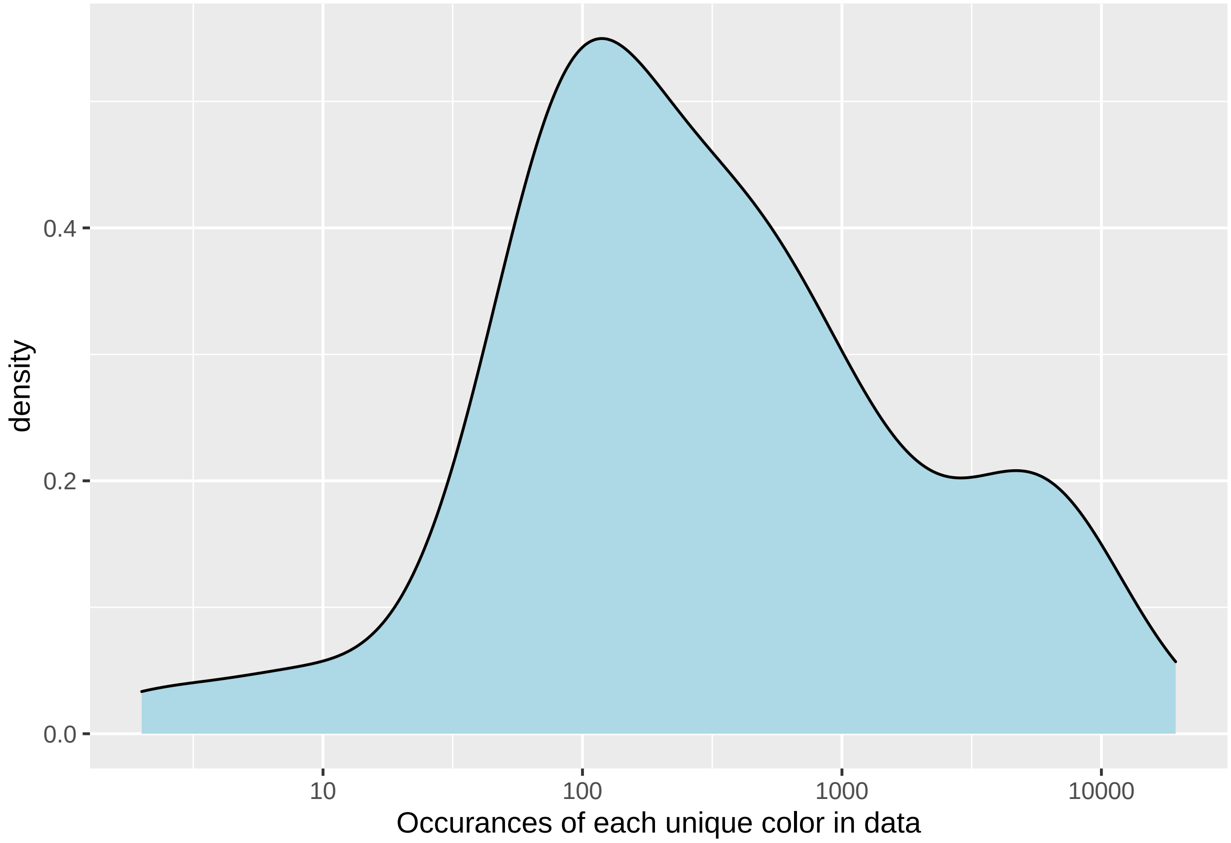
With this distribution, we have a good amount of dogs in each bin, while still being separated by age in a meaningful way.

Color

The 'Color' column includes some interesting data. While some occurrences are just single colors (i.e. "White" or "Black") many are combinations, (i.e. "White/Tan" or "Black/White"). Sometimes colors appear in reverse order (i.e. "Black/White" and "White/Black). The distinction between these two cases isn't made clear, but for our purposes we will assume the first color to be dominant.

In order to simplify the data, we will be removing the second color from any combinations which appear less than 50 times in the data. Additionally, combinations of the same color (i.e. "White/White") will be reduced to one color. Finally, we will be excluding any colors that appear less than 10 times in the data.

```
tmp <-
  aac %>%
  group_by(Color) %>%
  summarise(n=n())
aac <- aac %>%
  inner_join(tmp, by="Color") %>%
  mutate(a=str_extract(Color, "^.+?(?=/|$)"),
         b=str_extract(Color, "[^/]*$")) %>%
  mutate(Color=case_when(
    n<50 ~ a,
    a==b ~ a,
    a!=b ~ Color
  )) %>%
  select(-c(a,b,n))
aac %>%
  group_by(Color) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=n)) +
  geom_density(fill="light blue") +
  scale_x_log10() +
  xlab("Occurances of each unique color in data")
```



```
aac %>%
  group_by(Color) %>%
  summarise(occurances_in_data=n()) %>%
  arrange(desc(occurances_in_data)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Color       occurances_in_data
##    <chr>                    <int>
##  1 Black/White              19329
##  2 Brown/White               9315
##  3 White                      9168
##  4 Black                      8938
##  5 Tan/White                  8591
##  6 Tan                        7429
##  7 Brown                      6762
##  8 Tricolor                   6386
##  9 Black/Tan                  6304
## 10 White/Black                5867
```

Now, when we look at the distribution, we can see that the majority of colors appear around 100-200 times in the data. There are some colors, however, that appear significantly more than others. All 10 of the most common colors appear over 5000 times. In order to investigate further, we will check to see if these colors have any relationship with breeds.

The most common color, “Black/White” appears almost 20000 times. We will look at that color specifically to see the most common breed types for Black/White dogs.

```
breed_percentage <- aac %>%
  count(Breed) %>%
  arrange(desc(n)) %>%
  mutate(total=sum(n)) %>%
  mutate(percentage_total = (n/total)*100) %>%
  select(Breed, percentage_total)
aac %>%
  filter(Color == "Black/White") %>%
  group_by(Breed) %>%
  summarise(n=n()) %>%
  filter(n>250) %>%

  inner_join(breed_percentage, by = "Breed") %>%
  mutate(n_vs_percent = n/percentage_total) %>%
  arrange(desc(n_vs_percent)) %>%
  rename(proportional_frequency = n_vs_percent) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    Breed                     n percentage_total proportional_frequency
##    <chr>                 <int>            <dbl>                  <dbl>
##  1 Border Collie          1667             1.94                  859.
##  2 Pointer                 485             1.11                  438.
##  3 Siberian Husky          887             2.04                  435.
##  4 Labrador Retriever     5529            14.1                   393.
##  5 Staffordshire           278             1.12                  249.
##  6 Pit Bull               3140            15.2                   207.
##  7 Australian Cattle Dog   617             3.28                  188.
##  8 Boxer                   315             2.07                  152.
##  9 Chihuahua Shorthair    1214            11.3                   108.
## 10 German Shepherd         319             7.11                   44.8
```

```
  #To ensure proper proportional evaluation
aac %>%

  group_by(Breed) %>%
  summarise(n=n()) %>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Breed                     n
##    <chr>                 <int>
##  1 Pit Bull              24812
##  2 Labrador Retriever    23028
##  3 Chihuahua Shorthair   18406
##  4 German Shepherd       11631
##  5 Australian Cattle Dog  5356
##  6 Dachshund              3863
##  7 Boxer                  3386
##  8 Siberian Husky         3332
##  9 Border Collie          3174
## 10 Miniature Poodle       2544
```

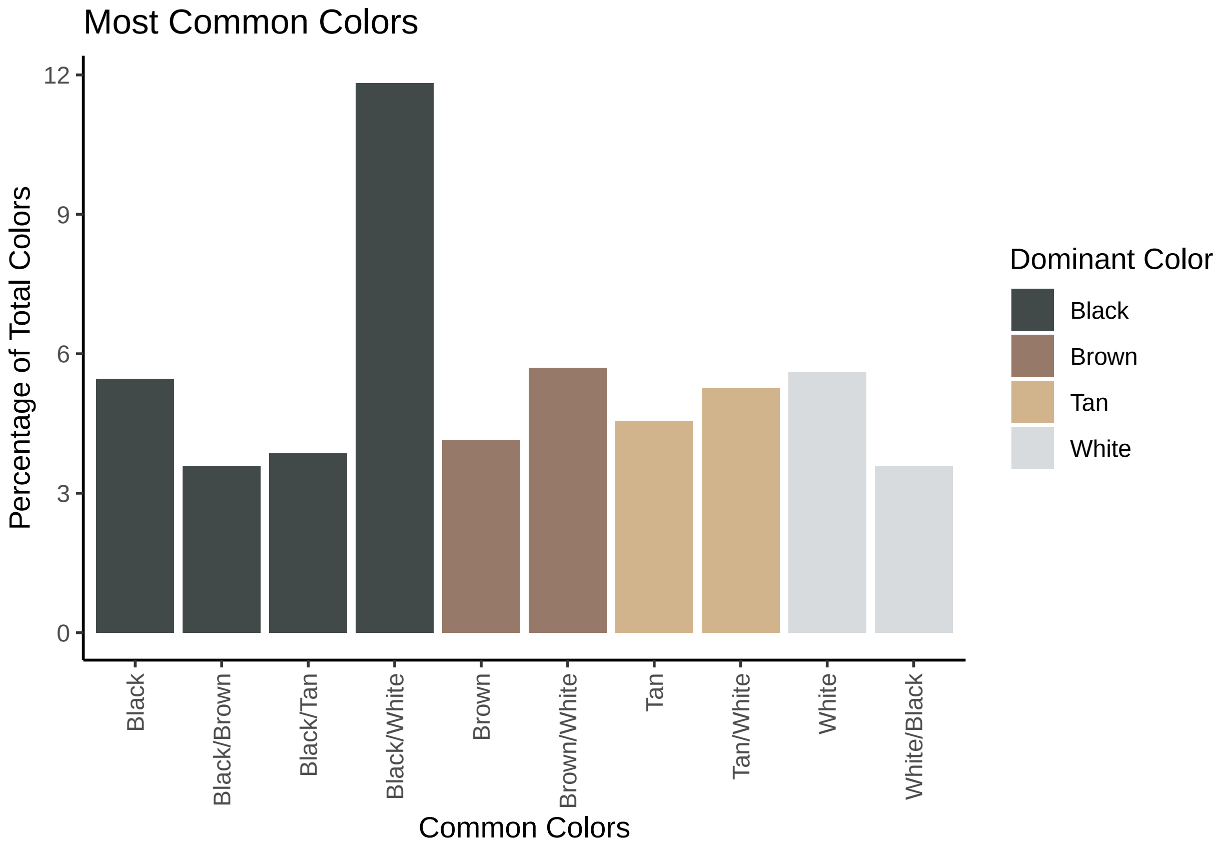
We can see that of the most common Black/White breeds (found by placing their occurrences against their percentages of all dogs), 7 appear in the most common breeds for the entire dataset. Lets look at some of the other most common colors.

```
color_percentage <- aac %>%
  count(Color) %>%
  mutate(total=sum(n)) %>%
  mutate(percentage_total = (n/total)*100) %>%
  select(Color, percentage_total)
aac %>%

  group_by(Color) %>%
  summarise(n=n()) %>%
  filter(Color != "Tricolor") %>%

  inner_join(color_percentage, by = "Color") %>%
  arrange(desc(percentage_total)) %>%
  head(10) %>%

  ggplot(aes(x = Color, y = percentage_total, fill = str_extract(Color, "^.+?(?=/|$)"))) +
  geom_col() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_fill_manual(values=c('#424949', '#967969', '#D2B48C', '#D7DBDD')) +
  labs(x = 'Common Colors', y = 'Percentage of Total Colors', title = 'Most Common Colors')+
  guides(fill=guide_legend(title="Dominant Color"))
```

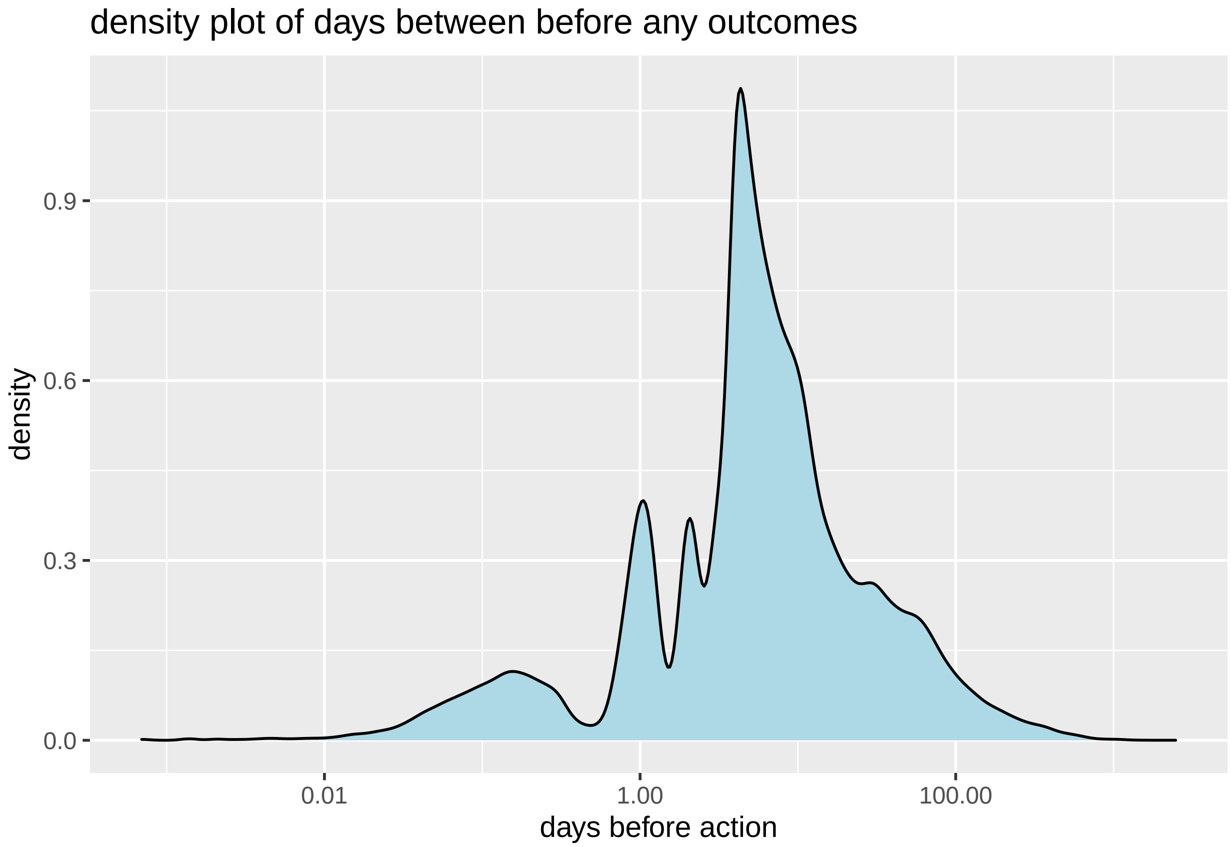


Black/White is the largest color by a significant margin, but all 10 of the most common colors are some combination of Black, Brown, and White tones. These colors reflect the most common colors for all dogs, so there isn't anything to be inferred about their large frequency in our dataset. Color percentages in the shelter remain consistent with those typically observed outside of the shelter.

Time Since Last Action

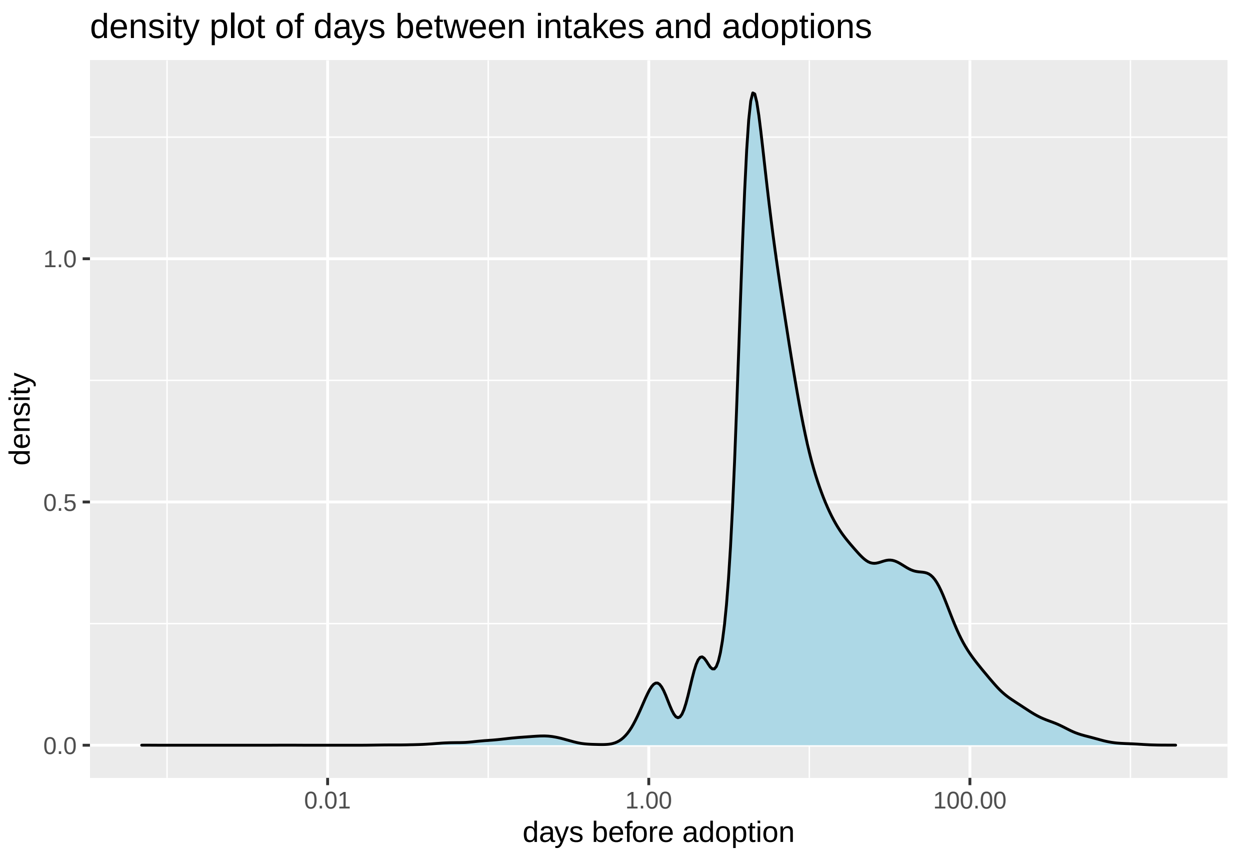
First, let's look at the distribution of times before outcomes throughout the data

```
aac %>%
  filter(!(is.na(Days.From.Last.Action)) & Action=="outcome") %>%
  ggplot(aes(x=Days.From.Last.Action)) +
  geom_density(fill="light blue") +
  scale_x_log10(labels=label_comma()) +
  xlab("days before action") +
  ggtitle("density plot of days between before any outcomes")
```



We can make some observations about this, but they frankly don't mean much if we don't specify the type of action. Let's try filtering for only adoptions.

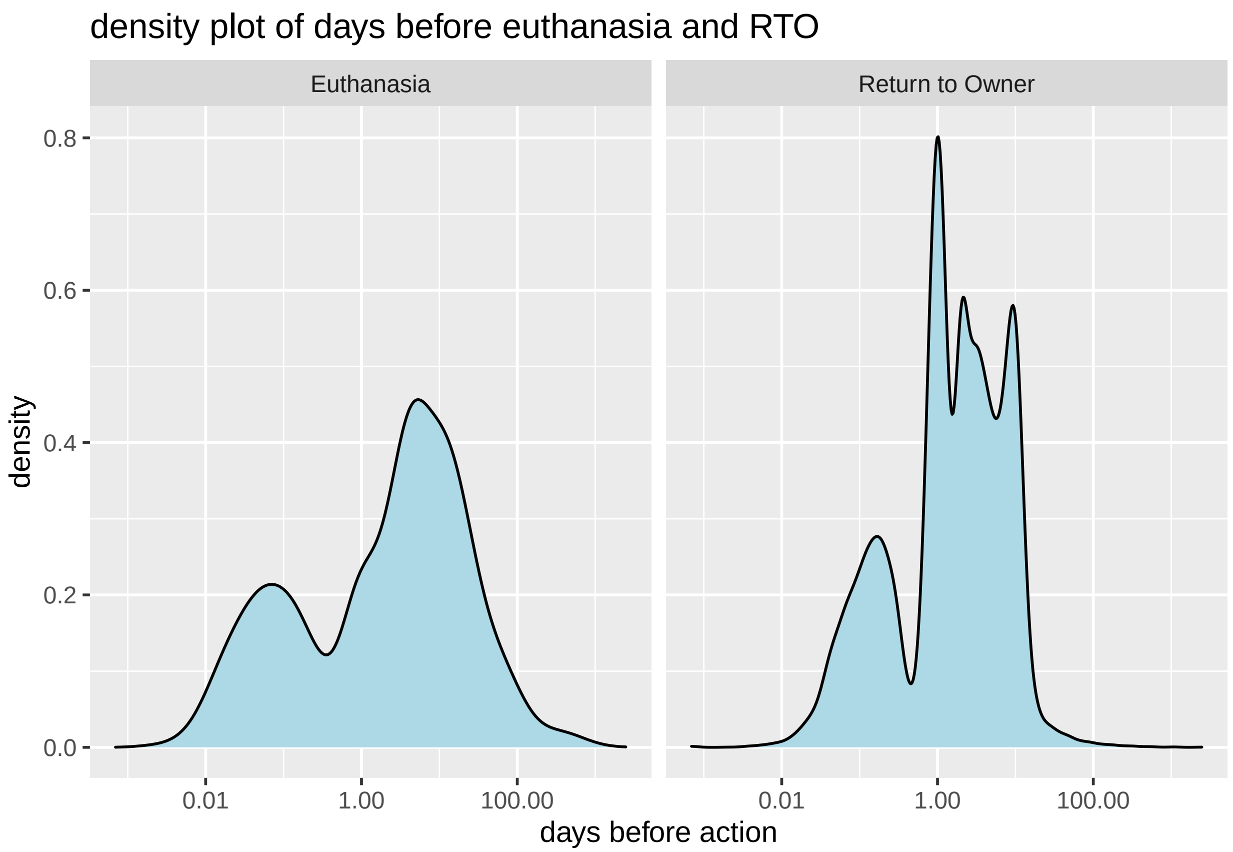
```
aac %>%
  filter(!(is.na(Days.From.Last.Action)) & Type=="Adoption") %>%
  ggplot(aes(x=Days.From.Last.Action)) +
  geom_density(fill="light blue") +
  scale_x_log10(labels=label_comma()) +
  xlab("days before adoption") +
  ggtitle("density plot of days between intakes and adoptions")
```



A similar trend shows up even stronger; we can see a spike of adoption times around 5-7 days, and nearly all the rest of the times are between that and 100 days. The spike seems very sudden for such a statistic, which could possibly illustrate something about the center's adoption process; perhaps most dogs need to be at the center for at least a few days before they can be put up for adoption.

Let's look at the same plot, but filtered for `Euthanasia` and `Return to Owner` outcomes; they represent popular but very distinct outcomes for the dogs, and so could lend themselves to some conclusion.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action)) & Type %in% c("Euthanasia", "Return to Owner")) %>%
  ggplot(aes(x=Days.From.Last.Action)) +
  geom_density(fill="light blue") +
  scale_x_log10(labels=label_comma()) +
  xlab("days before action") +
  ggtitle("density plot of days before euthanasia and RTO") +
  facet_wrap(~ Type)
```



A unique trend appears in both of these plots, in the form of a small bump close to 0 days. This likely reflects how both of these outcomes are typically not dependent on other people; that is to say, the center can usually figure out a dog's owner and return them or deem a stray as necessary for euthanasia very soon after the intake. However, adoptions are dependent on the dog actually being adopted, which, at a point, the center has no control over.

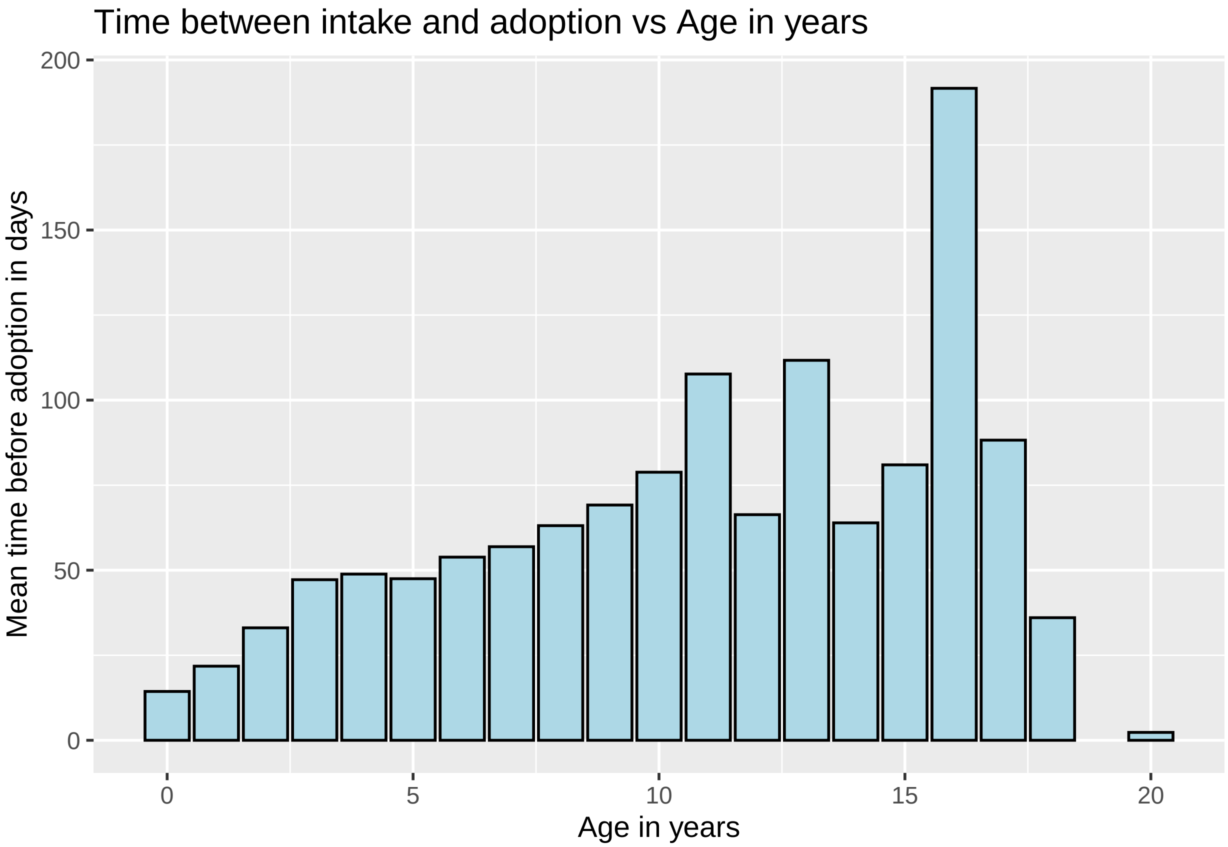
This column will serve as our response variable, so more analysis of it will be in the next section.

Section 2.2: Multivariate Exploratory Analysis

Age vs Time Before Adoption

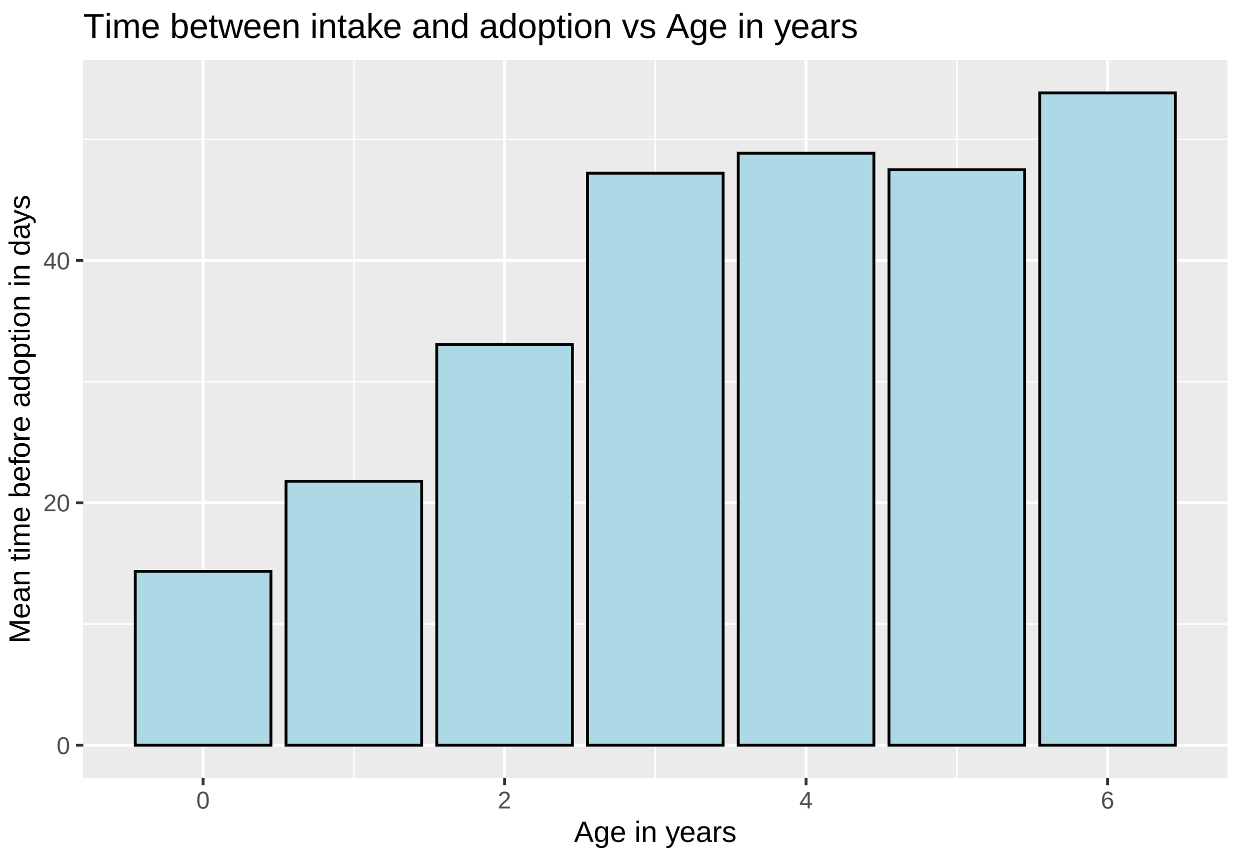
It feels obvious that dogs that are younger would get adopted quicker (they're cuter, they will probably live for longer, etc), but let's see if that trend reveals itself in the data.

```
aac %>%
  filter(Type=="Adoption", !(is.na(Days.From.Last.Action))) %>%
  mutate(Age=round(Age)) %>%
  group_by(Age) %>%
  summarise(a=mean(Days.From.Last.Action)) %>%
  ggplot(aes(x=Age, y=a)) +
  geom_col(color="black", fill="light blue") +
  ylab("Mean time before adoption in days") +
  xlab("Age in years") +
  ggtitle("Time between intake and adoption vs Age in years")
```



While there does seem to be an upward trend up until the 12-year-old dogs, it should be noted that there are few entries of dogs that are older than 5 or 6 in the data, and those entries would be considered outliers for age, as seen in the analysis of the `Age` column. Let's limit the plot to dogs that are 6 or younger to get a better picture of what this distribution would normally look like.

```r
aac %>%
  filter(Type=="Adoption", !(is.na(Days.From.Last.Action))) %>%
  filter(Age<=6) %>%
  mutate(Age=round(Age)) %>%
  group_by(Age) %>%
  summarise(a=mean(Days.From.Last.Action)) %>%
  ggplot(aes(x=Age, y=a)) +
  geom_col(color="black", fill="light blue") +
  ylab("Mean time before adoption in days") +
  xlab("Age in years") +
  ggtitle("Time between intake and adoption vs Age in years")
```

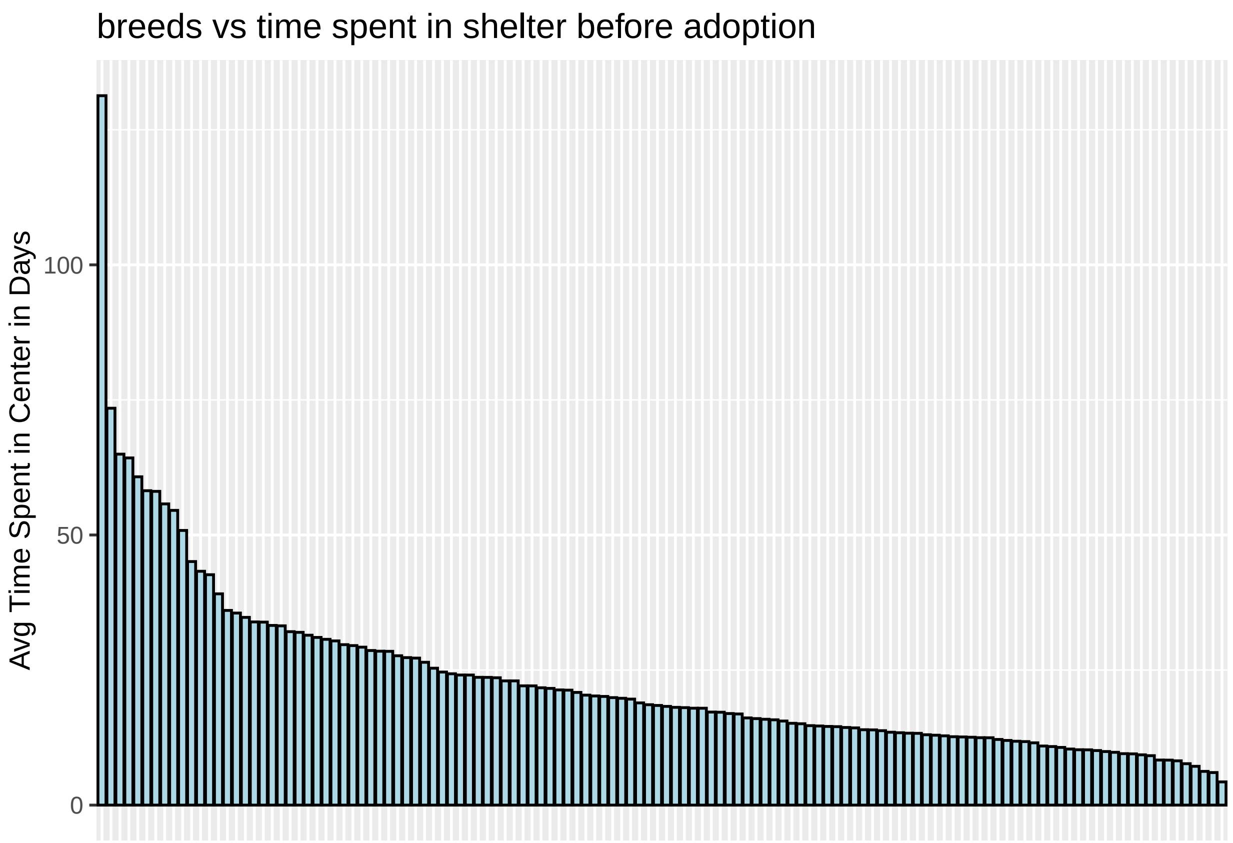


While a a left skew still holds up, we do see a slight plateu from 3-years-old and onwards. This could very well reflect how dogs tend to reach mid-adolescence around 2-3 years old; they seem to stop looking like puppies and look more like adults around this age.

Breed vs Time Before Adoption

It may be interesting to see if certain breeds spend more time in the shelter than others before being adopted. Since we will be taking an average of the time the breed spends in the shelter, we will only be looking at breeds with at least 15 adoption entries.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Breed) %>%
  filter(n()>=15) %>%
  summarise(a=mean(Days.From.Last.Action)) %>%
  ggplot(aes(x=fct_rev(fct_reorder(Breed, a)), y=a)) +
  geom_col(fill="light blue", color="black")  +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  ylab("Avg Time Spent in Center in Days") +
  ggtitle("breeds vs time spent in shelter before adoption")
```



We see a skew, though not as intensely as the distribution of breeds throughout the data. There are some that stick out, so let's look at the ones at the top.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Breed) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange(desc(mean_days_before_adoption)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Breed                            mean_days_before_adoption
##    <chr>                                                <dbl>
##  1 Dogo Argentino                                       131. 
##  2 Bulldog                                               73.5
##  3 Cane Corso                                            65.9
##  4 American Staffordshire Terrier                        65.0
##  5 American Pit Bull Terrier                             64.3
##  6 Staffordshire                                         60.8
##  7 Pit Bull                                              58.2
##  8 American Bulldog                                      58.1
##  9 Dogue De Bordeaux                                     55.8
## 10 Beauceron                                             54.6
```

If you're unfamiliar with dog breeds then the `Dogo Argentino` may seem random, but its appearance actually closely resembles that of a Pit Bull.

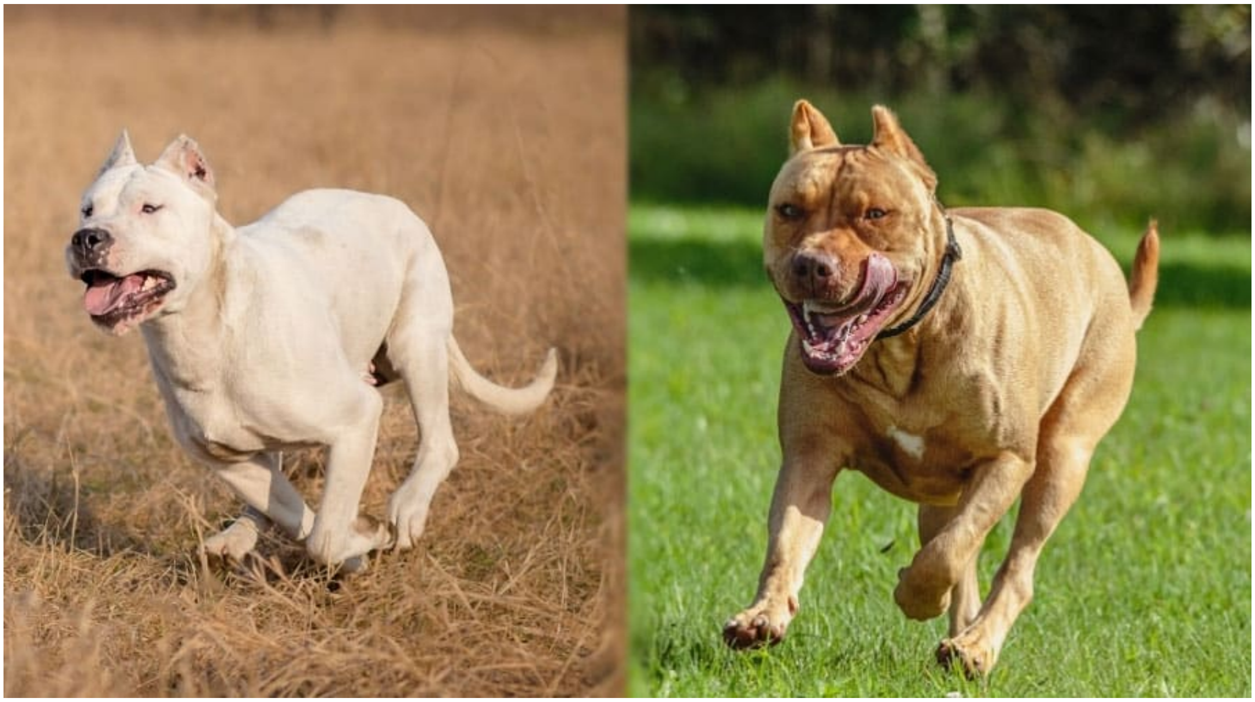


Figure 1: a Dogo Argentino on the left, and a Pit Bull on the right

While the two breeds are actually not related genetically, they are often mistaken for each other. Thus, it suffers from the same bad rap that the Pit Bull has of being a dangerous breed. **In fact, 9/10 of the breeds in that top 10 are relatives of the Pit Bull, or breeds that look similar**. Only the Beauceron looks distinct from a Pit Bull, and it likely has such a high average time because it looks similar to the Doberman, another breed with an "aggressive" reputation.

Looking into the breeds at the tail end of the graph may give us some insight, as well.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Breed) %>%
  filter(n()>=15) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange((mean_days_before_adoption)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Breed                    mean_days_before_adoption
##    <chr>                                        <dbl>
##  1 Havanese                                      4.30
##  2 Australian Terrier                            6.03
##  3 Scottish Terrier                              6.24
##  4 Norwich Terrier                               7.18
##  5 Lhasa Apso                                    7.66
##  6 Tibetan Spaniel                               8.20
##  7 Swedish Vallhund                              8.32
##  8 Cocker Spaniel                                8.34
##  9 Wire Hair Fox Terrier                         9.15
## 10 Pomeranian                                  9.31
```

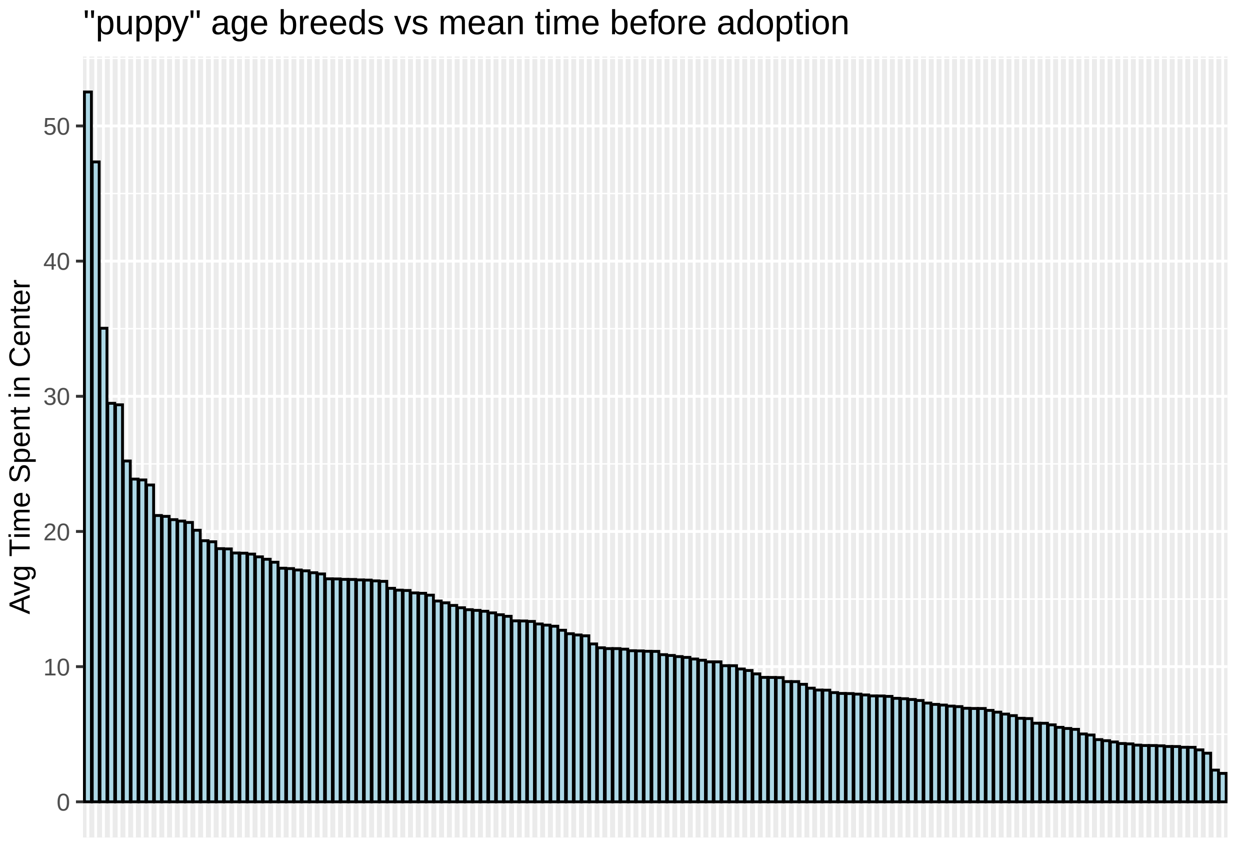
Another trend can be seen in the tail end of the graph: all of these breeds that tend to get adopted quickest could be classified as “small”, and tend to look cute.

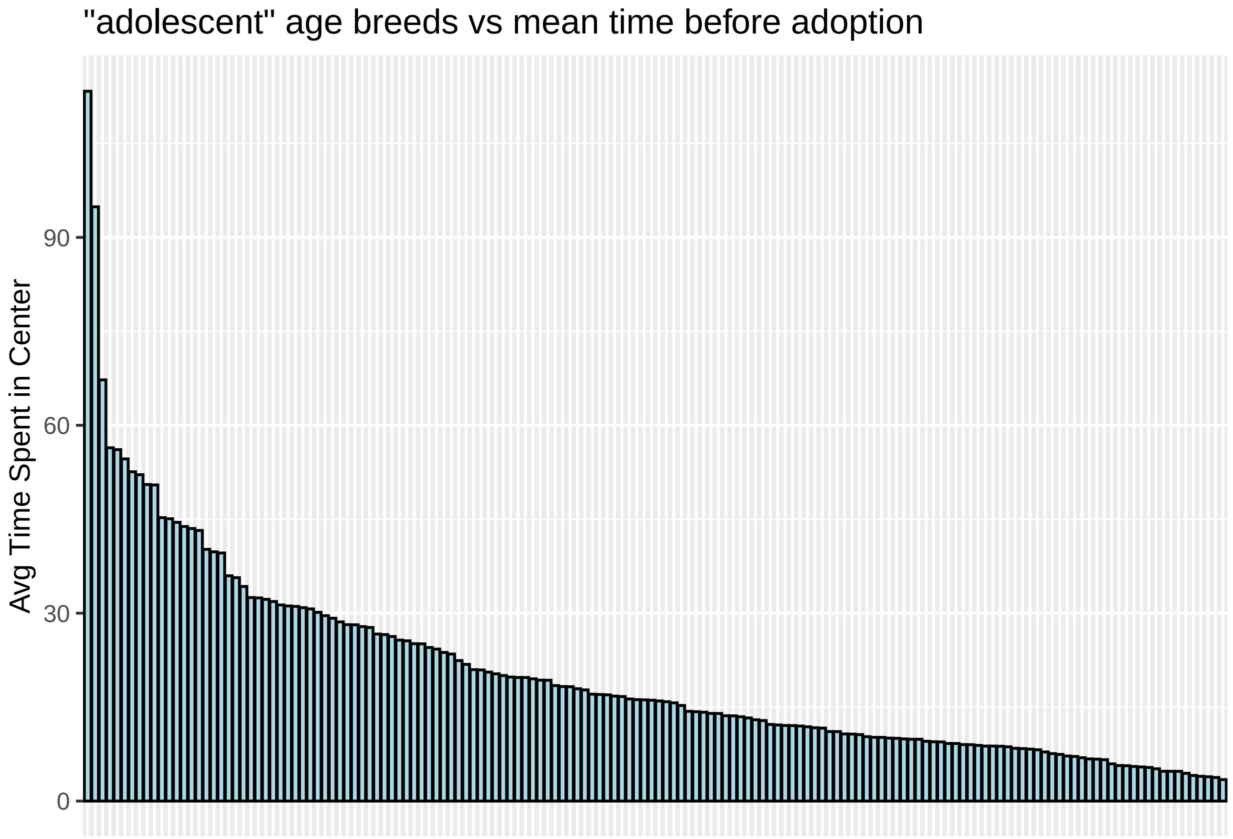
Figure 2: a Wire Hair Fox Terrier, a dog breed often compared to a stuffed animal

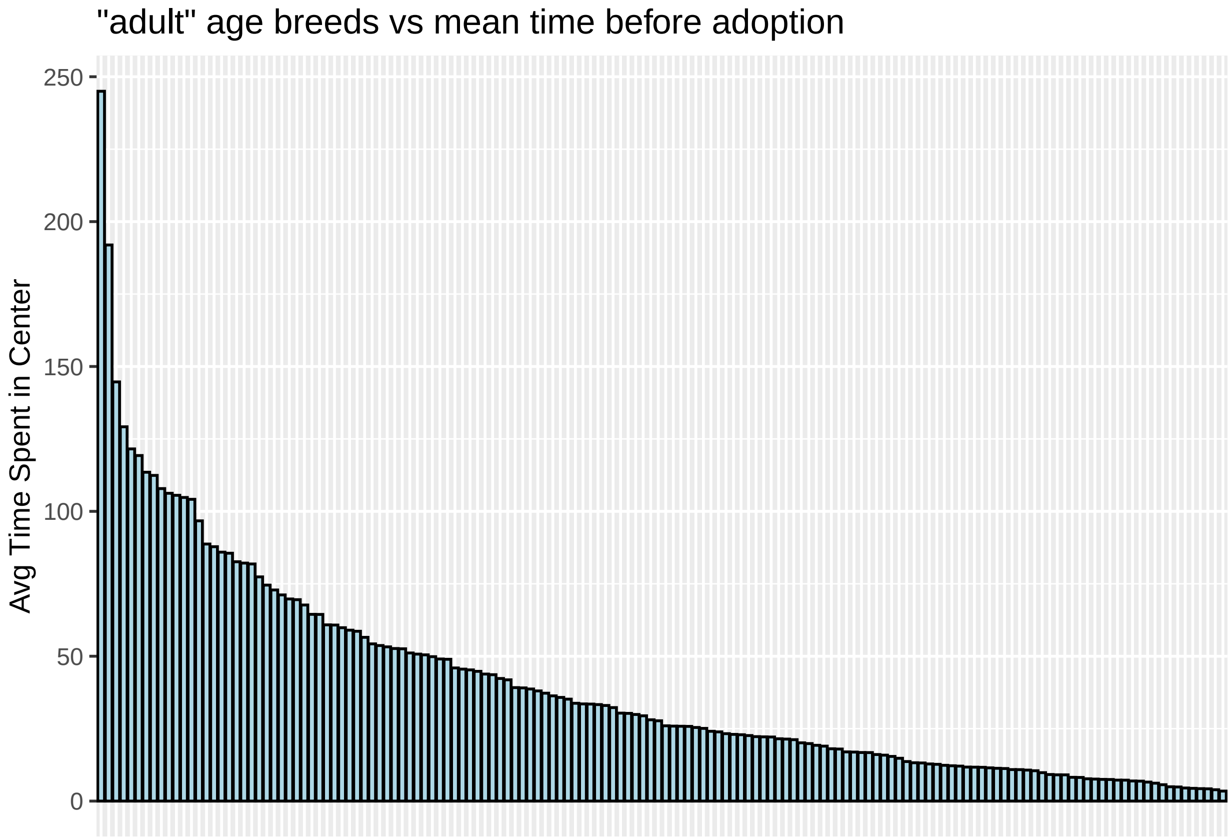
In hindsight, it may be unsurprising that the cuteness of a dog influences how quickly it gets adopted, though to see this trend manifest so strongly in our data like this is still useful.

It may be interesting to see if our distribution is different if we factor in the age bins we created earlier; for younger dogs, the breed may not matter as much as for older dogs.

```r
for(b in c("puppy", "adolescent", "adult"))
{
print(
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  group_by(Breed) %>%
  filter(n()>=15) %>%
  ungroup() %>%
  filter(Action=="outcome", Type=="Adoption", Age.Bin==b) %>%
  group_by(Breed) %>%
  summarise(a=mean(Days.From.Last.Action)) %>%
  ggplot(aes(x=fct_rev(fct_reorder(Breed, a)), y=a)) +
  geom_col(fill="light blue", color="black")  +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  ylab("Avg Time Spent in Center") +
  ggtitle(paste("\"", b, "\" age breeds vs mean time before adoption", sep=""))
)
}
```




"adolescent" age breeds vs mean time before adoption
Avg Time Spent in Center
90
60
30
0



From the graphs, the distributions all seem the same as before filtering for age. Let's see if the head or tail of the graphs are any different.

```
for(b in c("puppy", "adolescent", "adult"))
{
print(
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption", Age.Bin==b) %>%
  group_by(Breed) %>%
  filter(n()>=10) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange(desc(mean_days_before_adoption)) %>%
  head(10) %>%
  kable(caption=b)
)
}
```

```
##
##
## Table: puppy
##
## |Breed                           | mean_days_before_adoption|
## |:-------------------------------|-------------------------:|
## |American Foxhound               |                  29.36970|
## |Chinese Sharpei                 |                  25.21429|
## |Staffordshire                 |                23.87519|
## |American Pit Bull Terrier     |                23.81184|
## |Dachshund Longhair            |                21.17979|
## |American Staffordshire Terrier |                20.87416|
## |Shih Tzu                      |                19.31733|
## |Australian Kelpie             |                19.23899|
## |Miniature Poodle              |                18.72937|
## |Flat Coat Retriever           |                18.70844|
##
##
## Table: adolescent
##
## |Breed                         | mean_days_before_adoption|
## |:-----------------------------|-------------------------:|
## |American Bulldog              |                56.40506|
## |Bulldog                       |                56.11499|
## |Pit Bull                      |                54.63226|
## |Staffordshire                 |                52.58259|
## |American Pit Bull Terrier     |                52.11391|
## |English Coonhound             |                50.54007|
## |Chinese Sharpei               |                50.47334|
## |Black Mouth Cur               |                45.25659|
## |Mastiff                       |                45.07599|
## |American Staffordshire Terrier |                43.84029|
##
##
## Table: adult
##
## |Breed                         | mean_days_before_adoption|
## |:-----------------------------|-------------------------:|
## |Flat Coat Retriever           |               191.93921|
## |American Pit Bull Terrier     |               144.67631|
## |American Staffordshire Terrier |               129.18849|
## |Black Mouth Cur               |               119.27305|
## |Pit Bull                      |               113.50047|
## |Pointer                       |               112.41215|
## |Staffordshire                 |               106.23652|
## |Bulldog                       |               104.79504|
## |American Bulldog              |               104.15629|
## |Great Dane                    |                96.72774|
```

Confirming our earlier analysis of age vs time before adoption, dogs that are younger tend to get adopted much faster than those that are older. However, the top dogs, as it were, in each age category appear to be different than when we did not categorise by age. While the `adolescent` and `adult` categories look fairly similar to our uncategorised table, as the Pit Bull and its relatives occupy most of the spots, the `puppy` category is notable.

While the American Staffordshire Terrier, American Pit Bull Terrier, and Staffordshire are Pit Bull relatives, the rest of the breeds don't seem to have such an obvious explanation. The American Foxhound, the puppy breed that tends to take the *longest* to be adopted, is an average looking, mild-mannered dog that is usually very good with children and families. Additionally, the Shih Tzu and Mini Poodle are often lauded as being exceptionally cute dog breeds, which one would think is even moreso when the dog is a puppy, but they both appear to take longer, on average, than other puppy breeds to be adopted.

To be honest, I have no confident explanation for this. I can only assume that the only reason dogs such

Figure 3: a Shih Tzu puppy and Mini Poodle puppies

as Shih Tzus find themselves in the shelter so young is because they have some fundamental problem, such as being especially aggressive or difficult to train; alternatively, the strays of these breeds that come to the shelter may have been victims of the wild, and thus lose the cuteness factor that would normally mean that they are adopted quickly. Furthermore, the explanation could also go in the opposite direction; the Bulldogs and Pit Bull relatives that normally spend a long time in shelter are seen as much cuter when they are puppies, and so they are adopted quicker if they are young but as they mature, they look more intimidating and so are adopted less.

There isn't any more analysis to do here and I think this explanation is long enough, so let's now turn to the breeds that *do* get adopted quickly.

```
for(b in c("puppy", "adolescent", "adult"))
{
print(
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption", Age.Bin==b) %>%
  group_by(Breed) %>%
  filter(n()>=10) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange((mean_days_before_adoption)) %>%
  head(10) %>%
  kable(caption=b)
)
}
```

```
##
##
## Table: puppy
##
## |Breed                 | mean_days_before_adoption|
## |:---------------------|-------------------------:|
## |Pbgv                  |                  4.093388|
## |Cocker Spaniel        |                  4.601447|
## |Schnauzer Giant       |                  5.019711|
## |English Coonhound     |                  5.427194|
## |Bruss Griffon         |                  5.509491|
## |Shetland Sheepdog     |                  5.694742|
## |Standard Schnauzer    |                  5.810972|
## |Wire Hair Fox Terrier |                  5.815821|
## |St                    |                  6.158681|
## |Great Dane            |                  6.375024|
##
##
## Table: adolescent
##
## |Breed              | mean_days_before_adoption|
## |:------------------|-------------------------:|
## |Pomeranian         |                  3.959537|
## |Havanese           |                  4.088258|
## |Scottish Terrier   |                  4.435139|
## |Bruss Griffon      |                  4.751346|
## |Lhasa Apso         |                  5.158905|
## |Norwich Terrier    |                  5.525024|
## |West Highland      |                  5.668155|
## |Miniature Poodle   |                  6.925214|
## |Dachshund Wirehair |                  7.124483|
## |Dachshund Longhair |                  7.195547|
##
##
## Table: adult
##
## |Breed              | mean_days_before_adoption|
## |:------------------|-------------------------:|
## |Papillon           |                  6.178914|
## |Manchester Terrier |                  7.703145|
## |Dachshund Wirehair |                  9.813999|
## |Dachshund Longhair |                 10.484414|
## |Cairn Terrier      |                 10.861398|
## |Pomeranian         |                 11.237541|
## |English Bulldog    |                 11.474826|
## |Cocker Spaniel     |                 11.707433|
## |Lhasa Apso         |                 11.738836|
## |Maltese            |                 12.074201|
```

Once again, while our `adolescent` and `adult` tables look similar to that that are uncategorised, the `puppy` table looks notable. While not all of these dogs are particularly small or cute, most of them are actually rather rare breeds in the US. While most dogs look cute as puppies and don't start to really show the features of their breed until they are older, adoptors are likely enticed by the fact that a breed is rare. To illustrate, let's look at how many of each breed have ever been present at the center. We'll be removing duplicate `Animal.ID`s since we just want to look at each individual dog's breed, and not count a single dog multiple times. Another table denoting the occurrences of the most common breeds is provided to compare.

```r
tmp <-
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption", Age.Bin=="puppy") %>%
  group_by(Breed) %>%
  filter(n()>=10) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange((mean_days_before_adoption)) %>%
  head(10)

ad <-
  aac %>%
  filter(Breed %in% tmp$Breed)
ad <-
  ad[!duplicated(ad$Animal.ID),]
ad <-
  ad %>%
  group_by(Breed) %>%
  summarise(dogs_with_breed=n())

inner_join(tmp, ad, by="Breed")
```

```
## # A tibble: 10 x 3
##    Breed                 mean_days_before_adoption dogs_with_breed
##    <chr>                                     <dbl>           <int>
##  1 Pbgv                                       4.09              69
##  2 Cocker Spaniel                             4.60             281
##  3 Schnauzer Giant                            5.02              29
##  4 English Coonhound                          5.43              58
##  5 Bruss Griffon                              5.51              71
##  6 Shetland Sheepdog                          5.69              72
##  7 Standard Schnauzer                         5.81              95
##  8 Wire Hair Fox Terrier                      5.82             115
##  9 St                                         6.16              71
## 10 Great Dane                                 6.38             228
```

```
ad2 <-
  aac[!duplicated(aac$Animal.ID),]
ad2 %>%
  group_by(Breed) %>%
  summarise(dogs_with_breed=n()) %>%
  arrange(desc(dogs_with_breed)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Breed                 dogs_with_breed
##    <chr>                           <int>
##  1 Pit Bull                         9844
##  2 Labrador Retriever               9796
##  3 Chihuahua Shorthair              8303
##  4 German Shepherd                  4873
##  5 Australian Cattle Dog            2300
##  6 Dachshund                        1762
##  7 Boxer                            1372
##  8 Border Collie                    1357
##  9 Siberian Husky                   1345
## 10 Miniature Poodle                 1130
```

While breeds such as the Cocker Spaniel don't appear to be exceptionally rare, especially compared to something like the Petite Basset Griffon Vendéen (or Pbgv), the American Kennel Club notes that they are still highly desireable for their gentle attitude and beauty as a breed.

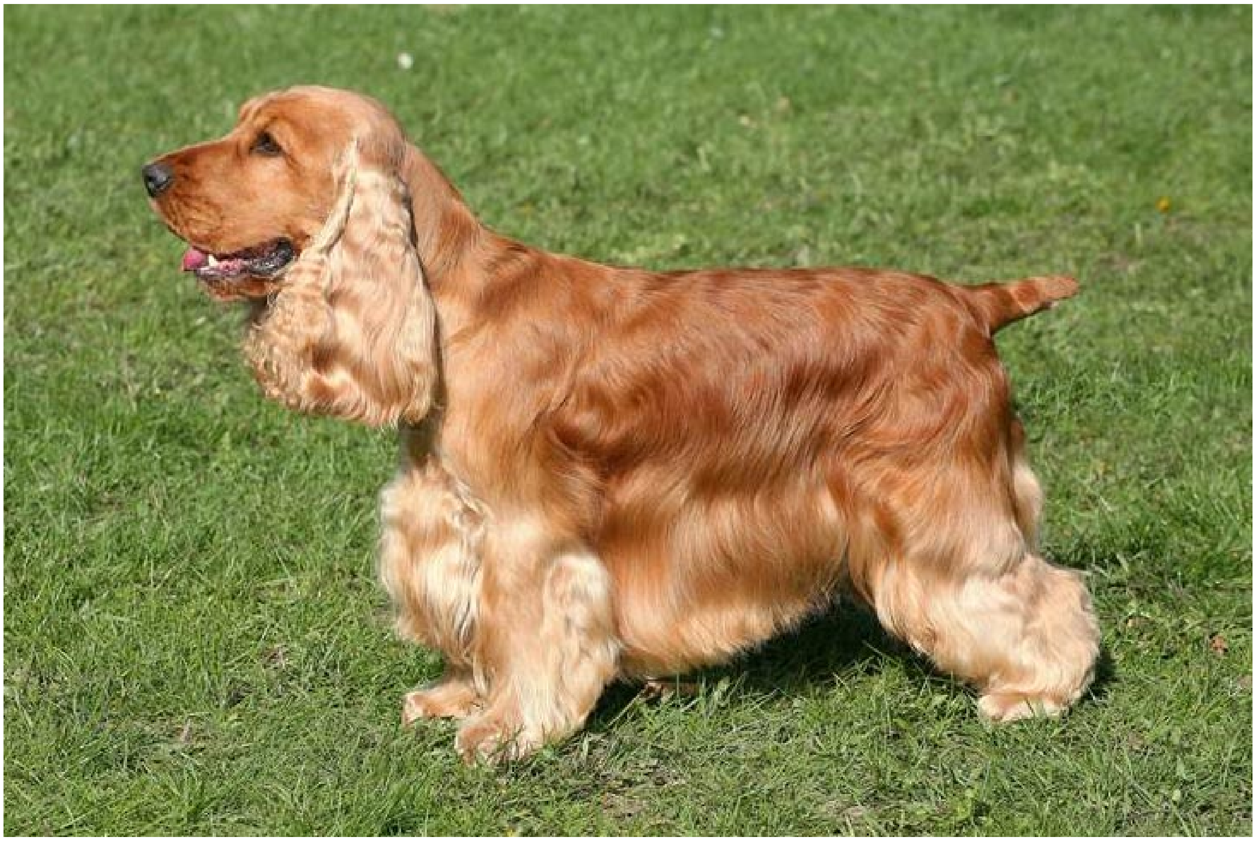
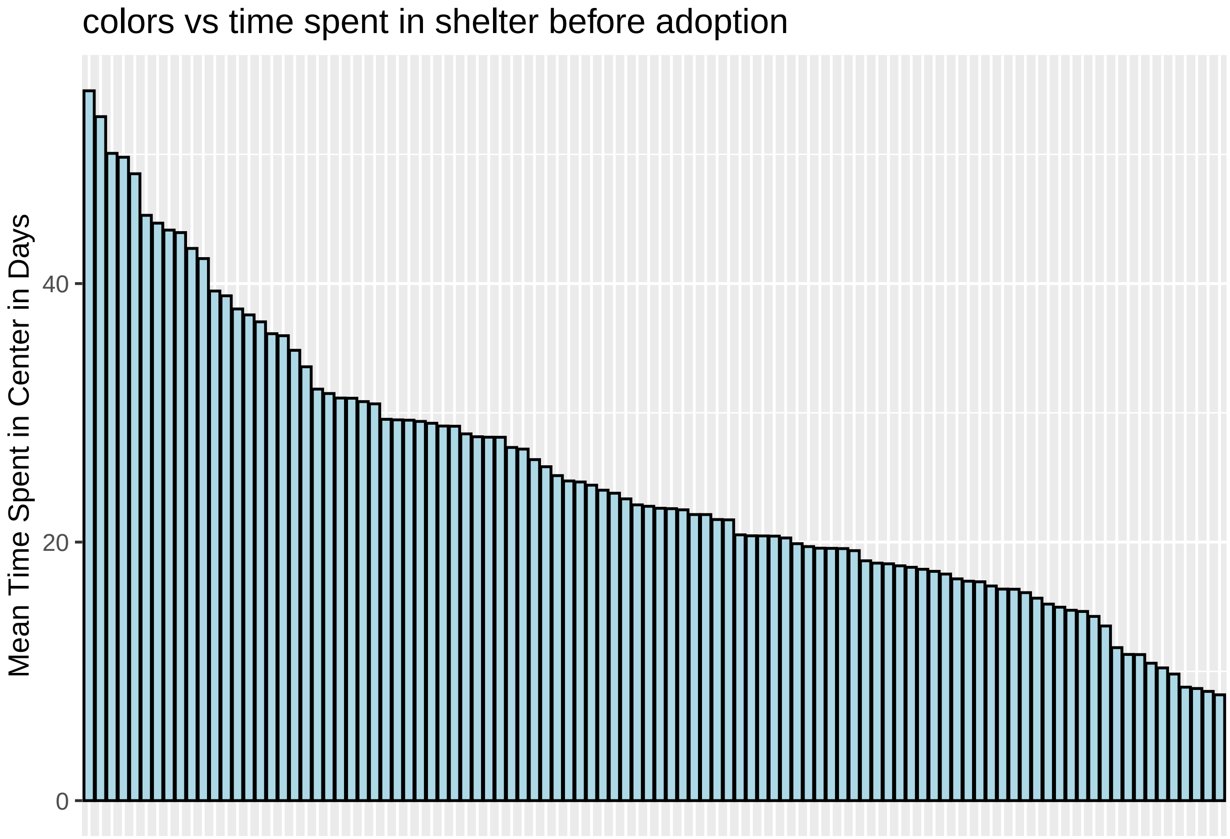


Figure 4: an adult Cocker Spaniel

Color vs Time Before Adoption

First, let's see how each color compares to each other for their mean time before adoption.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Color) %>%
  filter(n()>=15) %>%
  summarise(a=mean(Days.From.Last.Action)) %>%
  ggplot(aes(x=fct_rev(fct_reorder(Color, a)), y=a)) +
  geom_col(fill="light blue", color="black")  +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  ylab("Mean Time Spent in Center in Days") +
  ggtitle("colors vs time spent in shelter before adoption")
```



While there are no colors that stick out, there's seems to be a trend, as some colors have clearly longer mean times than others. Let's look at those at the tail and head.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Color) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange(desc(mean_days_before_adoption)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Color                  mean_days_before_adoption
##    <chr>                                      <dbl>
##  1 Tricolor/Brown Brindle                     148.
##  2 Black Brindle/White                         54.9
##  3 White/Blue                                  52.9
##  4 Blue/White                                  50.1
##  5 Blue Tiger/White                            49.8
##  6 White/Chocolate                             48.5
##  7 Black Tiger                                 47.9
##  8 Blue Tiger                                  47.7
##  9 White/Yellow                                47.0
## 10 Blue                                        45.3
```

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Color) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange((mean_days_before_adoption)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Color        mean_days_before_adoption
##    <chr>                            <dbl>
##  1 Agouti                            2.95
##  2 Blue Smoke                        4.00
##  3 Orange                            4.11
##  4 Calico                            4.49
##  5 Tan/Silver                        4.72
##  6 Gray/Brown                        5.42
##  7 White/Orange                      7.54
##  8 Silver                            8.19
##  9 Black/Cream                       8.45
## 10 Liver                             8.47
```

While it may be hard to draw conclusions from colors alone, it may be illuminating to see if these colors correlate with certain breeds, thus revealing the breed to be what is causing the trend here.

```
tmp <-
  aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Type=="Adoption") %>%
  group_by(Color) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange(desc(mean_days_before_adoption)) %>%
  head(10)

tmp2 <-
aac %>%
  filter(Color %in% tmp$Color) %>%
  group_by(Breed, Color) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'Breed'. You can override using the
## `.groups` argument.
```

```
tmp3 <-
  aac %>%
  filter(Color %in% tmp$Color) %>%
  group_by(Breed, Color) %>%
  summarise(n=n()) %>%
  ungroup() %>%
  group_by(Color) %>%
  summarise(n=max(n))
```

```
## `summarise()` has grouped output by 'Breed'. You can override using the
## `.groups` argument.
```

```
left_join(tmp3, tmp2, by=c("Color", "n")) %>%
  relocate(Breed, .before="n") %>%
  rename("most_common_breed"="Breed",
         "occurences_of_color+breed"="n")
```

```
## # A tibble: 10 x 3
##    Color                  most_common_breed  `occurences_of_color+breed`
##    <chr>                  <chr>                                    <int>
##  1 Black Brindle/White    Pit Bull                                   232
##  2 Black Tiger            Pit Bull                                     4
##  3 Blue                   Pit Bull                                   685
##  4 Blue Tiger             Pit Bull                                    20
##  5 Blue Tiger/White       Pit Bull                                    69
##  6 Blue/White             Pit Bull                                  3635
##  7 Tricolor/Brown Brindle Rat Terrier                                 66
##  8 White/Blue             Pit Bull                                   439
##  9 White/Chocolate        Pit Bull                                    79
## 10 White/Yellow           Labrador Retriever                          30
```

While Pit Bulls are very present in the data regardless of color, some colors that are almost unique to them, such as `Blue Tiger`, are associated with longer adoption times than most other colors. It feels safe to say that color itself has minimal effect on adoption times, and the real impact is coming from the dog's breed (…that being the Pit Bull).

Let's look at the same table for the colors that are adopted the quickest

```
tmp <-
  aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Type=="Adoption") %>%
  group_by(Color) %>%
  filter(n()>10) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange((mean_days_before_adoption)) %>%
  head(10)

tmp2 <-
aac %>%
  filter(Color %in% tmp$Color) %>%
  group_by(Breed, Color) %>%
  summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’Breed’. You can override using the
## ‘.groups‘ argument.
```

```
tmp3 <-
  aac %>%
  filter(Color %in% tmp$Color) %>%
  group_by(Breed, Color) %>%
  summarise(n=n()) %>%
  ungroup() %>%
  group_by(Color) %>%
  summarise(n=max(n))
```

```
## ‘summarise()‘ has grouped output by ’Breed’. You can override using the
## ‘.groups‘ argument.
```

```
left_join(tmp3, tmp2, by=c("Color", "n")) %>%
  relocate(Breed, .before="n") %>%
  rename("most_common_breed"="Breed",
         "occurences_of_color+breed"="n")
```

```
## # A tibble: 10 x 3
##    Color        most_common_breed   ‘occurences_of_color+breed‘
##    <chr>        <chr>                                     <int>
##  1 Apricot      Miniature Poodle                             99
##  2 Black/Cream  German Shepherd                              42
##  3 Black/Red    German Shepherd                              35
##  4 Brown/Cream  Siberian Husky                               14
##  5 Cream/Black  German Shepherd                              32
##  6 Gray/Brown   Miniature Schnauzer                          18
##  7 Red/Brown    Labrador Retriever                           18
##  8 Silver       Miniature Schnauzer                          98
##  9 Tan/Silver   Yorkshire Terrier                            72
## 10 White/Orange Brittany                                      8
```

Following with the previous table, most of the colors that are adopted the quickest are associated with breeds that are adopted quickly, such as the Mini Poodle and Mini Schnauzer. However, we do see some deviations from this trend, such as in the three colors with their most common breed as the German Shepherd. While not exceptionally low, we can see that German Shepherds certainly seem to be on the low end of mean adoption times.

```
aac %>%
  filter(!(is.na(Days.From.Last.Action))) %>%
  filter(Action=="outcome", Type=="Adoption") %>%
  group_by(Breed) %>%
  filter(n()>=15) %>%
  summarise(mean_days_before_adoption=mean(Days.From.Last.Action)) %>%
  arrange(mean_days_before_adoption) %>%
  filter(Breed=="German Shepherd")
```

```
## # A tibble: 1 x 2
##   Breed           mean_days_before_adoption
##   <chr>                               <dbl>
## 1 German Shepherd                      23.0
```

Section 3: Data Modeling

For modeling, we'll be comparing variables against the mean time before being adopted. As such, we will need to modify our dataset somewhat.

```
aac_m <-
  aac %>%
  filter(Type=="Adoption", !(is.na(Days.From.Last.Action)))

split_aac <- initial_split(aac_m, prop=0.85)
training <- training(split_aac)
testing <- testing(split_aac)
```

Breed Coefficients

LM and p-values

```
breed_v_time <- lm(Days.From.Last.Action~Breed, data=training)
coefs <- tidy(breed_v_time)
tbl <- coefs[order(coefs$p.value, decreasing=FALSE), ]
filter(tbl, p.value<=0.05)
```

```
## # A tibble: 8 x 5
##   term                                estimate std.error statistic    p.value
##   <chr>                                  <dbl>     <dbl>     <dbl>      <dbl>
## 1 BreedDogo Argentino                    132.       27.5      4.80 0.00000159
## 2 BreedAmerican Pit Bull Terrier          58.1      23.3      2.50 0.0125    
## 3 BreedAmerican Staffordshire Terrier     57.5      23.4      2.46 0.0140    
## 4 BreedStaffordshire                      56.5      23.1      2.45 0.0145    
## 5 BreedBulldog                            59.4      25.9      2.29 0.0219    
## 6 BreedAmerican Bulldog                   52.9      23.2      2.28 0.0225    
## 7 BreedPit Bull                           51.7      22.9      2.25 0.0243    
## 8 BreedBeauceron                          64.0      28.4      2.25 0.0243    
```

```
anova(breed_v_time)
```

```
## Analysis of Variance Table
##
## Response: Days.From.Last.Action
##              Df    Sum Sq Mean Sq F value    Pr(>F)    
## Breed       193   8227663   42630  11.595 < 2.2e-16 ***
## Residuals 33030 121439792    3677                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Since Breed is categorical, there are >300 p-values in this model; we've filtered for just those that are less than p=0.05, and they're all Pit Bulls or Pit Bull relatives. Not terribly surprising, but it does highlight the low association of a dog's breed and time before adoption except in particular cases. As for its predictions, it tends to predict quite high times for these breeds, mostly in the 50-60 day range.

R-squared

```
summary(breed_v_time) %>%
  glance() %>%
  pull(r.squared) %>%
  round(3)
```

```
## [1] 0.063
```

0.06 is quite a small R-squared value, indicating high variability in adoption times that is unexplained by the dog's Breed, but let's see how it stacks up to our other variables.

Color

LM and p-values

```
color_v_time <- lm(Days.From.Last.Action~Color, data=training)
coefs <- tidy(color_v_time)
tbl <- coefs[order(coefs$p.value, decreasing=FALSE), ]

head(tbl, 5)
```

```
## # A tibble: 5 x 5
##   term                            estimate std.error statistic p.value
##   <chr>                              <dbl>     <dbl>     <dbl>   <dbl>
## 1 ColorTricolor/Brown Brindle        145.       76.0      1.91  0.0565
## 2 ColorBlack Brindle/White            57.9      44.3      1.31  0.191
## 3 ColorBlue/White                     47.9      43.9      1.09  0.275
## 4 ColorWhite/Blue                     47.4      44.3      1.07  0.285
## 5 ColorBlue Tiger/White               48.9      45.8      1.07  0.286
```

```
anova(color_v_time)
```

```
## Analysis of Variance Table
##
## Response: Days.From.Last.Action
##              Df    Sum Sq Mean Sq F value    Pr(>F)
## Color       118   2197305 18621.2  4.8361 < 2.2e-16 ***
## Residuals 33105 127470150  3850.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

While the overall p-value for Color is very small, individual p-values for colors are all quite large, with only one less than 0.1. It was noted that any trend seen in the colors of the dogs could likely be attributed to certain colors being more common with certain breeds, so it tracks that color on its own would not have a terribly strong relationship to adoption times.

There is little point in analysing the coefficients for this variable, as the other statistics tell us that any estimate would likely be inaccurate.

R-squared

```
summary(color_v_time) %>%
  glance() %>%
  pull(r.squared) %>%
  round(3)
```

```
## [1] 0.017
```

An R-squared of 0.017 is, frankly, miniscule, going along with the fact that color and adoption times did not seem to have a very strong relationship at all, as shown by the p-values.

Age Coefficients

```
age_v_time <- lm(Days.From.Last.Action~Age, data=training)
summary(age_v_time)
```

```
##
## Call:
## lm(formula = Days.From.Last.Action ~ Age, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -142.87  -19.07  -13.57   -1.10 1825.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.4553     0.4279   38.46   <2e-16 ***
## Age           6.4372     0.1401   45.96   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 60.58 on 33222 degrees of freedom
## Multiple R-squared:  0.05977,    Adjusted R-squared:  0.05975
## F-statistic:  2112 on 1 and 33222 DF,  p-value: < 2.2e-16
```

```
summary(age_v_time) %>%
  glance() %>%
  pull(r.squared) %>%
  round(3)
```

```
## [1] 0.06
```

While we have a very small p-value, indicating a relationship between our independent and response variables, our R-squared value is also quite low, reflecting quite a bit of variability in adoption times when comparing them to ages. The coefficients tell us that the model predicts an increase in adoption time of about 7 days for each year older a dog is.

Full Model

```
full_model <-
  lm(Days.From.Last.Action~Breed+Age+Color, data=training)
coefs <- tidy(full_model)
tbl <- coefs[order(coefs$p.value, decreasing=FALSE), ]
filter(tbl, p.value<=0.05)
```

```
## # A tibble: 7 x 5
##   term                                estimate std.error statistic    p.value
##   <chr>                                  <dbl>     <dbl>     <dbl>      <dbl>
## 1 Age                                     6.97     0.140     49.9  0
## 2 BreedDogo Argentino                   123.      26.6        4.64 0.00000348
## 3 BreedAmerican Pit Bull Terrier         53.8     22.5        2.39 0.0167
## 4 BreedStaffordshire                     51.3     22.3        2.30 0.0216
## 5 BreedAmerican Staffordshire Terrier    51.6     22.6        2.28 0.0226
## 6 BreedPit Bull                          47.5     22.2        2.14 0.0321
## 7 BreedAmerican Bulldog                  46.9     22.4        2.09 0.0366
```

```
anova(full_model)
```

```
## Analysis of Variance Table
##
## Response: Days.From.Last.Action
##              Df    Sum Sq Mean Sq   F value  Pr(>F)
## Breed       193   8227663   42630   12.4709 < 2e-16 ***
## Age           1   8449185 8449185 2471.6833 < 2e-16 ***
## Color       118    487874    4135    1.2095 0.06079 .
## Residuals 32911 112502733    3418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
full_model %>%
  glance() %>%
  pull(r.squared) %>%
  round(3)
```

```
## [1] 0.132
```

The p-value for Color appears to be quite a bit greater than that for Breed and Age, though it was already noted that the relationship between adoption times and color is exceedingly weak. Our overall R-squared is only 0.159, which, while higher than for the variables on their own, is still very low. However, the low p-values still tell us that there is a relationship present between our variables; albeit, a relationship that is likely quite weak, as only about 16% of variability in adoption times is explained by the variables.

Full Model Testing

```
tmp <-
testing %>%
  filter(Breed %in% training$Breed, Color %in% training$Color)

pred_test <-
tmp %>%
  mutate(prediction=predict(full_model, tmp),
         diff=abs(Days.From.Last.Action-prediction)) %>%
  select(Breed, Days.From.Last.Action, prediction, diff)

head(pred_test)
```

```
## # A tibble: 6 x 4
##   Breed                 Days.From.Last.Action prediction  diff
##   <chr>                                <dbl>      <dbl> <dbl>
## 1 Labrador Retriever                   23.1       136.  113.
## 2 Labrador Retriever                   58.0        81.1  23.1
## 3 German Shepherd                      15.7        81.2  65.5
## 4 German Shepherd                      87.7        73.7  14.0
## 5 Jack Russell Terrier                  3.87       66.6  62.7
## 6 Labrador Retriever                    3.83      119.  115.
```

```
summarise(pred_test, diff_mean=mean(diff),
          testing_sd=sd(Days.From.Last.Action),
          stddevs=diff_mean/testing_sd)
```

```
## # A tibble: 1 x 3
##   diff_mean testing_sd stddevs
##       <dbl>      <dbl>   <dbl>
## 1      29.1       66.8   0.436
```

The mean of the difference between our model's predictions and the true adoption times for our testing data was about 30 days, which was within 0.45 standard deviations of the testing data's adoption times.

While that may seem like a low number of std. deviations, we think this model could fit a bit better to certain parts the data than others.

Fitting to Subsets

We seemed to have the best fit with certain breeds in the dataset, so let's try subsetting and see if we get a better fit. Those breeds that worked the best were those that were related to the Pit Bull, so we'll just test on those breeds.

```
t_breeds <- c("Dogo Argentino", "Pit Bull", "Staffordshire", "American Staffordshire Terrier", "American

aac_p <-
aac_m %>%
  filter(Breed %in% t_breeds)
```

```r
split_aac_p <- initial_split(aac_p, prop=0.85)
training_p <- training(split_aac_p)
testing_p <- testing(split_aac_p)

model_p <- lm(Days.From.Last.Action~Breed+Age+Color, data=training_p)
coefs_p <- tidy(model_p)
tbl_p <- coefs[order(coefs_p$p.value, decreasing=FALSE), ]
filter(tbl_p, p.value<=0.05)
```

```
## # A tibble: 4 x 5
##   term                                estimate std.error statistic    p.value
##   <chr>                                  <dbl>     <dbl>     <dbl>      <dbl>
## 1 BreedAmerican Bulldog                   46.9      22.4      2.09 0.0366
## 2 BreedDogo Argentino                    123.       26.6      4.64 0.00000348
## 3 BreedAmerican Staffordshire Terrier     51.6      22.6      2.28 0.0226
## 4 BreedAmerican Pit Bull Terrier          53.8      22.5      2.39 0.0167
```

```r
anova(model_p)
```

```
## Analysis of Variance Table
##
## Response: Days.From.Last.Action
##             Df   Sum Sq Mean Sq  F value    Pr(>F)
## Breed        6   148938   24823   3.0893  0.005074 **
## Age          1  6908320 6908320 859.7627 < 2.2e-16 ***
## Color       81   578293    7139   0.8885  0.752012
## Residuals 5444 43743340    8035
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
model_p %>%
  glance() %>%
  pull(r.squared) %>%
  round(3)
```

```
## [1] 0.149
```

While our R-squared was greater (though negligably), our p-value for breed went up, indicating that there is less confidence in the relationship between breed and adoption times. This makes sense considering that, when filtering to only breeds that, coincidentally, tend to have greater adoption times, any sort of relationship becomes less clear; however, the variance in the times becomes lesser, relfecting the increased R-squared value.

Conclusion

Insights

Overall, our models were largely unsuccessful in being useful for predictive purposes. They do give us insight, however; the low (<0.15) R-squared values despite low p-values indicate that, while a relationship is certainly present in the variables, it is very weak, and a large majority of the variance in the data is unaccounted for.

The most likely reason for this outcome is the small span of our variables when trying to answer quite a complex question. There are far more forces at play besides the attributes of the dog itself when one considers adopting, which would span far beyond the scope of this dataset (i.e. the kind of place or economic situation that would allow for people to own pets).

Additionally, it is worth noting that the rather large size of the dataset likely contributed to those low p-values. With enough data points, a trend became apparent, albeit a rather weak trend, illustrated by our also low R-squared values.

Improvments

Pairing this dataset with others could likely lead to much deeper analysis, allowing us to consider variables beyond the scope of the Austin Animal Center itself. For example, a dataset for home prices in Austin over time could lead to interesting analysis, as the `DateTime` column in our data would allow us to see if higher home prices may lead to less adoptions; additionally, we could analyse parts besides adoption times, such as the age and breeds of the dogs coming in/out of the shelter compared to home prices.

As far as our data modeling, a different, nonlinear approach may have lent itself to better results. Especially in a variable such as `Age`, as we earlier observed a plateau of adoption times after the dogs were older than 3 years old.

Reflection

Despite what may seem like little substantial results in our last data modeling section, we would say that, overall, our analysis of this dataset could be considered “successful”; that is to say, we uncovered quite a few trends that in hindsight may seem somewhat obvious, but we would certainly not be able to say them with as much confidence as we do now after our analysis. Pit Bulls are known for the aggressive reputation that they carry, and thus many may not want to adopt them due to the fear of being attacked by their own dog, but to see this fact manifest so strongly within our data was still quite interesting and surprising. Even though breeds such as the Dogo Argentino and Bulldogs are, at this point, almost completely separate from Pit Bulls genetically, the fact that they are treated the same by adopters confirms the belief that looks play the biggest role in deciding which dog people will adopt (and, frankly, looks play the biggest role when people judge things, in general). This same idea applies, somewhat, to the fact that age played such a noticeable role, as well. While a younger dog may be desirable because they live for longer, “puppies” are regarded as such because they are simply seen as cuter to most adopters.

While our exploratory analysis may illuminate trends in one way, our data models must not be brushed away. When considering whether the variables we analysed truly had any effect on the time taken for adoption, we confident in saying that while our variables did influence the time taken for adoption for the dogs, it was not a strong effect.