

Stage 2: Exploratory Data Analysis

Isaac Slagel and Jack Welsh

4/18/2019

EDA Main Report

Introduction

Our project plans to look into trend in animal treatment in Dallas animal shelters. Using a dataset containing information on all recent animals brought into the shelter, we are going to try to answer some of the following questions. What traits affect a dog's probability of adoption? Are there seasonal trends in the adoption of different animals? Are animals who enter the animal shelter with chips more likely to be returned to their owner? On the table below you can see some of the variables we are working with in the dataset.

Dogs

Dogs are the most common animal winding up in the animal shelters which form our dataset. From our total of 44,194 dogs, 35% were adopted, 30% were returned to owners, 16% were trasfered, and 15% were euthanized. The most common breed of dog in our dataset is the pitbull, with 10,033 being admitted to animal shelters in between 10/01/2017 and 04/03/2019. We are especially interested in how pitbulls are treated within our datasets as these dogs have an infamous reputation of being overly aggressive. Is it possible that this reputation results in a lower adoption rate for pitbulls that other breeds?

Pitbulls

To look into how pitbulls are handled within animal shelters, we decided to explore how the outcomes of pitbulls may differ from non-pitbulls. Table 1 describes the differences in outcome rates for non-pitbulls compared to pitbulls.

Table 1: Pitbull Outcomes		
Outcome	Pitbull (%)	Non-Pitbull (%)
Adoption	32.10	36.04
Euthanized	28.67	10.43
Returned to owner	22.03	31.01
Transfer	10.98	17.74
Foster	2.16	1.57
Other	1.83	0.99
Dead on arrival	0.88	0.83
Treatment	0.79	0.95
Died	0.52	0.39
Missing	0.04	0.06

We see that pitbulls, while only adopted at a slightly lower rate (~5%), are euthanized at well over double the rate of other dogs. Further, we see that other dogs have a much higher chance of being trasfered to another facility or returned to their owner. It would be valuble to get a better understanding of this relationship, but we likely need to control for things like chip status and intake condition. Moving foward we would like to look into some binomial regression to look into the outcomes of pitbulls. To do this we plan to make a more general outcome variable of Dead or Alive status (to reduce the dimension of the current outcome_type variable). In fitting this model we plan to control for intake status (with which we still need to do some work on string analysis), chip_status, and month.

Dogs vs Cats

In Minnesota and Iowa, our home states, animal shelters often report having more difficulty dealing with stray cats. There are lots of problems that feral cats cause including environmental damage and spread of disease. We were interested if we can see these kinds of differences in our data. Do cats spend longer in animal shelters? Are they euthanized at higher rates? To describe some of these differences we looked we created a summary graph of different characteristics of dogs and cats. Since our dataset contains observations of 44194 dogs and only 12891 cats, we created this chart in terms of proportion of each animal matching a certain criteria.

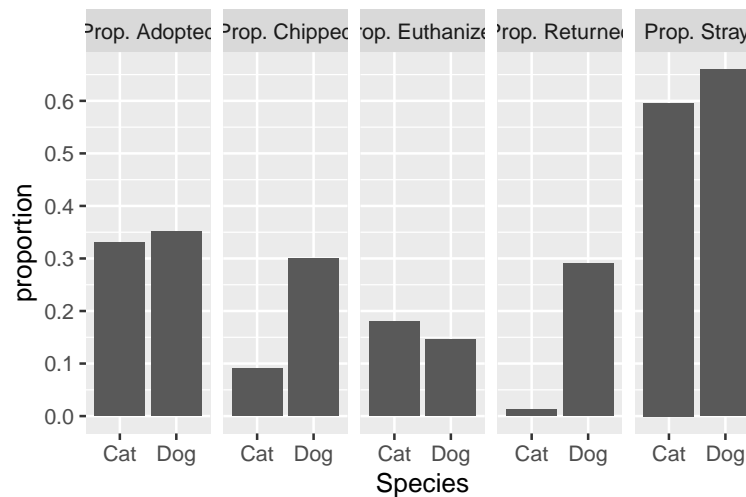


Figure 1: Cat vs Dog Characteristics

Interestingly enough, we see that the difference between proportions of cats and dogs adopted is not very pronounced. A slightly lower percentage of cats are adopted and a slightly higher percent of cats are euthanized compared to dogs. However, we do see that very few cats entering humane societies have scannable chips, and an even lower proportion are returned to their owner. Additionally it seems appears that fewer cats are brought into the shelter as strays. We may use this data to do another binomial analysis to piece apart differences between cat and dog adoptions. We may do a similar thing as the pitbull analysis with the creation of a dead or alive outcome variable. We would want to consider controlling for in this analysis would be breed, intake condition, and chip status.

Time of Year

Another question that we had in our Stage 1 report was what seasonal trends exists in adoptions. One could imagine a situation where more puppies are adopted in December or more rabbits are adopted in early spring. To explore this we did some plotting to see how many animals were adopted in each months.

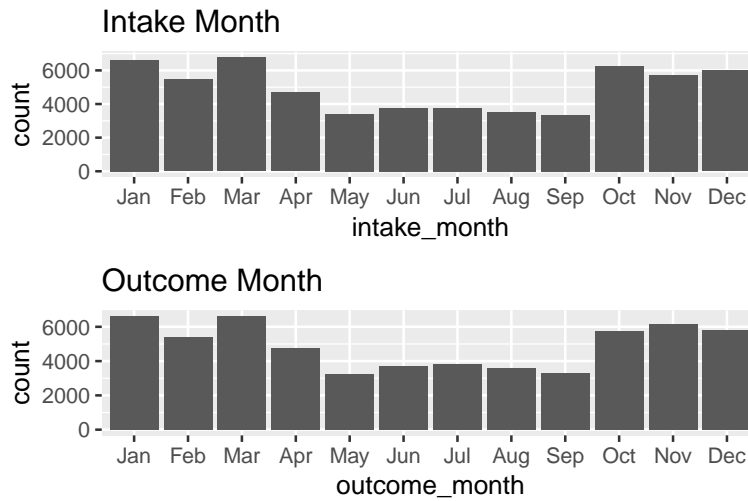


Figure 2: Animal Adoptions by Month

These trends surprised us. We see a large drop in adoptions during the summer but a spike in the fall, winter, and early spring. These trends were true for both Dog and Cat adoptions (see appendix). For rabbits, many were adopted in February and July. While we are unable to put together a rigorous defense for why these trends are occurring, we see that we may need to control for it in our binomial analysis.

Annotated Appendix and References

Variables

Table 2: Description of Variables

Variable Name	Variable Role	Variable Type	Range of Values	Units
animal_breed	explanatory	categorical	296 unique breeds	NA
animal_origin	explanatory	categorical	4 sources of shelter animals	NA
animal_type	explanatory	categorical	5 species of animal	NA
chip_status	explanatory	bianary	(0,1)	NA
intake_type	potential confounder	categorical	how animal came to be at the shelter	NA
outcome_type	reponse	categorical	how animals was removed from shelter	NA
intake_condition	potential confounder	categorical	keyword description of animal status at intake	NA
outcome_condition	potential confounder	categorical	keyword description of animal status at outcome	NA
intake_date	response	date	(2017-10-01, 2019-04-03)	y-m-d
outcome_date	response	date	(2017-10-01, 2019-04-03)	y-m-d

Citations

1. Lepper, M., Kass, P. H., & Hart, L. A. (2002). Prediction of adoption versus euthanasia among dogs and cats in a California animal shelter. *Journal of Applied Animal Welfare Science*, 5(1), 29-42.
2. Posage, J. M., Bartlett, P. C., & Thomas, D. K. (1998). Determining factors for successful adoption of dogs from an animal shelter. *Journal of the American Veterinary Medical Association*, 213(4), 478-482.
3. Lampe, R., & Witte, T. H. (2015). Speed of dog adoption: Impact of online photo traits. *Journal of applied animal welfare science*, 18(4), 343-354.

Exploring animal_type

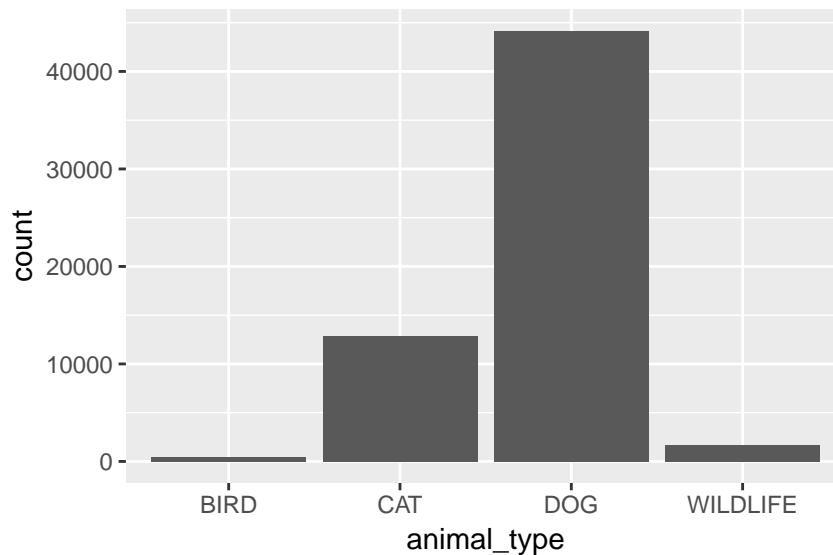
```
# animal type
adoptions %>%
  group_by(animal_type) %>%
  count(sort = TRUE)

## # A tibble: 6 x 2
## # Groups:   animal_type [6]
##   animal_type     n
##   <chr>         <int>
## 1 DOG           44194
## 2 CAT           12891
## 3 WILDLIFE       1696
## 4 BIRD            485
## 5 LIVESTOCK        33
## 6 D                1
```

```

adoptions %>%
  filter(!animal_type%in%c("D", "LIVESTOCK")) %>%
  ggplot(aes(x=animal_type))+
  geom_bar()

```



We see that most observations in our dataset were of dogs, followed by cats. It seems there is 1 mislabeled dog which we will take care of in our data cleaning file.

Exploring animal_breed

```

## Which dog breeds are in our data?
adoptions %>%
  filter(animal_type == "DOG") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=15)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)

```

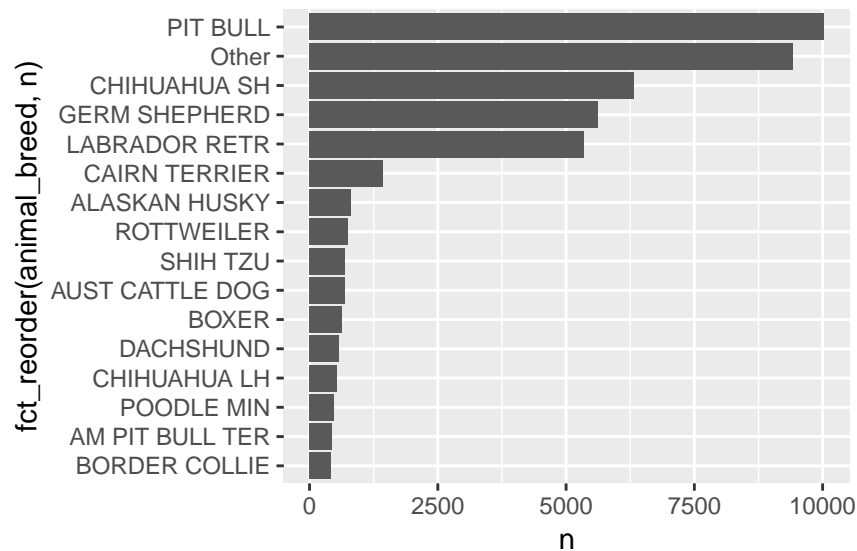
```

## # A tibble: 16 x 2
## # Groups:   animal_breed [16]
##   animal_breed      n
##   <fct>          <int>
## 1 PIT BULL       10033
## 2 Other          9423
## 3 CHIHUAHUA SH   6322
## 4 GERM SHEPHERD  5625
## 5 LABRADOR RETR  5353
## 6 CAIRN TERRIER  1430
## 7 ALASKAN HUSKY   812
## 8 ROTTWEILER      751
## 9 SHIH TZU        696
## 10 AUST CATTLE DOG 683
## 11 BOXER          629
## 12 DACHSHUND       581
## 13 CHIHUAHUA LH    528
## 14 POODLE MIN      476
## 15 AM PIT BULL TER 433

```

```
## 16 BORDER COLLIE      419
```

```
adoptions %>%
  filter(animal_type == "DOG") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=15)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)%>%
  ggplot(aes(x=fct_reorder(animal_breed,n), y=n))+
  geom_bar(stat="identity")+
  coord_flip()
```



We have a huge variety of dog breeds in our data. The pmost common breed is the pitbull, which we are planning to use for our analysis. There are many other dog breeds that are seen as agresive (rottweilers and boxers) that we may make into a new variable `aggressive`.

```
## Which cat breeds are in our data?
```

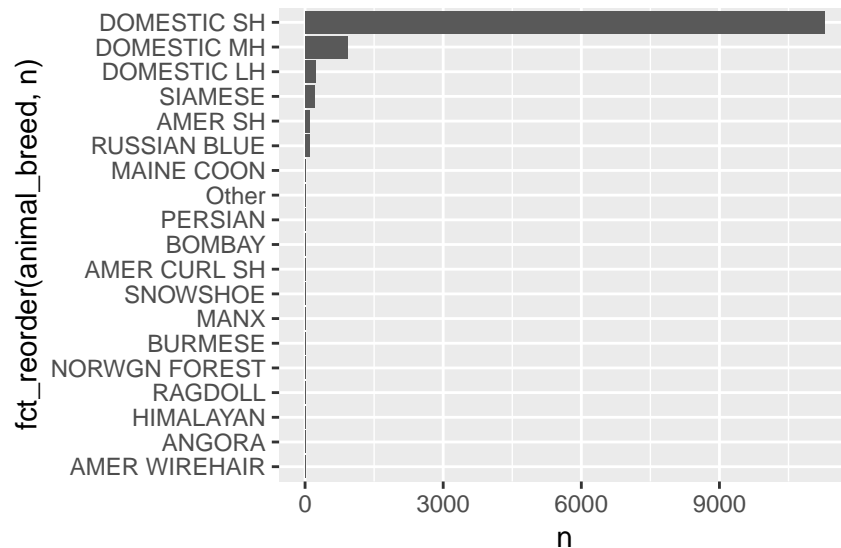
```
adoptions %>%
  filter(animal_type == "CAT") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=10)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)
```

```
## # A tibble: 11 x 2
## # Groups:   animal_breed [11]
##   animal_breed      n
##   <fct>         <int>
## 1 DOMESTIC SH  11273
## 2 DOMESTIC MH   916
## 3 DOMESTIC LH   219
## 4 SIAMESE       214
## 5 AMER SH       101
## 6 RUSSIAN BLUE   96
## 7 Other         33
## 8 MAINE COON     19
## 9 PERSIAN        8
## 10 AMER CURL SH   6
## 11 BOMBAY        6
```

```

adoptions %>%
  filter(animal_type == "CAT") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=15)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)%>%
  ggplot(aes(x=fct_reorder(animal_breed,n), y=n))+
  geom_bar(stat="identity")+
  coord_flip()

```



Almost all of the cats in our data were domestic short hair cats.

```

## Which types of wildlife are in our data?
adoptions %>%
  filter(animal_type == "WILDLIFE") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=10)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)

```

```

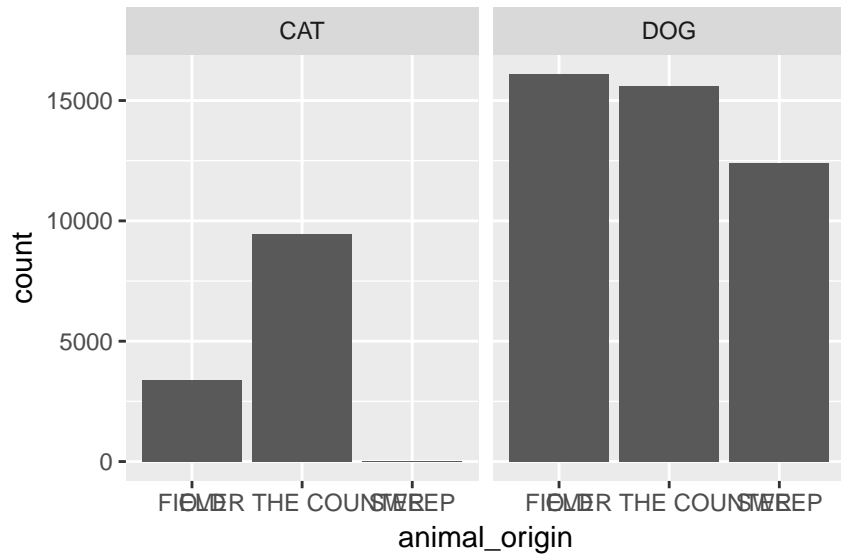
## # A tibble: 11 x 2
## # Groups:   animal_breed [11]
##   animal_breed      n
##   <fct>          <int>
## 1 RACCOON         471
## 2 OPOSSUM         404
## 3 TURTLE          190
## 4 GUINEA PIG      148
## 5 RABBIT SH       136
## 6 HAMSTER         98
## 7 SQUIRREL        82
## 8 Other           68
## 9 BAT             59
## 10 FOX            23
## 11 SKUNK           17

```

This variable is more for fun. Look at all these fun creatures they get to deal with at the animal shelter. We are kind of interested in rabbits so we will talk about that later.

Exploring animal origin

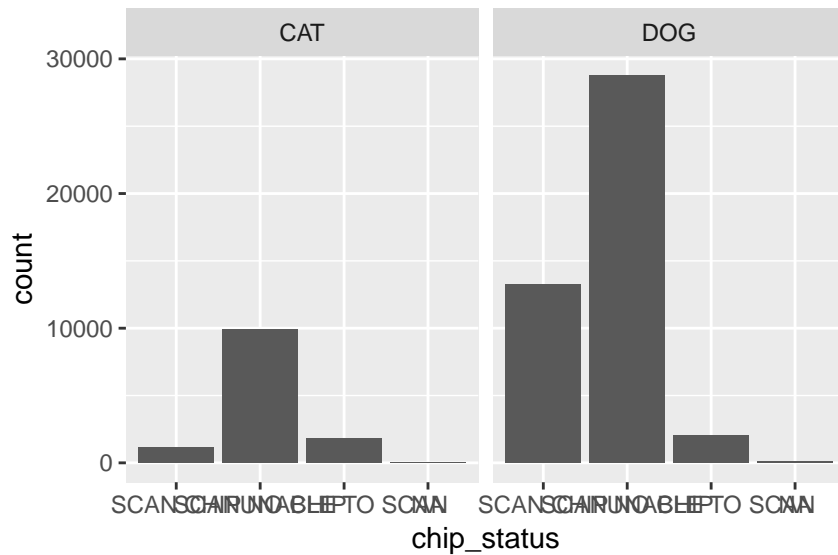
```
adoptions %>%  
  filter(dog == 1 | cat == 1) %>%  
  filter(!is.na(animal_origin)) %>%  
  ggplot(aes(x=animal_origin))+  
  geom_bar()+  
  facet_grid(~animal_type)
```



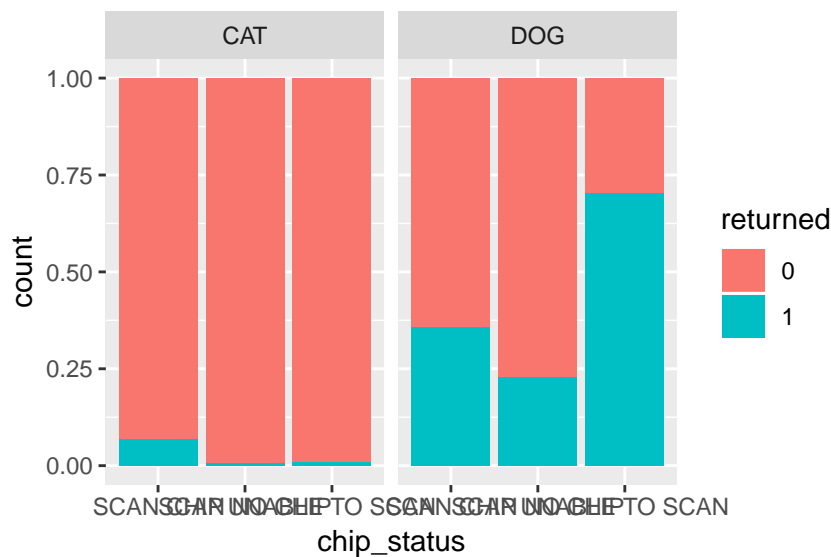
Dogs and cats can come from a 3 possible origins: field, over the counter, and sweep. As of this moment, we are not sure what the difference between field and sweep are. However, it is interesting that only dogs are brought in from sweeps.

Exploring chip status

```
## General distribution of variable  
adoptions %>%  
  filter(dog == 1 | cat == 1) %>%  
  ggplot(aes(x=chip_status))+  
  geom_bar()+  
  facet_grid(~animal_type)
```

```
## Relationship with returned to owner status
adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  filter(!is.na(chip_status)) %>%
  mutate(returned = as.factor(ifelse(outcome_type == "RETURNED TO OWNER", 1, 0))) %>%
  ggplot(aes(x=chip_status, fill = returned)) +
  geom_bar(position = "fill") +
  facet_grid(~animal_type)
```



We see that most cats and dogs do not have chips. However, a larger proportion of dogs have chips than cats. A higher proportion of dogs are returned to owners than cats in all three categories of chip status.

Exploring intake_subtype

```
## Which dog breeds are in our data?
adoptions %>%
  mutate(intake_subtype = fct_lump(intake_subtype, n=10)) %>%
  group_by(intake_subtype) %>%
```

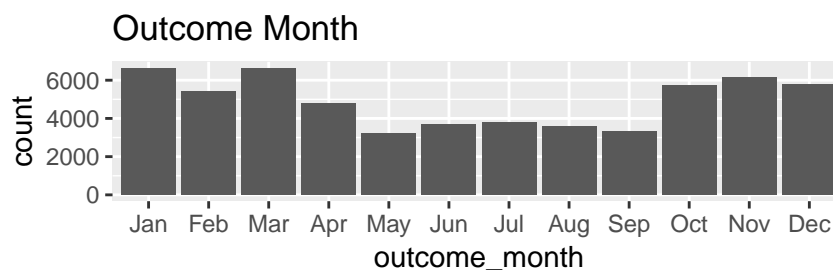
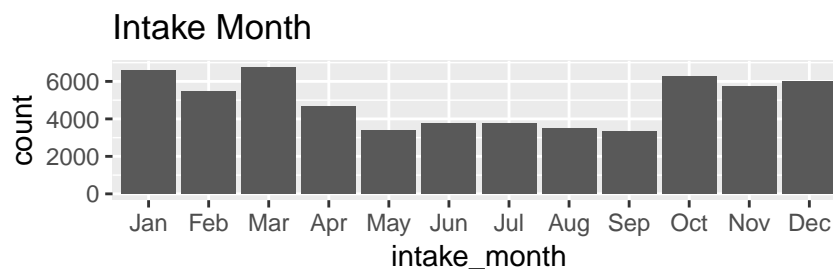
```
count(sort = TRUE)
```

```
## # A tibble: 11 x 2
## # Groups:   intake_subtype [11]
##   intake_subtype      n
##   <fct>          <int>
## 1 AT LARGE        28889
## 2 GENERAL         15495
## 3 CONFINED         5598
## 4 POSSIBLY OWNED   2085
## 5 Other           1813
## 6 QUARANTINE       1454
## 7 RETURN30        1312
## 8 INJURED          972
## 9 KEEP SAFE        781
## 10 UNINJURED        474
## 11 EUTHANASIA REQUESTED 427
```

This general intake listing does not give us a lot of information. However we do see that a few animals may have been returned (non-independent observation), and some animals are requested to be euthanized upon admission.

Exploring intake_date and outcome_date

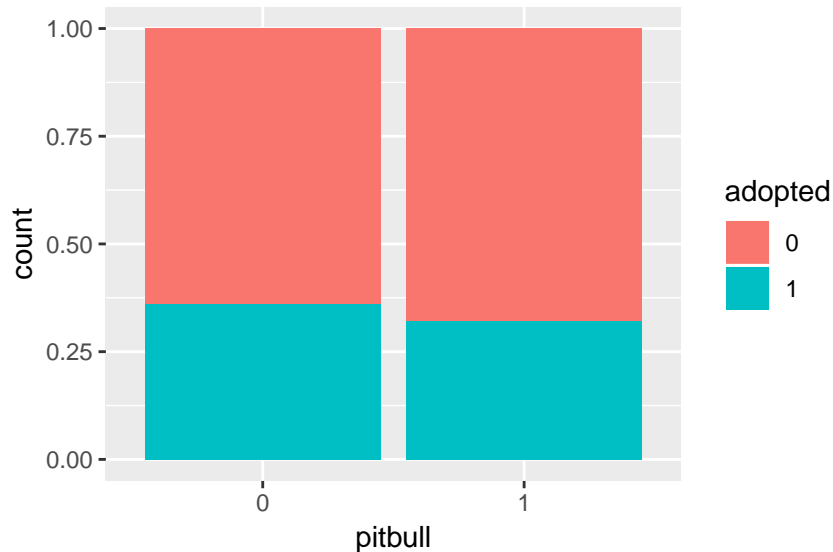
```
intake <- adoptions %>%
  mutate(intake_month = month(intake_date, label = TRUE)) %>%
  ggplot(aes(x=intake_month))+
  geom_bar()+ggtitle("Intake Month")
outcome <- adoptions %>%
  filter(!is.na(outcome_date)) %>%
  mutate(outcome_month = month(outcome_date, label = TRUE)) %>%
  ggplot(aes(x=outcome_month))+
  geom_bar()+ ggtitle("Outcome Month")
gridExtra::grid.arrange(intake, outcome)
```



Fewer dogs seem to be adopted in the summer compared to other months.

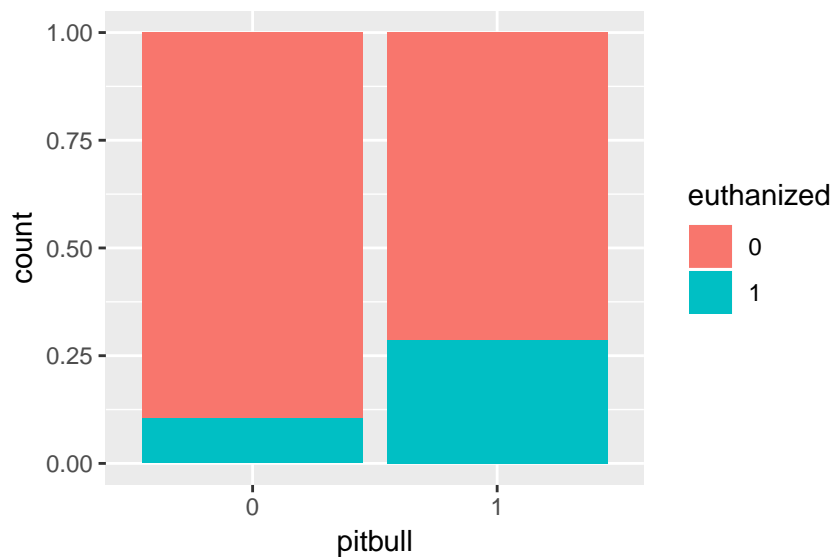
Pitbulls and Adoptions

```
adoptions %>%  
  filter(dog == 1) %>%  
  mutate(adopted = as.factor(adopted),  
         pitbull = as.factor(pitbull)) %>%  
  ggplot(aes(x = pitbull, fill = adopted))+  
  geom_bar(position = "fill")
```



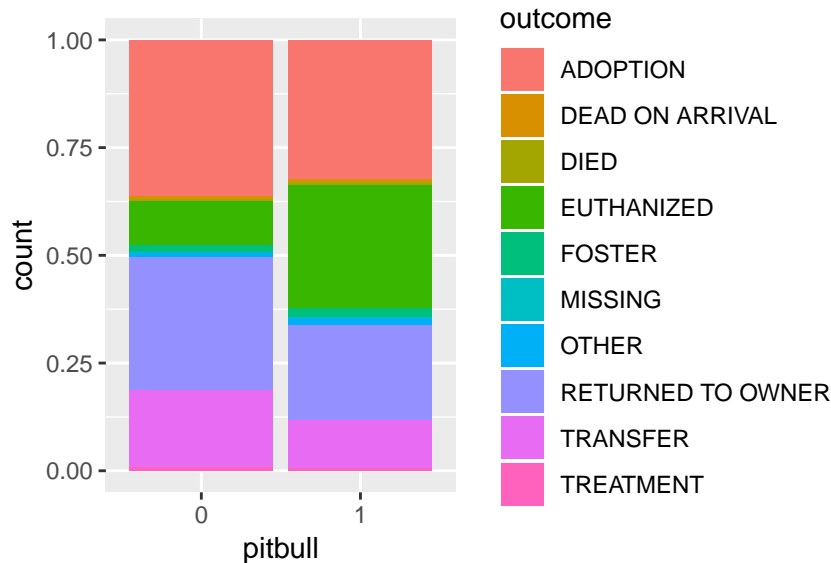
There does not seem to be a large difference between the adoption rates of pitbull and non-pitbulls. We should try and filter out some of the non-adopted dogs though, not all dogs come to the shelter to be adopted.

```
adoptions %>%  
  filter(dog == 1) %>%  
  mutate(euthanized = as.factor(euthanized),  
         pitbull = as.factor(pitbull)) %>%  
  ggplot(aes(x = pitbull, fill = euthanized))+  
  geom_bar(position = "fill")
```



While pitbulls are not adopted at a higher rate, they are euthanized at a higher rate than other dogs. Let's look at all possible outcomes to see how this is possible.

```
adoptions %>%
  filter(dog == 1) %>%
  mutate(outcome = as.factor(outcome_type),
         pitbull = as.factor(pitbull)) %>%
  ggplot(aes(x = pitbull, fill = outcome)) +
  geom_bar(position = "fill")
```



```
adoptions %>%
  filter(dog == 1) %>%
  mutate(outcome = as.factor(outcome_type),
         pitbull = as.factor(pitbull)) %>%
  group_by(pitbull, outcome_type) %>%
  count(name = "freq") %>%
  group_by(pitbull) %>%
  mutate(freq = freq / sum(freq)) %>%
  spread(outcome_type, freq)
```

```
## # A tibble: 2 x 11
## # Groups:   pitbull [2]
##   pitbull ADOPTION `DEAD ON ARRIVAL` DIED EUTHANIZED FOSTER MISSING
##   <fct>      <dbl>          <dbl> <dbl>    <dbl> <dbl> <dbl>
## 1 0          0.360          0.00831 0.00386    0.104 0.0157 6.15e-4
## 2 1          0.321          0.00877 0.00518    0.287 0.0216 3.99e-4
## # ... with 4 more variables: OTHER <dbl>, `RETURNED TO OWNER` <dbl>,
## #   TRANSFER <dbl>, TREATMENT <dbl>
```

A far higher percentage of pitbulls are euthanized than other breeds. Other dogs are adopted, returned to owners, and transferred at a higher rate.

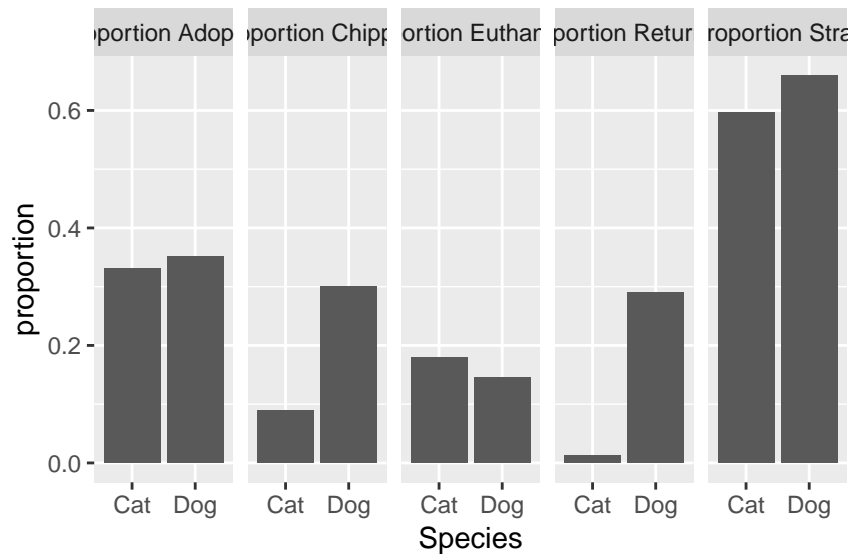
Differences between dogs and cats

```
adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  mutate(Species = ifelse(dog == 1, "Dog", "Cat")) %>%
```

```

mutate(returned = ifelse(outcome_type == "RETURNED TO OWNER", 1, 0)) %>%
group_by(Species) %>%
summarise(count = n(),
           meanDays = mean(shelter_days, na.rm = TRUE),
           `Proportion Euthanized` = sum(euthanized)/count,
           `Proportion Adopted` = sum(adopted)/count,
           `Proportion Stray` = sum(stray)/count,
           `Proportion Returned` = sum(returned)/count,
           `Proportion Chipped` = sum(chip_status == "SCAN CHIP", na.rm = TRUE)/count)%>%
gather(key = key, value = proportion, 4:8) %>%
ggplot(aes(x=Species, y= proportion)) +
  geom_bar(stat = "identity")+
  facet_grid(cols = vars(key))

```



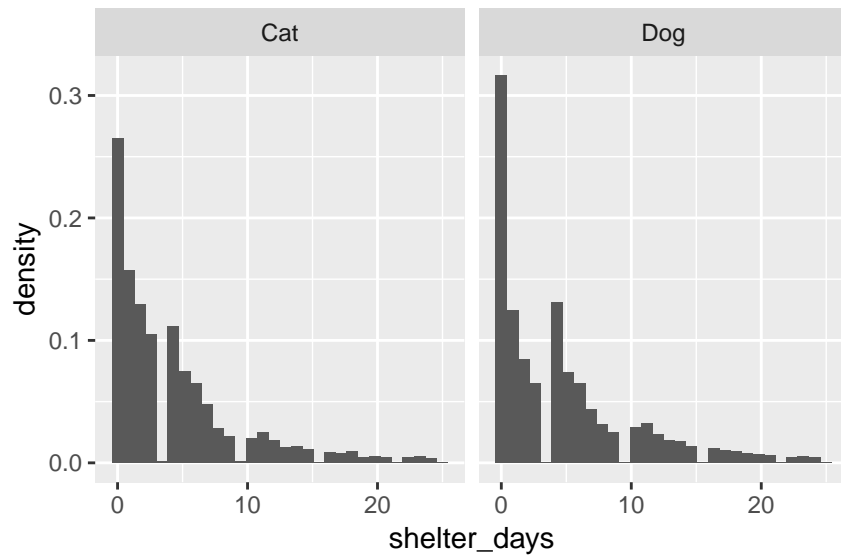
Cats are returned at a much lower rate than dogs. Cats have a much lower proportion with scannable chips than dogs.

```

adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  filter(shelter_days < 25) %>%
  mutate(DogCat = ifelse(dog == 1, "Dog", "Cat")) %>%
  ggplot(aes(x=shelter_days, y=..density..)) +
  geom_histogram()+
  facet_grid(~DogCat)

```

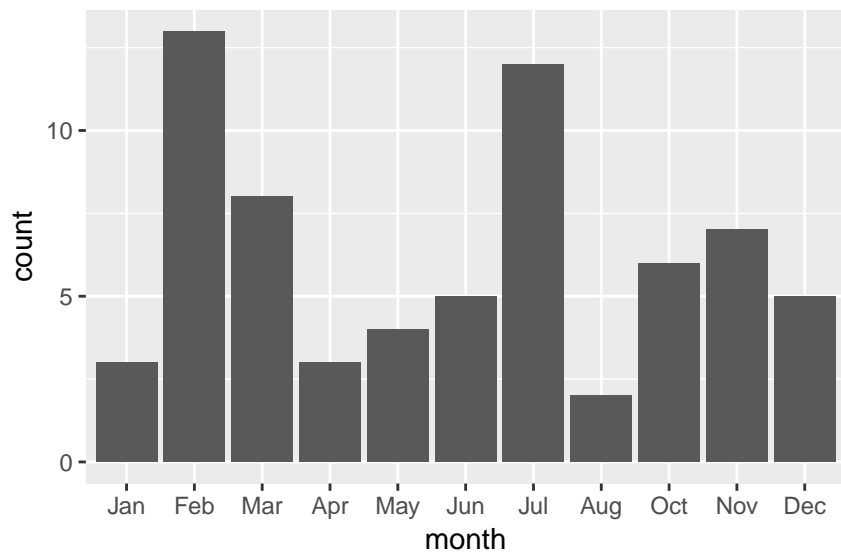
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



Dogs and cats show similar trends in the number of days until adoptions.

What about rabbits?

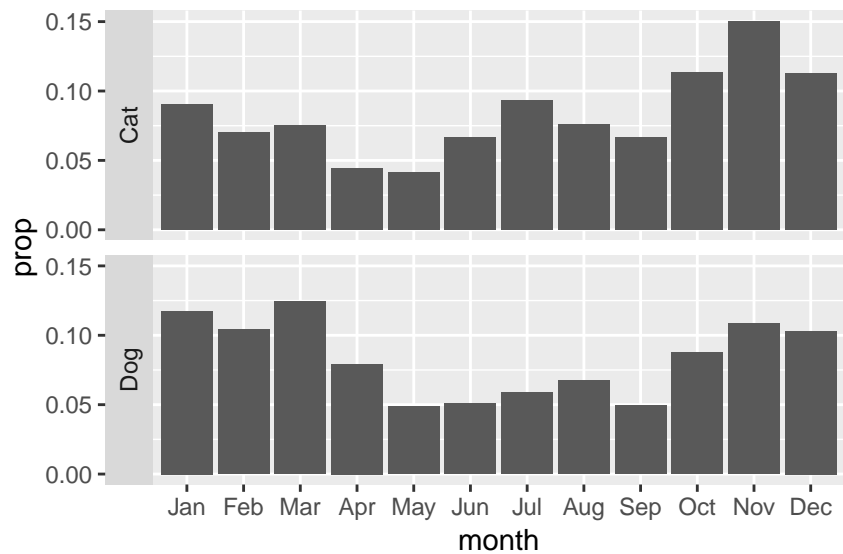
```
adoptions %>%
  filter(animal_breed == "RABBIT SH") %>%
  filter(adopted == 1) %>%
  mutate(month = lubridate::month(outcome_date, label = TRUE)) %>%
  ggplot(aes(x=month)) +
  geom_bar()
```



Alright, so maybe it is not the case that more rabbits are adopted over easter. Are there trends in dog adoptions?

```
adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  mutate(DogCat = ifelse(dog == 1, "Dog", "Cat")) %>%
  filter(adopted == 1, !is.na(month)) %>%
  mutate(month = lubridate::month(outcome_date, label = TRUE)) %>%
```

```
ggplot(aes(x=month, y = ..prop.., group = 1)) +  
  geom_bar() +  
  facet_grid(rows = vars(DogCat), switch = "both")
```



These both seem to follow the trend of fewer adoptions in the summer.