Stage 2: Exploratory Data Analysis

Isaac Slagel and Jack Welsh

4/18/2019

2. EDA Report. In no more than 3 pages, summarize the main findings of your exploratory analysis, referring to specific plots and summary statistics where necessary. In addition, describe your plans for building models to address important research questions, including which variables will be important to consider in light of your exploratory analyses.

• No more than 3 pages • Begin with a short paragraph introducing your project and primary research questions. (This introduction will be expanded into several paragraphs for the final paper.) • Use your graphical and numerical summaries to tell a story, supporting your conclusions with summary statistics. Weave numerical summaries seamlessly into your text, and refer to graphs where appropriate. • Include at least 2 interesting plots (hopefully more!). Name each plot (e.g. Figure 1) so they are easily referred to in your report, and format the figures neatly within your report (without taking up too much space). Exploratory plots can be loose with captions and axis labels, but for your final paper it is essential that your figures have meaningful captions and axis labels! • Preview directions you plan to go with modeling. What models will you begin by fitting, and what variables will be involved. • Write well! Complete sentences, good flow, proper grammar, the works...

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.

Variable Name	Variable Role	Variable Type	Range of Values	Units of Measurement
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	categorial	how animal came to be at the shelter	NA
outcome_type	reponse	categorial	how animals was removed from shelter	NA
intake_condition	potential confounder	categorial	keyword description of animal status at intake	NA
outcome_condition	potential confounder	categorial	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

What's the deal with the 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 admited 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 agressive. 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.

Annotated Appendix and References

(b) Your Annotated Appendix and References section should include these elements: • Definitions of important variables and the source of the data. • R scripts and (commented) output so that I can trace how you constructed your final data set, what the results of your exploratory data analyses were, and what plots you generated. • A short annotation – one or two sentences – on what each analysis shows. • Tables and figures that are informative but were not referenced specifically in the main report. Include a short annotation – one or two sentences on what they show. • A citation for each reference article (in APA format or something similar) you included in your proposal. Also include a link, if appropriate. Remember that you must have the entire paper and not just an abstract, and at least two must be from peer-reviewed journals.

Exploring animal_breed

```
# animal type
adoptions %>%
  filter(animal_type=="DOG") %>%
  group_by(outcome_type) %>%
  summarise(count = n(), prop = count/44194)
```

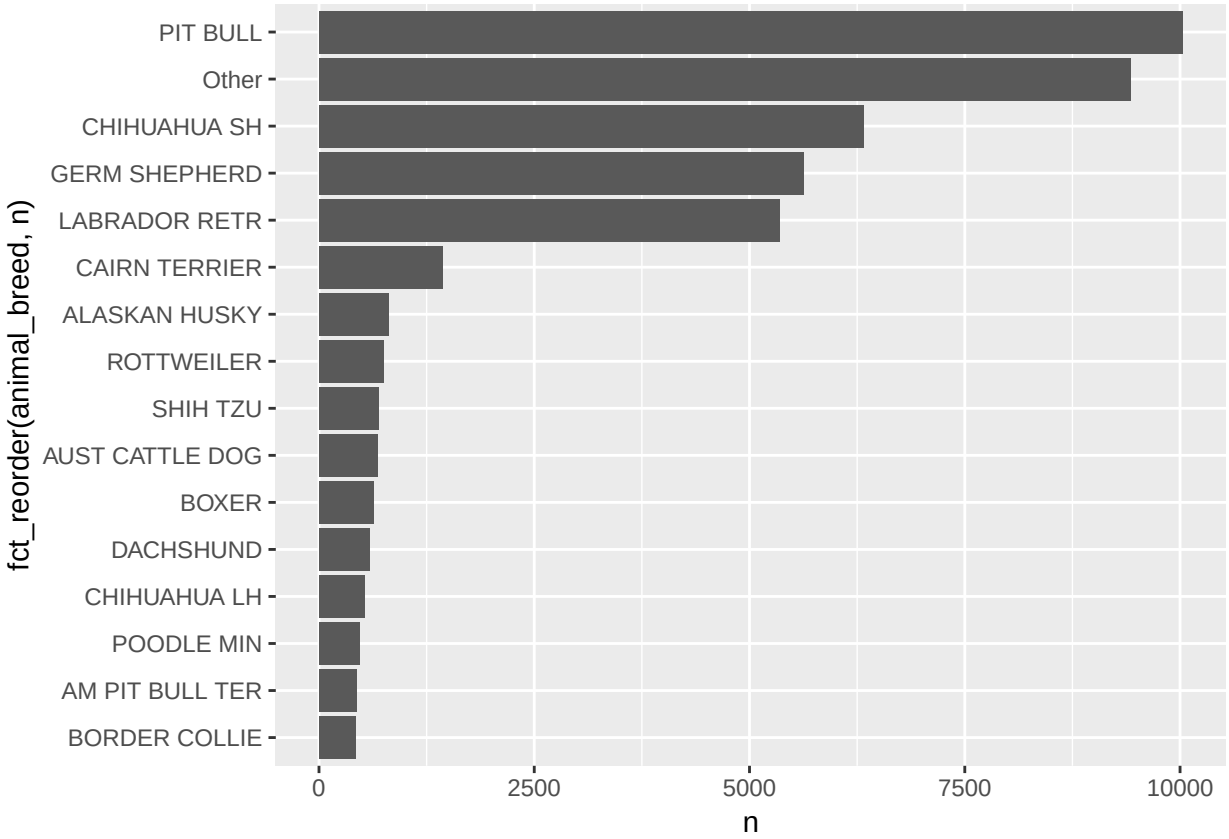
```
## # A tibble: 10 x 3
##    outcome_type      count     prop
##    <chr>             <int>    <dbl>
##  1 ADOPTION          15531 0.351
##  2 DEAD ON ARRIVAL     372 0.00842
##  3 DIED                184 0.00416
##  4 EUTHANIZED         6438 0.146
##  5 FOSTER              754 0.0171
##  6 MISSING              25 0.000566
##  7 OTHER               522 0.0118
##  8 RETURNED TO OWNER 12803 0.290
##  9 TRANSFER           7163 0.162
## 10 TREATMENT           402 0.00910
```

```
## 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()
```

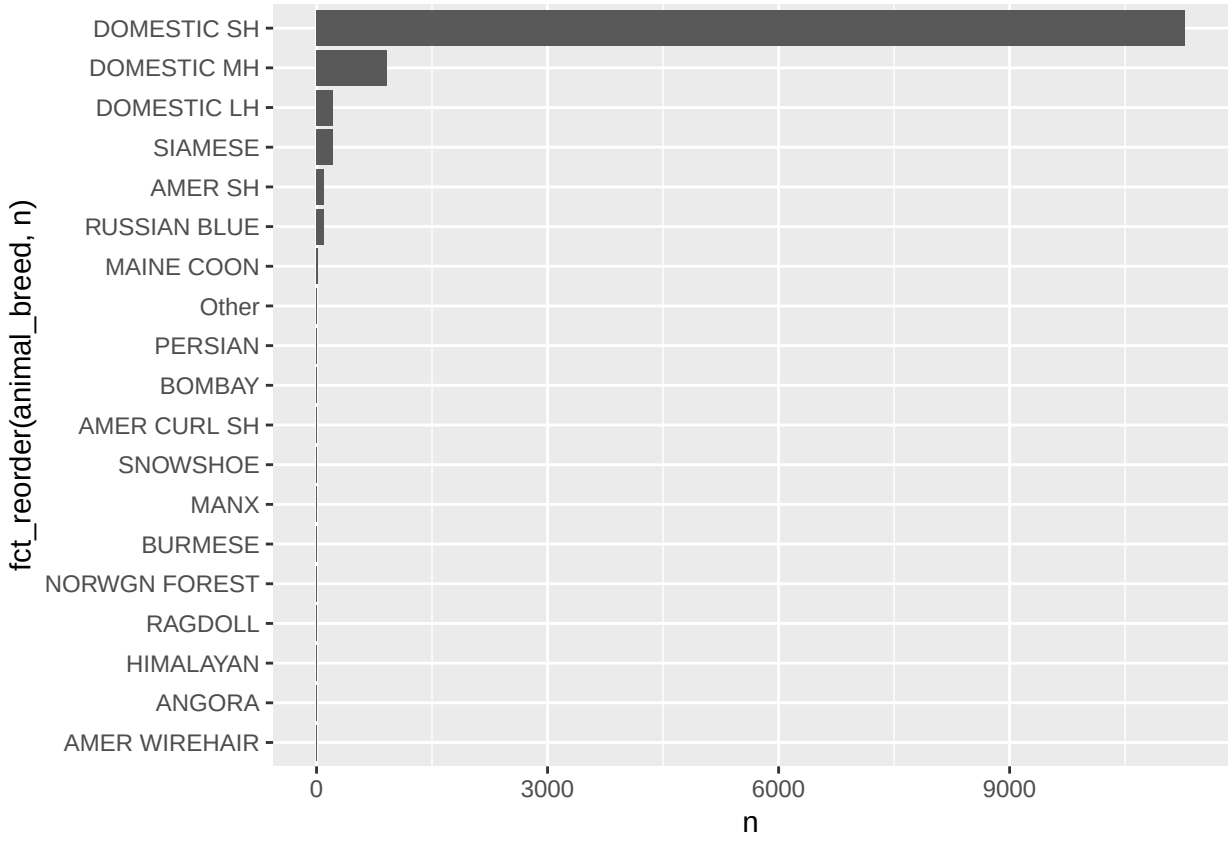



```
## Which cat breeds are in our data?
adoptions %>%
  filter(animal_type == "CAT") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=10)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)
```

```
## # 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()
```



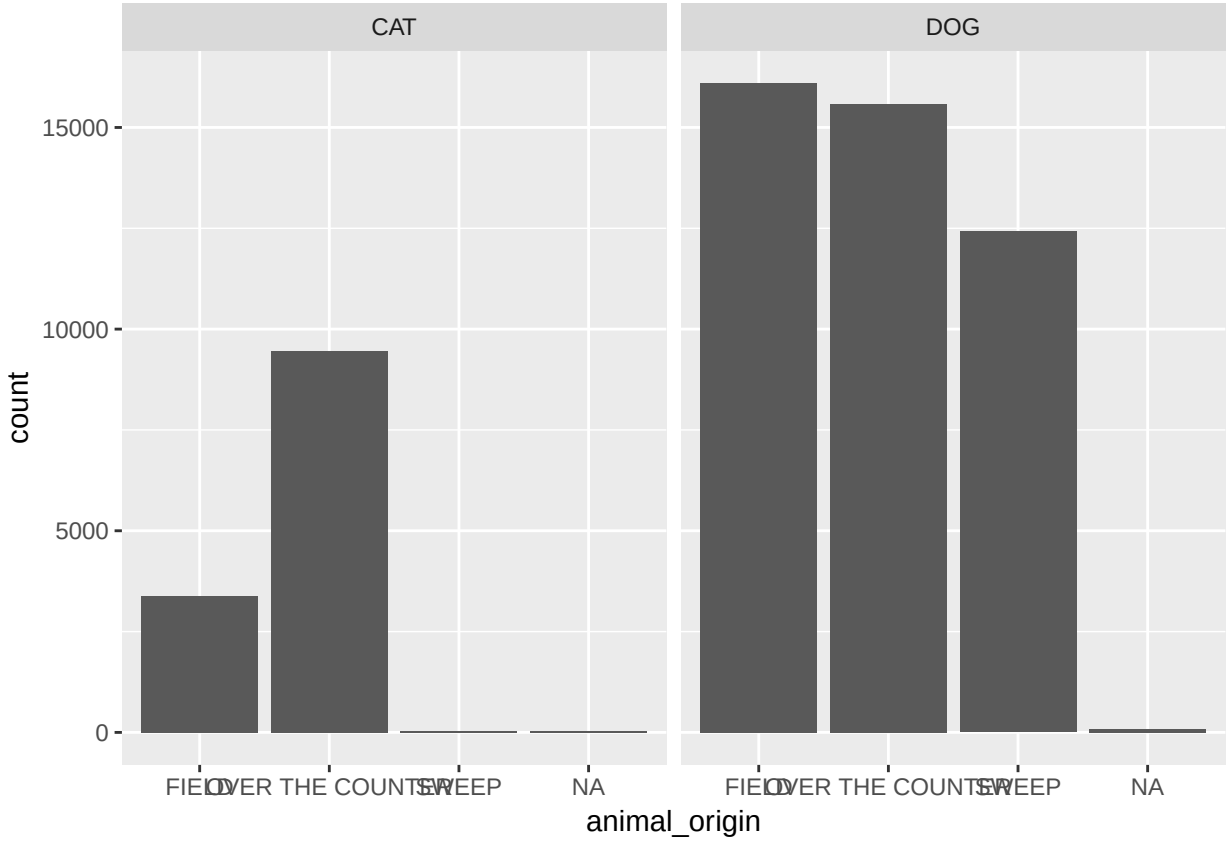

```
## Which types of wildlife are in our data?
adoptions %>%
  filter(animal_type == "WILDLIFE") %>%
  mutate(animal_breed = fct_lump(animal_breed, n=10)) %>%
  group_by(animal_breed) %>%
  count(sort = TRUE)
```

```
## # 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
```

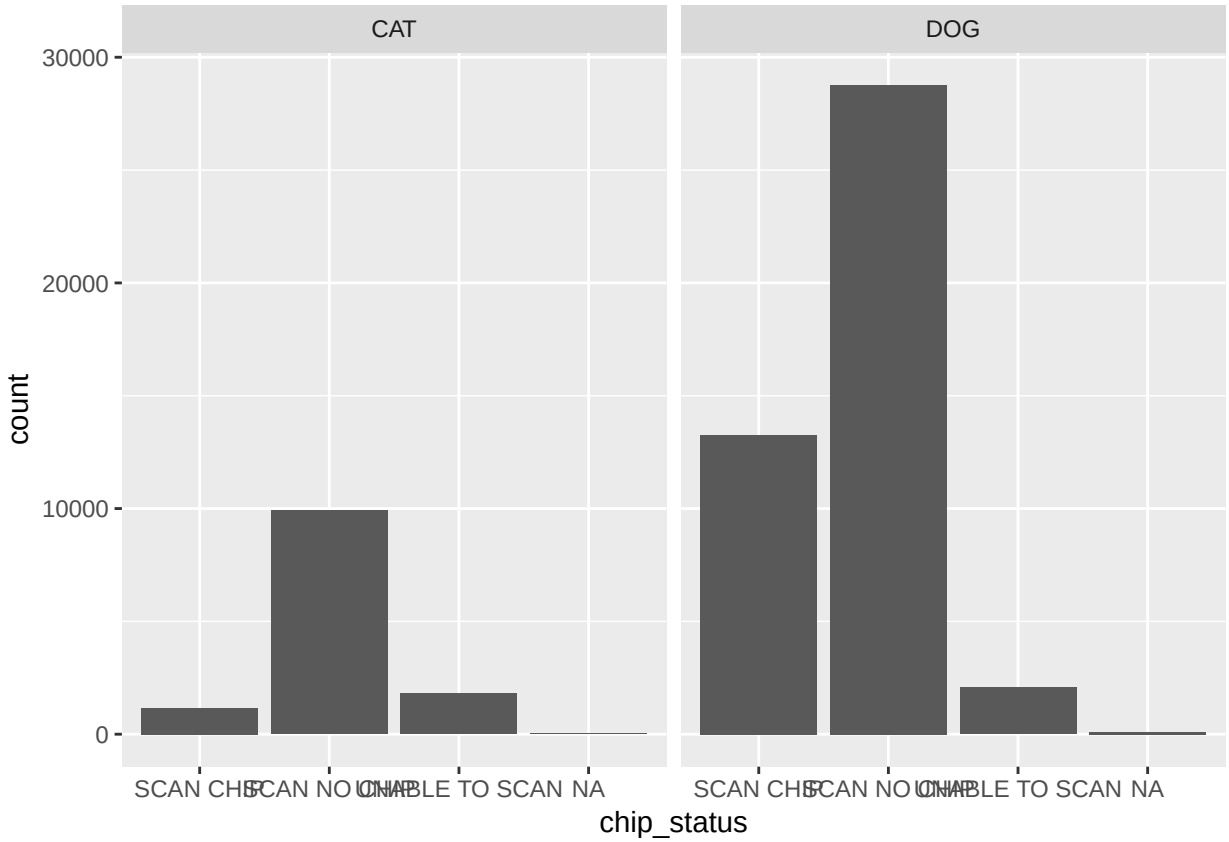
Exploring animal origin

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

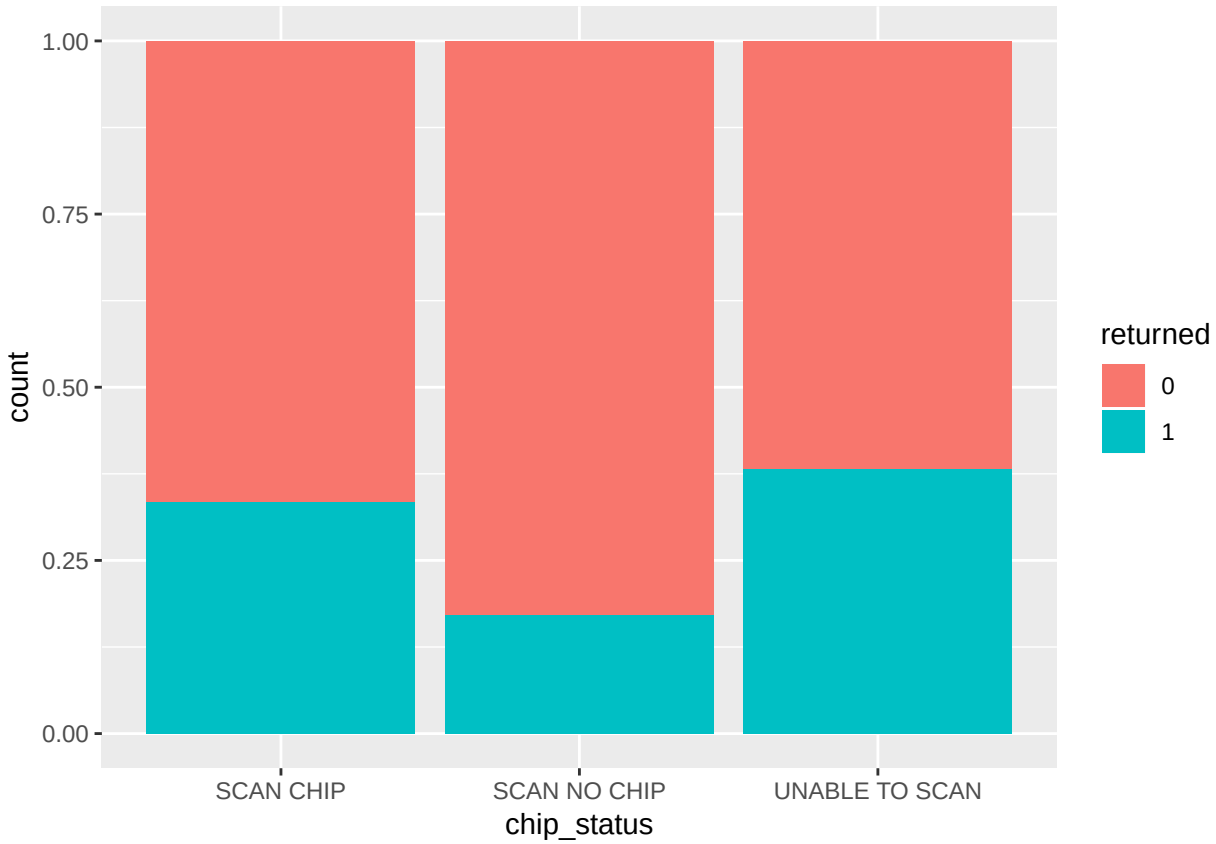


Exploring chip status

```
## 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")
```



Exploring intake_subtype

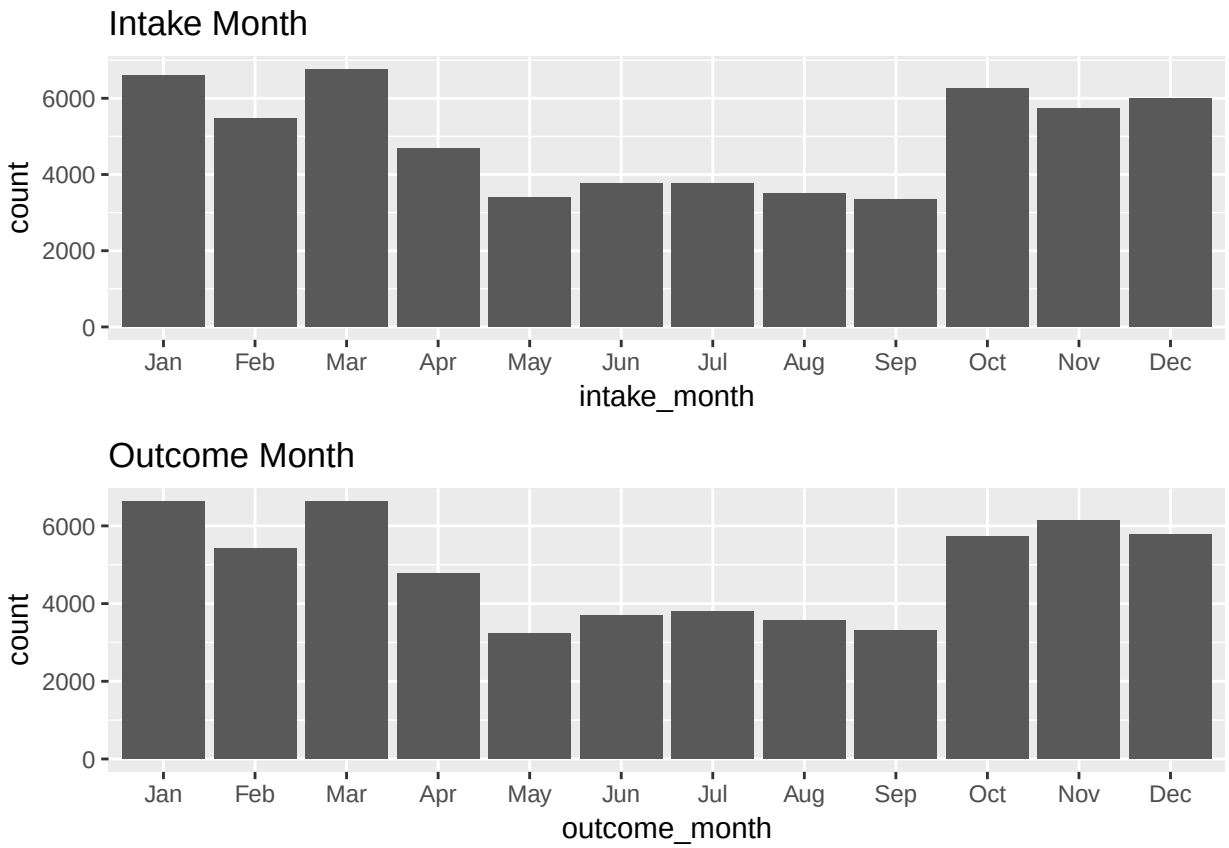
```
## Which dog breeds are in our data?
adoptions %>%
  mutate(intake_subtype = fct_lump(intake_subtype, n=15)) %>%
  group_by(intake_subtype) %>%
  count(sort = TRUE)
```

```
## # A tibble: 16 x 2
## # Groups:   intake_subtype [16]
##    intake_subtype           n
##    <fct>                <int>
##  1 AT LARGE             28889
##  2 GENERAL              15495
##  3 CONFINED              5598
##  4 POSSIBLY OWNED        2085
##  5 QUARANTINE            1454
##  6 RETURN30              1312
##  7 INJURED                972
##  8 KEEP SAFE              781
##  9 Other                  588
## 10 UNINJURED              474
## 11 EUTHANASIA REQUESTED   427
## 12 DEAD ON ARRIVAL        401
## 13 HEART WORM             401
## 14 OTHER                  179
## 15 RETURN                 128
```

```
## 16 WEB                    116
```

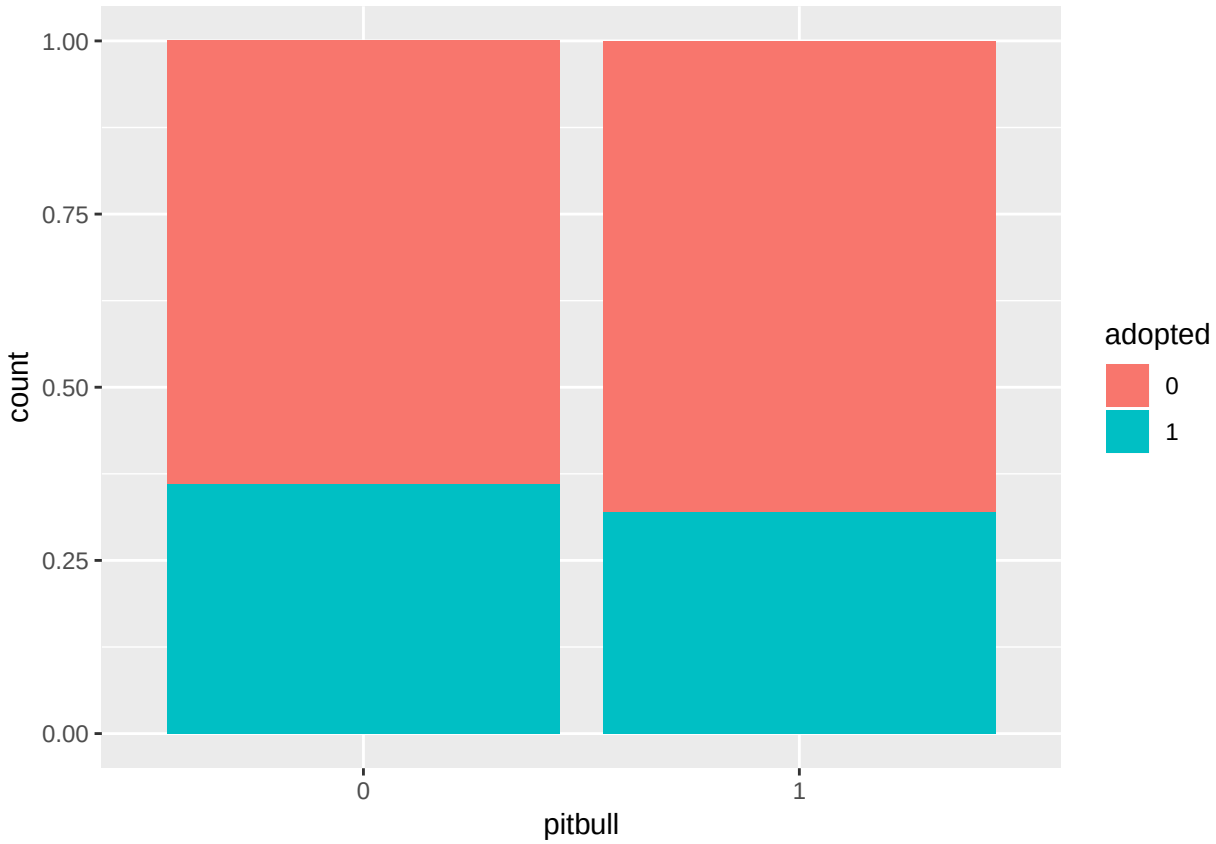
Exploring intake_date and outcome_date

```
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)
```



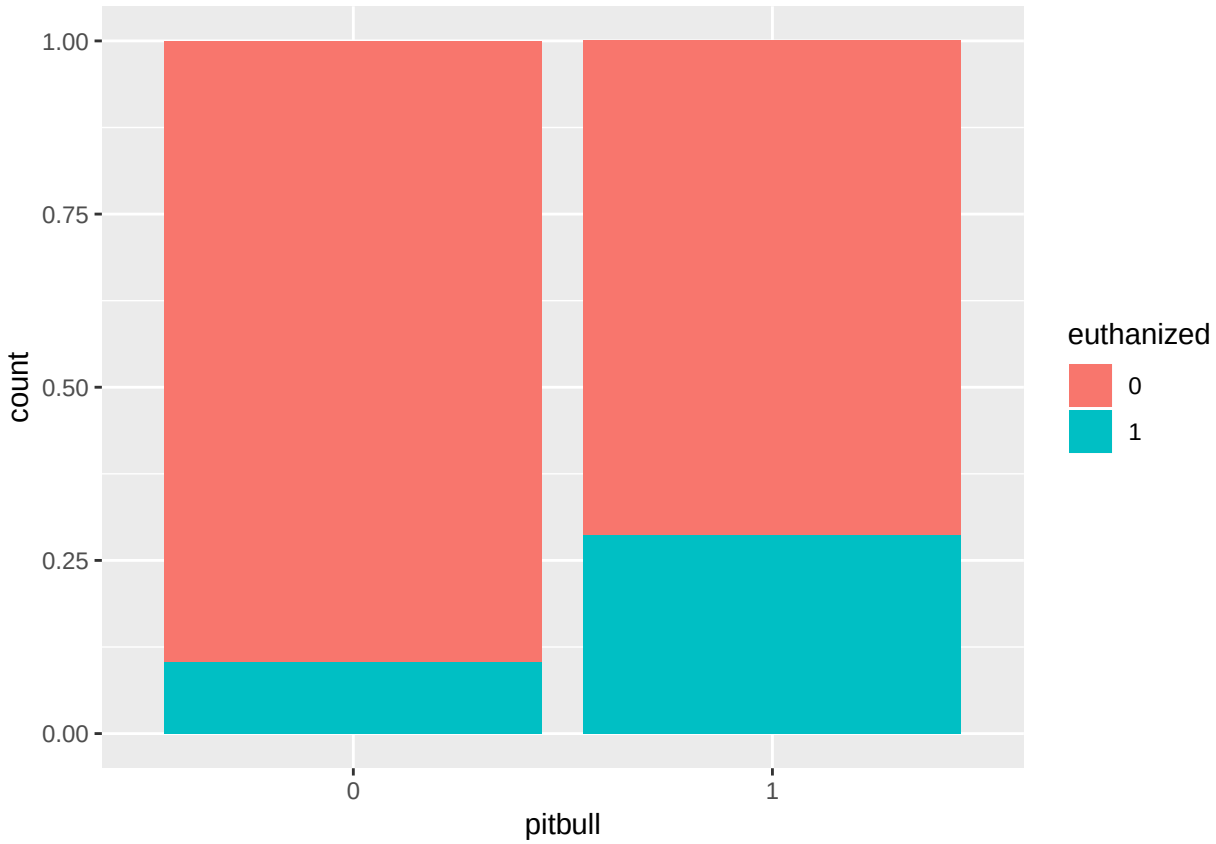
Pitbulls and Adoptions

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



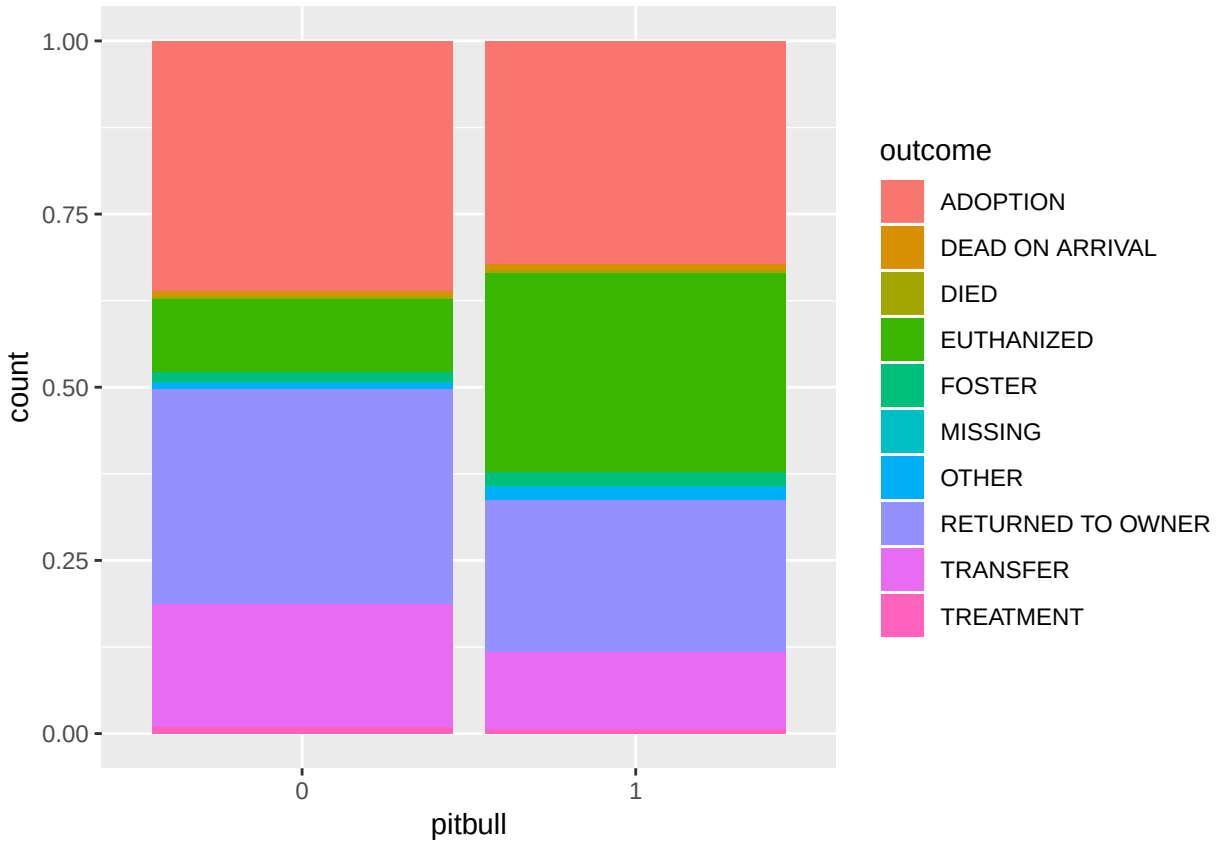
There does not seem to be a large difference between the adoption rates of pitbull and non-pitbulls. We should try and filter out some of the non-adopted dogs though, not all dogs come to the shelter to be adopted.

```
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, the are euthanized at a higher rate than other dogs. Lets 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) %>%
  kable()
```

pitbull	ADOPTION	DEAD ON ARRIVAL	DIED	EUTHANIZED	FOSTER	MISSING	OTHER	RE
0	0.3603524	0.0083136	0.0038641	0.104271	0.0157197	0.0006147	0.0098943	
1	0.3210406	0.0087711	0.0051829	0.286654	0.0216286	0.0003987	0.0183395	

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

Differences between dogs and cats

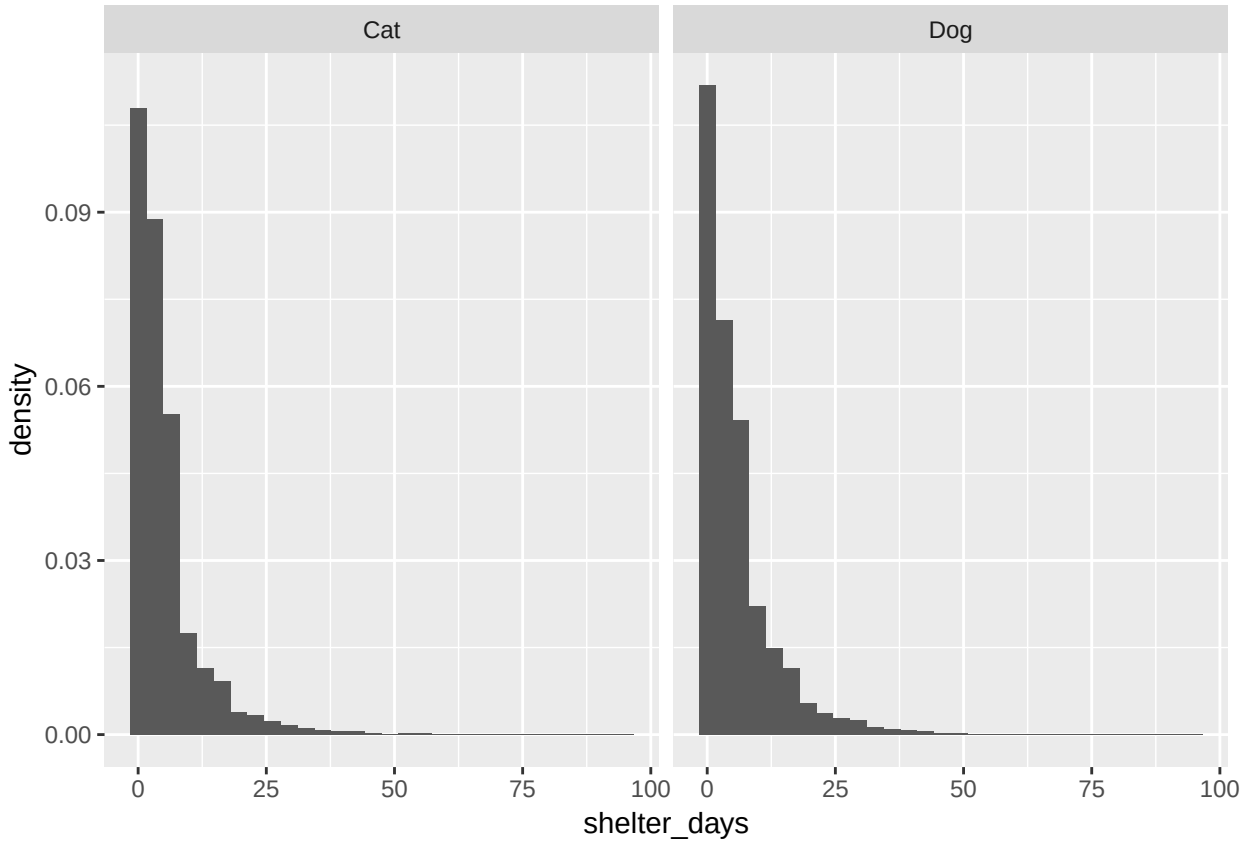
```
adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  mutate(DogCat = ifelse(dog == 1, "Dog", "Cat")) %>%
  group_by(DogCat) %>%
  summarise(count = n(),
            meanDays = mean(shelter_days, na.rm = TRUE),
            prop_euth = sum(euthanized)/count,
            prop_adopt = sum(adopted)/count,
```

```
            prop_stray = sum(stray)/count)
```

```
## # A tibble: 2 x 6
##   DogCat count meanDays prop_euth prop_adopt prop_stray
##   <chr>  <int>    <dbl>     <dbl>      <dbl>      <dbl>
## 1 Cat    12891     5.07     0.179      0.330      0.596
## 2 Dog    44194     5.70     0.146      0.351      0.660
```

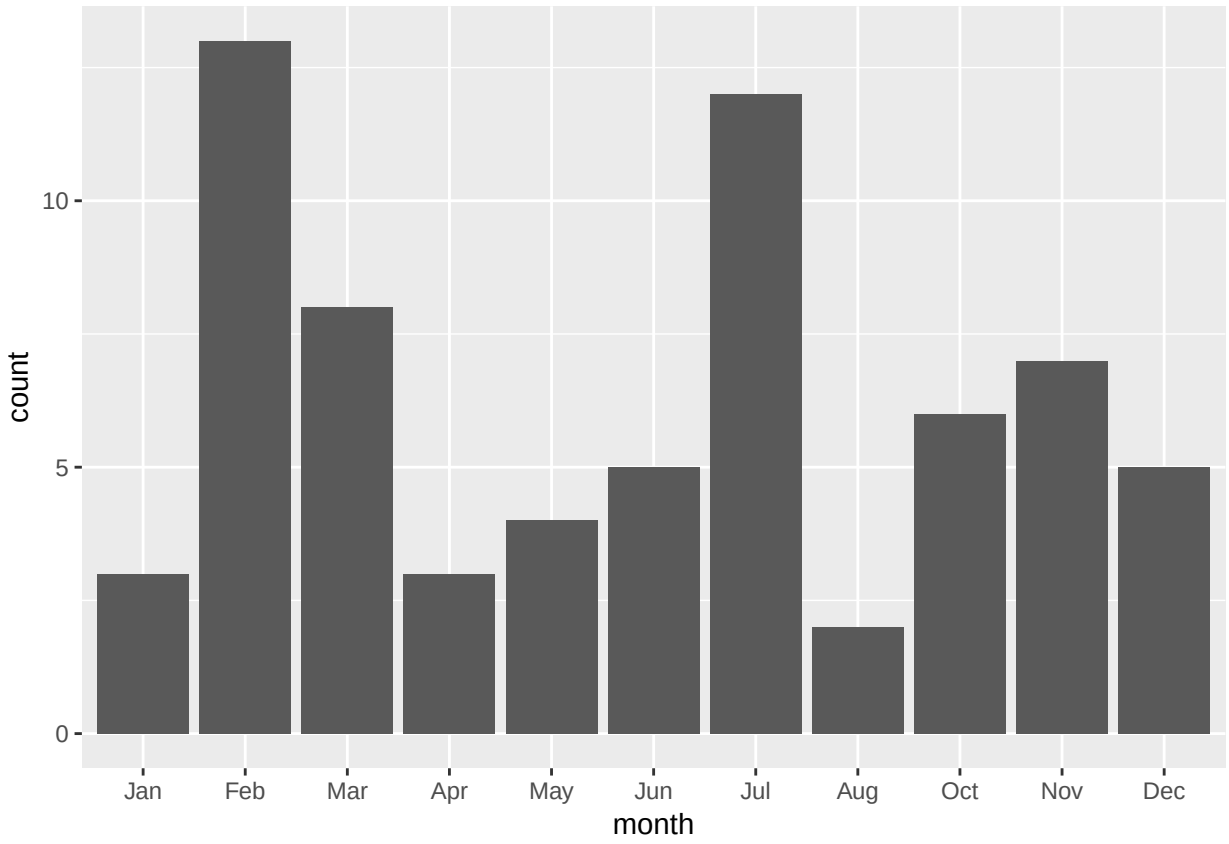
```
adoptions %>%
  filter(dog == 1 | cat == 1) %>%
  filter(shelter_days < 100) %>%
  mutate(DogCat = ifelse(dog == 1, "Dog", "Cat")) %>%
  ggplot(aes(x=shelter_days, y=..density..)) +
  geom_histogram()+
  facet_grid(~DogCat)
```

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



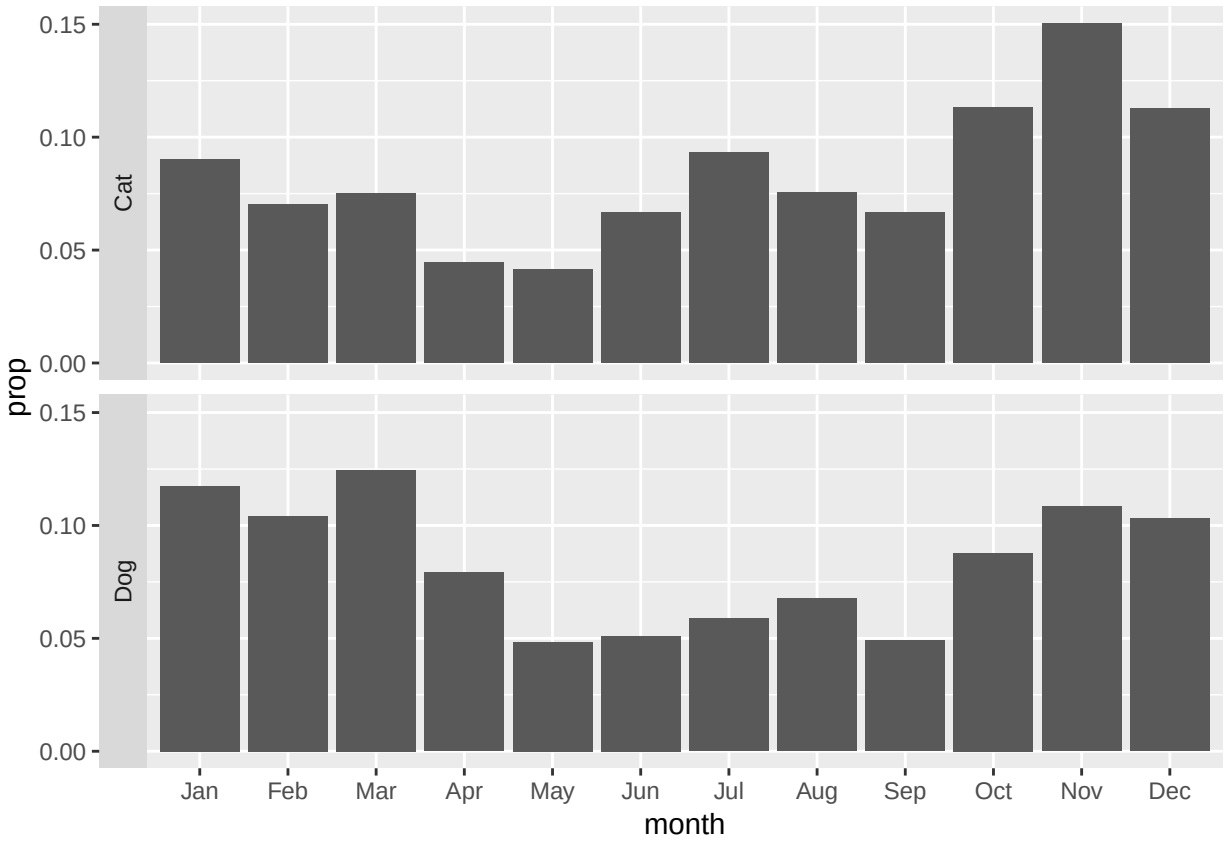
What about rabbits?

```
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 rabits 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")
```



1. Exploratory data analysis. Calculate or prepare appropriate numeric and graphical summaries for all relevant variables. Summarize the data using methods that are appropriate to the data type.

(a) Most likely you will need to "clean" your dataset first. Make note of any problematic data and observations that need to be removed. Consider implications about any decisions you make about missing data.

(b) Descriptive statistics (5 number summaries for continuous variables; tables of counts and proportions for categorical variables) for all relevant variables in your data set. Plot continuous variables using histograms.

(c) Explore the relationships between important pairs of variables both graphically and numerically. Depending on the type of your response and explanatory variables, you may consider graphs such as boxplots, scatterplots, cdplots, spaghetti plots, and segmented bar charts, and you may consider summary statistics (like mean, median, standard deviation) by group, correlations, regression equations, and two-way tables with proportions. Exploratory plots can be loose with titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels!

A few extra details on the EDA Report:

(a) The Main Body of your report should follow these guidelines:

(b) General hints for your Stage II report: • Aim your report at audience familiar with 272-level statistics, but may be a little rusty. Also, they have no specific knowledge on your research topic, but they have the ability to catch on quickly. Explain your terms clearly. • Show your work and output in an Appendix so your analyses can be reconstructed, if necessary.

• Give concise but precise statements interpreting summary statistics, etc. – in the context of your data set and research questions you pose. Avoid vague terms like “this data”, “these results”, etc. Also avoid cryptic variable names that you may have used in R.