Foundations of Data Analysis

The University of Texas at Austin

R Tutorials: Week 2

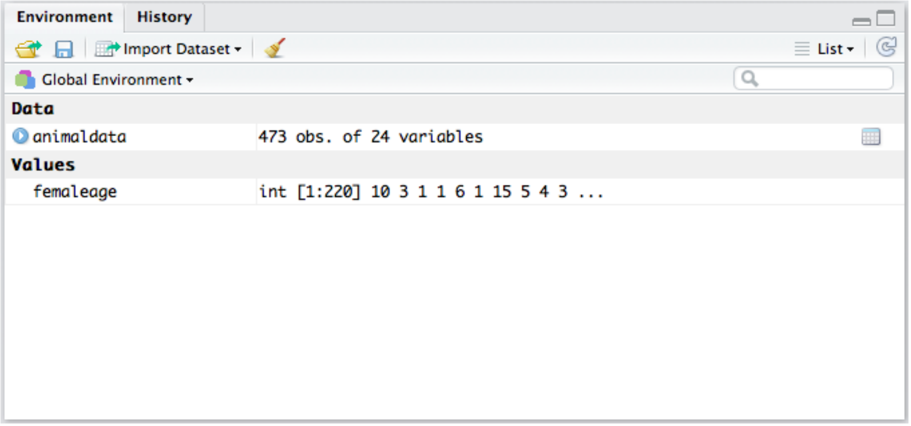
Histograms by Groups

In this R tutorial, we're going to build on the previous tutorial where we learned how to make a histogram of a single numeric variable. And we're going to actually see how we can create a histogram for different groups.

So again using our animal shelter data set, we are going to create two histograms of the age of intake variable by splitting up the data set into male animals and female animals using the gender variable. So I'll create an object that contains the age of intake values for our animals where the sex variable is equal to female and then I can do the same thing for the males. So let's call this "femaleage", and then we'll use the assignment key and say I want to take the variable "age at intake", and then I'm going to use those brackets to index this variable. And I only want to pull the rows under the condition that the sex variable equals, "Female".

```
femaleage <- animaldata$Age.Intake[animaldata$Sex ==
    "Female"]
```

So I can see here after submitting it in my workspace, or environment, that I have this vector female age and it has 220 cases.



Data
animaldata 473 obs. of 24 variables
Values
femaleage int [1:220] 10 3 1 1 6 1 15 5 4 3 ...

Figure 1: Importing the AnimalData CSV file.

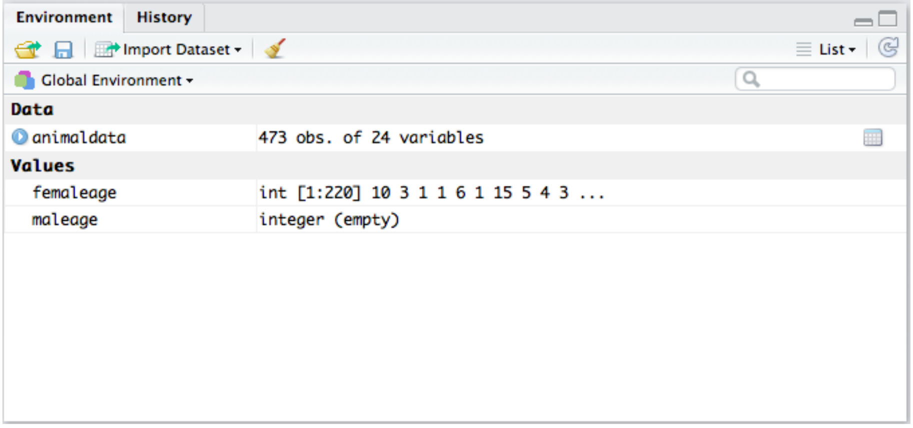
Which is what I expect because I have 220 females in my data set, so it looks like that worked correctly. Now let's do the same thing for male. I'm going to call my object "maleage" and again take the age of intake variable only where the sex is now "Male".

```
maleage <- animaldata$Age.Intake[animaldata$Sex ==
    "Male"]
```

So let's say for instance that I forgot that R was case sensitive and I just put lowercase "male" for the condition. What do you think would happen? Well let's try it and see.

```
maleage <- animaldata$Age.Intake[animaldata$Sex ==
    "male"]
```

If I run this, R won't give me any sort error (go ahead and try it) – but if I check up in my environment I see that it's actually an empty vector.



Data

animaldata 473 obs. of 24 variables

Values

femaleage int [1:220] 10 3 1 1 6 1 15 5 4 3 ...

maleage integer (empty)

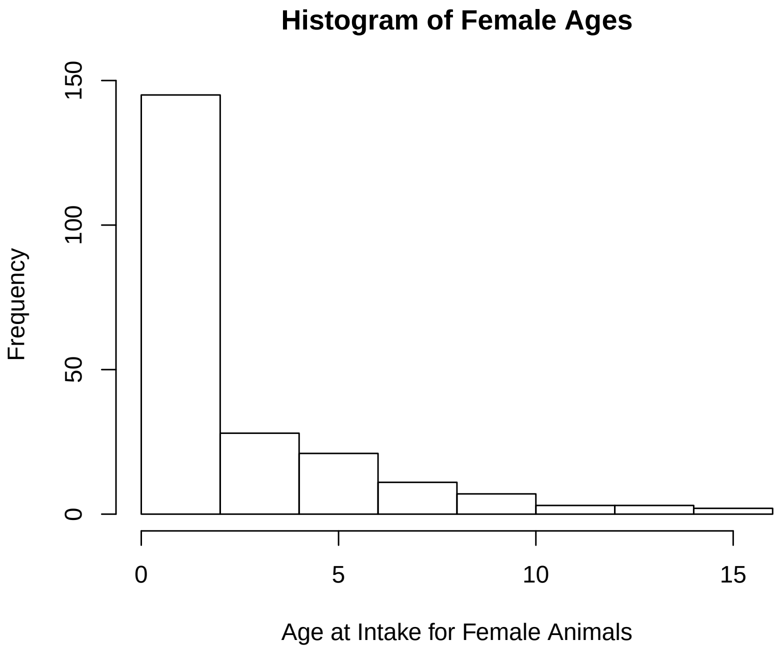
Figure 2: Importing the AnimalData CSV file.

There aren't any values in there and that's because R doesn't see any rows where this condition is true because lowercase 'm" to it completely different than an uppercase "M." So if I were to check my workspace I would have caught that. I could go back and change and make sure that this is the correct condition with an uppercase "M," rerun it, and now I see the correct 253 cases in my male age vector.

Before moving on, make sure you've got "maleage" as an object with 253 cases.

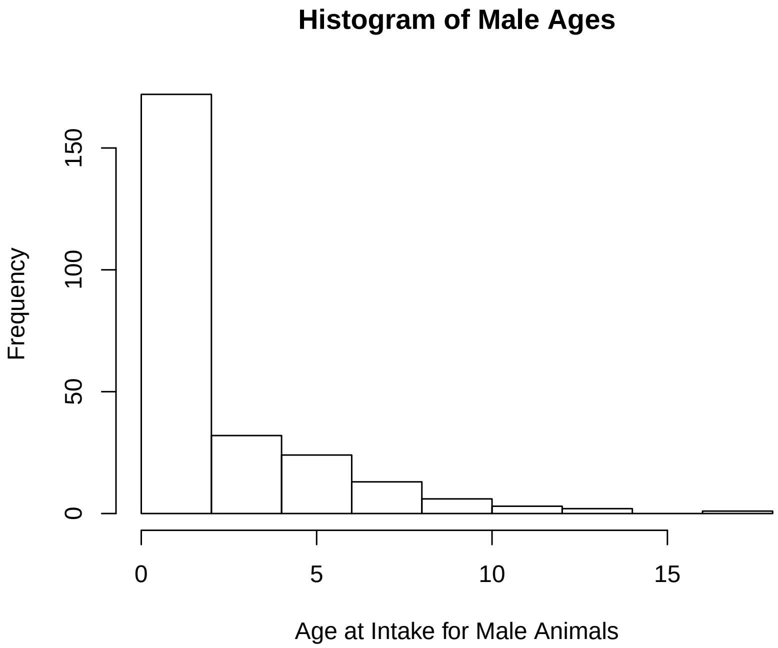
So now the only thing left to do would be to make a histogram of my female ages and then one for my male ages. So right now this "femaleage" vector is an object all by itself. It's not within the animal data set so I don't need to call it from a data frame. I just basically give it the object name then I can again change my axis and main title to say "Histogram of Female Ages". And my x label can – again – be something like age of intake of female animals or whatever describes that variable. Now I can run that and I see that the distribution of females is also very positively skewed.

```
hist(femaleage, main = "Histogram of Female Ages",
    xlab = "Age at Intake for Female Animals")
```



And now let's make a similar one for the male ages just to compare the distribution and see if it is similar to the females.

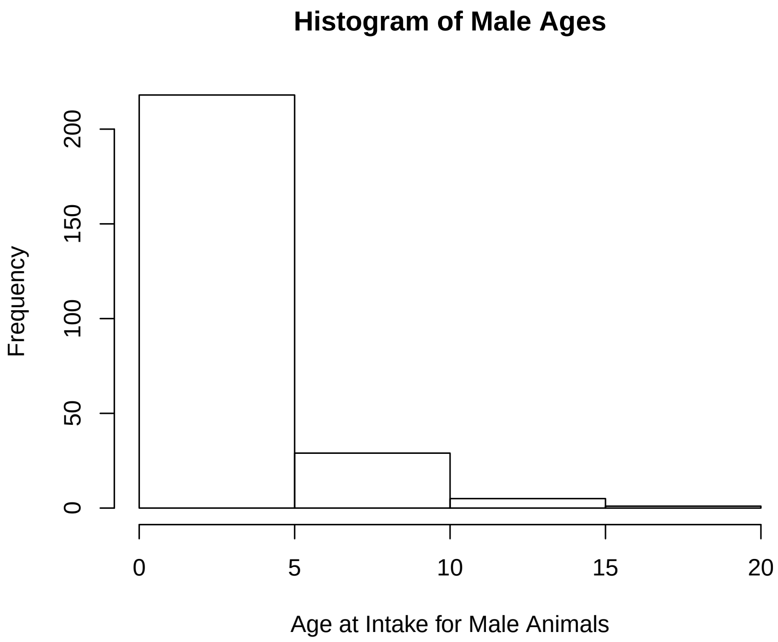
```r
hist(maleage, main = "Histogram of Male Ages", xlab = "Age at Intake for Male Animals")
```



And we see a very similar shape. So there's one very, very old dog that entered the shelter when he was over 15 years old, but for the majority of the male animals they entered the shelter when they were less than five years old. One final option that I'd like to mention for the histogram is that you can change the number of bins that R uses to make its histogram.

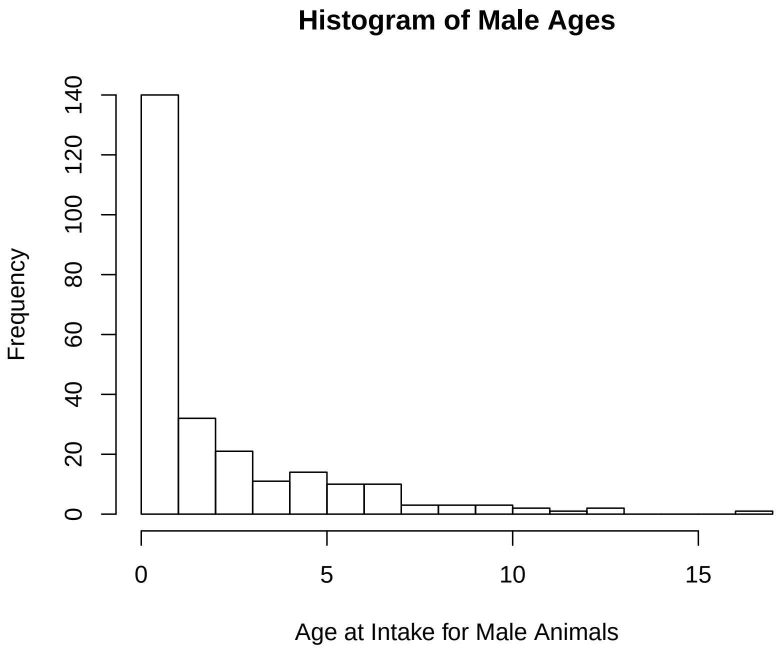
So in general you'd like to have between 5 and 15 bins. Usually R will give you a pretty good number of bins just by the default but you can adjust it. So let's go back up here in our male histogram of ages and add one more option to our line. So I'm going to add another comma– and the option here is called breaks– and I'm going to say let's give me 5 breaks and see what happens.

```
hist(maleage, main = "Histogram of Male Ages", xlab = "Age at Intake for Male Animals",
    breaks = 5)
```



Now we see fewer events than we did before. So there aren't exactly 5 bins here but R will give you the closest approximation it can while still capturing all the data. So changing the number of bins in your histogram can change the shape of the distribution that you see with the histogram. Although in this case we still see this general skew, we do see less definition. And on the flip side, what if we were to change the brakes to say 15? If we ran that, now we'd see a lot of definition.

```
hist(maleage, main = "Histogram of Male Ages", xlab = "Age at Intake for Male Animals",
    breaks = 15)
```



We would see that there is one case way far away from the other cases that we might want to go investigate which we would have missed if we had just stuck with our original default histogram. So sometimes it's a good idea to adjust the number of breaks in your histogram just to see if there's anything that jumps out at you with more or fewer bins. If we wanted to go find out the information from this animal here that was much older than any of the other animals when it entered the shelter, we can do that with a function called "**which()**" and **which()** will pull out a case from your data frame that follows a certain condition.

So first let's see what the maximum age of intake was. We can do that with the "**max()**" function if we ask for the max of male age that's just going to pull out the value that is the largest from that vector.

```
max(maleage)
```

```
## [1] 17
```

So 17 is the oldest age for the male animals. Let's just make sure that that's larger than any of our female animals.

```
max(femaleage)
```

```
## [1] 15
```

So here we see the max age of the females was only 15. So overall – across all animals in our data set – 17 is the oldest animal and it happened to be a male.

But let's go in and see if we can find out the other characteristics of this animal. Like whether it was a dog or cat, and whether it was sick, injured, or healthy. So what we can do for that is call the "**which()**" function and say give me which record in my animal data set where the age dot intake equals 17.

```
which(animaldata$Age.Intake == 17)
```

```
## [1] 415
```

Now if I run this I'm going to get a number that might not make sense right away. It just gives me the number 415. So what this is actually telling you is that in row number – or record number – 415 this age dot intake variable is equal to 17, so that's the animal that I want. Now in order to call that, I can ask for "animaldata" – and remember to index it, I can give it my square brackets and then give it a "row comma column". So the row I want is literally row 415. And let's say I want to know every single variable associated with that animal, so I'm just going to – again – leave the column space blank and that will return all columns.

```r
animaldata[which(animaldata$Age.Intake == 17), ]
```

```
##     Impound.No Intake.Date     Intake.Type Animal.Type Neutered.Status
## 415 K12-020475    11/11/12 Owner Surrender         Dog        Neutered
##      Sex Age.Intake       Condition      Breed Aggressive Independent
## 415 Male         17 Injured or Sick Dachshund           Y           N
##     Intelligent Loyal Social Good.with.Kids Max.Life.Expectancy Max.Weight
## 415           N     Y      Y              N                  14         28
##     Dog.Group  Color Weight    Lab.Test Outcome.Date      Outcome.Type
## 415     Hound Brown     6.5 No Lab Test     11/15/12 Humane Euthanasia
##     Days.Shelter
## 415            4
```

So this will display the animal who was 17 years old when he entered the shelter and we can see that it was a dog. And that he actually was injured or sick when he came into the shelter.