Lab8 Guide

2022-10-26

Tasks

- Go over the acorn size case study from Lab 7
- Learn conditional statements (if/else)
- Use if/else to complete Exercise 2 (Lab 7)
- Learn about `apply` and `replicate`
- Use these tools to solve Exercise 3 (Lab 7)
- Extra practice with conditional statements

Case Study: Introduction (from Lab 7)

It is thought that the size of a plant's seed may have some effect on the geographic range of a plant. In fact, a positive correlation is believed to exist between acorn size and the geographical range of the North American Oaks. The idea behind this theory is that larger acorns will be carried away by larger animals who in turn have a wider territorial range. Aizen and Patterson studied 39 species of oak trees to examine this correlation.

Case Study: Protocol (1/2)

Fifty species of oaks are found growing in the United States, 80% of which are accounted for in the Atlantic and California regions. The 28 oaks in the Atlantic region and the 11 oaks in the California region were used in this study. Acorn size was expressed as a volume, using measurements of specific nut lengths and widths to estimate the acorn volume as the volume of an ellipsoid. The areas of the geographical range for each species were obtained from the available literature.

Case Study: Protocol (2/2)

The range of species number 11 of the California region is unusual in that it does not include any land on the continental United States.

This particular species of oak grows only on the Channel Islands of Southern Cax`lifornia (see the map) and the island of Guadalupe off the coast of Baja California. The area of the Channel Islands is 1014 sq. km and the area of the island of Guadalupe is 265 sq. km.

Case Study: Data

The data collected by Aizen & Patterson is provided in the file `acorn.txt`. The file `Readme_acorn.txt` describes what is in the data file.

Case Study: Questions

Look at the data file and the readme file and answer:

```
Species Region  Range Acorn size  Tree height
Quercus alba L. Atlantic  24196 1.4 27
Quercus bicolor Willd.  Atlantic  7900  3.4 21
Quercus macrocarpa Michx. Atlantic  23038 9.1 25
Quercus prinoides Willd.  Atlantic  17042 1.6 3
Quercus Prinus L. Atlantic  7646  10.5  24
Quercus stellata Wang.  Atlantic  19938 2.5 17
Quercus virginiana Mill Atlantic  7985  0.9 15
Quercus Michauxii Nutt. Atlantic  8897  6.8 0.3
Quercus lyrata Walt.  Atlantic  8982  1.8 24
Quercus Laceyi Small. Atlantic  233 0.3 11
Quercus Chapmanii Sarg. Atlantic  1598  0.9 15
Quercus Durandii Buckl. Atlantic  1745  0.8 23
Quercus Muehlenbergii Engelm  Atlantic  17042 2 24
Quercus ilicifolia Wang.  Atlantic  4082  1.1 3
Quercus incana Bartr. Atlantic  3775  0.6 13
Quercus falcata Michx.  Atlantic  13688 1.8 30
Quercus laevis Walt.  Atlantic  3978  4.8 9
Quercus laurifolia Michx. Atlantic  5328  1.1 27
Quercus marilandica Muenchh.  Atlantic  18480 3.6 9
Quercus nigra L.  Atlantic  10161 1.1 24
Quercus palustris Muenchh.  Atlantic  8643  1.1 23
Quercus Phellos L.  Atlantic  9920  3.6 27
Quercus rubra L.  Atlantic  28389 8.1 24
Quercus velutina Lam. Atlantic  21067 3.6 23
Quercus imbricaria Michx. Atlantic  14870 1.8 18
Quercus myrtifolia Willd. Atlantic  2540  0.4 9
Quercus texana Buckl. Atlantic  829 1.1 9
Quercus coccinea Muenchh. Atlantic  8992  1.2 4
Quercus Douglasii Hook. & Arn California  559 4.1 18
Quercus dumosa Nutt.  California  433 1.6 6
Quercus Engelmannii Greene  California  259 2 17
Quercus Garryana Hook.  California  1061  5.5 20
Quercus lobata Nee  California  870 5.9 30
```

Case Study Questions

- how many variables are present in the data file?
  Six: Species, Region,Range, Acorn size, Tree height.
- which ones are numerical? (name the subtype)
  Continuous: Acorn size, Tree height. Continuous in principle but could be treated as discrete here: Range.
- which ones are categorical (name the subtype)
  Nominal: Species, Region. Ordinal: None.

Case Study: Questions

- how are the columns separated in the data file?

Let's look at the function `read.table`:

```
#read.table("acorn.txt") #fails
?read.table
```

"If sep = " (the default for read.table) the separator is 'white space', that is one or more spaces, tabs, newlines or carriage returns."

Case Study: Questions

- how are the columns separated in the data file?
  Tab!

Let's try tab!

```
#read.table("acorn.txt") #fails
read.table("acorn.txt", sep="\t") #works!
#"\t" is tab in linux/programming world.
read.delim("acorn.txt") #also works.
#see ?read.delim. It's default is sep="\t"
readr::read_table("acorn.txt") #does not work;
#expects white space
```

Case Study: Questions

- Read in the data file
- Check its dimensions
- Check its structure
- Do the classes of each column seem appropriate to you? Now is a good time to change them.
- Check that there are no missing values.

Case Study: Questions

Read in the data file; check its dimensions.
39 rows, 5 columns

```
acorn<-read.delim("input_files/acorn.txt") #Read in the data file
dim(acorn) #39 rows, 5 columns
```

```
## [1] 39  5
```

```
nrow(acorn) #39 rows
```

```
## [1] 39
```

```
ncol(acorn) #5 columns
```

```
## [1] 5
```

Case Study: Questions

Check its structure

```
#check its structure
str(acorn)
```

```
## 'data.frame':    39 obs. of  5 variables:
##  $ Species    : chr  "Quercus alba L." "Quercus bicolor Willd." "Quercus macrocarpa Michx."
##  $ Region     : chr  "Atlantic" "Atlantic" "Atlantic" "Atlantic" ...
##  $ Range      : int  24196 7900 23038 17042 7646 19938 7985 8897 8982 233 ...
##  $ Acorn.size : num  1.4 3.4 9.1 1.6 10.5 2.5 0.9 6.8 1.8 0.3 ...
##  $ Tree.height: num  27 21 25 3 24 17 15 0.3 24 11 ...
```

Case Study: Questions

Do the classes of each column seem appropriate to you? Now is a good time to change them.
Species and Region could be better handled as factors (not ordered).

```
library(dplyr)
acorn <- acorn %>%
  mutate(Species=factor(Species), #transform to factor
         Region=factor(Region)) #transform to factor
```

Case Study: Questions

```
summary(acorn) # Check that there are no missing values. no NAs show up in summaries
```

```
##                         Species          Region       Range           Acorn.size      Tree.he
##  Quercus agrifolia Nee.    : 1   Atlantic  :28   Min.   :   13.0   Min.   : 0.300   Min.   :
##  Quercus alba L.           : 1   California:11   1st Qu.:  827.5   1st Qu.: 1.100   1st Qu.:
##  Quercus bicolor Willd.    : 1                   Median : 5328.0   Median : 1.800   Median :
##  Quercus Chapmanii Sarg.   : 1                   Mean   : 7882.6   Mean   : 3.341   Mean   :
##  Quercus chrysolepis Liebm.: 1                   3rd Qu.:11924.5   3rd Qu.: 4.450   3rd Qu.:
##  Quercus coccinea Muenchh. : 1                   Max.   :28389.0   Max.   :17.100   Max.   :
##  (Other)                   :33
```

Case Study: Questions

```
#2
table(is.na(acorn)) #195 FALSE
```

```
##
## FALSE
##   195
```

Case Study: Questions

```
table(na.omit(acorn) == acorn) #if all are TRUE, there are no NAs
```

```
##
## TRUE
##  195
```

Case Study: Questions

```
#4
anyNA(acorn) #if FALSE, all good.
```

```
## [1] FALSE
```

Case Study: Questions

- What type of visualization would be appropriate in order to investigate the relationship between tree range and acorn size?
- Who are the response and explanatory variables here?
- Make that plot.

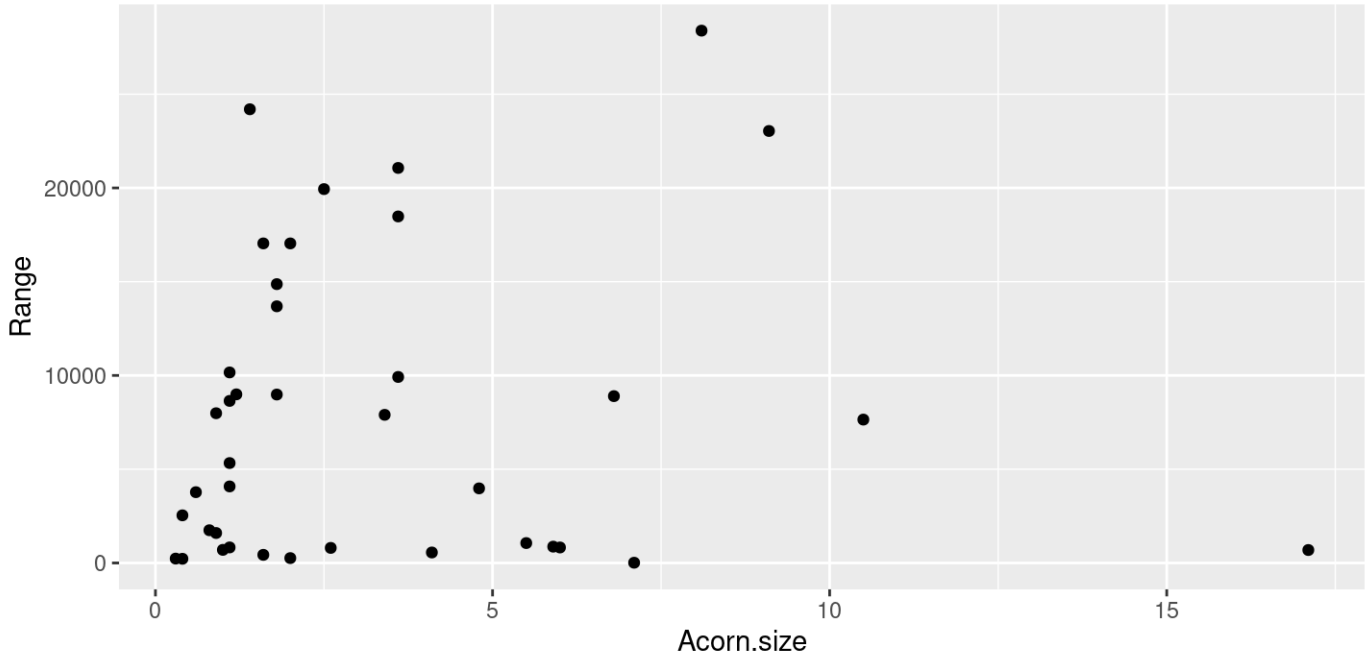
Case Study: Questions

- What type of visualization would be appropriate in order to investigate the relationship between tree range and acorn size?
  Tree range is numerical and acorn size is also numerical. Therefore, a scatter plot.

- Who are the response and explanatory variables here?
  The idea is to investigate the idea that larger acorns get dispersed further by larger animals, so acorn size (explanatory) and range (response).

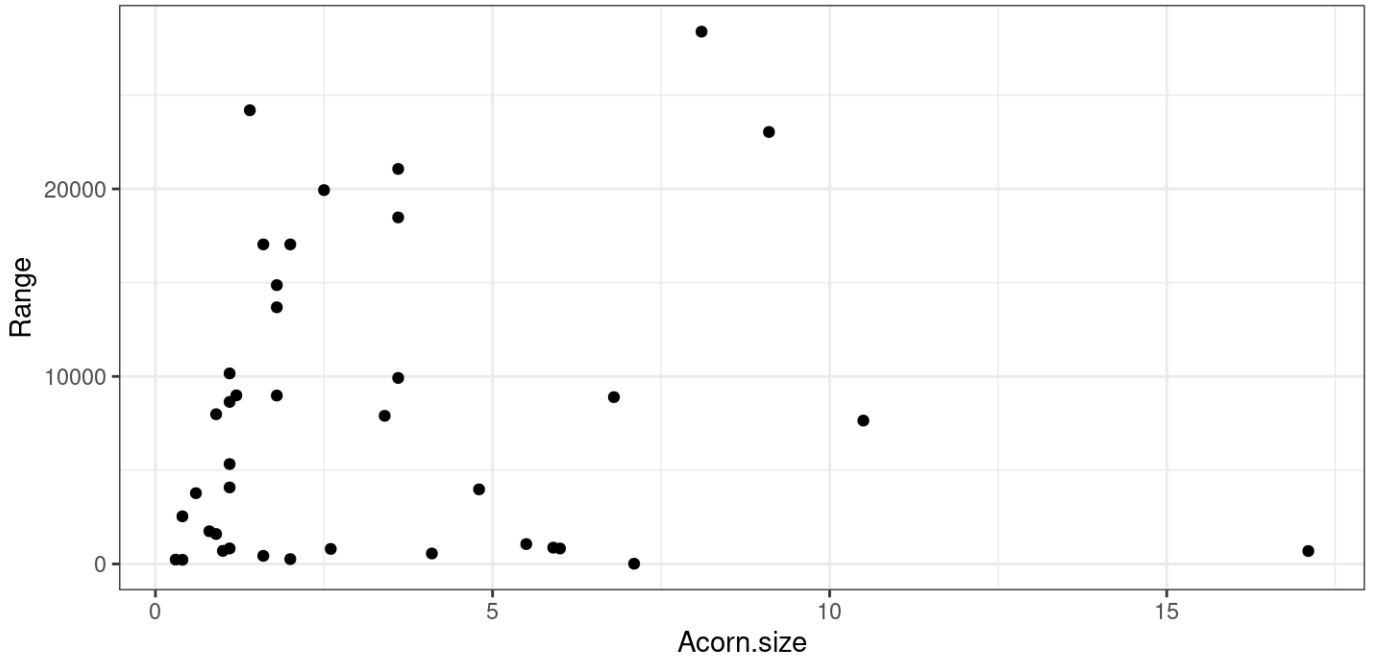
Case Study: Questions

```
library(ggplot2) #load ggplot package
ggplot(acorn, aes(x=Acorn.size, y=Range)) +  # basic scatter plot
  geom_point()
```



Case Study: Questions

```
library(ggplot2) #load ggplot package
ggplot(acorn, aes(x=Acorn.size, y=Range)) + geom_point() +
  theme_bw()  # white background
```



Case Study: Questions

Examine the summary statistics for tree range.

- What are the mean and the standard deviation?
- What do these values tell you about the likely shape of the distribution?

Case Study: Questions

```
mean(acorn$Range) #mean
```

```
## [1] 7882.564
```

```
sd(acorn$Range) #standard deviation
```

```
## [1] 8054.823
```

Case Study: Questions

```
library(dplyr)
acorn %>%
  summarise(mean(Range), sd(Range)) #produce two summaries at once
```

```
##   mean(Range) sd(Range)
## 1    7882.564  8054.823
```

Case Study: Questions

```
library(dplyr)
acorn %>%
  summarise(MeanRange=mean(Range), SDRange=sd(Range)) #name summaries
```

```
##   MeanRange  SDRange
## 1  7882.564 8054.823
```

$$SD >> Mean$$

Case Study: Questions

```
summary(acorn) # mean, median, quartiles, min, max
```

```
##                         Species          Region       Range           Acorn.size      Tree.he
##  Quercus agrifolia Nee.     : 1   Atlantic  :28   Min.   :   13.0   Min.   : 0.300   Min.   :
##  Quercus alba L.            : 1   California:11   1st Qu.:  827.5   1st Qu.: 1.100   1st Qu.:
##  Quercus bicolor Willd.     : 1                   Median : 5328.0   Median : 1.800   Median :
##  Quercus Chapmanii Sarg.    : 1                   Mean   : 7882.6   Mean   : 3.341   Mean   :
##  Quercus chrysolepis Liebm.: 1                    3rd Qu.:11924.5   3rd Qu.: 4.450   3rd Qu.:
##  Quercus coccinea Muenchh.  : 1                   Max.   :28389.0   Max.   :17.100   Max.   :
##  (Other)                    :33
```

Case Study: Questions

```
library(dplyr)
acorn %>%
  select(Range) %>% #select col Range
  summary() #summary()
```

```
##      Range
##  Min.   :   13.0
##  1st Qu.:  827.5
##  Median : 5328.0
##  Mean   : 7882.6
##  3rd Qu.:11924.5
##  Max.   :28389.0
```

$$Median << Mean$$

Case Study: Questions

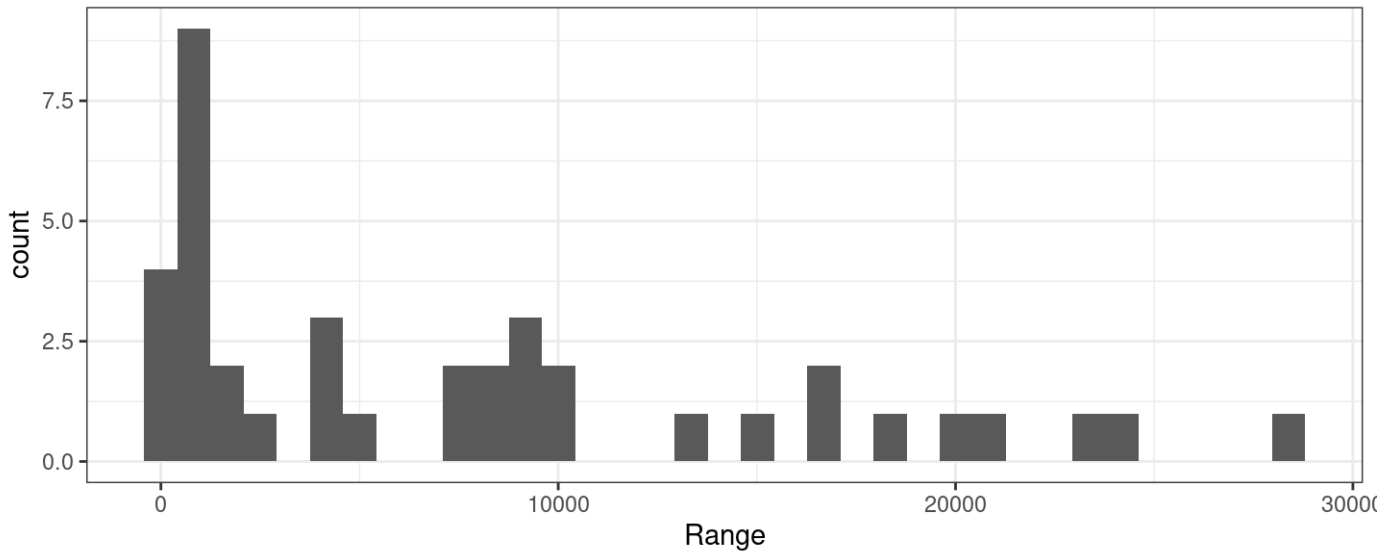
What are the mean and standard deviation?
7882.564,8054.823

What do these values tell you about the likely shape of the distribution?
This suggests an enormous spread or range of the data. Also, 50% of the data are contained within 827 and 11924.5 and 75% of the data have values below 11924.5. Further, the median (5328) is much lower than the mean (7882.6), suggesting a strong right (positive) skew.

Case Study: Questions

We could also plot the distribution for Range:

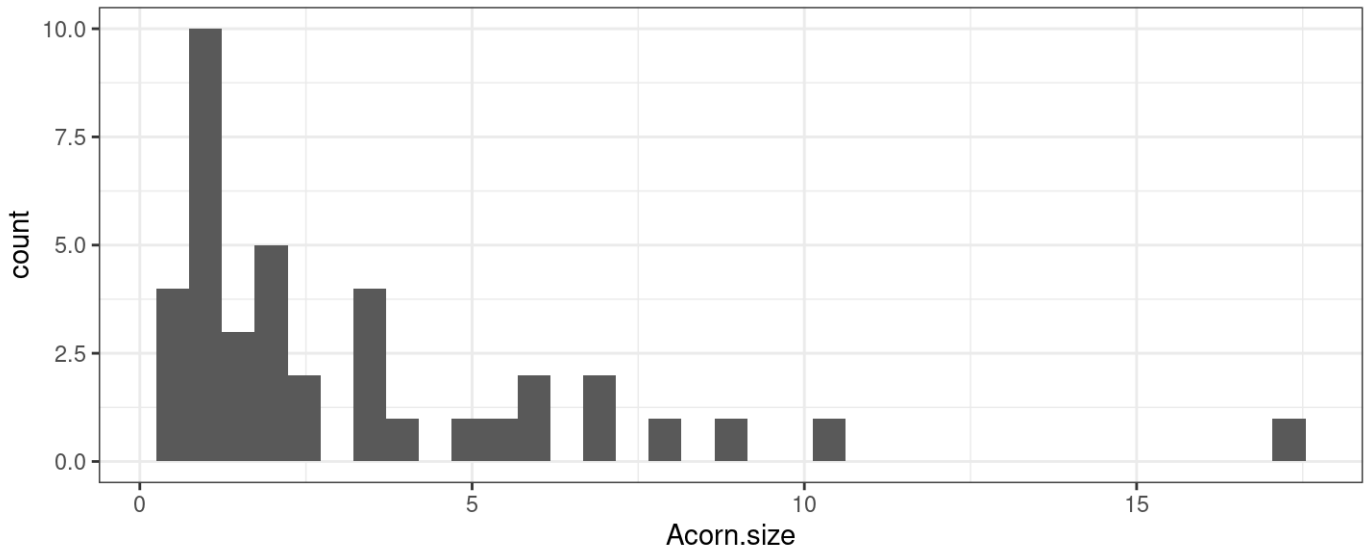
```
# place the mean (7882.564) and median (5328) in this distribution
ggplot(acorn, aes(x=Range))  + #define aesthetics
  geom_histogram(bins=35) + #define layer
  theme_bw() #white background
```



Case Study: Questions

We could also plot the distribution for Acorn size:

```
# place the mean (3.341026) and median (1.8) in this distribution
ggplot(acorn, aes(x=Acorn.size))  +
  geom_histogram(bins=35) +
  theme_bw()
```



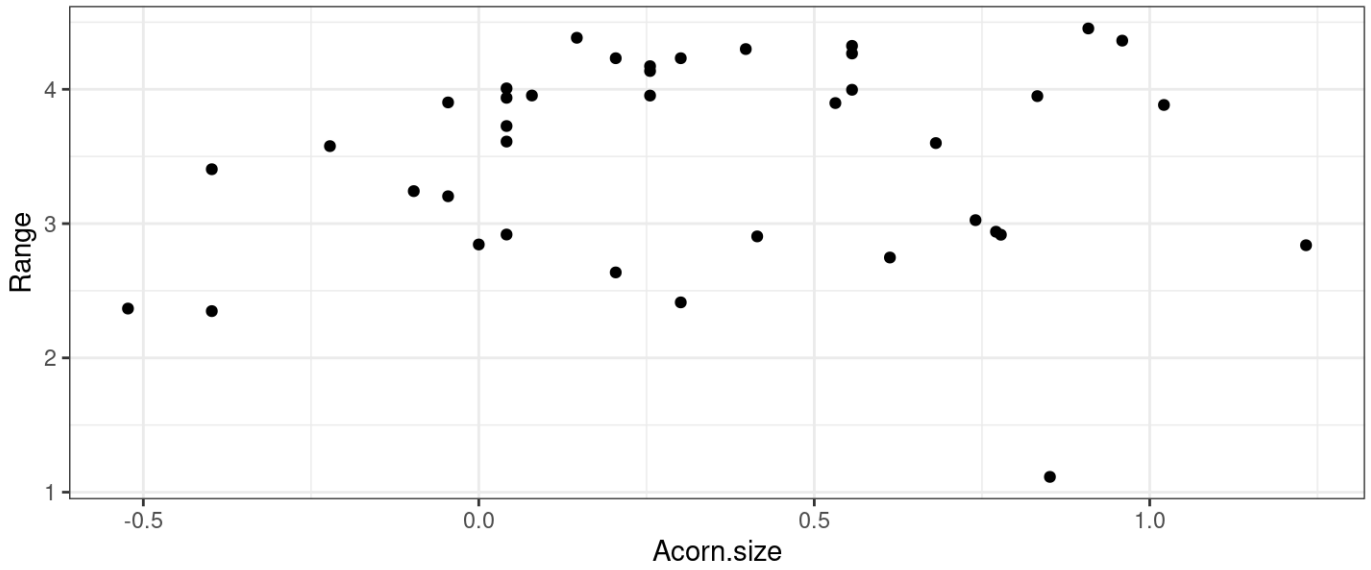
Case Study: Questions

Transform the data using the log transformation on both the range and the size.

The problem did not specify if the log base should be *e* or 10, but let's start with 10.

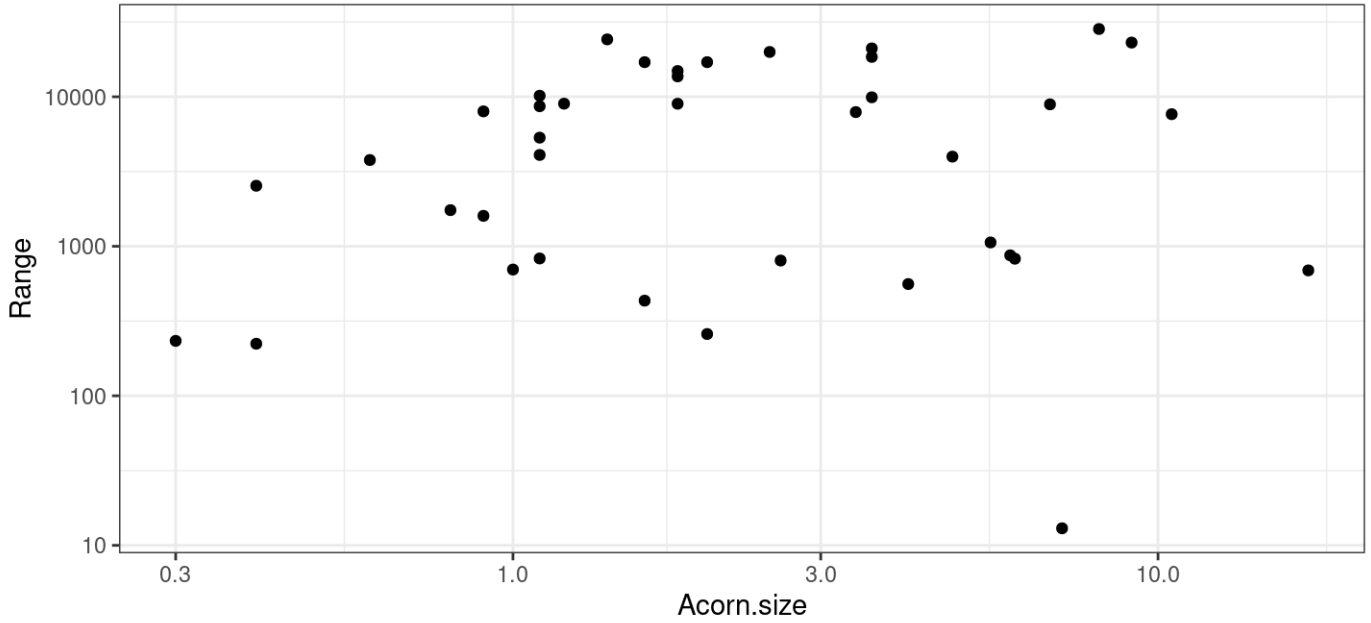
Case Study: Questions

```
acorn2<- acorn %>% #option 1
  mutate(Range=log(Range, base=10), #log10 would also work
           Acorn.size=log(Acorn.size, base=10))
ggplot(acorn2, aes(x=Acorn.size, y=Range)) +
  geom_point() +
  theme_bw()  # white background
```



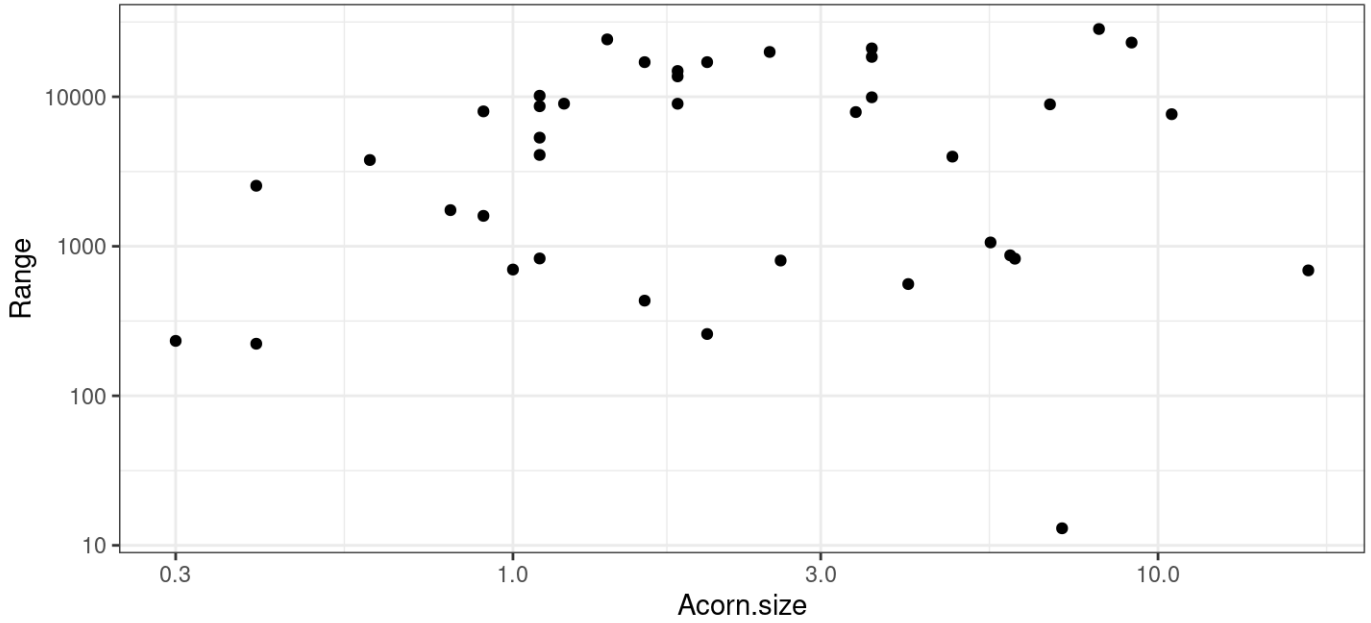
Case Study: Questions

```
ggplot(acorn, aes(x=Acorn.size, y=Range)) + #option 2
  geom_point() +
  scale_y_continuous(trans="log10") +
  scale_x_continuous(trans="log10") +
  theme_bw()  # white background
```



Case Study: Questions

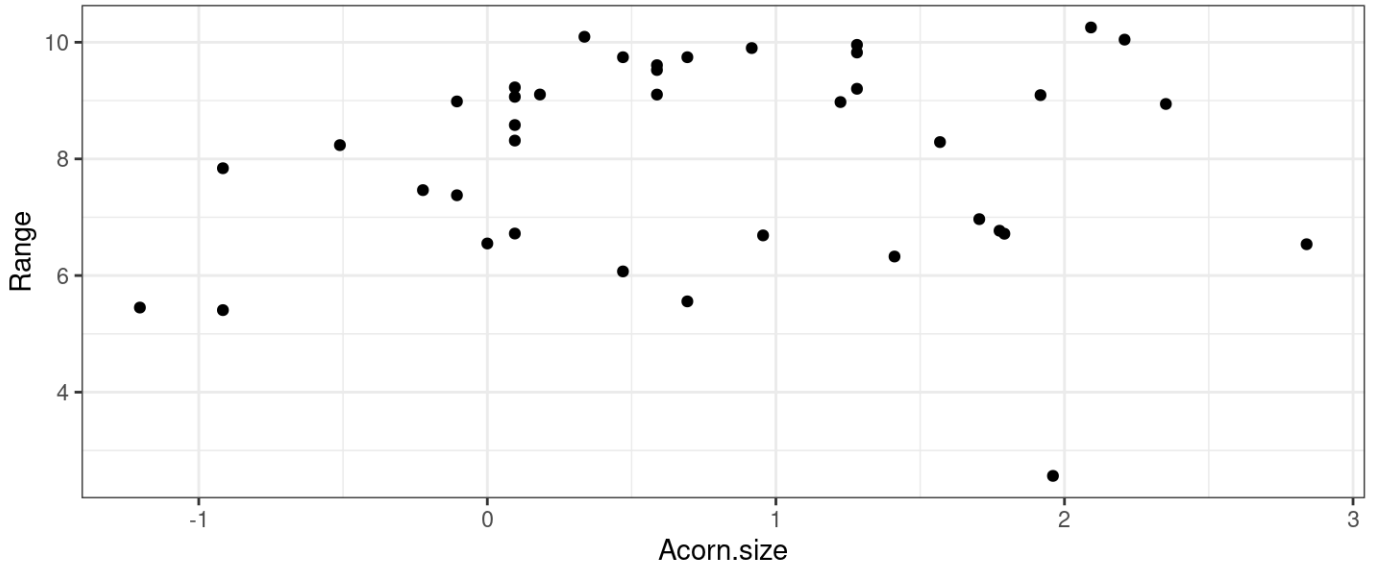
```
ggplot(acorn, aes(x=Acorn.size, y=Range)) + #option 3
  geom_point() +
  scale_x_log10() + #built-in ggplot log10 option
  scale_y_log10() +
  theme_bw()  # white background
```



Case Study: Questions

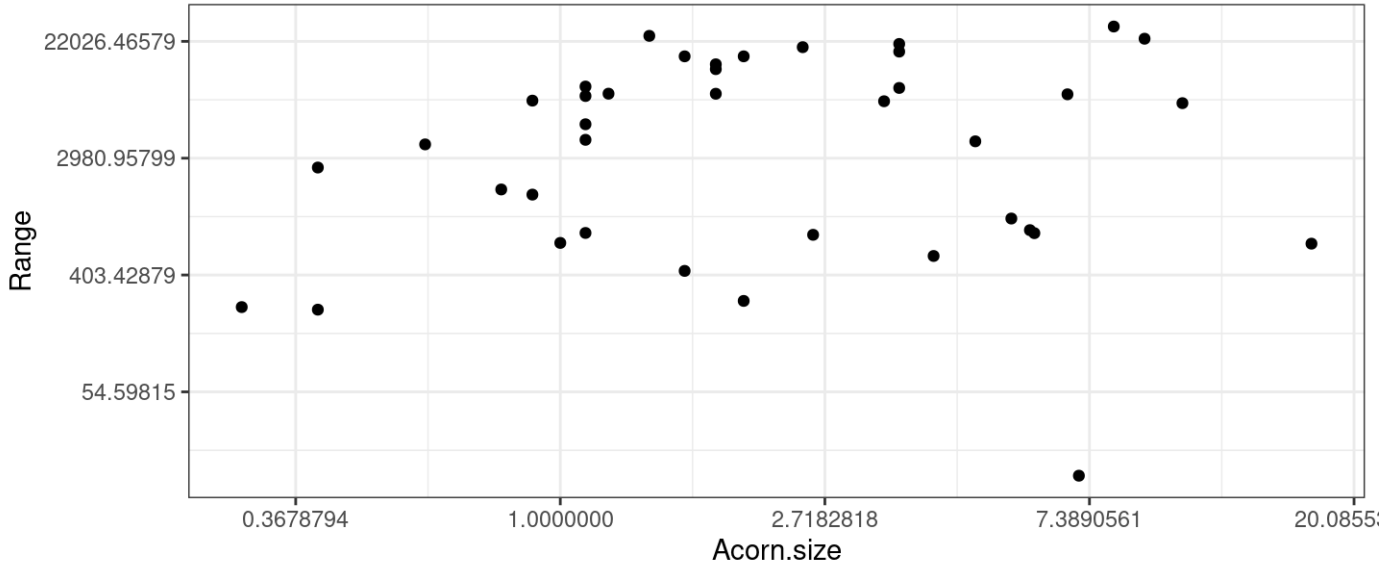
Now make a scatter plot of Ln(range) vs. Ln(acorn size).

```
acorn3<- acorn %>% #option 1
  mutate(Range=log(Range), #default base is e (2.718281828459)
           Acorn.size=log(Acorn.size))
ggplot(acorn3, aes(x=Acorn.size, y=Range)) +
  geom_point() + theme_bw()  #white background
```



Case Study: Questions

```
ggplot(acorn, aes(x=Acorn.size, y=Range)) + #option 2
  geom_point() +
  scale_y_continuous(trans="log") + #convert x to log scale
  scale_x_continuous(trans="log") +
  theme_bw()  # white background
```



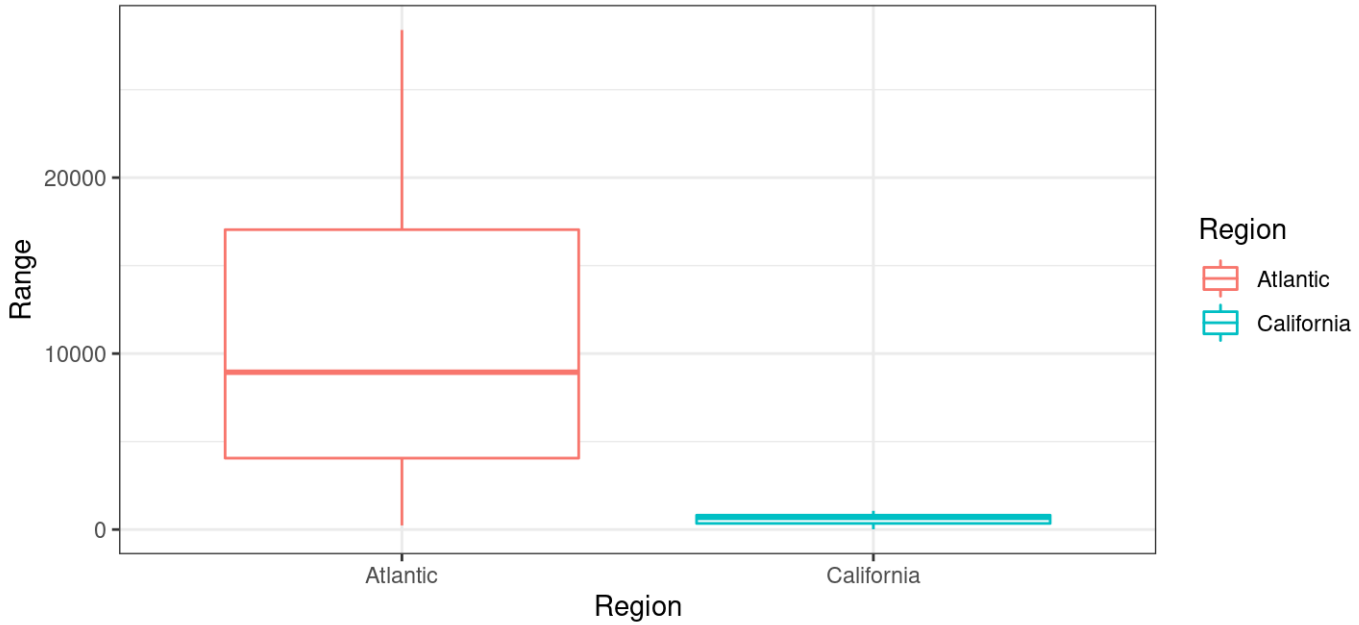
Case Study: Questions

How did the pattern change? Does it surprise you? Do you see any obvious reason that might help explain the correlation?

Since both distributions (Range and Acorn size) have a longer than normal right-tail, a log transformation might be suggested to dampen the influence of the largest observations. Also, relationships between the transformed variables would then be interpretable on a relative scale. It now becomes apparent there is a positive relationship between the two. In other words, there was a relationship between the two all along, but it was not LINEAR.

Case Study: Questions

```
#boxplot of Range by Region
ggplot(acorn, aes(x=Region, y=Range, color=Region)) + #color by Region
  geom_boxplot() + theme_bw() #white background
```



Case Study: Questions

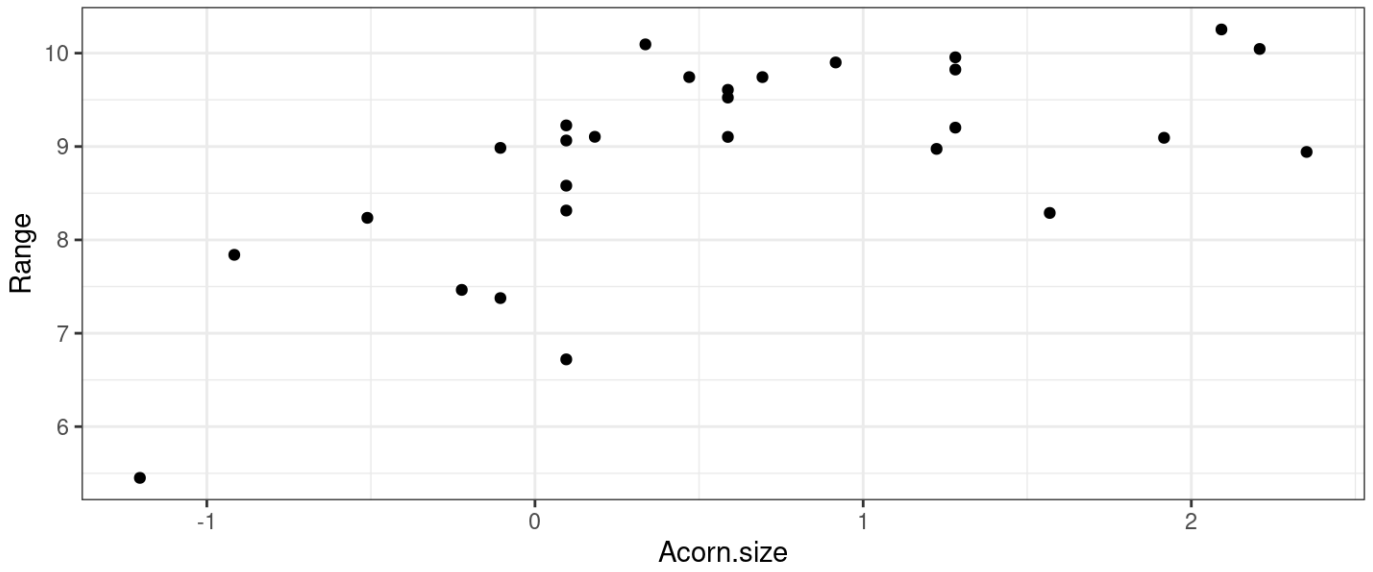
Compare boxplots of tree range by geographical region in order to investigate the relationship between tree range and region. What do you learn?

The comparative boxplots show that the bigger ranges are found in the Atlantic region, the smaller ranges are found in the California region. The range/scales are so different it is hard to see both boxplots with a shared y-axis.

Case Study: Questions

Make an appropriate plot of Ln(range) vs. Ln(acorn size) for the Atlantic region.

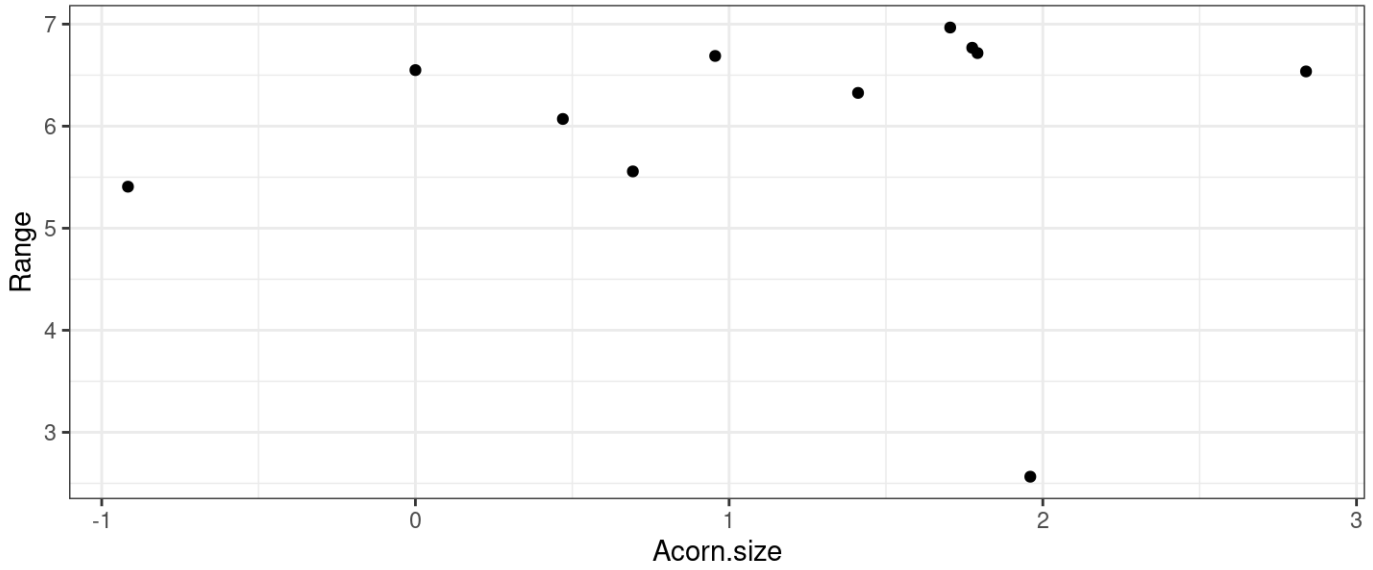
```
acorn3 %>% #this object has ln-transformed Range and Acorn.size
  filter(Region=="Atlantic") %>% #Atlantic region
  ggplot(aes(x=Acorn.size, y=Range)) + #does not require input because of %>%
  geom_point() + theme_bw()  # white background
```



Case Study: Questions

Make an appropriate plot of Ln(range) vs. Ln(acorn size) for the California region.

```
acorn3 %>% #this object has ln-transformed Range and Acorn.size
  filter(Region =="California") %>% #California region
  ggplot(aes(x=Acorn.size, y=Range)) + #does not require input because of %>%
  geom_point() + theme_bw()  #white background
```



Case Study: Questions

Make an appropriate plot of Ln(range) vs. Ln(acorn size) for the Atlantic region. Is the correlation (visually) any better than that found in Question 1?

You didn't really have to calculate the correlation but just visually assess it. The scatter plot for the Atlantic Region shows a moderate positive association.

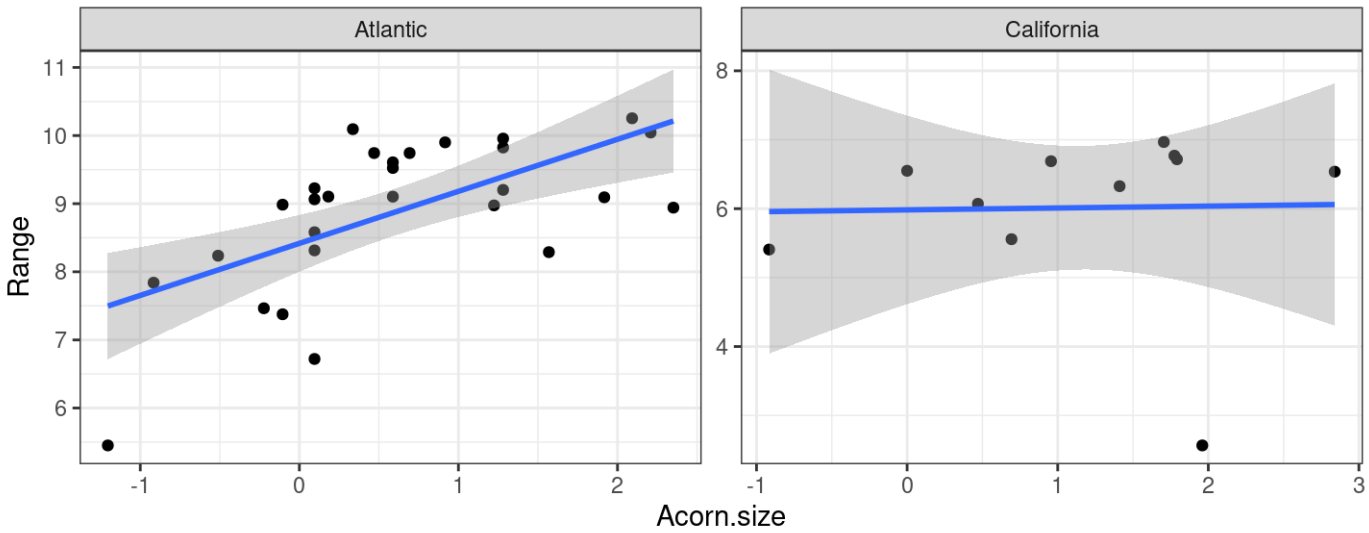
Make an appropriate plot of Ln(range) vs. Ln(acorn size) for the California region. What is the correlation? Why do you think that the correlation is so low?

You didn't really have to calculate the correlation but just visually assess it. It seems pretty low here.This low correlation is due to the presence of the outlier at the bottom right of the plot. The rest of the data seem to be associated in a stronger manner.

Case study: Convince yourself of this

Add a linear regression line for both plots

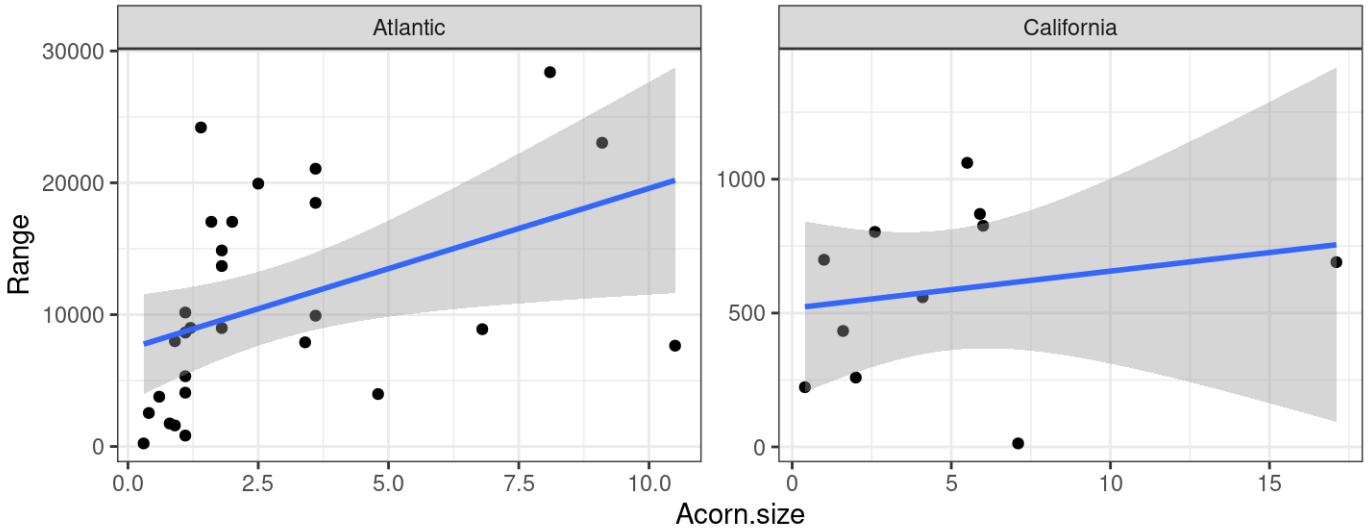
```
acorn3 %>% ggplot(aes(x=Acorn.size, y=Range)) + geom_point() +
  stat_smooth(method="lm", formula='y~x')+
  theme_bw() +  # white background
  facet_wrap(vars(Region),nrow = 1,scales="free") #splits plots by column Region
```



Case study: Convince yourself of this

Same plots without log transformation

```
acorn %>% ggplot(aes(x=Acorn.size, y=Range)) + geom_point() +
  stat_smooth(method="lm", formula='y~x') +
  theme_bw() + facet_wrap(vars(Region),nrow = 1,scales="free")
```



Linear regression

We will learn about this when discussing *linear regressions*, but this was just a start at building this intuition.

Ok, next activity!

Exercise 2 (Lab 7): Taking your first function to the next level!

In Lab 6, you created a simple function that takes in a user defined value and converts into something else. Specifically, the function convert a temperature in Celsius to the corresponding temperature in Fahrenheit. You were then asked to make a similar function that converts temperature in Fahrenheit to Celsius. If you didn't complete that activity, you should do so before moving forward. This is under Lab6_Activity in Rstudio Cloud but also posted on Moodle.

Taking your first function to the next level!

- Combine both functions into one. Basically, you want the function to have two arguments: temperature in Celsius and temperature in Fahrenheit. A few things to consider:
- only one of the two arguments should be provided by the user
- how does the function decide what to do based on the user input?
- what happens if the user does not understand the function and provides both arguments?
- Apply your function to the “Temp” column to the `airquality` dataset from the `datasets` package.

Conditional statements

Before solving this, let's learn about conditional statements.

```
if(3>1){ #logical statement
  print("TRUE.")
}else{
  print("FALSE.")
}
```

```
## [1] "TRUE."
```

Conditional Statements:

Another example

```
x<-2
if (class(x)=="character"){
  print ("This will execute...")
} else
{
  print ("but this will not.")
}
```

```
## [1] "but this will not."
```

Play around with this for a bit!

Task 1: Combine both functions into one.

Celsius to Fahrenheit function (from lab 6):

```
#function to convert Fahrenheit to Celsius
FtoC<-function(tf){ #one argument:tf
tc<-(tf-32)/1.8 #conversion F to C
return(tc) #output of the function
}
```

Fahrenheit to Celsius function (from lab 6):

```
#function to convert Fahrenheit to Celsius
CtoF<-function(tc){ #one argument:tf
tf<-(tc * 1.8) + 32 #conversion C to F
return(tf) #output of the function
}
```

Task 1: Combine both functions into one.

```
FtoC(tf=104) # test
```

```
## [1] 40
```

```
CtoF(tc=40) # test
```

```
## [1] 104
```

```
CtoF(tc=c(0,25,42,100)) # test
```

```
## [1]  32.0  77.0 107.6 212.0
```

```
FtoC(tf=c(32,77,107.6,212)) #test
```

```
## [1]   0  25  42 100
```

Task 1: Combine both functions into one.

Setting the stage: (this is not real code)

```
temp_conv<-function(parameters){ #define function and parameters
  if(temp is in Celsius){ #if temp is in Celsius
    convert to fahrenheit
  } else { #othersise
    convert fahrenheit to celsius
  }
```

Task 1: Combine both functions into one.

Real code:

```
temp_conv<-function(t=30, unit="C"){ #define function and parameters
  if(unit=="C"){
    t2<-(t * 1.8) + 32 #converto to Fahrenheit
  } else { #otherwise
    t2<-(t-32)/1.8 #convert to Celsius
  }
  return(t2)
}
```

Task 1: Combine both functions into one.

```
temp_conv(t=45, unit="C") #test
```

```
## [1] 113
```

```
temp_conv(t=45, unit="F") #test
```

```
## [1] 7.222222
```

```
temp_conv(t=110, unit="F") #test
```

```
## [1] 43.33333
```

```
temp_conv(t=110, unit="C") #test
```

```
## [1] 230
```

Functions applied to vectors

```
temp_conv(t=c(0,25,42,100), unit="C") #freezing, room, very hot, boiling water
```

```
## [1]  32.0  77.0 107.6 212.0
```

```
temp_conv(t=c(32,77,107.6,212), unit="F") #freezing, room, very hot, boiling water
```

```
## [1]   0  25  42 100
```

Let's look at the tasks again:

- only one of the two arguments should be provided by the user.
  Done.
- how does the function decide what to do based on the user input.
  parameter "unit"
- what happens if the user does not understand the function and provides both arguments?
  Not possible in the way we set up this function.

Next task: Apply your function

Apply your function to the “Temp” column to the `airquality` dataset from the `datasets` package.

- load the dataset
- check its structure

```
data(airquality) #load dataset
str(airquality) # temp seems to be in Fahrenheit
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

Next task: Apply your function

- Use your function to convert temperatures from Fahrenheit to Celsius

```
airquality_tempC<-temp_conv(t=airquality$Temp, unit="F")
summary(airquality$Temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   72.00   79.00   77.88   85.00   97.00
```

```
summary(airquality_tempC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.33   22.22   26.11   25.49   29.44   36.11
```

The end!