Lab 1:Intro to R and Data Description

Melissa Pulley

2024-01-26

You will need to submit the CSV file, and your code. You can answer the questions by annotating your answers in the code, or, if you prefer by uploading a word document

These boxes will inform you of things you need to submit or questions you need to answer!

If you struggle with R today, **Don't worry** it will get better and easier! Feel free to reach to any of your instructors for help, or meet with us at some point.

My recommendation is: try to work your way through this lab. Work with peers, and you can look for things online, or ask questions. If you are very lost, check the end of this document for help and resources.

1 Introduction

Some of you may have heard of R, used R, and are even experts in the R coding language, while others might not even have heard of it. My aim in this class is for you all to be functional, proficient users of R and acquire the skills to (1) perform commonly-used analyses in ecology and related fields and (2) have a good statistical foundation to explore and understand approaches that are less common but may eventually be needed in your work.

Today we will begin with the first task you might be asked to do if you work in a lab as a research assistant or technician: data entry! You might think data entry is easy but it takes some thought as to how you would enter data from a data sheet that came straight from the field/lab into an Excel spreadsheet on your computer.

Hopefully, the short presentation was useful!

2 Data entry

Download and open the files, Crab_GTM.jpg, Crab_VCR.jpg, Crab_NIB.jpg, and Crab_PIE.jpg. These files comprise data that researchers have collected by sampling and measuring Atlantic marsh fiddler crabs (*Minuca pugnax*) at four marshes along the US Atlantic coast. The air temperature (from nearby weather stations) and latitude are also recorded on the data sheets.

For the following section, I encourage you to discuss, and work with your peers! In general, you are allowed to discuss and help each other, but the answer and documents submitted MUST be yours!

Enter data of the location, latitude of location, air temperature of location, size (carapace width) of all

individual fiddler crabs into an Excel spreadsheet. Think about how you would enter the information and share/discuss your thoughts with your neighbors.

Enter data of the location, latitude of location, air temperature of location, size (carapace width) of all individual fiddler crabs into an Excel spreadsheet. Think about how you would enter the information and share/discuss your thoughts with your neighbors.

Save the spreadsheet as a .csv file named **FiddlerCrabData.csv** into a new directory named **EEB411_560**.

Upload the *.csv* file to Canvas as part of your assignment (2pts)

Make a (mental) note of where (in the directory) you saved your file!

3 Using R

3.1 Importing Data

You can import data of various formats into R; they include data tables in the form of .dbf, .csv, and .xlsx files or even spatial data such as vectors (.shp) or rasters (.nc).

But the most common type of data files imported into R are probably .csv files, which you have created in the last section.

You can import the dataset using the function read.csv()

Now, let's import the dataset!

You will need to submit your code, so, if you are using RStudio write your code in the source window (figure 2). This way your code is saved. Your RStudio might look different. Just make sure it is the source window and not the console!!!

The way **I** did it, was by typing the following in the **code editor (or source window)**: Do not write code directly in the console!!!!!

```
read.csv("~/projects/EEB411_560/L2/FiddlerCrabData.csv")
```

but, the way **YOU** will do it, is slightly different. You need to tell R where your file is located. Yours might look something like:

```
read.csv('C:/Users/mpulley/FiddlerCrabData.csv') # if windows or
read.csv('~/FiddlerCrabData.csv') #if windows saved in "my documents"
```

How do you find the file path? For mac, open Finder, click to the folder where you saved your data, find your file, right-click (or ctrl-click) it, and then click on GetInfo. **I am not a Mac user, so be patient with me**

For windows, find the file on explorer and copy the path.

Please note that R requires a forward slash "/" while windows usually copies a backslash!

Once you have identified your path, you can import your file! Hopefully, after you import it, it looks something like this:

```
getwd()
```

```
## [1] "C:/Users/mpulley/EEB560"
```

```
read.csv("C:/Users/mpulley/EEB560//FiddlerCrabData.csv")
```

```
##    Site Latitude AirTemp Width
## 1   GTM     30.0  21.742 12.43
## 2   GTM     30.0  21.742 11.84
## 3   GTM     30.0  21.742 11.58
## 4   GTM     30.0  21.742 14.44
## 5   GTM     30.0  21.742  7.55
## 6   GTM     30.0  21.742 13.10
## 7   GTM     30.0  21.742 13.00
## 8   GTM     30.0  21.742 10.32
## 9   GTM     30.0  21.742 12.68
## 10  GTM     30.0  21.742 13.56
## 11  NIB     33.3  18.967 12.71
## 12  NIB     33.3  18.967 14.53
## 13  NIB     33.3  18.967 14.47
## 14  NIB     33.3  18.967 14.69
## 15  NIB     33.3  18.967 15.14
## 16  NIB     33.3  18.967  8.52
## 17  NIB     33.3  18.967 15.88
## 18  NIB     33.3  18.967 12.70
## 19  NIB     33.3  18.967  9.72
## 20  NIB     33.3  18.967 15.67
## 21  PIE     42.7  10.293 21.10
## 22  PIE     42.7  10.293 18.75
## 23  PIE     42.7  10.293 19.51
## 24  PIE     42.7  10.293 17.58
## 25  PIE     42.7  10.293 21.19
## 26  PIE     42.7  10.293 18.67
## 27  PIE     42.7  10.293 18.94
## 28  PIE     42.7  10.293 16.29
## 29  PIE     42.7  10.293 19.37
## 30  PIE     42.7  10.293 17.52
## 31  VCR     37.2  14.950 16.11
## 32  VCR     37.2  14.950 20.78
## 33  VCR     37.2  14.950 18.71
## 34  VCR     37.2  14.950 13.15
## 35  VCR     37.2  14.950 12.94
## 36  VCR     37.2  14.950 15.06
## 37  VCR     37.2  14.950 13.84
## 38  VCR     37.2  14.950 19.36
## 39  VCR     37.2  14.950 19.76
## 40  VCR     37.2  14.950 14.20
```

If it doesn't, now is the time for you to ask for help (either from the internet, from your peers, a TA or your instructor).

I hate to be repetitive... But! Write all the code in the editor (or in any code editor you use, like notepad++), you will have to submit it

This is also called a data frame in R. Typing the function read.csv() prints out the data frame but it doesn't store it in a format that can be further manipulated in R. What we want to do is to place this data in an R object. You can place any information (a number, a vector of numbers, a vector of character strings, data frames, matrices, spatial data etc.) in an R object using <-

```
CrabDat <- read.csv("C:/Users/mpulley/EEB560//FiddlerCrabData.csv")
```

now, there is an object called CrabDat. What happens when you run the following functions?

Answer: It prints the data.

```
print(CrabDat)
```

And what happens when you run

```
print(head(CrabDat)) #what is this?
```

What are those functions (print) and (head) doing?

Question! Describe the function print() and the function head(). 1 pt

Answer: It prints the first 6 rows of data with the data headers.

If you are not sure, try one of my favorite functions in R: the help function!

Run:

```
?head #?is a great function!
```

And see what it does. You can run that (?) with any function! It is super helpful!

allows us to annotate code. Nothing that's past the # is run!

3.2 Data exploration

Let's start with exploring the data! First, let's look at a summary!

```
summary(CrabDat)
```

Question! Based on the summary function, summarize your data in a couple lines and describe what this function (summary) does. 2 pts

Answer: This provides a summary of data per each column. For the character column (size), we are told that the class is a character and the length of the data. For numeric data including AirTemp, Latitude, and Width, we are given summary statistics including the min and max, the median, the mean, and 1st and 3rd quartiles.

You can also explore the dimensions of the data frame using the function dim(). Try this:

```
dim(CrabDat)
```

What do you think each number represents?

Answer: 40 is the number of rows and 4 is the number of columns.

Why are these numbers important? The number of rows give you information on the sample size, i.e., the total number of observations or "data points". The number of columns tell you how many variables, i.e, different information, recorded for each observation or data point.

Also try this:

```
nrow(CrabDat) # what is this? Is the name of the function intuitive?
ncol(CrabDat) # what is this? Is the name of the function intuitive?
```

To see the names of columns, i.e., names of variables recorded, you can do this:

```
names(CrabDat)
```

names(CrabDat) is a vector. A vector is a one-dimensional series of elements. In this case, you can see four elements – each of them the name of a column in CrabDat. CrabDat has 4 columns/variables so it makes sense to get 4 column/variable names.

To retrieve a specific element of a vector (i.e., xth element), you'd need to add [x] to add the end of names(CrabDat), therefore for the 4th element of names(CrabDat), try:

```
names(CrabDat)[4] # does it make sense to you?
names(CrabDat) # check the previous answer against this. Makes sense?
```

If you want to rename one column, try

```
names(CrabDat)[3] = "air_temp"

names(CrabDat)[3]
#previously air_temp in my example
```

The square brackets [] also apply to other types of R objects, such as data frames. Unlike vectors, data frames are two-dimensional: i.e., there are columns as well as rows. Often each row is an observation, while columns describe the different types of data/variables recorded for each observation. The value of a one variable recorded for a given individual is an element.

See the abbreviated CrabDat data frame for a graphical explanation (figure 2)

To retrieve data for a specific row (i.e, the ith row), use [i,] immediately after CrabDat. So if you want to retrieve data for the first row (which is all the information available for observation #1), you can type this:

```
CrabDat[1,] # note position of i or 1 in this case. It is left of the comma.
```

To retrieve data for a specific column (i.e, the pth column), use [,p] immediately after CrabDat. So if you want to retrieve data for the first column (values of the first variable – site – recorded for all observations), you can type this:

```
CrabDat[,1] # note position of p or 1 in this case. It is right of the comma.
```

If you want to retrieve the value of the 3rd column belonging to the 6th observation/row (which is the value contained within the orange box above), you can type this:

```
CrabDat[6,3] # row position is left of comma, column pos. is right of comma
```

Write code to retrieve the size of the 10th observation. What is the size in mm? 1 pt

```
CrabDat[10,4]
```

```
## [1] 13.56
```

Answer: The size of the 10th crab is 13.56 mm.

Write code to retrieve the latitude of the 2nd observation. Write code to retrieve all data (site, latitude, size, and air temperature) from the third observation. 2 pts

Answers:

```
CrabDat[2,2] #Latitude of 2nd observation
```

```
## [1] 30
```

```
CrabDat[3,] #all data from third observation
```

```
##   Site Latitude AirTemp Width
## 3  GTM       30  21.742 11.58
```

You can extract multiple columns and rows of CrabDat if you want to. The idea is to concatenate the different numbers referring to the row and column positions you want, using c(). For example:

```
c(1,2,4) # this prints 1, 2, and 4
```

```
## [1] 1 2 4
```

Look (and run, the following three code chunks)

```
CrabDat[c(1,2,4), 3] # note that c(1,2,4) is left of the comma
```

Answer: This code extract the 1st, 2nd, and 4th rows (observations) of the 3 column, AirTemp

```
CrabDat[1:5, 3]
```

Answer: This code extracts the air temperature (column 3) from observations (rows) 1 through 5.

```
1:5
```

Answer: This code gives integers between 1 and 5.

Describe what each of the last three code chunks did1 pt

See above.

Another way of retrieving data for a specific column or variable is to use $VarName immediately after CrabDat, where VarName is the name of the column/variable. So to retrieve measurements of size for all fiddler crab individuals in the data set, you can type:

```
CrabDat$Width # since 'size' is the name of the variable you want
```

You can also do this to get the same information as above (size of first five individuals):

```
CrabDat$Width[1:5]
```

Question! Why is there no comma within the square brackets [ ]? 1 pt. It's ok to ask us for hints!!!

Answer: Because "CrabDat$Width" is only one column, it is not necessary to index by row and column, instead, we can extract by the index (or row) without needing to reference a column.

Question! Write code to retrieve the air temperature associated with the first 10 individuals. Does it make sense to you that the air temperature values are all the same along the first 10 individuals? Why? [2 pts]

```
CrabDat$AirTemp[1:10]
```

```
##  [1] 21.742 21.742 21.742 21.742 21.742 21.742 21.742 21.742 21.742 21.742
```

Answer: Yes, because all of the first 10 individuals were observed at the same site.

3.3 Beginning to describe and visualize data

Obviously, we are just scratching the surface of data processing in the previous section. These are the basic building blocks, which I hope will help you learn different ways of processing data in R in the future. We will come back to this in future tutorials and you'd realize that there is often many ways of doing the same thing in R.

Now let's move on to simple data description and visualization. We will briefly cover some of these topics in the future, but I want you all to get an intuition by doing first.

3.4 Types of data

There are two main data/variable types: (1) categorical or qualitative data; and (2) numerical or quantitative data.

In CrabDat, which variables (columns) are categorical and which ones are numeric?

```
summary(CrabDat)
```

```
##      Site              Latitude        AirTemp          Width
##  Length:40          Min.   :30.00   Min.   :10.29   Min.   : 7.55
##  Class :character   1st Qu.:32.48   1st Qu.:13.79   1st Qu.:12.88
##  Mode  :character   Median :35.25   Median :16.96   Median :14.61
##                     Mean   :35.80   Mean   :16.49   Mean   :15.18
##                     3rd Qu.:38.58   3rd Qu.:19.66   3rd Qu.:18.68
##                     Max.   :42.70   Max.   :21.74   Max.   :21.19
```

If a variable has information like mean, median, or max, it is numerical/quantitative data. The variable site is shown as a character variable. This is an example of categorical or qualitative data. It makes sense, as each code refers to a given site where crabs were sampled.

You can formally define site as a categorical (factor) variable using as.factor

```
CrabDat$Site <- as.factor(CrabDat$Site)
```

Now run a summary()of CrabDat again. What does it say under site now?

```
summary(CrabDat) # is there more information now? Yes
```

```
##   Site       Latitude        AirTemp          Width
##  GTM:10   Min.   :30.00   Min.   :10.29   Min.   : 7.55
##  NIB:10   1st Qu.:32.48   1st Qu.:13.79   1st Qu.:12.88
##  PIE:10   Median :35.25   Median :16.96   Median :14.61
##  VCR:10   Mean   :35.80   Mean   :16.49   Mean   :15.18
##           3rd Qu.:38.58   3rd Qu.:19.66   3rd Qu.:18.68
```

```
##           Max.   :42.70   Max.   :21.74   Max.   :21.19
```

See the new information!

3.5. Describing and visualizing different types of data

Categorical/qualitative data are data that do not have magnitude that can be measured by numbers (i.e., not quantitative). Categorical data include assignments of samples/individuals/observations into groups or categories. For example, the type of high-schools in the US can be classified into public vs. private high schools. There are two discrete categories: public or private. Therefore this is a categorical variable.

In CrabDat, site is categorical. We can summarize categorical variables by using table()

```
table(CrabDat$Site)
```

```
##
## GTM NIB PIE VCR
##  10  10  10  10
```

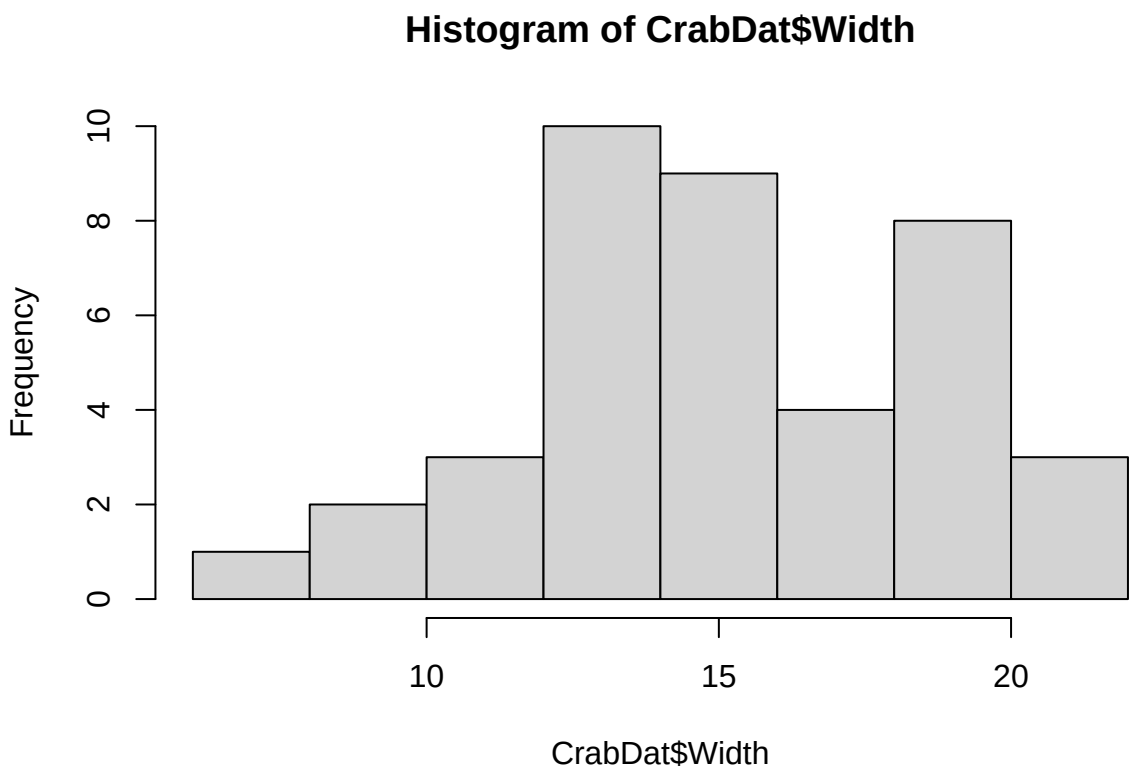
The printed output counts the number of observations belonging to each category of the site variable.

Numeric/quantitative data are data that have magnitude and they can be described numerically. Larger values of a given quantitative variable have a larger magnitude than smaller values of the same variable. For example, the mean ACT score of seniors in a given high school is a numeric/quantitative variable.

In CrabDat, size, air_temp, and latitude are numeric/quantitative variables.

For quantitative variables, we can plot a histogram to visualize the distribution (or spread) of values. In R, we can use hist()

```
hist(CrabDat$Width) # what do you think of the distribution
```

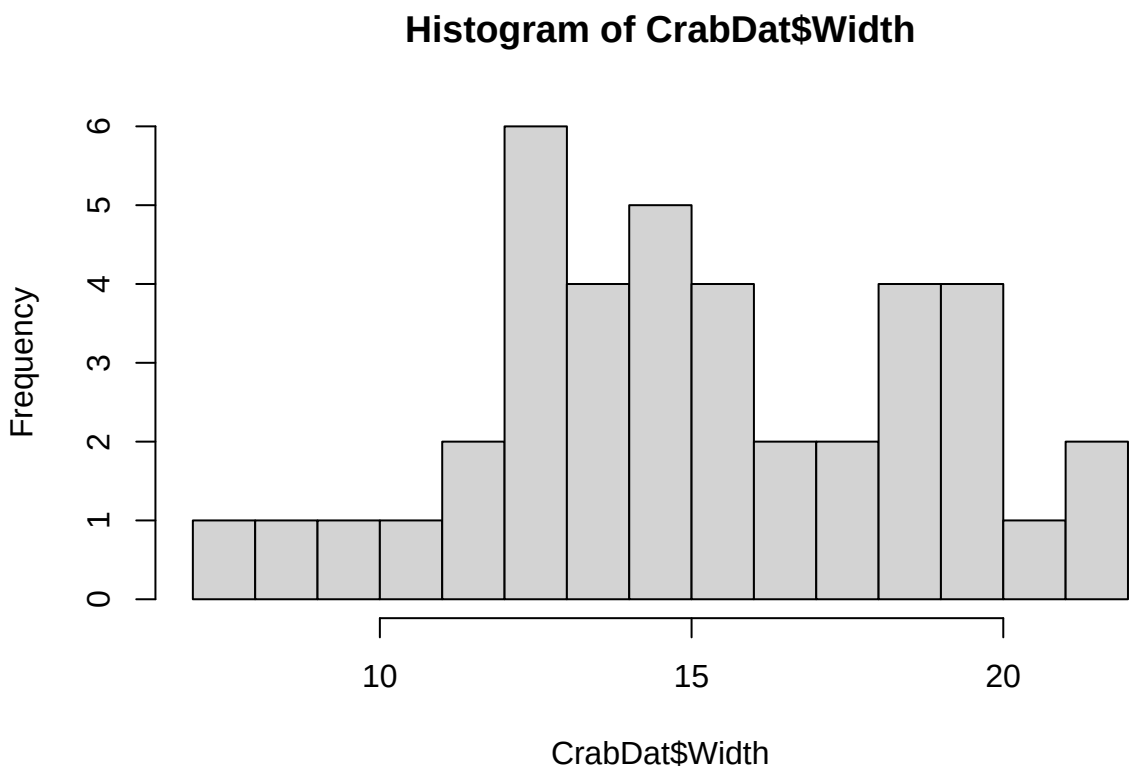


We can use breaks=n to change the number of breaks. Look at:

```
?hist
```

and then run:

```
hist(CrabDat$Width, breaks=10) # what do you think of the distribution
```



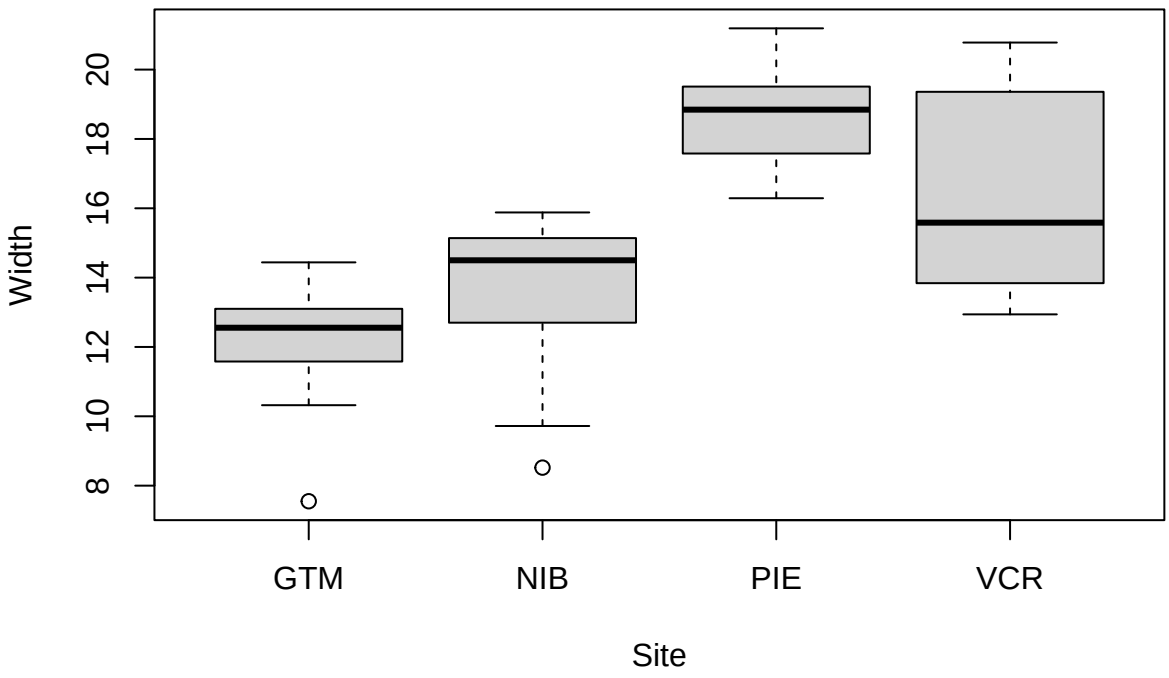
But replace x for an appropriate number.

Upload your plot with an appropriate number of breaks to Canvas. Justify why you chose those breaks. 1pt

Answer: I chose 10 breaks because I wanted to have a finer partition of the data so that I can see more detail at the center of the data.

You can also use boxplots to visualize the spread of a given across different categories of another categorical variables. Here, we want to visualize the distribution of individual crab sizes across different sites using boxplot().

```
boxplot(Width~Site, data=CrabDat) # what do you think?
```



Question! what do the lines and boxes represent? [1 pt]

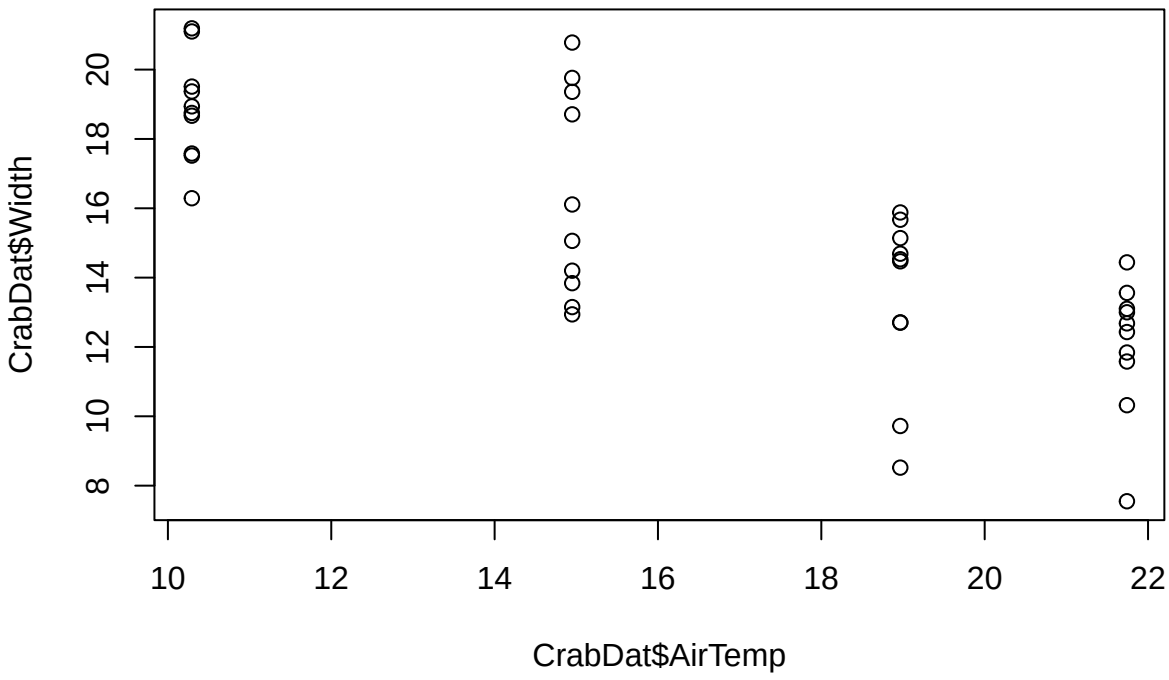
Answer: For one site, the dark line indicates the median width for that site. The box starts from the first quartile and extends to the third quartile to indicate the spread of the central data points. The thin lines indicate the min and max data points, with the exception of outliers, indicated by circles.

Last, for this tutorial, sometimes we want to visualize the relationship between two quantitative variables. The quickest way to do this is to make a dot plot (scatterplot) using plot()

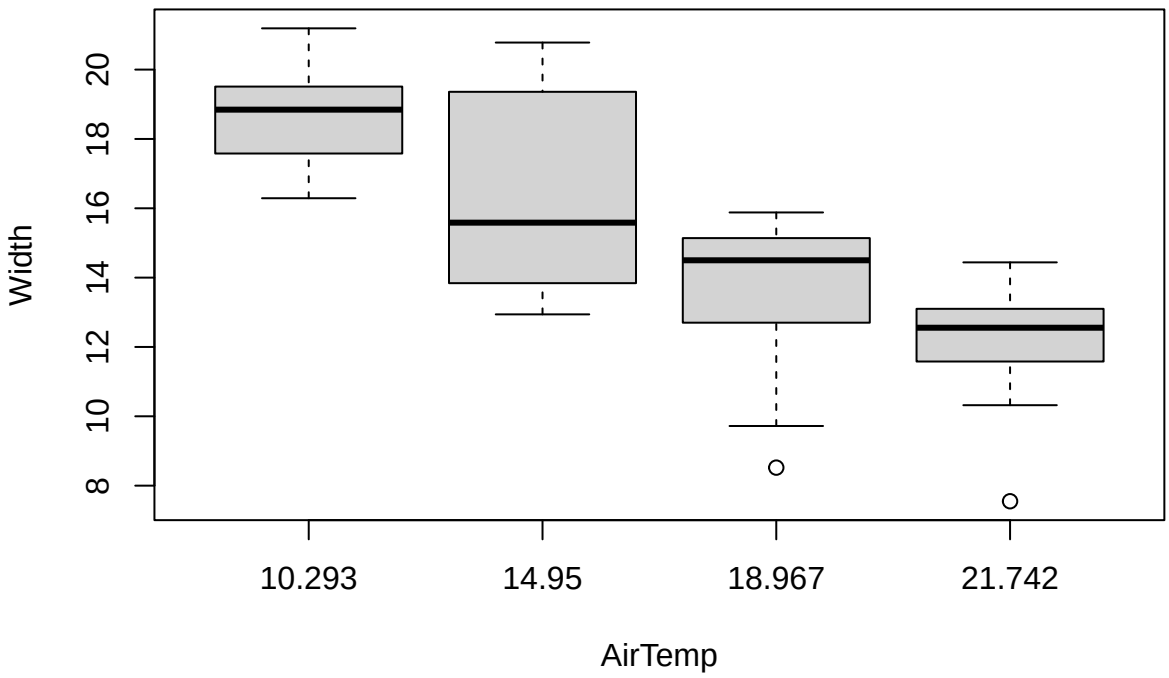
```
#plot(size~latitude, data=CrabDat) # what do you think?
```

Write code to make a plot of size (y-axis) against air temperature (x-axis). You decide what type of plot to present (dot plot or boxlplot), and explain why you made that choice. 2pts

```
plot(CrabDat$AirTemp,CrabDat$Width)
```




```
boxplot(Width~AirTemp, data=CrabDat)
```



Answer While I examined both a dot plot and a box plot, because there are only four possible temperatures, the box plot is more descriptive because it visualizes summary of the data distribution so that you can see median, quartiles etc for each location, identified by air temperature.

End of Lab 1

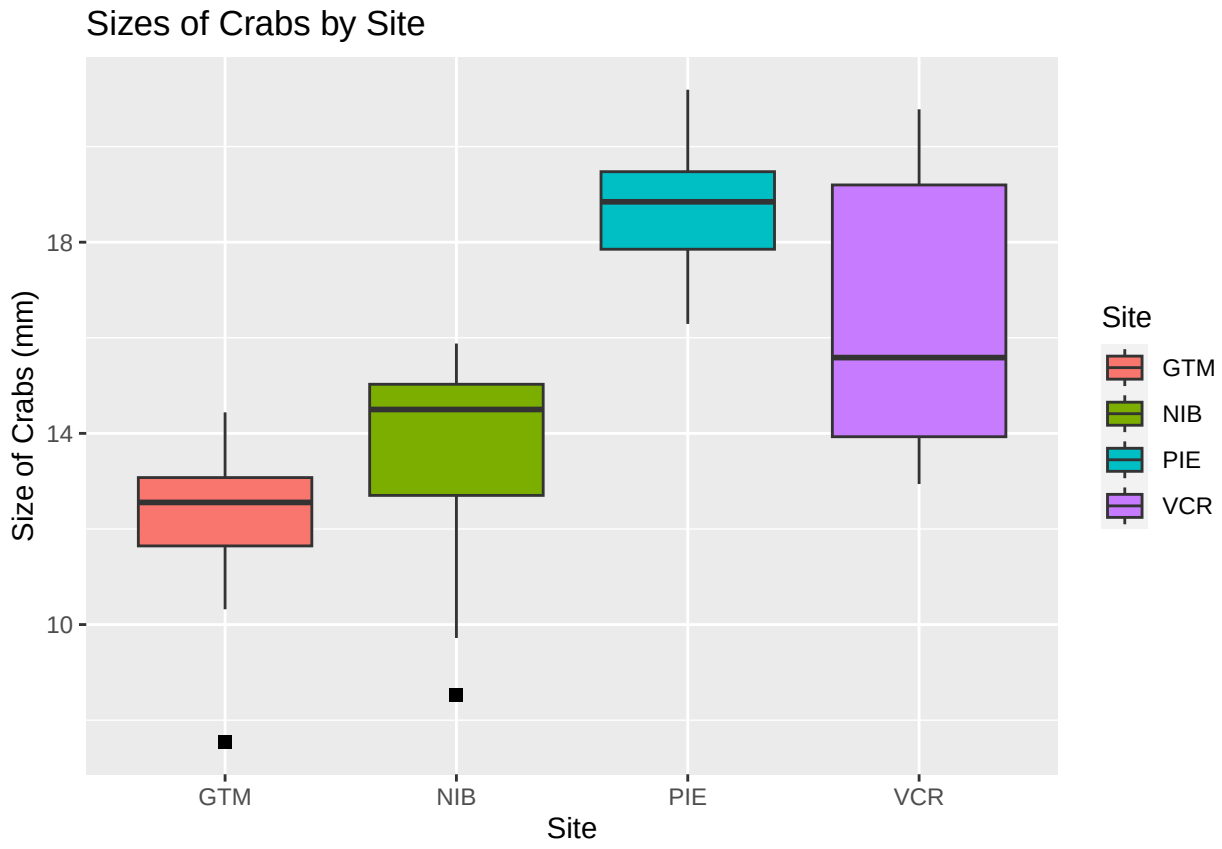
Make sure to check the resources and potential extra credits at the end of this document

---

4 Extra resources, Extra credits, and other information

EXTRA CREDIT (1 pt) CREATE A "PUBLICATION QUALITY PLOT". Research online, and figure out how to manipulate some of the graphical parameters (e.g. size of points, color, thickness of lines, etc.) Using the fiddler crab data, make a pretty plot that is of "publication quality" (i.e., a plot that could be on a research paper). Add an appropriate figure legend (read other papers to see what they include in a figure legend)

```
library(ggplot2)
ggplot(data=CrabDat, aes(x=Site,y=Width, fill=Site)) +
  geom_boxplot(outlier.colour="black", outlier.shape=15,
               outlier.size=2, notch=FALSE) +
    labs(title = "Sizes of Crabs by Site") +
        ylab("Size of Crabs (mm)") + xlab("Site")
```



Resources

These are hyperlinks

The base R cheat-sheet

Dr. Fordyce Intro to R video

An introduction to R

End of document