Lab #4 – EDA Visualization   **CSC 3220**   **Fall 2021**

Name: Caleb Lich, Jacob Blovins, Nathan Hynek, Michael Steets

GRADING

- 20 points per question (5*20 = 100 points)
- **NOTE**: You must show BOTH your R code and your visualizations.
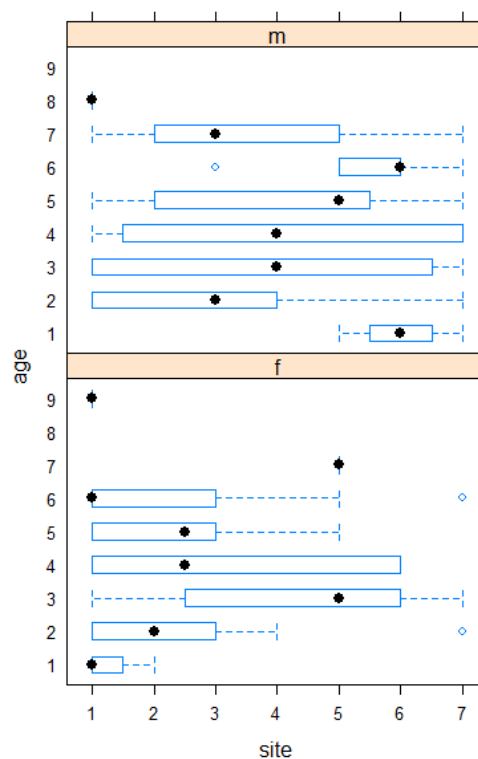
PRELIMINARIES

- Use the DAAG library and the lattice library

QUESTIONS

1. Use the lattice function `bwplot()` to display, for each combination of site and sex in the data frame **possum** (DAAG package), the distribution of ages. Show a two-panel, side-by-side **box plots** for the different sites, with one panel for male and one for female.

> bwplot(age~site | sex, possum)

2. Plot a **histogram** of the **earconch** measurements for the possum data. The distribution should appear bimodal (two peaks). This is a simple indication of clustering, possibly due to sex differences.

> hist(earconch)

Histogram of earconch
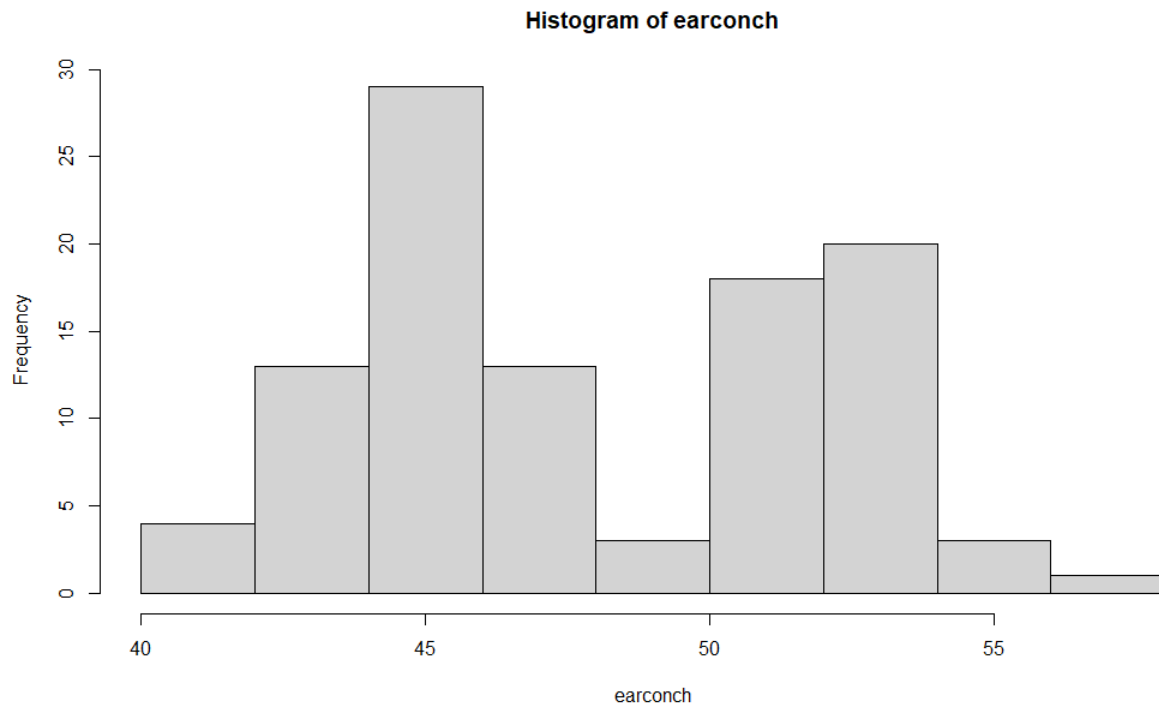
Show side-by-side **boxplots** of both the male and female earconch measurements (i.e., two box plots). How do these measurement distributions differ?

> boxplot(earconch~sex)
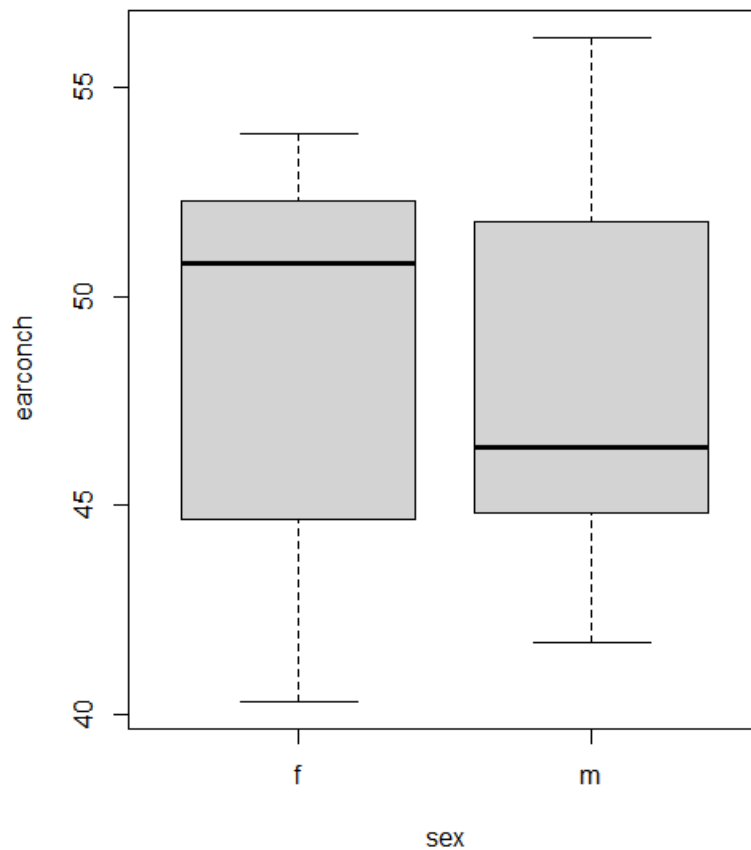
Now what if you had done two histograms – one for male and one for female. Can you predict what the corresponding histograms would look like? Plot them to check your answer.

We predict that the higher frequency in ear conches will be in female than male because the values are higher in the f sex for the box plot than the male

> hist(earconch [sex == "f"])
> hist(earconch [sex == "m"])

Histogram of earconch[sex == "f"]   Histogram of earconch[sex == "m"]
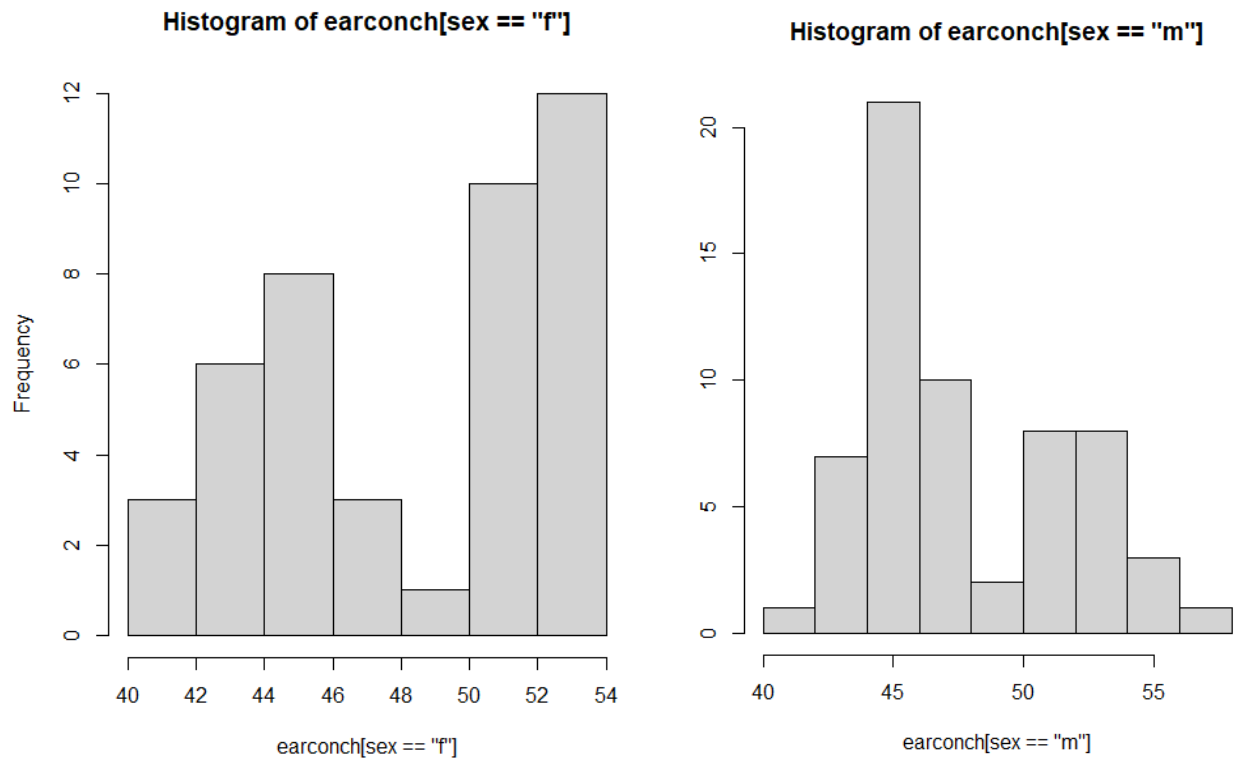
3. For the data frame **ais** (DAAG package), draw boxplots (again using **bwplot**) that show how the values of the hematological measure *red cell count* vary with the *sport* and *sex* of the athlete. (Again, separate each of the boxplots by sex, and put them side-by-side so they are easier to compare.)

> bwplot(rcc~sport | sex, ais)
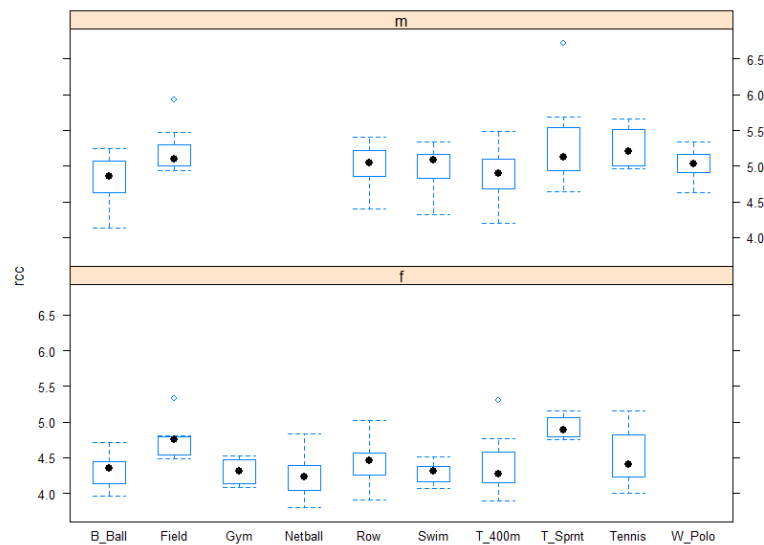
4. Using the data frame cuckoohosts, plot `clength` against `cbreadth`, and `hlength` against `hbreadth`, all on the same graph and using a different color to distinguish the first set of points (for the cuckoo eggs) from the second set (for the host eggs). Join the two points that relate to the same host species with a line. What does a line that is long, relative to other lines, imply? Here is code that you may wish to use:

```
> attach(cuckoohosts)
> plot(c(clength, hlength), c(cbreadth, hbreadth),
+       col=rep(1:2,c(12,12)))
> for(i in 1:12)lines(c(clength[i], hlength[i]),
+                      c(cbreadth[i], hbreadth[i]))
> text(hlength, hbreadth, abbreviate(rownames(cuckoohosts),8))
> detach(cuckoohosts)
```

That there are big variations between egg sizes. The ones with shorter lines mean that the size variance of eggs will be narrow. For example, wren eggs will vary very greatly in sizes while robin eggs will not as much.
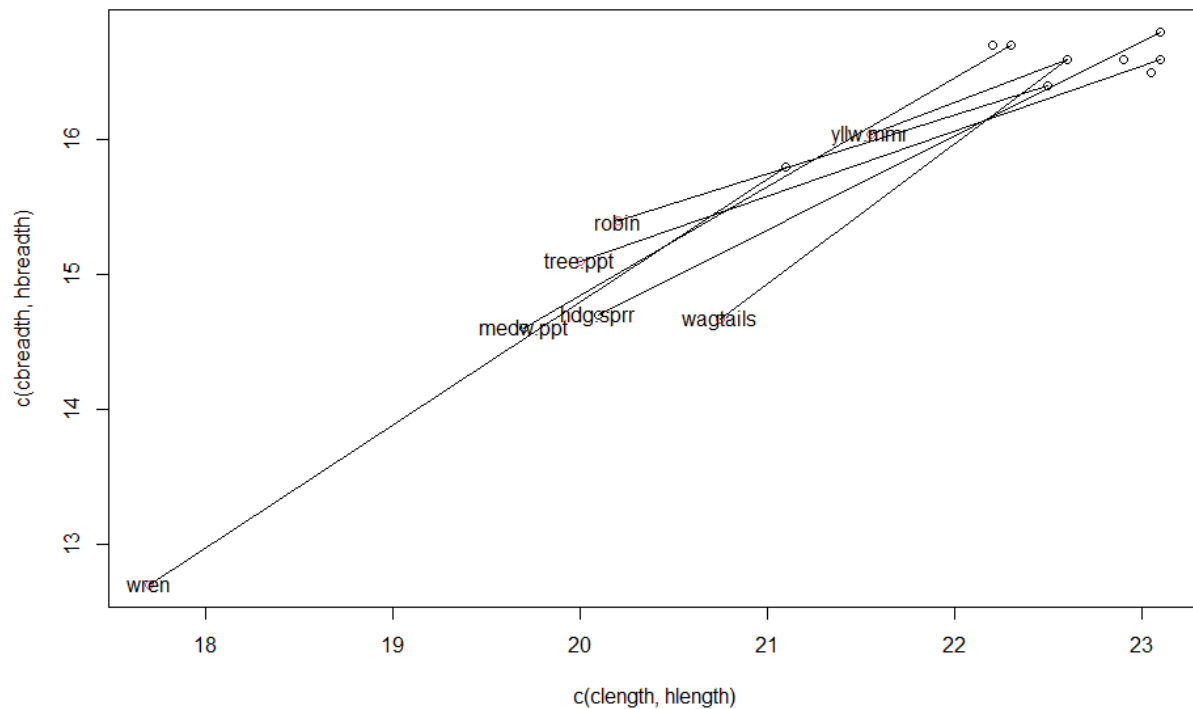
5. Examine the help for the function `mean()`, and use it to learn about the trimmed mean. For the total lengths of female **possums**, calculate the mean, the median, and the 10% trimmed mean. How does the 10% trimmed mean differ from the mean for these data? Under what circumstances will the trimmed mean differ substantially from the mean?

> mean(possum$totlngth [sex == "f"])
[1] 87.90698
> median(possum$totlngth [sex == "f"])
[1] 88.5
> mean(possum$totlngth [sex == "f"], trim = .1)
[1] 88.04286

If there was a lot of outliers, the regular mean would be very different from the trimmed mean. This would be because the mean takes the average of everything while the trimmed part cuts out that outlier.

(Submit your answers into the corresponding iLearn dropbox WITH ALL OF YOUR NAMES <u>in a PDF file</u>)