3. Worksheet: Basic R

Anna Lennon; Z620: Quantitative Biodiversity, Indiana University

24 January, 2025

OVERVIEW

This worksheet introduces some of the basic features of the R computing environment (http://www.r-project.org). It is designed to be used along side the **3. RStudio** handout in your binder. You will not be able to complete the exercises without the corresponding handout.

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom today, you must **push** this file to your GitHub repo, at whatever stage you are. This will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '3.RStudio' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create a **pull request** via GitHub. Your pull request should include this file (**3.RStudio_Worksheet.Rmd**) with all code blocks filled out and questions answered) and the PDF output of `Knitr` (**3.RStudio_Worksheet.pdf**).

The completed exercise is due on **Wednesday, January 22nd, 2025 before 12:00 PM (noon)**.

1) HOW WE WILL BE USING R AND OTHER TOOLS

You are working in an RMarkdown (.Rmd) file. This allows you to integrate text and R code into a single document. There are two major features to this document: 1) Markdown formatted text and 2) "chunks" of R code. Anything in an R code chunk will be interpreted by R when you *Knit* the document.

When you are done, you will *knit* your document together. However, if there are errors in the R code contained in your Markdown document, you will not be able to knit a PDF file. If this happens, you will need to review your code, locate the source of the error(s), and make the appropriate changes. Even if you are able to knit without issue, you should review the knitted document for correctness and completeness before you submit the Worksheet. Next to the `Knit` button in the RStudio scripting panel there is a spell checker button (`ABC`) button.

2) SETTING YOUR WORKING DIRECTORY

In the R code chunk below, please provide the code to: 1) clear your R environment, 2) print your current working directory, and 3) set your working directory to your '3.RStudio' folder.

```
#1
rm(list = ls())
#2
getwd()
```

```
## [1] "/cloud/project/QB2025_ALennon/Week1-RStudio"
```

```
#3
setwd("/cloud/project/QB2025_ALennon/Week1-RStudio")
```

3) USING R AS A CALCULATOR

To follow up on the pre-class exercises, please calculate the following in the R code chunk below. Feel free to reference the **1. Introduction to version control and computing tools** handout.

1) the volume of a cube with length, l, = 5 (volume = l^3)
2) the area of a circle with radius, r, = 2 (area = pi * r^2).
3) the length of the opposite side of a right-triangle given that the angle, theta, = pi/4. (radians, a.k.a. 45°) and with hypotenuse length sqrt(2) (remember: sin(theta) = opposite/hypotenuse).
4) the log (base e) of your favorite number.

```
print(paste("1)", 5^3))
```

```
## [1] "1) 125"
```

```
#2
r <- 2
print(paste ("2)", pi *r**2))
```

```
## [1] "2) 12.5663706143592"
```

```
#3
theta <- pi/4
hyp <- sqrt(2)
print(paste ("3)",(sin(theta))*hyp))
```

```
## [1] "3) 1"
```

```
#4
print(paste ("4)", log(2)))
```

```
## [1] "4) 0.693147180559945"
```

4) WORKING WITH VECTORS

To follow up on the pre-class exercises, please perform the requested operations in the R-code chunks below.

Basic Features Of Vectors

In the R-code chunk below, do the following: 1) Create a vector `x` consisting of any five numbers. 2) Create a new vector `w` by multiplying `x` by 14 (i.e., "scalar"). 3) Add `x` and `w` and divide by 15.

```
x <- c(1, 2, 3, 4, 5)
w <- x*14
print(w)
```

```
## [1] 14 28 42 56 70
```

```
print((x*w)/15)
```

```
## [1]  0.9333333  3.7333333  8.4000000 14.9333333 23.3333333
```

Now, do the following: 1) Create another vector (`k`) that is the same length as `w`. 2) Multiply `k` by `x`. 3) Use the combine function to create one more vector, `d` that consists of any three elements from `w` and any four elements of `k`.

```
k <- c(2, 22, 222, 2222, 22222)
print(x*k)
```

```
## [1]      2     44    666   8888 111110
```

```
print(w)
```

```
## [1] 14 28 42 56 70
```

```
print(k)
```

```
## [1]     2    22   222  2222 22222
```

```
d <- c(w[1:3], k[1:4])
print(d)
```

```
## [1]   14   28   42    2   22  222 2222
```

Summary Statistics of Vectors

In the R-code chunk below, calculate the **summary statistics** (i.e., maximum, minimum, sum, mean, median, variance, standard deviation, and standard error of the mean) for the vector (`v`) provided.

```
v <- c(16.4, 16.0, 10.1, 16.8, 20.5, NA, 20.2, 13.1, 24.8, 20.2, 25.0, 20.5, 30.5, 31.4, 27.1)

print(max(na.omit(v)))
```

```
## [1] 31.4
```

```
print(min(na.omit(v)))
```

```
## [1] 10.1
```

```
print(sum(na.omit(v)))
```

```
## [1] 292.6
```

```
print(mean(na.omit(v)))
```

```
## [1] 20.9
```

```
print(median(na.omit(v)))
```

```
## [1] 20.35
```

```
print(var(na.omit(v)))
```

```
## [1] 39.44
```

```
print(sd(na.omit(v)))
```

```
## [1] 6.280127
```

5) WORKING WITH MATRICES

In the R-code chunk below, do the following: Using a mixture of Approach 1 and 2 from the **3. RStudio** handout, create a matrix with two columns and five rows. Both columns should consist of random numbers. Make the mean of the first column equal to 8 with a standard deviation of 2 and the mean of the second column equal to 25 with a standard deviation of 10.

```
c1 <- c(rnorm(5, mean = 8, sd = 2))
c2 <- c(rnorm(5, mean = 25, sd = 10))
m1 <- cbind (c1, c2)
print(m1)
```

```
##             c1       c2
## [1,]  9.414854 26.42853
## [2,]  7.922644 28.72667
## [3,]  6.349440 20.33888
## [4,] 10.831494 23.96845
## [5,]  5.465372 25.87375
```

Question 1: What does the `rnorm` function do? What do the arguments in this function specify? Remember to use `help()` or type `?rnorm`.

> Answer 1: rnorm is a function that generates the normal distribution for a given data set. rnorm(n, mean = 0, sd =1) where n is the number of observations, mean is the vector of means, and sd is the vector of standard deviations

In the R code chunk below, do the following: 1) Load `matrix.txt` from the **3.RStudio** data folder as matrix `m`. 2) Transpose this matrix. 3) Determine the dimensions of the transposed matrix.

```
m <- as.matrix(read.table("data/matrix.txt", sep ="\t", header = FALSE))
print(m)
```

```
##       V1 V2 V3 V4 V5
##  [1,]  8  1  7  6  1
##  [2,]  5  5  2  4  1
##  [3,]  2  5  4  3  3
##  [4,]  3  2  5  1  4
##  [5,]  9  9  1  1  2
##  [6,] 11  8  1  8  8
##  [7,]  2  2  5  8  5
##  [8,]  3  3  6  7  6
##  [9,]  5  5  1  3  6
## [10,]  6  5  9  2  2
```

```
m2t <- t(m)
print(dim(m2t))
```

```
## [1]  5 10
```

Question 2: What are the dimensions of the matrix you just transposed?

> Answer 2: [5,10]

###Indexing a Matrix

In the R code chunk below, do the following: 1) Index matrix `m` by selecting all but the third column. 2) Remove the last row of matrix `m`.

```
#print(m)
mI <- m[ ,c(1:2, 4:5)]
```

```
#print(mI)
mIrm <-m[-10,]
print(mIrm)
```

```
##       V1 V2 V3 V4 V5
##  [1,]  8  1  7  6  1
##  [2,]  5  5  2  4  1
##  [3,]  2  5  4  3  3
##  [4,]  3  2  5  1  4
##  [5,]  9  9  1  1  2
##  [6,] 11  8  1  8  8
##  [7,]  2  2  5  8  5
##  [8,]  3  3  6  7  6
##  [9,]  5  5  1  3  6
```

6) BASIC DATA VISUALIZATION AND STATISTICAL ANALYSIS

Load Zooplankton Data Set

In the R code chunk below, do the following: 1) Load the zooplankton data set from the **3.RStudio** data folder. 2) Display the structure of this data set.

```
zoop <-read.table("/cloud/project/QB2025_ALennon/Week1-RStudio/data/zoop_nuts.txt", sep = "\t", header =

print(zoop)
```

```
##    TANK NUTS     TP     TN   SRP     TIN  CHLA     ZP
## 1    34    L  20.31  720.1  4.02  131.62  1.52 1.7808
## 2    14    L  25.55  750.5  1.56  141.10  4.00 0.4090
## 3    23    L  14.22  610.1  4.97  107.70  0.61 1.2014
## 4    16    L  39.11  760.9  2.89   71.28  0.53 3.3598
## 5    21    L  20.09  570.4  5.11   80.40  1.44 0.7332
## 6     5    L  15.75  680.5  4.68  135.77  1.19 0.9773
## 7    25    L  19.55  665.5  5.00   79.40  0.37 1.0999
## 8    27    L  16.19  660.8  0.10  100.91  0.72 2.2714
## 9    30    M  29.46 1770.4  7.90 1329.26  6.93 3.1633
## 10   28    M  37.88 2590.3  3.92 1163.64  0.94 1.8747
## 11   35    M  30.26 2110.9  4.45 1850.18  1.36 4.3802
## 12   36    M  36.94 2060.9  5.14  249.93 38.38 2.4051
## 13   12    M  34.73 1370.1  4.69  420.01 15.99 1.7079
## 14   22    M  26.00 2110.3  5.35 1466.70  0.95 4.0999
## 15   18    M  28.50 1760.4  7.15 1351.83  1.36 5.4430
## 16   15    M  35.33 1360.8  5.96 1036.27  2.13 4.2677
## 17   17    H  41.56 4130.1 20.34 3421.43  1.44 8.2084
## 18   10    H  53.50 4530.4 33.57 4042.10  0.93 4.2273
## 19   29    H  99.07 4410.9 11.57 3307.05  0.61 6.2381
## 20    6    H 128.04 4750.4 26.27 3686.17  1.27 8.5713
## 21   24    H  33.47 3410.4  9.32 2791.52  1.11 1.4240
## 22   19    H  52.41 3710.3  3.23 2890.73 17.59 2.9714
## 23    4    H  42.21 3690.4 12.71 3041.75  1.08 8.1509
## 24   11    H  77.65 4380.6 21.86 3041.75  1.08 8.3868
```

```
str(zoop)
```

```
## 'data.frame':    24 obs. of  8 variables:
##  $ TANK: int  34 14 23 16 21 5 25 27 30 28 ...
```

```
##  $ NUTS: chr  "L" "L" "L" "L" ...
##  $ TP  : num  20.3 25.6 14.2 39.1 20.1 ...
##  $ TN  : num  720 750 610 761 570 ...
##  $ SRP : num  4.02 1.56 4.97 2.89 5.11 4.68 5 0.1 7.9 3.92 ...
##  $ TIN : num  131.6 141.1 107.7 71.3 80.4 ...
##  $ CHLA: num  1.52 4 0.61 0.53 1.44 1.19 0.37 0.72 6.93 0.94 ...
##  $ ZP  : num  1.781 0.409 1.201 3.36 0.733 ...
```

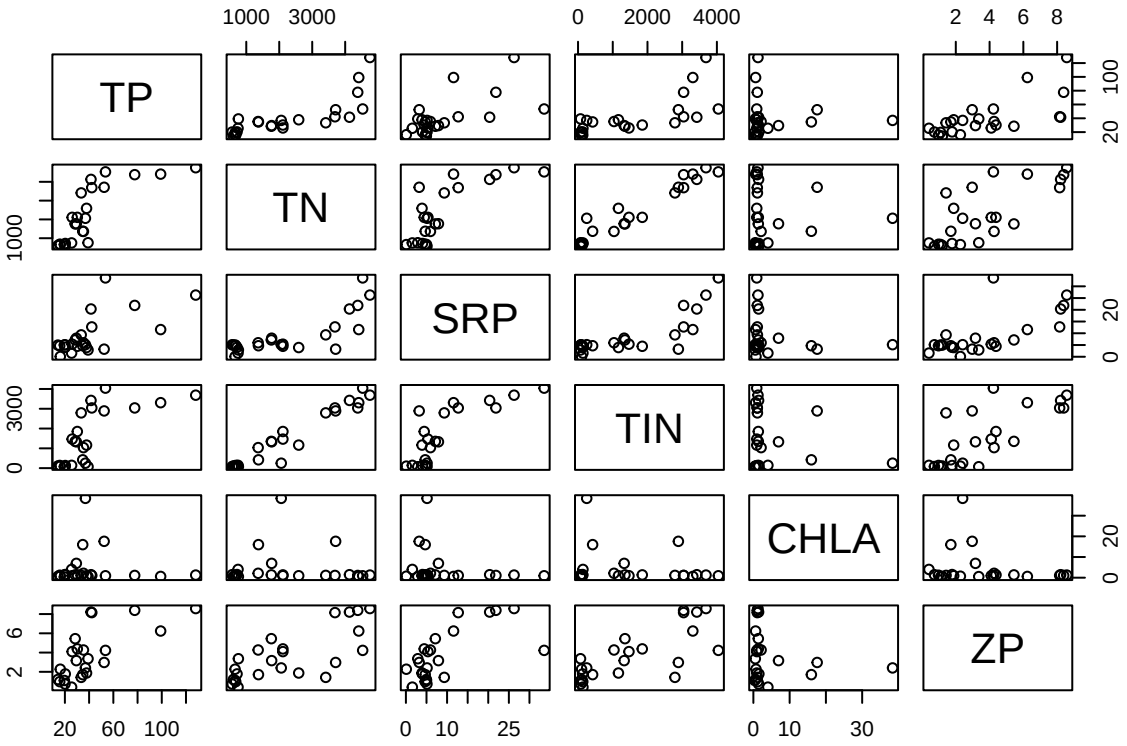
Correlation

In the R-code chunk below, do the following: 1) Create a matrix with the numerical data in the `meso` dataframe. 2) Visualize the pairwise **bi-plots** of the six numerical variables. 3) Conduct a simple **Pearson's correlation** analysis.

```
zoop.num <- zoop[, 3:8]
print(zoop.num)
```

```
##        TP     TN   SRP     TIN  CHLA     ZP
## 1   20.31  720.1  4.02  131.62  1.52 1.7808
## 2   25.55  750.5  1.56  141.10  4.00 0.4090
## 3   14.22  610.1  4.97  107.70  0.61 1.2014
## 4   39.11  760.9  2.89   71.28  0.53 3.3598
## 5   20.09  570.4  5.11   80.40  1.44 0.7332
## 6   15.75  680.5  4.68  135.77  1.19 0.9773
## 7   19.55  665.5  5.00   79.40  0.37 1.0999
## 8   16.19  660.8  0.10  100.91  0.72 2.2714
## 9   29.46 1770.4  7.90 1329.26  6.93 3.1633
## 10  37.88 2590.3  3.92 1163.64  0.94 1.8747
## 11  30.26 2110.9  4.45 1850.18  1.36 4.3802
## 12  36.94 2060.9  5.14  249.93 38.38 2.4051
## 13  34.73 1370.1  4.69  420.01 15.99 1.7079
## 14  26.00 2110.3  5.35 1466.70  0.95 4.0999
## 15  28.50 1760.4  7.15 1351.83  1.36 5.4430
## 16  35.33 1360.8  5.96 1036.27  2.13 4.2677
## 17  41.56 4130.1 20.34 3421.43  1.44 8.2084
## 18  53.50 4530.4 33.57 4042.10  0.93 4.2273
## 19  99.07 4410.9 11.57 3307.05  0.61 6.2381
## 20 128.04 4750.4 26.27 3686.17  1.27 8.5713
## 21  33.47 3410.4  9.32 2791.52  1.11 1.4240
## 22  52.41 3710.3  3.23 2890.73 17.59 2.9714
## 23  42.21 3690.4 12.71 3041.75  1.08 8.1509
## 24  77.65 4380.6 21.86 3041.75  1.08 8.3868
```

```
zoop.bip <- pairs(zoop.num) #output is biplots
```




```
zoop.cor <-cor(zoop.num)
print (zoop.cor)
```

```
##                TP           TN         SRP        TIN         CHLA         ZP
## TP    1.00000000  0.786510407  0.6540957  0.7171143 -0.016659593  0.6974765
## TN    0.78651041  1.000000000  0.7841904  0.9689999 -0.004470263  0.7562474
## SRP   0.65409569  0.784190400  1.0000000  0.8009033 -0.189148017  0.6762947
## TIN   0.71711434  0.968999866  0.8009033  1.0000000 -0.156881463  0.7605629
## CHLA -0.01665959 -0.004470263 -0.1891480 -0.1568815  1.000000000 -0.1825999
## ZP    0.69747649  0.756247384  0.6762947  0.7605629 -0.182599904  1.0000000
```

Question 3: Describe some of the general features based on the visualization and correlation analysis above?

> Answer 3: In general, there seems to be a high, positive correlation between the variables with notiable exceptions of weak negative correlations with the CHLA variable. This pattern is observed visually with the distribution of the vriables on the biplots being roughly linear spread across the plot. The CHLA variable has a clustered, left skewed pattern.

In the R code chunk below, do the following: 1) Redo the correlation analysis using the `corr.test()` function in the `psych` package with the following options: method = "pearson", adjust = "BH". 2) Now, redo this correlation analysis using a non-parametric method. 3) Use the print command from the handout to see the results of each correlation analysis.

```
#install.packages("psych", repos = "http://cran.rstudio.com/")
library(psych)

cor2 <- corr.test(zoop.num, method = "pearson", adjust = "BH")
print(cor2, digits =3)
```

```
## Call:corr.test(x = zoop.num, method = "pearson", adjust = "BH")
## Correlation matrix
##          TP     TN    SRP    TIN   CHLA     ZP
## TP    1.000  0.787  0.654  0.717 -0.017  0.697
```

```
## TN    0.787  1.000  0.784  0.969 -0.004  0.756
## SRP   0.654  0.784  1.000  0.801 -0.189  0.676
## TIN   0.717  0.969  0.801  1.000 -0.157  0.761
## CHLA -0.017 -0.004 -0.189 -0.157  1.000 -0.183
## ZP    0.697  0.756  0.676  0.761 -0.183  1.000
## Sample Size
## [1] 24
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         TP    TN   SRP   TIN  CHLA    ZP
## TP   0.000 0.000 0.001 0.000 0.983 0.000
## TN   0.000 0.000 0.000 0.000 0.983 0.000
## SRP  0.001 0.000 0.000 0.000 0.491 0.000
## TIN  0.000 0.000 0.000 0.000 0.536 0.000
## CHLA 0.938 0.983 0.376 0.464 0.000 0.491
## ZP   0.000 0.000 0.000 0.000 0.393 0.000
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
cor2.5 <- corr.test(zoop.num, method = "pearson")
print(cor2.5, digits =3)
```

```
## Call:corr.test(x = zoop.num, method = "pearson")
## Correlation matrix
##          TP     TN    SRP    TIN   CHLA     ZP
## TP    1.000  0.787  0.654  0.717 -0.017  0.697
## TN    0.787  1.000  0.784  0.969 -0.004  0.756
## SRP   0.654  0.784  1.000  0.801 -0.189  0.676
## TIN   0.717  0.969  0.801  1.000 -0.157  0.761
## CHLA -0.017 -0.004 -0.189 -0.157  1.000 -0.183
## ZP    0.697  0.756  0.676  0.761 -0.183  1.000
## Sample Size
## [1] 24
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         TP    TN   SRP   TIN  CHLA    ZP
## TP   0.000 0.000 0.003 0.001 1.000 0.001
## TN   0.000 0.000 0.000 0.000 1.000 0.000
## SRP  0.001 0.000 0.000 0.000 1.000 0.002
## TIN  0.000 0.000 0.000 0.000 1.000 0.000
## CHLA 0.938 0.983 0.376 0.464 0.000 1.000
## ZP   0.000 0.000 0.000 0.000 0.393 0.000
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
cor3 <- corr.test(zoop.num, method = "kendall", adjust = "BH")
print(cor3, digits =3)
```

```
## Call:corr.test(x = zoop.num, method = "kendall", adjust = "BH")
## Correlation matrix
##         TP    TN    SRP   TIN   CHLA     ZP
## TP   1.000 0.739  0.391 0.577  0.044  0.536
## TN   0.739 1.000  0.478 0.809  0.015  0.551
## SRP  0.391 0.478  1.000 0.563 -0.066  0.449
## TIN  0.577 0.809  0.563 1.000  0.044  0.548
## CHLA 0.044 0.015 -0.066 0.044  1.000 -0.051
## ZP   0.536 0.551  0.449 0.548 -0.051  1.000
```

```
## Sample Size
## [1] 24
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         TP    TN   SRP   TIN  CHLA    ZP
## TP   0.000 0.000 0.088 0.014 0.899 0.015
## TN   0.000 0.000 0.034 0.000 0.946 0.014
## SRP  0.059 0.018 0.000 0.014 0.899 0.046
## TIN  0.003 0.000 0.004 0.000 0.899 0.014
## CHLA 0.839 0.946 0.760 0.839 0.000 0.899
## ZP   0.007 0.005 0.028 0.006 0.813 0.000
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Question 4: Describe what you learned from `corr.test`. Specifically, are the results sensitive to whether you use parametric (i.e., Pearson's) or non-parametric methods? When should one use non-parametric methods instead of parametric methods? With the Pearson's method, is there evidence for false discovery rate due to multiple comparisons? Why is false discovery rate important?

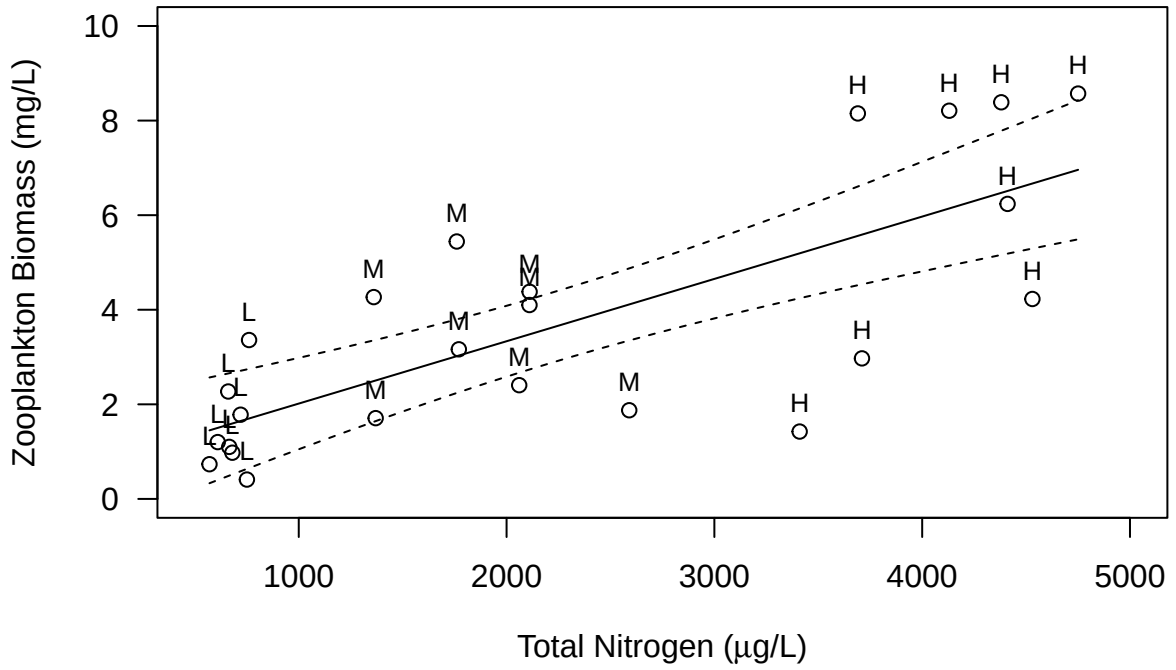
> Answer 4: The results are extremely sensitive to whether you utilize parameters. Traditionally, parametric methods are used for normally distributed data while non parametric are used for continuous data. The Benjamini and Hochberg corrected p-values reduced the potential of type-1 error (false discovery rate) when making multiple comparisons. With the Pearson method, there were multiple correlations being made which increases the probability of false discovery. Type I error is especially problematic as it is a false positive of signficance, skewing the results. ### Linear Regression

In the R code chunk below, do the following: 1) Conduct a linear regression analysis to test the relationship between total nitrogen (TN) and zooplankton biomass (ZP). 2) Examine the output of the regression analysis. 3) Produce a plot of this regression analysis including the following: categorically labeled points, the predicted regression line with 95% confidence intervals, and the appropriate axis labels.

```
#install.packages("ggplot2")
#install.packages ("corrplot")
#install.package.,    s("tidyverse")
meso <-read.table("/cloud/project/QB2025_ALennon/Week1-RStudio/data/zoop_nuts.txt", sep = "\t", header =
fitreg <- lm(ZP ~ TN, data = meso)
summary(fitreg)
```

```
##
## Call:
## lm(formula = ZP ~ TN, data = meso)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7690 -0.8491 -0.0709  1.6238  2.5888
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6977712  0.6496312   1.074    0.294
## TN          0.0013181  0.0002431   5.421 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.75 on 22 degrees of freedom
## Multiple R-squared:  0.5719, Adjusted R-squared:  0.5525
## F-statistic: 29.39 on 1 and 22 DF,  p-value: 1.911e-05
```

```
plot(meso$TN, meso$ZP, ylim = c(0, 10), xlim = c(500, 5000),
     xlab = expression(paste("Total Nitrogen (", mu,"g/L)")),
     ylab = "Zooplankton Biomass (mg/L)", las = 1)
text(meso$TN, meso$ZP, meso$NUTS, pos = 3, cex = 0.8)
newTN <- seq(min(meso$TN), max(meso$TN), 10)
regline <- predict(fitreg, newdata = data.frame(TN = newTN))
lines(newTN, regline)

conf95 <- predict(fitreg, newdata = data.frame(TN = newTN),
                  interval = c("confidence"), level = 0.95, type = "response")
matlines(newTN, conf95[, c("lwr", "upr")], type="l", lty = 2, lwd = 1, col = "black")
```



Question 5: Interpret the results from the regression model

> Answer 5: Based on the regression model, there appears to be a positive relationship between total nitrogen and zooplankton biomass. Additioanlly, the distribution of low, medium, and high are relatively evenly distributed with low being the most clustered around under 1000. One particular note, there are two "high" values that have a lower biomass compared to the other high values around 3000-4000 ug/L. It would be interesting to determine if these values are outliers.

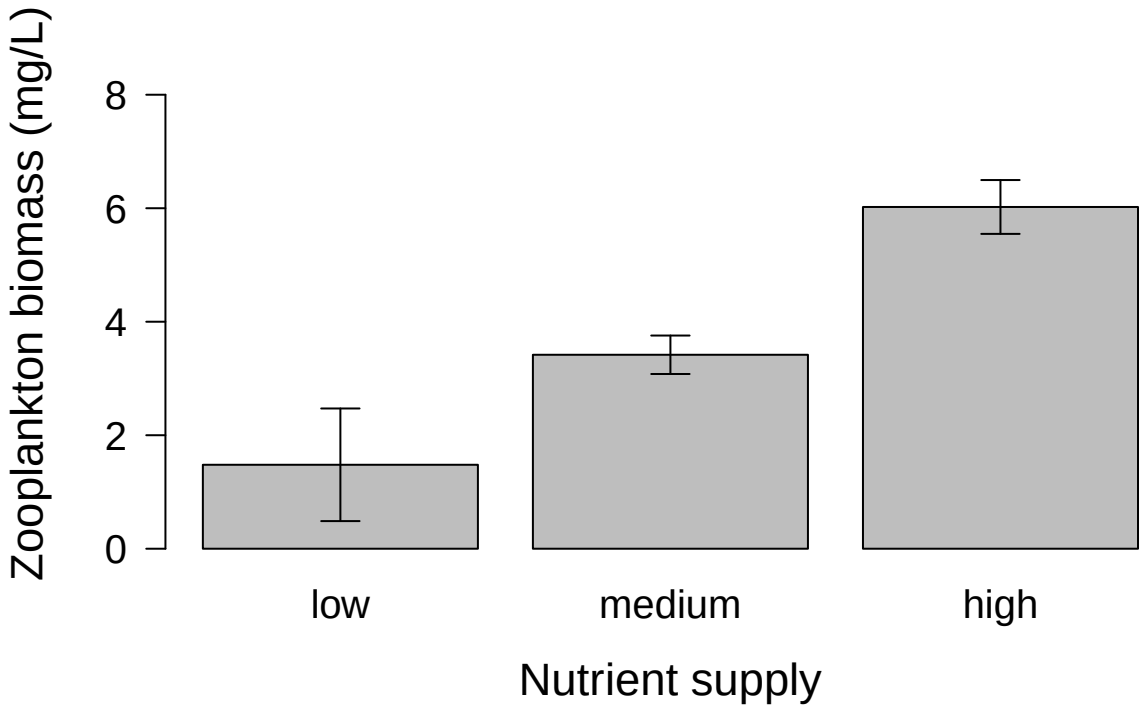
Analysis of Variance (ANOVA)

Using the R code chunk below, do the following: 1) Order the nutrient treatments from low to high (see handout). 2) Produce a barplot to visualize zooplankton biomass in each nutrient treatment. 3) Include error bars (+/- 1 sem) on your plot and label the axes appropriately. 4) Use a one-way analysis of variance (ANOVA) to test the null hypothesis that zooplankton biomass is affected by the nutrient treatment.

```
nuts <- zoop$NUTS
ZP <-zoop$ZP
nuts.order <- factor(nuts, levels = c("L", "M", "H"))
print(nuts.order)
```

```
##  [1] L L L L L L L M M M M M M M M H H H H H H H H
## Levels: L M H
```

```r
# Standard Errors
zp.means <- tapply(zoop$ZP, nuts.order, mean)
sem <- function(x) { sd(na.omit(x)) / sqrt(length(na.omit(x))) }
zp.sem <- tapply(ZP, nuts, sem)

# Plot
bp <- barplot(zp.means, ylim = c(0, round(max(ZP), digits = 0)), pch = 15, cex = 1.25, las = 1,
              cex.lab = 1.4, cex.axis = 1.25,
              xlab = "Nutrient supply", ylab = "Zooplankton biomass (mg/L)",
              names.arg = c("low", "medium", "high"))
#error bars
arrows(x0 = bp, y0 =zp.means, y1 = zp.means - zp.sem, angle =90, length = .1, lwd =1)
arrows(x0 = bp, y0 =zp.means, y1 = zp.means + zp.sem, angle =90, length = .1, lwd =1)
```




```r
#ANOVA
nitroAOV <- aov(ZP ~ nuts, data = meso)
summary(nitroAOV)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## nuts         2  83.15   41.58   11.77 0.000372 ***
## Residuals   21  74.16    3.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(nitroAOV)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ZP ~ nuts, data = meso)
##
## $nuts
```

```
##          diff        lwr        upr     p adj
## L-H -4.543175 -6.9115094 -2.1748406 0.0002512
## M-H -2.604550 -4.9728844 -0.2362156 0.0294932
## M-L  1.938625 -0.4297094  4.3069594 0.1220246
```

SYNTHESIS: SITE-BY-SPECIES MATRIX

In the R code chunk below, load the zoops.txt data set in your **3.RStudio** data folder. Create a site-by-species matrix (or dataframe) that does *not* include TANK or NUTS. The remaining columns of data refer to the biomass (µg/L) of different zooplankton taxa:

- CAL = calanoid copepods
- DIAP = *Diaphanasoma* sp.
- CYL = cyclopoid copepods
- BOSM = *Bosmina* sp.
- SIMO = *Simocephallus* sp.
- CERI = *Ceriodaphnia* sp.
- NAUP = naupuli (immature copepod)
- DLUM = *Daphnia lumholtzi*
- CHYD = *Chydorus* sp.

Question 6: With the visualization and statistical tools that we learned about in the **3. RStudio** handout, use the site-by-species matrix to assess whether and how different zooplankton taxa were responsible for the total biomass (ZP) response to nutrient enrichment. Describe what you learned below in the "Answer" section and include appropriate code in the R chunk.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
#install.packages("corrplot")
#read in zoops.txt as a data frame
zoops <- as.data.frame(read.table("/cloud/project/QB2025_ALennon/Week1-RStudio/data/zoops.txt", sep = "
print(zoops)
```

```
##    TANK NUTS   CAL  DIAP  CYCL BOSM   SIMO  CERI NAUP DLUM   CHYD
## 1     5    L  70.5   0.0  66.1  2.2  417.8 159.8  0.0  0.0  266.9
## 2    14    L  27.1  19.2 129.6  0.0    0.0  79.4  0.0  0.0  158.7
## 3    16    L   5.3   8.8  12.7  0.0   73.1 107.5  1.2  0.0 3158.2
## 4    21    L  79.2  17.9 141.3  3.4    0.0 199.0  0.0  0.0  298.5
## 5    23    L  31.4   0.0  11.0  0.0  482.0 101.9  0.0  0.0  580.2
## 6    25    L  22.7 285.1 153.0  0.0  241.5 135.5  1.2  6.6  262.4
## 7    27    L   0.0   2.3  11.0  0.0   73.1 185.0  1.6  0.0 2004.4
## 8    34    L  35.7  65.9 102.9  0.0    0.0 318.5  3.1  0.0 1260.7
## 9    12    M  74.8 178.7 266.5  0.0    0.0   1.9  0.0  0.0 1190.9
```

```
## 10   15    M   5.3   4.9  87.8  0.0 1099.2 136.4  1.4  0.0 2939.6
## 11   18    M  18.4   2.3  29.4  0.0  393.8 147.6  1.2  0.0 4857.3
## 12   22    M  14.0   2.3  37.7  0.0 1251.5  74.8  0.0  0.0 2725.5
## 13   28    M  14.0   2.3 132.9  0.0  818.6  98.1  1.2  0.0  814.5
## 14   30    M  48.8   2.3 107.9  2.2    9.0 132.7  0.0  0.0 2867.5
## 15   35    M   0.0   0.0  17.7  0.0  145.3  19.7  0.0  0.0 4201.6
## 16   36    M 292.0 269.5 373.4 10.7    0.0   8.5  1.2  0.0 1456.8
## 17    4    H   9.7   0.0  41.1  0.0 2397.8   9.4  0.0  0.0 5697.9
## 18    6    H   0.0   2.3   0.0  0.0  225.5  24.3  0.0  0.0 8323.2
## 19   10    H   5.3   0.0  86.2  0.0  465.9 527.7  1.2  0.0 3146.9
## 20   11    H  14.0   7.5  69.5  0.0  594.2  78.5  0.0  0.0 7629.2
## 21   17    H   0.0  24.4 101.2  0.0  313.6 176.6  0.0  0.0 7597.6
## 22   19    H   0.0   7.5 253.2  8.3    0.0 112.1  1.6  0.0 2594.8
## 23   24    H   5.3   2.3  96.2  0.0  786.6  76.6  0.0  0.0  463.0
## 24   29    H   0.0   2.3  66.1  0.0  826.7  85.1  0.0  0.0 5263.0
```

```
zoops.df.clean <- subset(zoops, select = -c(TANK, NUTS))

# Site column
zoops.df.clean$Site[1:24] <- 1:24
zoops.df.clean <- zoops.df.clean[, c("Site", setdiff(names(zoops.df.clean), "Site"))]

print(zoops.df.clean)
```

```
##    Site   CAL  DIAP  CYCL BOSM   SIMO  CERI NAUP DLUM   CHYD
## 1     1  70.5   0.0  66.1  2.2  417.8 159.8  0.0  0.0  266.9
## 2     2  27.1  19.2 129.6  0.0    0.0  79.4  0.0  0.0  158.7
## 3     3   5.3   8.8  12.7  0.0   73.1 107.5  1.2  0.0 3158.2
## 4     4  79.2  17.9 141.3  3.4    0.0 199.0  0.0  0.0  298.5
## 5     5  31.4   0.0  11.0  0.0  482.0 101.9  0.0  0.0  580.2
## 6     6  22.7 285.1 153.0  0.0  241.5 135.5  1.2  6.6  262.4
## 7     7   0.0   2.3  11.0  0.0   73.1 185.0  1.6  0.0 2004.4
## 8     8  35.7  65.9 102.9  0.0    0.0 318.5  3.1  0.0 1260.7
## 9     9  74.8 178.7 266.5  0.0    0.0   1.9  0.0  0.0 1190.9
## 10   10   5.3   4.9  87.8  0.0 1099.2 136.4  1.4  0.0 2939.6
## 11   11  18.4   2.3  29.4  0.0  393.8 147.6  1.2  0.0 4857.3
## 12   12  14.0   2.3  37.7  0.0 1251.5  74.8  0.0  0.0 2725.5
## 13   13  14.0   2.3 132.9  0.0  818.6  98.1  1.2  0.0  814.5
## 14   14  48.8   2.3 107.9  2.2    9.0 132.7  0.0  0.0 2867.5
## 15   15   0.0   0.0  17.7  0.0  145.3  19.7  0.0  0.0 4201.6
## 16   16 292.0 269.5 373.4 10.7    0.0   8.5  1.2  0.0 1456.8
## 17   17   9.7   0.0  41.1  0.0 2397.8   9.4  0.0  0.0 5697.9
## 18   18   0.0   2.3   0.0  0.0  225.5  24.3  0.0  0.0 8323.2
## 19   19   5.3   0.0  86.2  0.0  465.9 527.7  1.2  0.0 3146.9
## 20   20  14.0   7.5  69.5  0.0  594.2  78.5  0.0  0.0 7629.2
## 21   21   0.0  24.4 101.2  0.0  313.6 176.6  0.0  0.0 7597.6
## 22   22   0.0   7.5 253.2  8.3    0.0 112.1  1.6  0.0 2594.8
## 23   23   5.3   2.3  96.2  0.0  786.6  76.6  0.0  0.0  463.0
## 24   24   0.0   2.3  66.1  0.0  826.7  85.1  0.0  0.0 5263.0
```

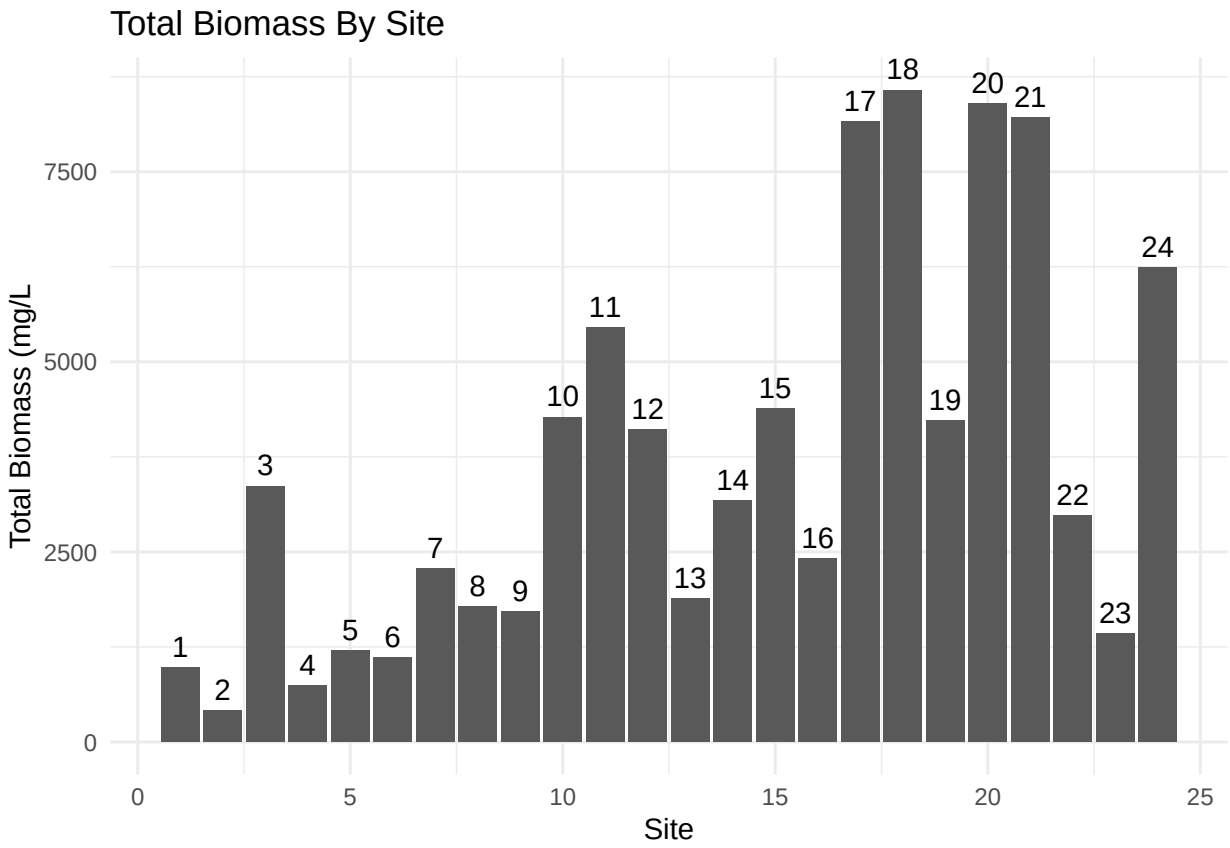
```
#total Biomass column
zoops.df.clean$Total.Biomass <- rowSums(zoops.df.clean[, -1])

# Print the updated dataframe
print(zoops.df.clean)
```

```
##    Site   CAL  DIAP  CYCL BOSM   SIMO  CERI NAUP DLUM   CHYD Total.Biomass
## 1     1  70.5   0.0  66.1  2.2  417.8 159.8  0.0  0.0  266.9         983.3
## 2     2  27.1  19.2 129.6  0.0    0.0  79.4  0.0  0.0  158.7         414.0
## 3     3   5.3   8.8  12.7  0.0   73.1 107.5  1.2  0.0 3158.2        3366.8
## 4     4  79.2  17.9 141.3  3.4    0.0 199.0  0.0  0.0  298.5         739.3
## 5     5  31.4   0.0  11.0  0.0  482.0 101.9  0.0  0.0  580.2        1206.5
## 6     6  22.7 285.1 153.0  0.0  241.5 135.5  1.2  6.6  262.4        1108.0
## 7     7   0.0   2.3  11.0  0.0   73.1 185.0  1.6  0.0 2004.4        2277.4
## 8     8  35.7  65.9 102.9  0.0    0.0 318.5  3.1  0.0 1260.7        1786.8
## 9     9  74.8 178.7 266.5  0.0    0.0   1.9  0.0  0.0 1190.9        1712.8
## 10   10   5.3   4.9  87.8  0.0 1099.2 136.4  1.4  0.0 2939.6        4274.6
## 11   11  18.4   2.3  29.4  0.0  393.8 147.6  1.2  0.0 4857.3        5450.0
## 12   12  14.0   2.3  37.7  0.0 1251.5  74.8  0.0  0.0 2725.5        4105.8
## 13   13  14.0   2.3 132.9  0.0  818.6  98.1  1.2  0.0  814.5        1881.6
## 14   14  48.8   2.3 107.9  2.2    9.0 132.7  0.0  0.0 2867.5        3170.4
## 15   15   0.0   0.0  17.7  0.0  145.3  19.7  0.0  0.0 4201.6        4384.3
## 16   16 292.0 269.5 373.4 10.7    0.0   8.5  1.2  0.0 1456.8        2412.1
## 17   17   9.7   0.0  41.1  0.0 2397.8   9.4  0.0  0.0 5697.9        8155.9
## 18   18   0.0   2.3   0.0  0.0  225.5  24.3  0.0  0.0 8323.2        8575.3
## 19   19   5.3   0.0  86.2  0.0  465.9 527.7  1.2  0.0 3146.9        4233.2
## 20   20  14.0   7.5  69.5  0.0  594.2  78.5  0.0  0.0 7629.2        8392.9
## 21   21   0.0  24.4 101.2  0.0  313.6 176.6  0.0  0.0 7597.6        8213.4
## 22   22   0.0   7.5 253.2  8.3    0.0 112.1  1.6  0.0 2594.8        2977.5
## 23   23   5.3   2.3  96.2  0.0  786.6  76.6  0.0  0.0  463.0        1430.0
## 24   24   0.0   2.3  66.1  0.0  826.7  85.1  0.0  0.0 5263.0        6243.2
```

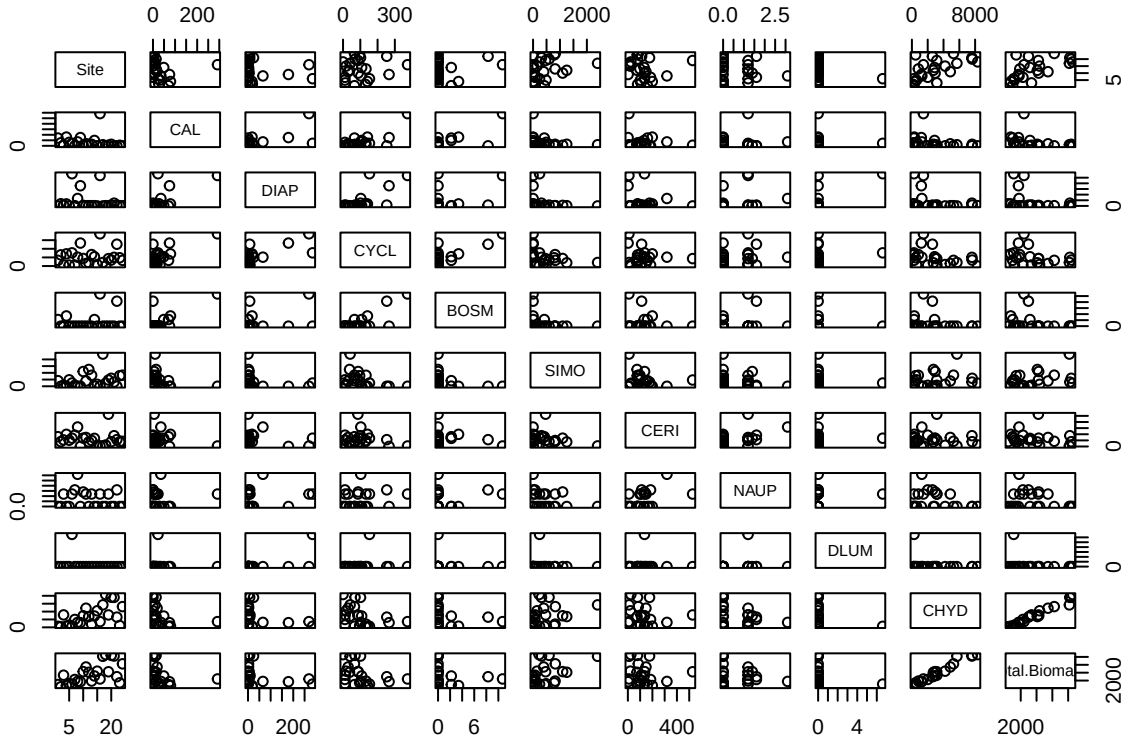
```
str(zoops.df.clean)
```

```
## 'data.frame':    24 obs. of  11 variables:
##  $ Site         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CAL          : num  70.5 27.1 5.3 79.2 31.4 22.7 0 35.7 74.8 5.3 ...
##  $ DIAP         : num  0 19.2 8.8 17.9 0 ...
##  $ CYCL         : num  66.1 129.6 12.7 141.3 11 ...
##  $ BOSM         : num  2.2 0 0 3.4 0 0 0 0 0 0 ...
##  $ SIMO         : num  417.8 0 73.1 0 482 ...
##  $ CERI         : num  159.8 79.4 107.5 199 101.9 ...
##  $ NAUP         : num  0 0 1.2 0 0 1.2 1.6 3.1 0 1.4 ...
##  $ DLUM         : num  0 0 0 0 0 6.6 0 0 0 0 ...
##  $ CHYD         : num  267 159 3158 298 580 ...
##  $ Total.Biomass: num  983 414 3367 739 1206 ...
```

```
#Graphical Visualization
zoop.bp <-ggplot(zoops.df.clean, aes (x= Site, y = Total.Biomass)) +geom_bar(stat = "identity") + labs
print(zoop.bp)
```




```
pairs(zoops.df.clean)
```



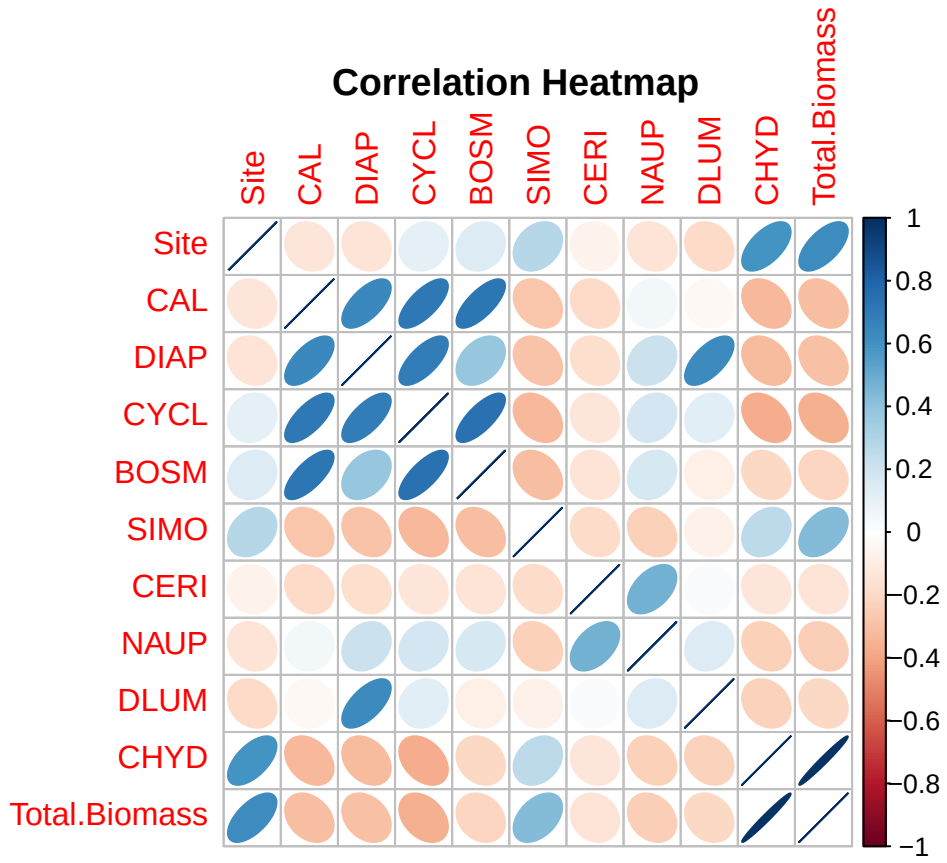

```
#Pearson's correlation
zoop.cor <- corr.test(zoops.df.clean, method = "pearson", adjust = "BH")
```

```r
print(zoop.cor, digits = 3)
```

```
## Call:corr.test(x = zoops.df.clean, method = "pearson", adjust = "BH")
## Correlation matrix
##                  Site    CAL   DIAP   CYCL   BOSM   SIMO   CERI   NAUP   DLUM
## Site           1.000 -0.135 -0.141  0.110  0.146  0.283 -0.067 -0.143 -0.196
## CAL           -0.135  1.000  0.643  0.712  0.728 -0.271 -0.191  0.058 -0.034
## DIAP          -0.141  0.643  1.000  0.694  0.381 -0.287 -0.172  0.217  0.637
## CYCL           0.110  0.712  0.694  1.000  0.747 -0.325 -0.132  0.186  0.125
## BOSM           0.146  0.728  0.381  0.747  1.000 -0.308 -0.141  0.179 -0.086
## SIMO           0.283 -0.271 -0.287 -0.325 -0.308  1.000 -0.183 -0.237 -0.077
## CERI          -0.067 -0.191 -0.172 -0.132 -0.141 -0.183  1.000  0.475  0.020
## NAUP          -0.143  0.058  0.217  0.186  0.179 -0.237  0.475  1.000  0.148
## DLUM          -0.196 -0.034  0.637  0.125 -0.086 -0.077  0.020  0.148  1.000
## CHYD           0.596 -0.322 -0.314 -0.369 -0.206  0.262 -0.135 -0.238 -0.224
## Total.Biomass  0.627 -0.307 -0.299 -0.355 -0.214  0.431 -0.141 -0.244 -0.207
##                 CHYD Total.Biomass
## Site           0.596         0.627
## CAL           -0.322        -0.307
## DIAP          -0.314        -0.299
## CYCL          -0.369        -0.355
## BOSM          -0.206        -0.214
## SIMO           0.262         0.431
## CERI          -0.135        -0.141
## NAUP          -0.238        -0.244
## DLUM          -0.224        -0.207
## CHYD           1.000         0.981
## Total.Biomass  0.981         1.000
## Sample Size
## [1] 24
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                  Site   CAL  DIAP  CYCL  BOSM  SIMO  CERI  NAUP  DLUM  CHYD
## Site          0.000 0.630 0.630 0.683 0.630 0.451 0.801 0.630 0.599 0.013
## CAL           0.528 0.000 0.006 0.001 0.001 0.477 0.599 0.819 0.893 0.420
## DIAP          0.511 0.001 0.000 0.002 0.303 0.451 0.609 0.573 0.006 0.420
## CYCL          0.608 0.000 0.000 0.000 0.001 0.420 0.630 0.599 0.642 0.323
## BOSM          0.496 0.000 0.066 0.000 0.000 0.420 0.630 0.599 0.757 0.573
## SIMO          0.181 0.199 0.175 0.122 0.143 0.000 0.599 0.540 0.779 0.494
## CERI          0.757 0.371 0.421 0.538 0.510 0.393 0.000 0.105 0.925 0.630
## NAUP          0.505 0.789 0.309 0.385 0.403 0.265 0.019 0.000 0.630 0.540
## DLUM          0.359 0.876 0.001 0.560 0.688 0.722 0.925 0.491 0.000 0.573
## CHYD          0.002 0.125 0.136 0.076 0.334 0.216 0.528 0.263 0.293 0.000
## Total.Biomass 0.001 0.145 0.156 0.088 0.315 0.036 0.512 0.251 0.332 0.000
##               Total.Biomass
## Site                  0.007
## CAL                   0.420
## DIAP                  0.430
## CYCL                  0.347
## BOSM                  0.573
## SIMO                  0.178
## CERI                  0.630
## NAUP                  0.540
## DLUM                  0.573
## CHYD                  0.000
```

```
## Total.Biomass         0.000
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

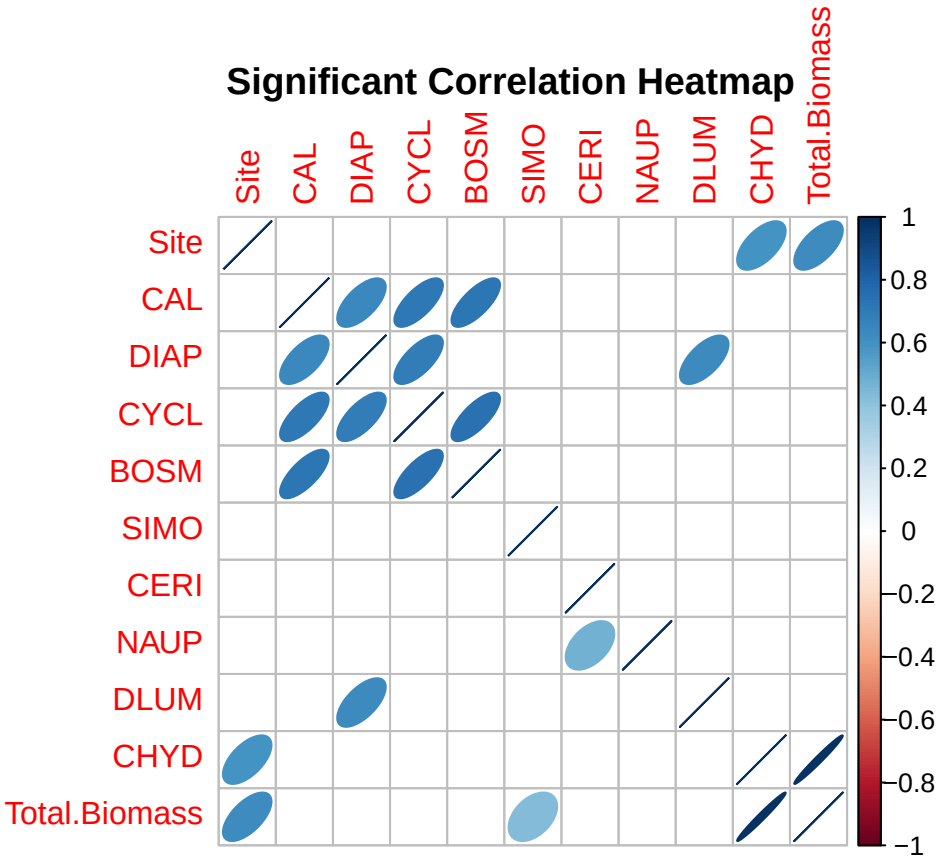
```
require("corrplot")
corrplot(zoop.cor$r, method = "ellipse")
title("Correlation Heatmap")
```



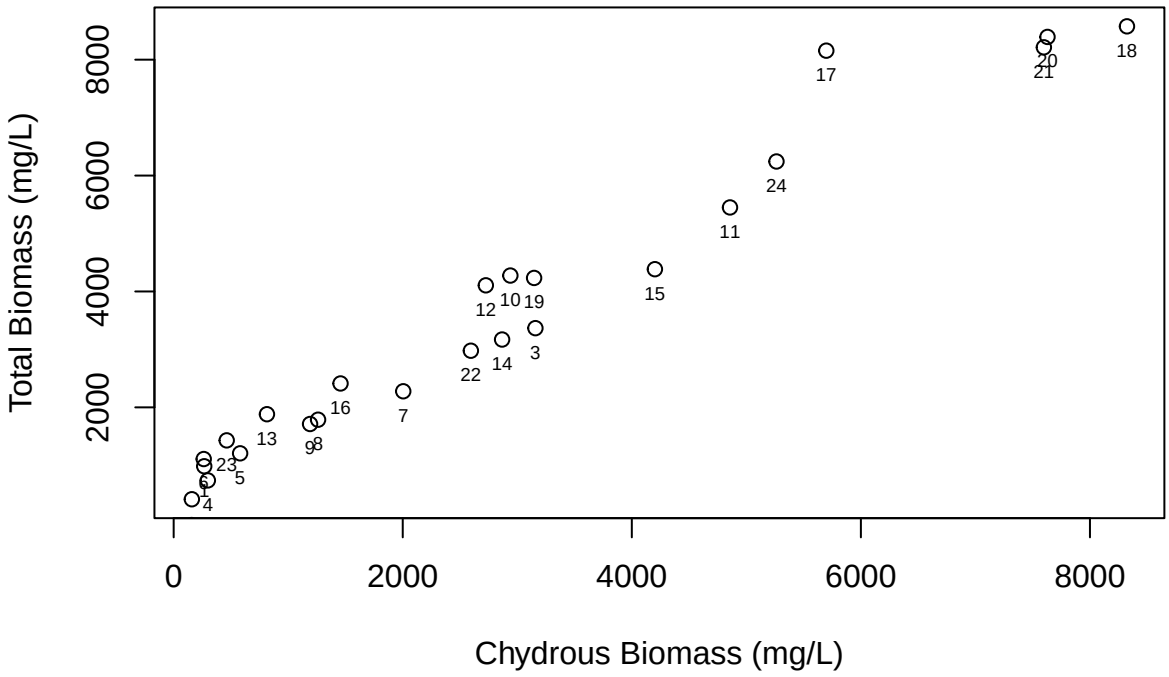

```
require(psych)

cor_matrix <- zoop.cor$r
p_values <- zoop.cor$p
significance_level <- 0.05


cor_df <- as.data.frame(as.table(cor_matrix))
p_df <- as.data.frame(as.table(p_values))

significant_results <- cbind(cor_df, p_value = p_df$Freq)

colnames(significant_results) <- c("Variable1", "Variable2", "Correlation", "P_Value")

# Filter to keep only significant values
significant_results_df <- significant_results[significant_results$P_Value < significance_level, ]

# View the significant correlations
print(significant_results_df)
```

```
##          Variable1     Variable2 Correlation      P_Value
## 1             Site          Site   1.0000000 0.000000e+00
## 10            CHYD          Site   0.5963341 2.101254e-03
## 11  Total.Biomass          Site   0.6270983 1.039662e-03
## 13             CAL           CAL   1.0000000 0.000000e+00
## 14            DIAP           CAL   0.6425998 7.087491e-04
## 15            CYCL           CAL   0.7118997 9.558847e-05
## 16            BOSM           CAL   0.7280986 5.501968e-05
## 24             CAL          DIAP   0.6425998 6.459572e-03
## 25            DIAP          DIAP   1.0000000 0.000000e+00
## 26            CYCL          DIAP   0.6943602 1.670415e-04
## 31            DLUM          DIAP   0.6366939 8.221273e-04
## 35             CAL          CYCL   0.7118997 1.314342e-03
## 36            DIAP          CYCL   0.6943602 1.837457e-03
## 37            CYCL          CYCL   1.0000000 0.000000e+00
## 38            BOSM          CYCL   0.7466915 2.778257e-05
## 46             CAL          BOSM   0.7280986 1.008694e-03
## 48            CYCL          BOSM   0.7466915 7.640207e-04
## 49            BOSM          BOSM   1.0000000 0.000000e+00
## 61            SIMO          SIMO   1.0000000 0.000000e+00
## 66  Total.Biomass          SIMO   0.4309401 3.552289e-02
## 73            CERI          CERI   1.0000000 0.000000e+00
## 74            NAUP          CERI   0.4745400 1.912999e-02
## 85            NAUP          NAUP   1.0000000 0.000000e+00
## 91            DIAP          DLUM   0.6366939 6.459572e-03
## 97            DLUM          DLUM   1.0000000 0.000000e+00
## 100           Site          CHYD   0.5963341 1.284100e-02
## 109           CHYD          CHYD   1.0000000 0.000000e+00
## 110 Total.Biomass          CHYD   0.9811118 3.447275e-17
## 111           Site Total.Biomass   0.6270983 7.147677e-03
## 120           CHYD Total.Biomass   0.9811118 1.896001e-15
## 121 Total.Biomass Total.Biomass   1.0000000 0.000000e+00
```

```r
#Signficant values Graphical Representation
significant_cor_matrix <- cor_matrix
significant_cor_matrix[!(p_values < significance_level)] <- NA
require("corrplot")
corrplot(significant_cor_matrix, method = "ellipse", na.label = " ")
title("Significant Correlation Heatmap")
```




```
#Linear Regression
CHYD <- zoops.df.clean$CHYD
ZP <- zoops.df.clean$Total.Biomass


plot(CHYD, ZP,
     xlab = "Chydrous Biomass (mg/L)",
     ylab = "Total Biomass (mg/L)",
     main = "Chydrous Biomass as a Predictor for Total Biomass")
text(CHYD, ZP, zoops.df.clean$Site, pos = 1, cex =.6)
```



Answer 6: For this dataset, multiple steps were done to determine the impact of taxa on total biomass. A site by species dataframe was created along with a new variable, Total.Biomass, which was the total biomass at each site. A bar plot, Total Biomass By Site, represents the total biomass per site to highlight the overall variation of biomass at different sites. A Pearson's correlation was conducted to determine the signifcant correlations (Correlation Heatmap). From the Pearson correlation, signficant results (p-value $< .05$), were separated (Signficant Correlation Heatmap). All signficant correlations were positive and the only taxa that had a signficant, positive correlation on total biomass was the CHYD (Chydrous) taxa (Chydrous Biomass as a Predictor for Total Biomass).

SUBMITTING YOUR WORKSHEET

Use Knitr to create a PDF of your completed **3.RStudio_Worksheet.Rmd** document, push the repo to GitHub, and create a pull request. Please make sure your updated repo include both the PDF and RMarkdown files.

This assignment is due on **Wednesday, January 22nd, 2025 at 12:00 PM (noon)**.