Exercise 1: Identify any outlying observations in this dataset using Mahalanobis distance*. How can you tell they are outlying observations?

```
X <-
structure(c(0.29, 0.61, 0.3, 0.94, 0.81, 0.88, 0.71, 0.82, 0.88,
0.93, 0, 0.12, 0.64, 0.49, 0.18, 0.28, 0.75, 0.82, 0.84, 0.21,
0.67, 0.45, 0.79, 0.69, 0.47, 0.23, 0.97, 0.2, 0.88, 0.53, 0.86,
0.85, 0.63, 0.43, 0.82, 0.7, 0.33, 0.77, 0.05, 0.8, 0.26, 0.54,
0.55, 0.82, 0.99, 0.78, 0.68, 0.73, 0.57, 0.67, 0.06, 0.36, 0.46,
0.94, 0.58, 0.98, 0.44, 0.12, 0.53, 0.47, 0.6, 0.41, 0.69, 0.24,
0.36, 0.05, 0.82, 0.74, 0.19, 0.77, 0.13, 0.76, 0.31, 0.11, 0.64,
0.36, 0.31, 0.26, 0.35, 0.63, 0.92, 0.06, 0.86, 0.52, 0.21, 0.76,
0.71, 0.5, 0.37, 0.71, 1.324, 1.46, 0.982, 1.472, 1.772, 1.744,
1.082, 1.638, 0.948, 1.9, 0.338, 1.32, 1.214, 1.384, 1.368, 1.198,
1.482, 1.654, 1.566, 0.912, 0.44, 0.936, 1.388, 1.696, 1.2, 1.384,
1.454, 0.412, 1.446, 1.188, 0.94, 0.51, 0.26, 0.54, 0.11, 0.69,
0.72, 0.16, 0.6, 0.91, 0.8, 0.62, 0.93, 0.29, 0.17, 0.54, 0.52,
0.37, 0.32, 0.06, 0.94, 0.22, 0.2, 0.68, 0.18, 0.2, 0.09, 0.4,
0.9, 0.3), .Dim = c(30L, 5L), .Dimnames = list(NULL, c("x1",
"x2", "x3", "x4", "x5")))
```

Solution:
Recall that the Mahalanobis distance is a measure of distance from an observation to the center of the data, de-correlated by the estimated covariance matrix,

$$D^2 = (\hat{x} - \hat{\mu})^T S^{-1} (\hat{x} - \hat{\mu})$$

We can quickly compute this distance over all observations $\hat{x}$ using the Mahalanobis() function in r. Doing so we get the following,
Code:

```
mahal <- mahalanobis(X, center = colMeans(X), cov = cov(X))
> mahal
 [1]  5.971155  2.355915  2.521097  2.785808  3.162355  4.865521
 [7]  3.498117  3.567076  6.134994  7.074793  9.485738 19.856193
[13]  3.082716  3.620258  4.578934  2.081030  1.004770  1.804360
[19]  1.877088  4.885187 14.401283  4.915442  4.662472  3.313876
[25]  2.394622  4.681316  5.559891  5.742372  3.101340  2.014282


## Plotting chiSquared 95% cuttoff.
hist(mahal, n=10, freq=FALSE, xlab = 'Mahalanobis Distance',
     main = 'Density of Mahalanobis Dist.')

xseq <- seq(0,45,length=300)
lines(xseq, dchisq(xseq, df=ncol(X)), col=2, lwd=2)
abline(v = qchisq(p = 0.95 , df = ncol(X)))

## Pulling the values and indices for outliers
Values <- mahal[mahal > qchisq(p = 0.95 , df = ncol(X))]
```

```
#[1] 19.85619 14.40128

Index <- which(mahal > qchisq(p = 0.95 , df = ncol(X)))
#[1] 12 21

## Plotting pair plot with outliers highlighted
color <- rep(1, nrow(X))
color[Index] <- 2
pairs(X, col=color, pch=19)
```

Figure 1: Visualizing the χ^2 95% cutoff.

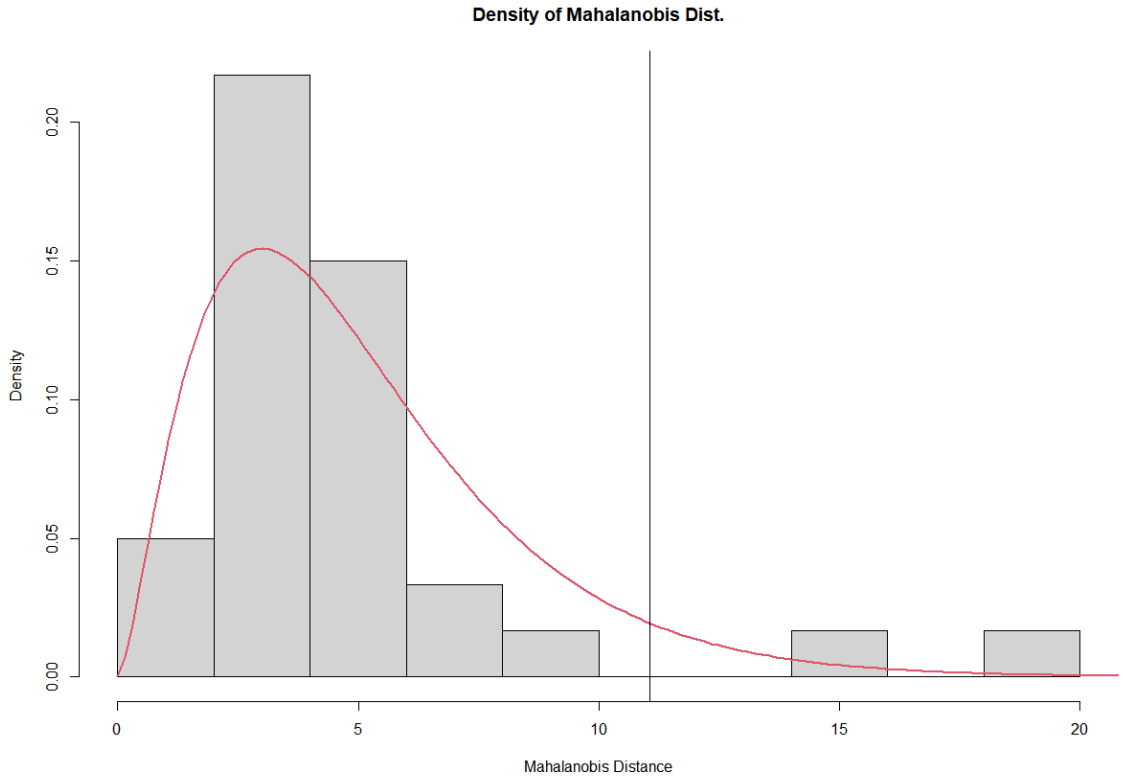
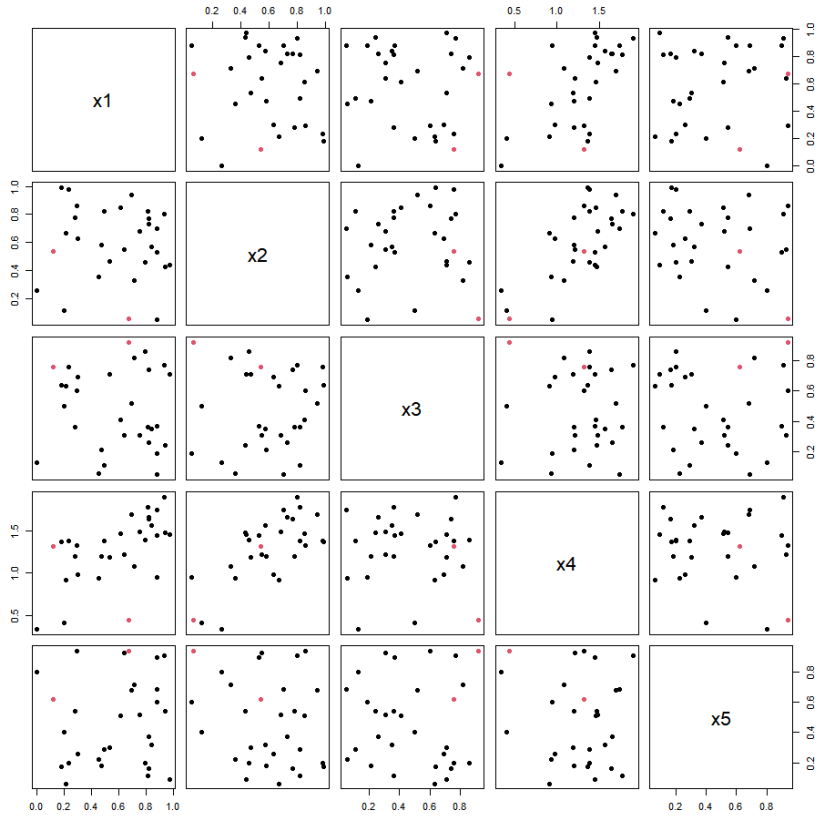


Figure 2: Pair Plot with Outliers Highlighted



Exercise 2: Beginner's approach to ANOSIM:

a Consider the following observations (species counts):

GROUP	sp1	sp2	sp3	sp4	sp5	sp6	sp7
A	0	0	1	6	1	2	0
A	0	4	3	8	3	9	0
B	1	1	1	0	0	0	11
B	8	3	0	0	0	0	0

Compute the Bray-Curtis **dissimilarity** (BC) between each pair of observations and make a 4x4 "distance" matrix out of the values, where the entry in the *ith* row and *jth* column is the Bray-Curtis dissimilarity index between the *ith* and *jth* observations, computed as

$$BC = \frac{\sum_{s=1}^{7} |n_{ik} - n_{jk}|}{\sum n_{ik} + n_{jk}}$$

where n_{ik} is the count for the ith observation and kth species. So, for the first two rows, $BC = (0+4+2+2+2+7+0)/(0+4+4+14+4+11+0) = 0.4594$. This will

give the same answer as the approach described in class to compute the Bray-Curtis dissimilarity index.

Solution:
We can quickly compute the Bray-Curtis dissimilarity index by using the vegdist() function from the vegan package. Doing so we get the following,
Code:

```
data <- c(0, 0, 1, 6, 1, 2, 0,
          0, 4, 3, 8, 3, 9, 0,
          1, 1, 1, 0, 0, 0, 11,
          8, 3, 0, 0, 0, 0, 0)

DataMatrix <- matrix(data, nrow = 4, ncol = 7, byrow = TRUE)
vegdist(DataMatrix, method = 'bray')
     1         2         3
2 0.4594595
3 0.9166667 0.9024390
4 1.0000000 0.8421053 0.8400000
```

b Now replace each "distance" with its rank (shortest distance is rank 1, second shortest is 2, etc.).

Solution:
Doing so we get the following,

	1	2	3
2	1		
3	5	4	
4	6	3	2

c Let rw = average rank within groups A and B. Let rb = average rank between groups A and B
Then the ANOSIM statistic is:

$$R = (rb - rw)/(N * (N-1)/4).$$

ANOSIM is a very robust relative of MANOVA, where we will check to see if two groups are distinct, which we will determine by comparing R to the null distribution of R that we get by permuting the observations (mixing up groups A and B).

Solution:
Computing rw we get the following,

$$rw = \frac{1+2}{2} = \frac{3}{2}.$$

Computing rb we get the following,

$$rb = \frac{5+4+6+3}{4} = \frac{9}{2}.$$

Computing the ANOSIM statistic with $N = 4$ for the number of samples.

$$R = \frac{\frac{9}{2} - \frac{3}{2}}{\frac{4*3}{4}} = 1$$

Interpreting this statistic we would say that there is more similarity between observations inside groups than observations outside of the groups.

d BONUS: Perform a permutation test of Ho: groups are same vs Ha: groups are different, using the ANOSIM statistic.

Solution:
There are only 2 more permutations possible, which produce a different ANOSIM statistic. The first being one where we swap observation 2 and 3, and the second where we swap observations 2 and 4. Computing the ANOSIM statistic with observations 2 and 3 swapped we get the following,

$$rw = \frac{5+3}{2} = 4.$$

Computing rb we get the following,

$$rb = \frac{1+6+4+2}{4} = \frac{13}{4}.$$

Computing the ANOSIM statistic with $N = 4$ for the number of samples.

$$R = \frac{\frac{13}{4} - 4}{\frac{4*3}{4}} = -.25.$$

Computing the ANOSIM statistic with observations 2 and 4 swapped we get the following,

$$rw = \frac{6+4}{2} = 5.$$

Computing rb we get the following,

$$rb = \frac{1+5+3+2}{4} = \frac{11}{4}.$$

Computing the ANOSIM statistic with $N = 4$ for the number of samples.

$$R = \frac{\frac{11}{4} - 5}{\frac{4*3}{4}} = -.75.$$

Given that we have so little data I'm not sure how we would approximate the null dist. and compute a worthwhile p-value. I would say that after computing the other permutations, they seem to corroborate the idea that groups are different. Since we got a negative test statistic every time we swapped observations between groups, this implies that there is a greater difference among observations inside groups that across groups.

Exercise 3: Similarity measures and dissimilarity measures for presence- absence data:

obs	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8
1	1	1	1	0	0	0	0	0
2	0	0	1	1	1	0	1	0
3	0	0	0	0	1	1	1	1

a Compute the simple matching index between each pair of observations.

Solution:
Recall that the simple matching index between a pair of observations is the following,

$$s(x_1, x_2) = \frac{\text{\# of matching presence + \# of matching absence}}{\text{\# of predictors}}$$

Computing the simple matching index between each pair we get the following,

$$s(x_1, x_2) = \frac{1+2}{8} = \frac{3}{8}.$$

$$s(x_1, x_3) = \frac{0+1}{8} = \frac{1}{8}.$$

$$s(x_2, x_3) = \frac{2+2}{8} = \frac{1}{2}.$$

b Compute the Dice-Sorensen index between each pair of observations.

Solution:
Recall that the Dice-Sorensen index is computed by the following,

$$s(x_1, x_2) = \frac{2 * \text{\# of matching presence}}{2 * \text{\# of matching presence + \# only presence in } x_1 + \text{\# only presence in } x_2}$$

Computing the Dice-Sorensen index for between each pair of observations we get the following,

$$s(x_1, x_2) = \frac{2(1)}{2(1) + 2 + 3} = \frac{2}{7}.$$

$$s(x_1, x_3) = \frac{2(0)}{2(0) + \ldots} = 0.$$

$$s(x_2, x_3) = \frac{2(2)}{2(2) + 2 + 2} = \frac{1}{2}.$$

c Horseshoe effect. Take the Dice-Sorensen (DS) index and compute $1 - DS$ for each pair of observations. This is a dissimilarity index. The three observations seem to follow a steadily changing species composition. Let $1 - DS_{ij}$ be the distance from observation i to observation j. Sketch a plot with three points, where the ith and jth points are roughly separated by distance $1 - DS_{ij}$. Trace a line from the 1st to 2nd point and then on to the 3rd point. Does this form a straight line? The tendency for these plots to be bent is called the horseshoe effect and will become important later in the class.

Solution:
Computing the $1 - DS$ dissimilarity index,

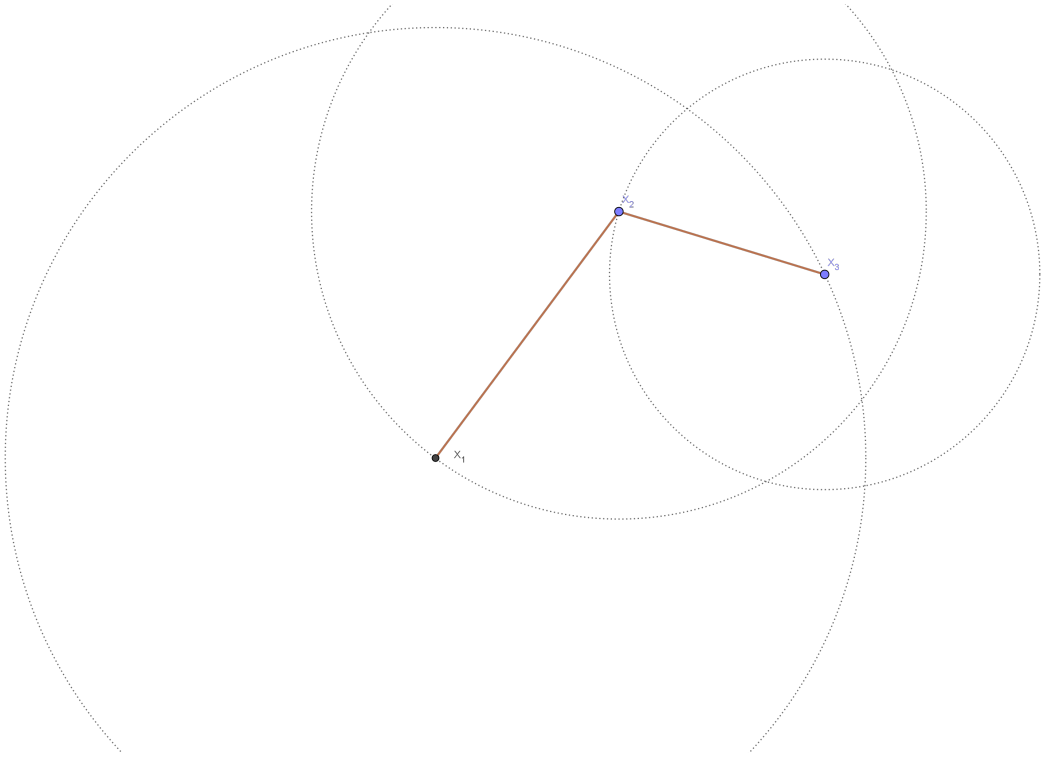
$$1 - s(x_1, x_2) = 1 - \frac{2}{7} = \frac{5}{7}.$$

$$1 - s(x_1, x_3) = 1 - 0 = 1.$$

$$1 - s(x_2, x_3) = 1 - \frac{1}{2} = \frac{1}{2}.$$

I'm a little confused as to how the plot you're describing should be constructed. So I made a geometric construction in geogebra. I started with a circle with radius 1 and center x_1 then put x_3 on that circle. On x_3 I constructed a circle with radius $1/2$ then placed x_2 on that circle. On x_2 I constructed a circle with radius $5/7$ then moved x_2 so that it's circle would intersect x_1. Connecting the points in order we do see some sort of $\cap$ or horseshoe.

Figure 3: $1 - DS$ Horseshoe Plot.



Exercise 4: Consider the following set of observations:

obs	sp1	sp2	sp3	sp4
a	1	0	0	1
b	1	1	1	1
c	0	1	1	0

Find the Dice-Sorensen similarity between each pair, then compute 1 - DC as ”distances”. Show that 1 - DC isn’t really a valid distance, as it does not follow the triangle inequality.

Solution:
Computing the $1 - DC$ dissimilarity index,

$$1 - s(a, b) = 1 - \frac{2(2)}{2(2) + 0 + 2} = \frac{1}{3}.$$

$$1 - s(a, c) = 1 - \frac{2(0)}{2(0) + \ldots} = 1.$$

$$1 - s(b, c) = 1 - \frac{2(2)}{2(2) + 2 + 0} = \frac{1}{3}.$$

Recall the triangle equality. If D is a valid distance measure for the observations a, b, c then it must follow that,

$$D(a, c) \leq D(a, b) + D(b, c).$$

Considering $1 - DC$ as a distance measure we get,

$$1 \leq \frac{1}{3} + \frac{1}{3}$$

Which clearly does not follow the triangle inequality.

Exercise 5: Consider one of the following 'distance' measures: $1-(simplematchingindex)$, $1 - (Dice - Sorensen)$, or $1 - (Jaccard)$. (They are really dissimilarity measures).

a Does the one you chose to test have the property that if the distance between observations is ZERO, then the observations have to have exactly the same species composition?

Solution:
For this problem we will consider the following dissimilarity measure $1-(simplematchingindex)$. Recall that for two observations we it is computed by the following,

$$s(x_1, x_2) = 1 - \frac{\#\text{ of matching presence} + \#\text{ of matching absence}}{\#\text{ of predictors}}$$

To prove that $1-(simplematchingindex)$ has the property that if the distance between observations is ZERO, then the observations have to have exactly the same species composition we will use contradiction. Consider two observations x_1 and x_2 with different species compositions such that $s(x_1, x_2) = 0$. Recall the formula for the $1 - (simplematchingindex)$ dissimilarity measure,

$$s(x_1, x_2) = 1 - \frac{\#\text{ of matching presence} + \#\text{ of matching absence}}{\#\text{ of predictors}} = 0$$

Note that in order for this computation to result in $s(x_1, x_2) = 0$ it follows that,

$$\#\text{ of matching presence} + \#\text{ of matching absence} = \#\text{ of predictors}$$

Clearly this is a contradiction since x_1 and x_2 with different species compositions so there must be at least one predictor where the observations do not match.

b Does the one you chose have the property that if observations are identical (same species composition), then they are distance zero apart?

Solution:
Suppose two observations x_1 and x_2 with the same species compositions, therefore everywhere we have a presence in one observations the other observation must have the same presence, and similarly with absences. Thus we can make the following observation,

$$\#\text{ of matching presence} + \#\text{ of matching absence} = \#\text{ of predictors}$$

As described in the previous section, it follows from here that the dissimilarity index $1 - (simplematchingindex)$ will be zero.