Homework 4

Jasmine Williamson

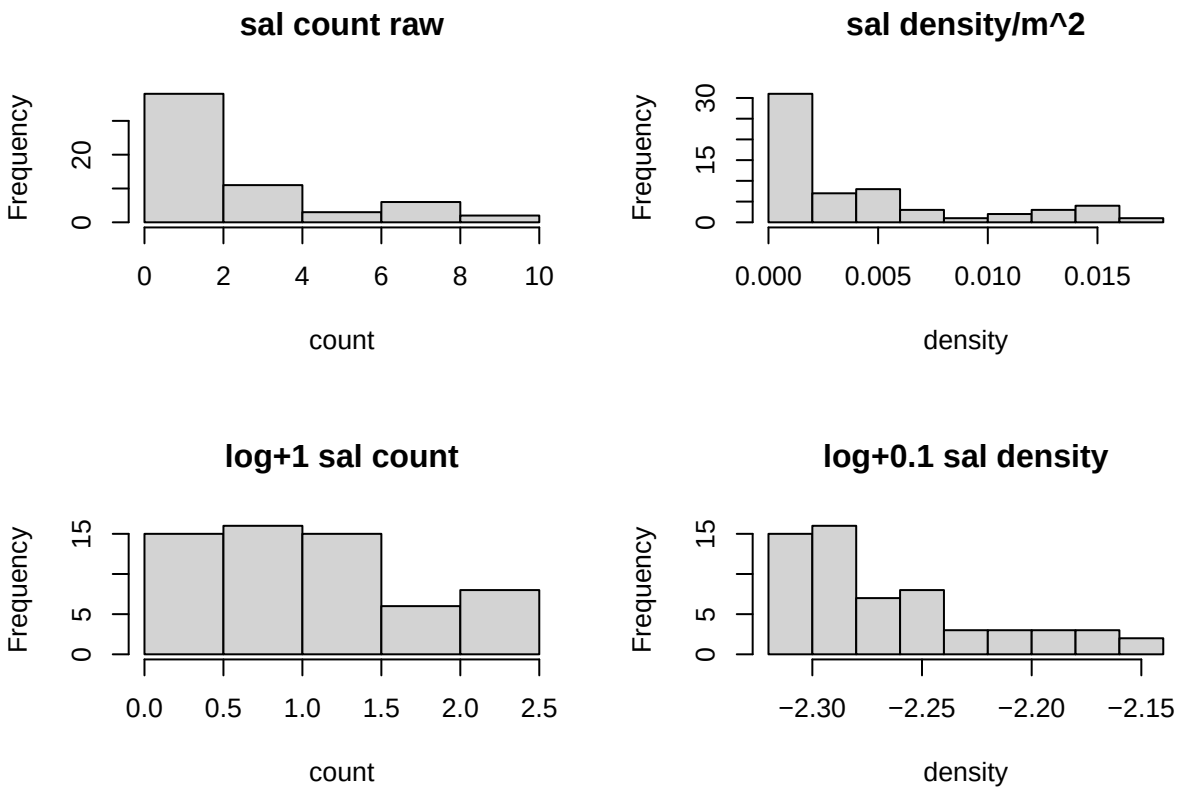
2024-10-21

Standardizing and Transforming the data.

```
#env data
  env_std <- decostand(env_cont, "standardize") #Z-scores the data in each column
  env_subset_std <- decostand(env_subset, "standardize") #Z-scores the data in each column

#standardizing salamanders by sampling area
  sal_dens <- sals
  for(i in 1:nrow(sals)){
      sal_dens[i,] <- sals[i,]/567
    }

#transforming and standardizing the data as needed
  log_sal_cou <- log(sals + 1)
  log_sal_dens <- log(sal_dens + .1)
```



###Question 1) Explain the different linkage methods (single, complete, average, and Ward's) used in agglomerative hierarchical clustering. How does each method affect the shape and size of the resulting clusters?

Single-linkage methods use the minimum distance between points to create clusters. Often results in a chain effect.
Complete-linkage methods use the maximum distance between points to create clusters. Produces more compact sperical clusters.
Intermediate methods use a compromise between single and complete to create clusters, ex: the *average method* uses the average distance between points. Dendrograms show the hierarchy of clusters and can be more intuitive.
Wards methods choose the lowest variance to create clusters. This results in compact clusters with little variance.

###Question 2) Discuss the implications of choosing different association coefficients in agglomerative hierarchical clustering. How do these choices impact the clustering results?

The association matrix influences how clusters are formed. They emphasize different parts of the data, so the one you choose has to match your objectives. Some matrices are sensitive to outliers, data scaling, zeros, etc, and these things can cause issues that will impact the clustering. The shape, size, and makeup of the clusters can be affected pretty substantially with different matrices.

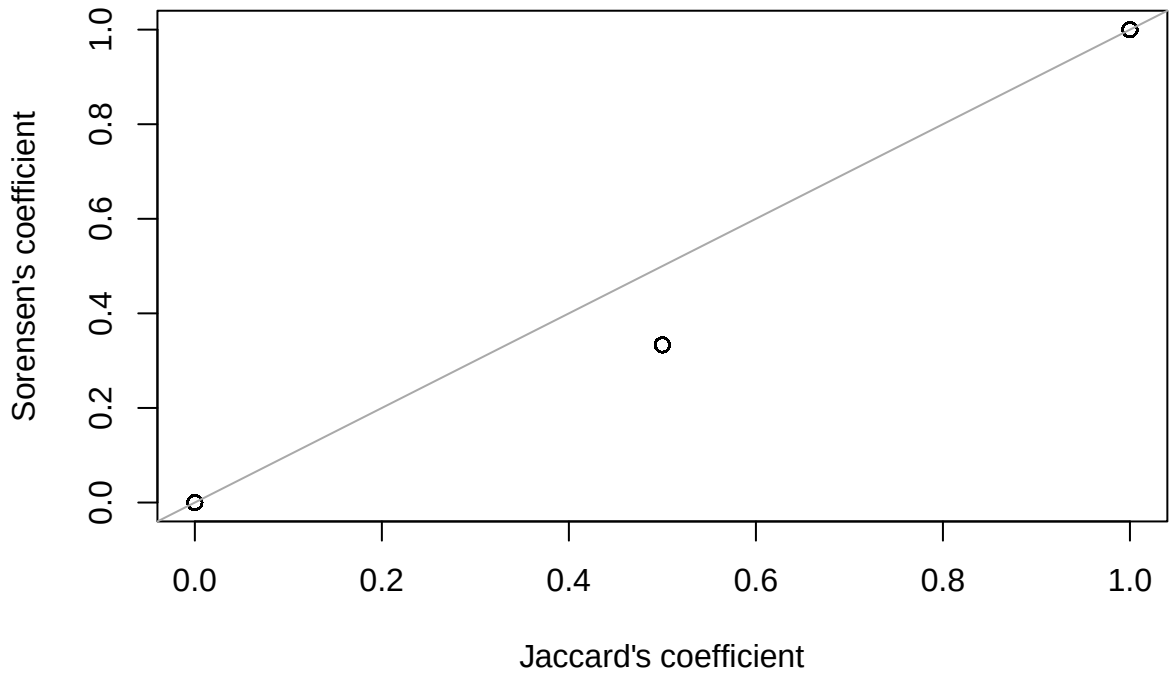
Salamander pres/abs: Jaccard or Sorensen association methods. The plot is weird, I cant figure out why there are only three points. Is this bad? Also, the coldiss functions for these will not work. I get lots of warnings that I cant figure out as well.

```
## Warning in vegdist(sal_occ, method = "jaccard"): you have empty rows: their dissimilarities may be
##                  meaningless in method "jaccard"
```

```
## Warning in vegdist(sal_occ, method = "jaccard"): missing values in results
```

```
## Warning in vegdist(sal_occ, method = "bray"): you have empty rows: their dissimilarities may be
##                   meaningless in method "bray"
```

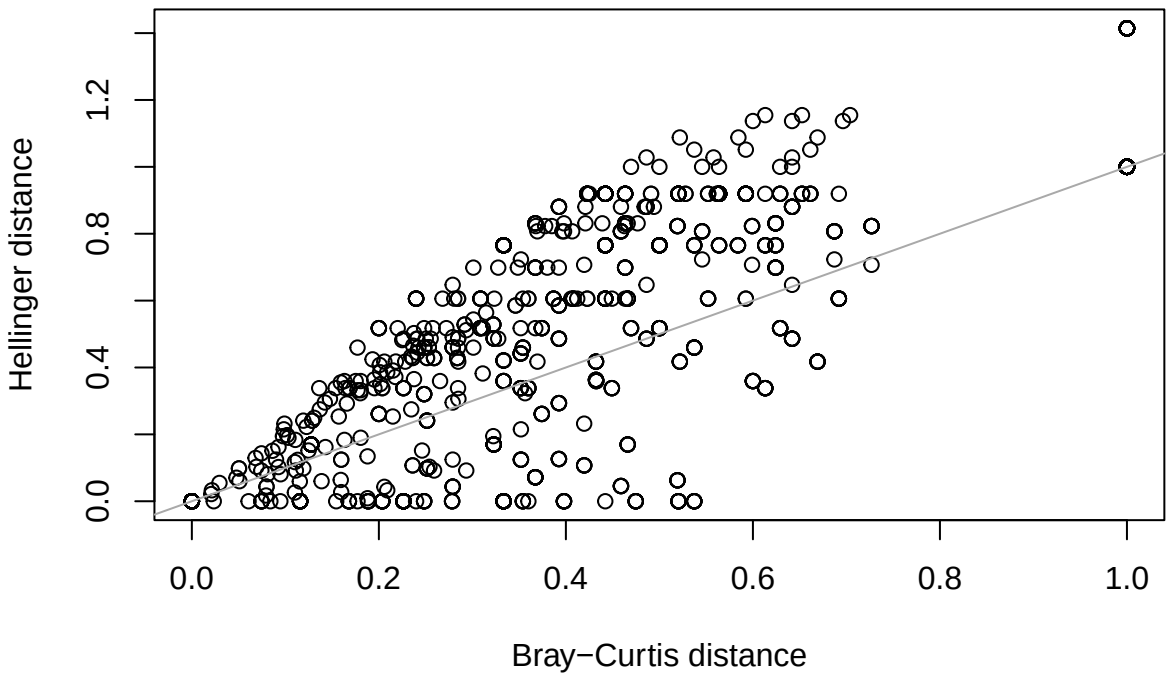
```
## Warning in vegdist(sal_occ, method = "bray"): missing values in results
```



Salamander abundance: Bray curtis or hellinger. (coldiss functions still wont work for these, Error in if (max(D) > 1) D <- D/max(D) : missing value where TRUE/FALSE needed)

```
## Warning in vegdist(log_sal_cou, method = "bray"): you have empty rows: their dissimilarities may be
##                   meaningless in method "bray"
```

```
## Warning in vegdist(log_sal_cou, method = "bray"): missing values in results
```



Env data: Euclidean distance method

```
env.euc <- vegdist(env_std, metric="euclidean")
```

```
## Warning in vegdist(env_std, metric = "euclidean"): results may be meaningless because data have nega
##                   in method "bray"
```

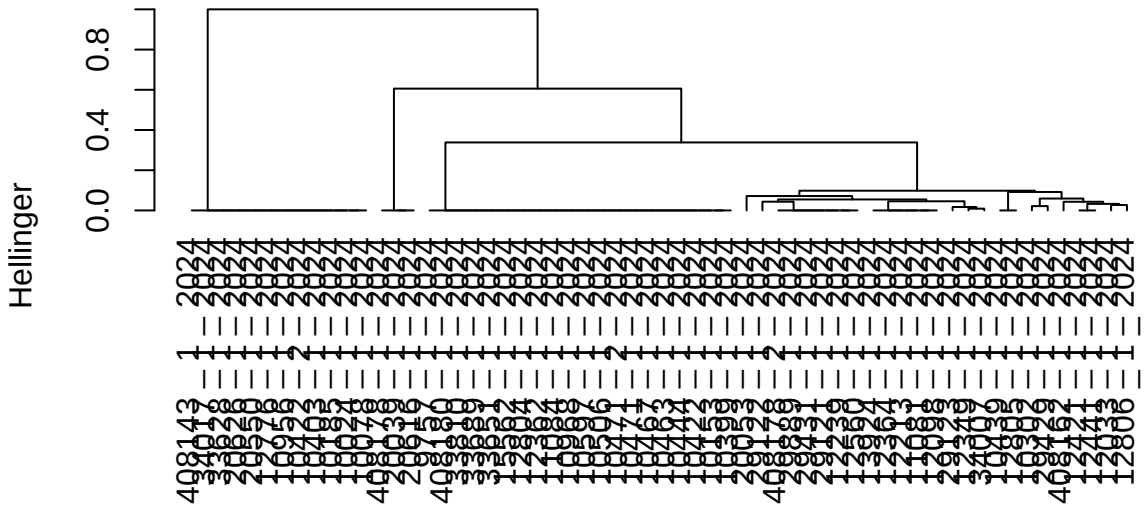
```
env.euc.subs <- vegdist(env_subset_std, metric="euclidean")
```

```
## Warning in vegdist(env_subset_std, metric = "euclidean"): results may be meaningless because data ha
##                   in method "bray"
```

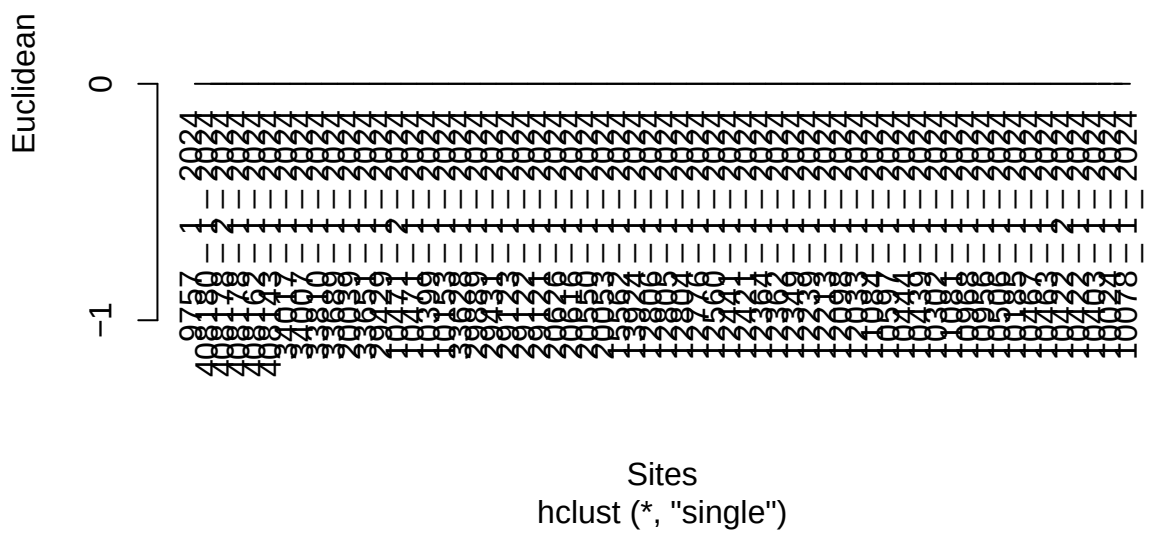
###Question 3) Try each of the agglomerative hierarchical clustering methods with your own dataset. Which one appears to perform best? What is the optimal number of clusters? Verify this using the appropriate non-statistical and statistical approaches.

Single Linkage I have spent most of today working with this and I'm really struggling to get my data to work with any of these. Issues with the species data makes sense because I only have two specie,s but not sure why the env data isnt working. Even when I subsetted things and used a smaller number of sites and variables it looks like poo. Is it just because I have so many sites?

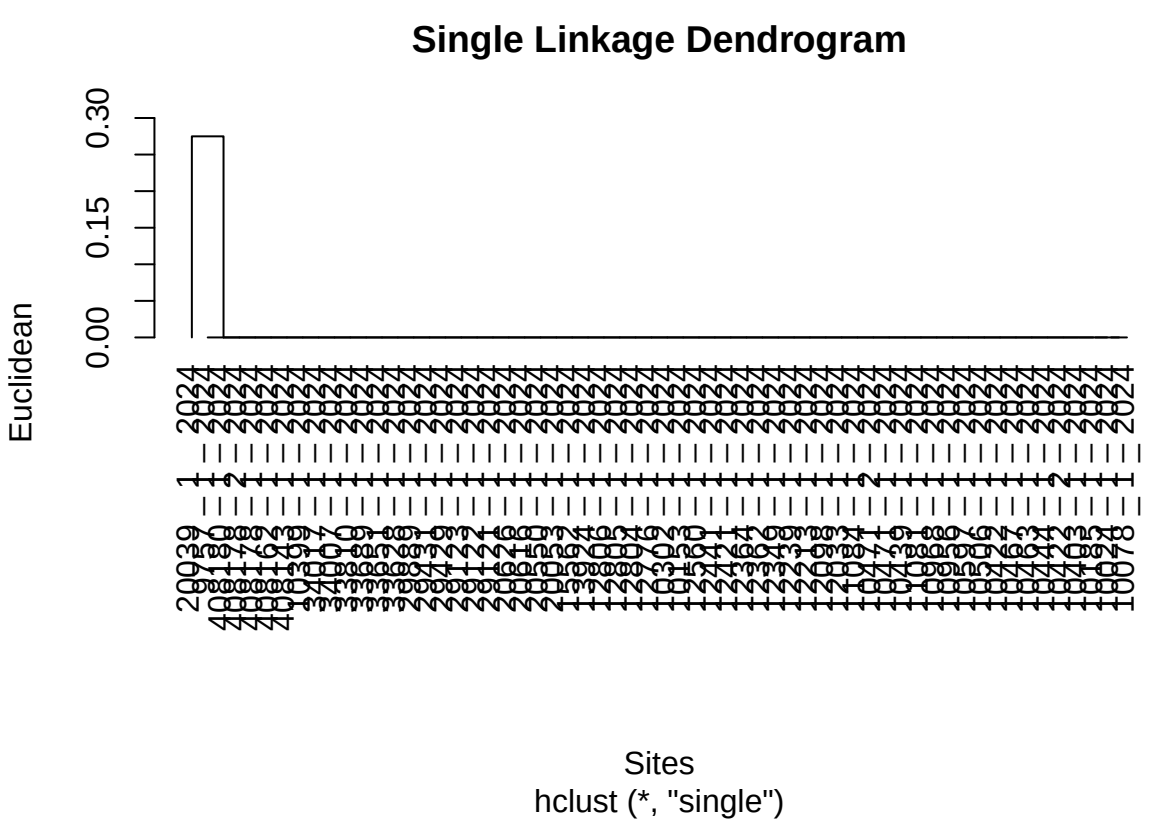
```
sals.sin <- hclust(sal.hel, method = "single")
        plot(sals.sin, main="Single Linkage Dendrogram",
             xlab="Sites",
             ylab="Hellinger",
             hang=-1)
```




```
env.sin <- hclust(env.euc, method = "single")
        plot(env.sin, main="Single Linkage Dendrogram",
             xlab="Sites",
             ylab="Euclidean",
             hang=-1)
```

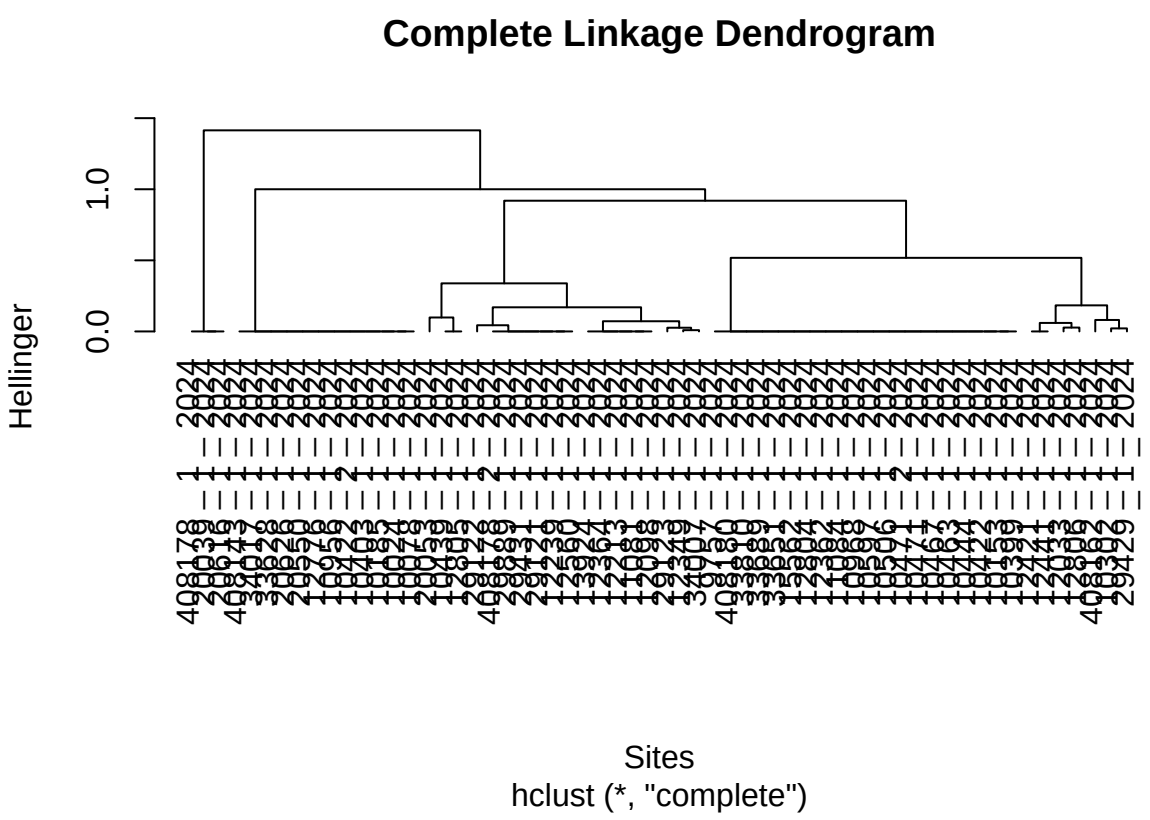



```
env.sin2 <- hclust(env.euc.subs, method = "single")
        plot(env.sin2, main="Single Linkage Dendrogram",
             xlab="Sites",
             ylab="Euclidean",
             hang=-1)
```

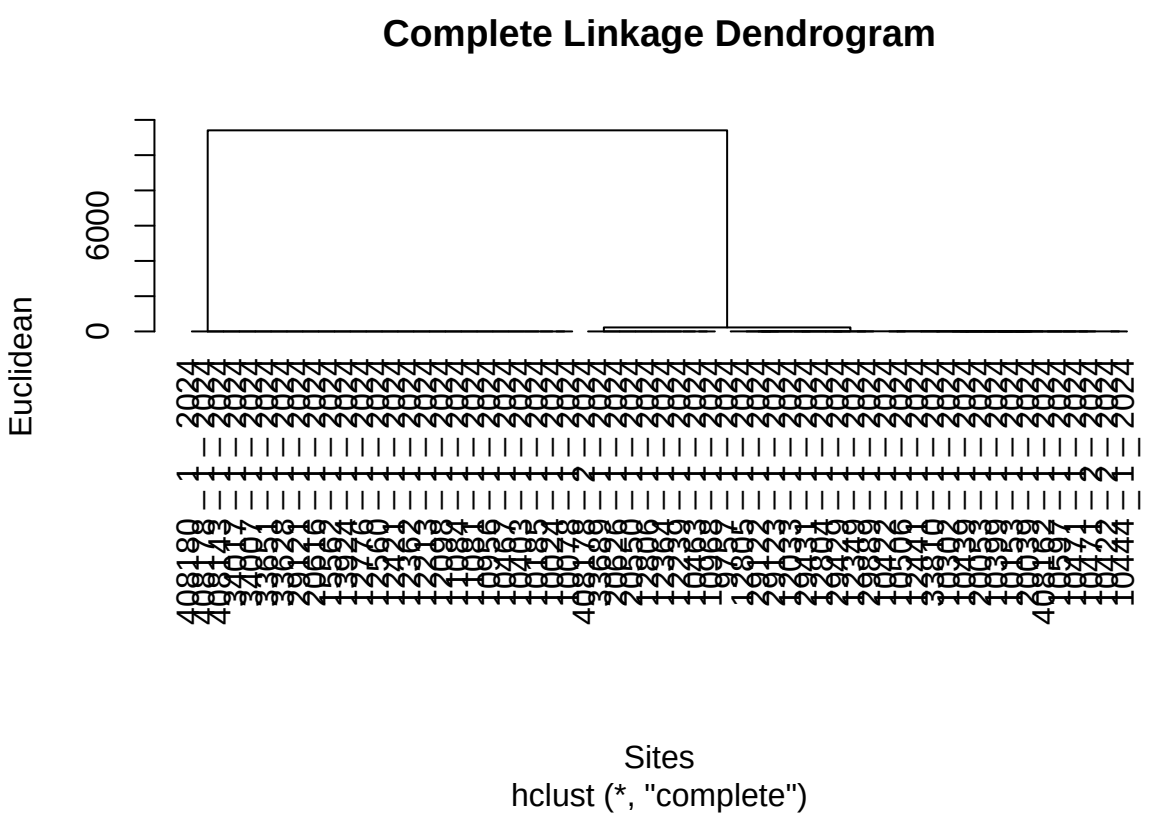


Complete Linkage - these might look slightly better

```
salcl.com <- hclust(sal.hel, method = "complete")
        plot(salcl.com, main="Complete Linkage Dendrogram",
             xlab="Sites",
             ylab="Hellinger",
             hang=-1)
```

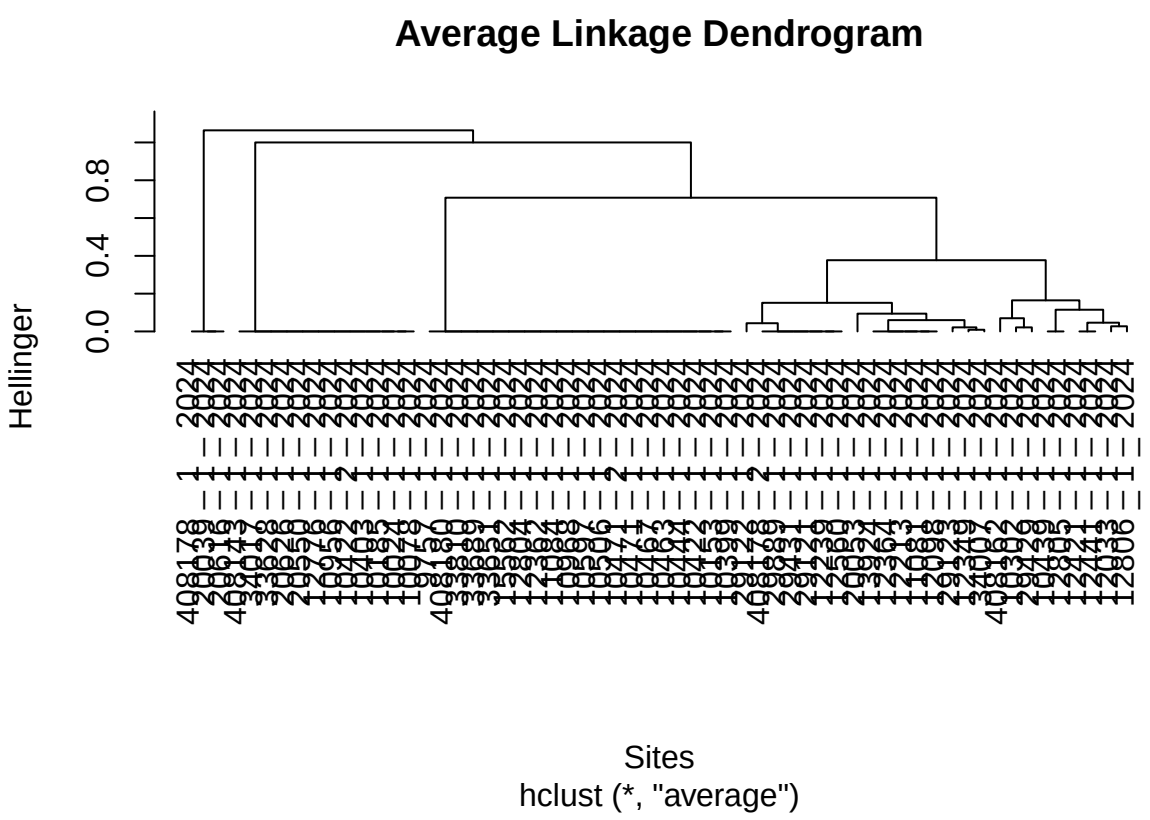



```
env.com <- hclust(env.euc.subs, method = "complete")
        plot(env.com, main="Complete Linkage Dendrogram",
             xlab="Sites",
             ylab="Euclidean",
             hang=-1)
```

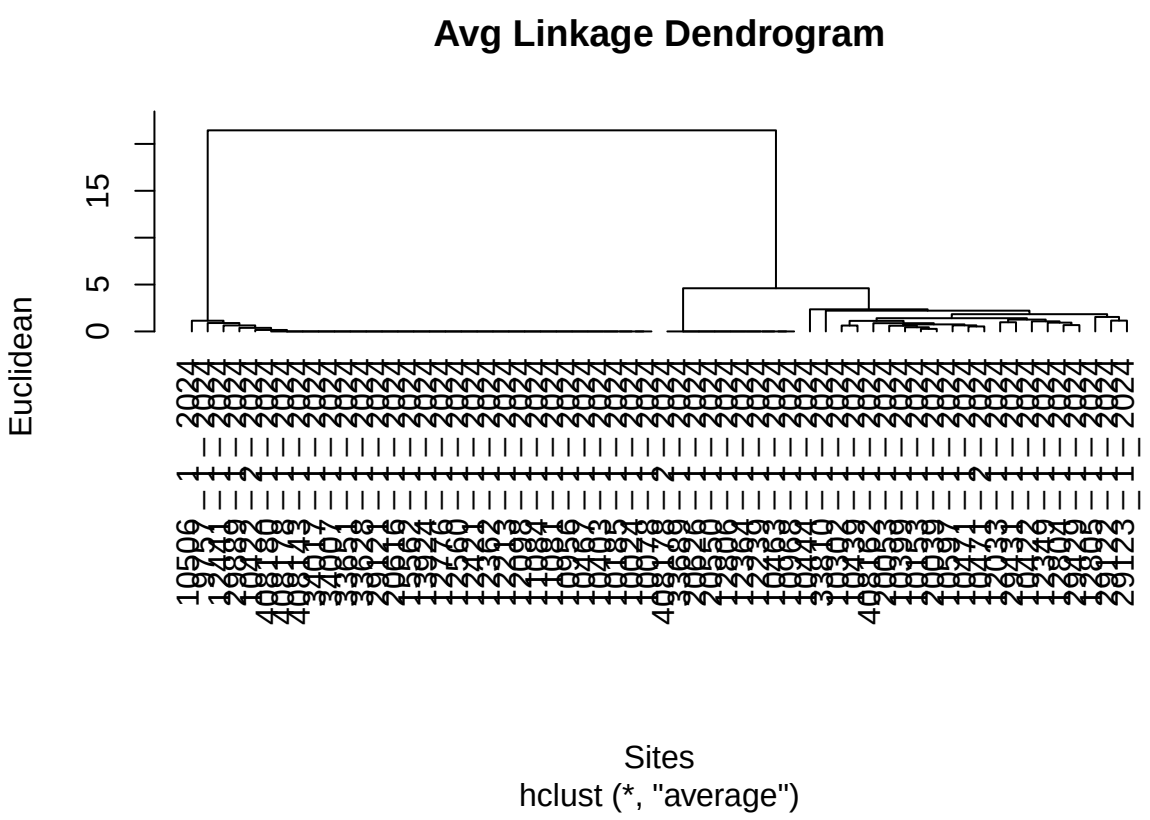


Average Linkage

```
salcl.ave <- hclust(sal.hel, method = "average")
        plot(salcl.ave, main="Average Linkage Dendrogram",
             xlab="Sites",
             ylab="Hellinger",
             hang=-1)
```

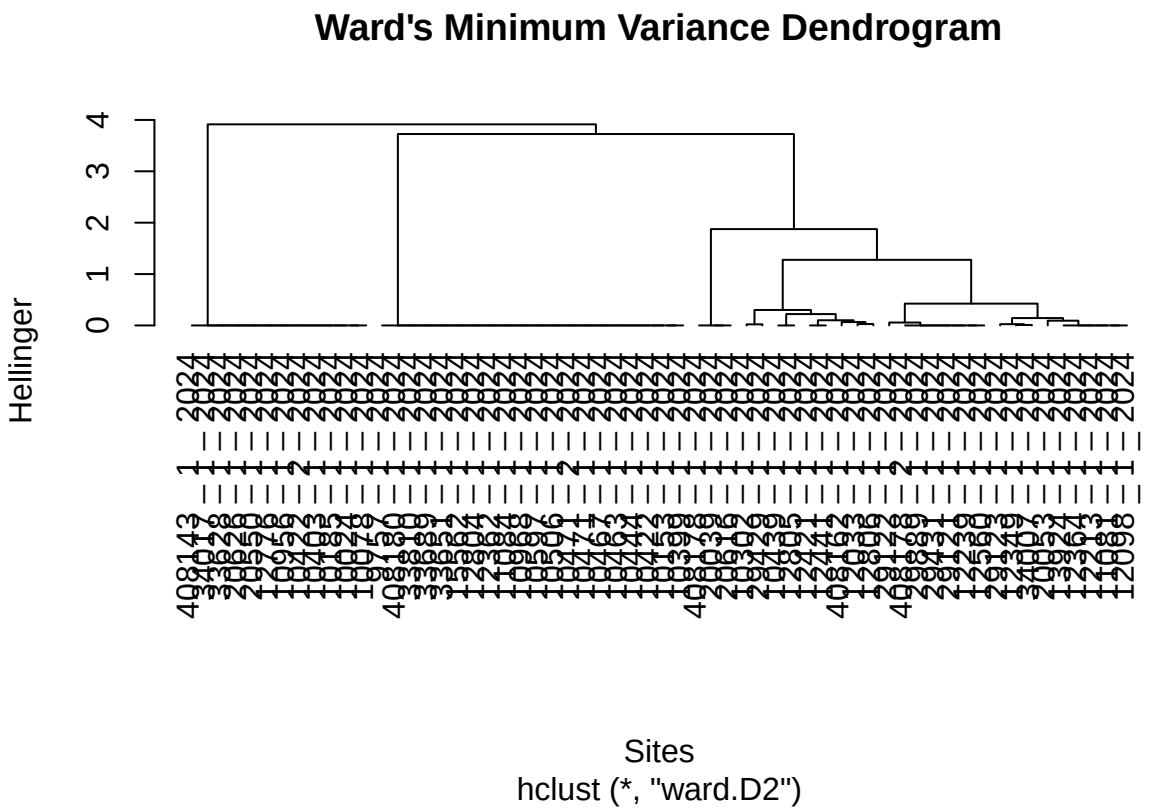



```
env.ave <- hclust(env.euc.subs, method = "average")
        plot(env.ave, main="Avg Linkage Dendrogram",
             xlab="Sites",
             ylab="Euclidean",
             hang=-1)
```

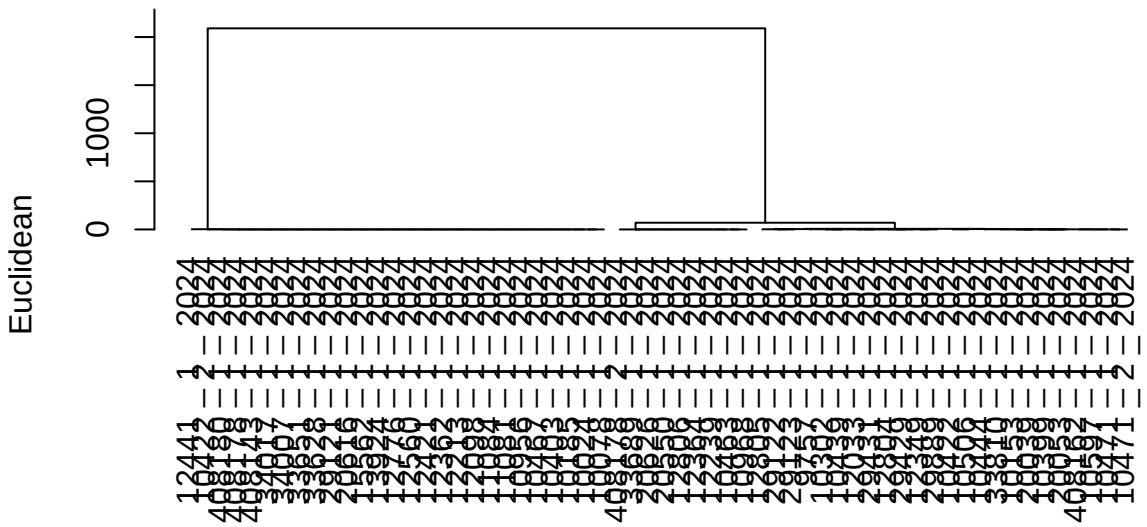


Wards

```
salcl.ward <- hclust(sal.hel, method = "ward.D2")
        plot(salcl.ward, main="Ward's Minimum Variance Dendrogram",
                xlab="Sites",
            ylab="Hellinger",
            hang=-1)
```




```
envcl.ward <- hclust(env.euc.subs, method = "ward.D2")
        plot(envcl.ward, main="Complete Linkage Dendrogram",
             xlab="Sites",
             ylab="Euclidean",
             hang=-1)
```



###Question 4) Do you see any evidence that noise and outliers are impacting your results? If so, how could you treat the data differently to account for these sources of error?

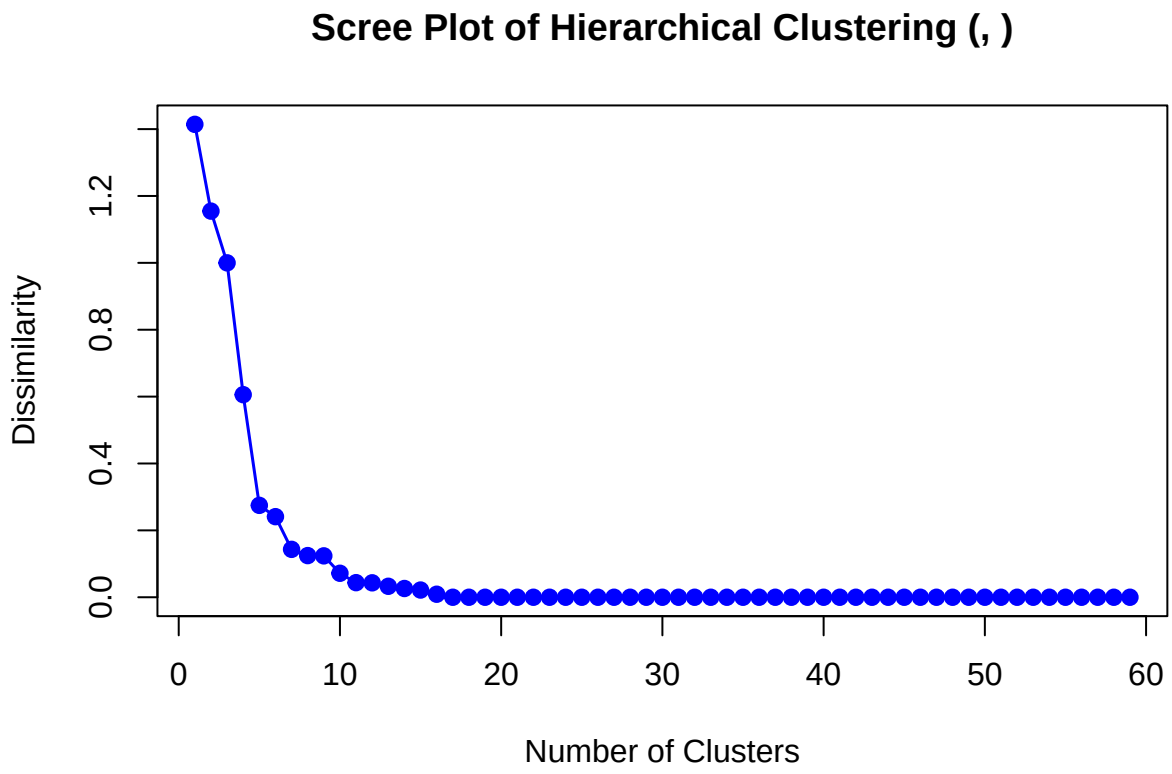
I'm really not sure what I'm looking at in terms of errors here. There are some sites that are so different from the others that its really stretching the y-axis. There are so many connections that its hard to parse out, so I subsetted the data to use only one year of sites and a subset of the env data but it didnt seem to fix things.

###Question 5) Discuss the advantages and disadvantages of using divisive hierarchical clustering over agglomerative hierarchical clustering in ecological data analysis. When you run a divisive analysis (DIANA) using your data, how do the results differ from the agglomerative clustering output?

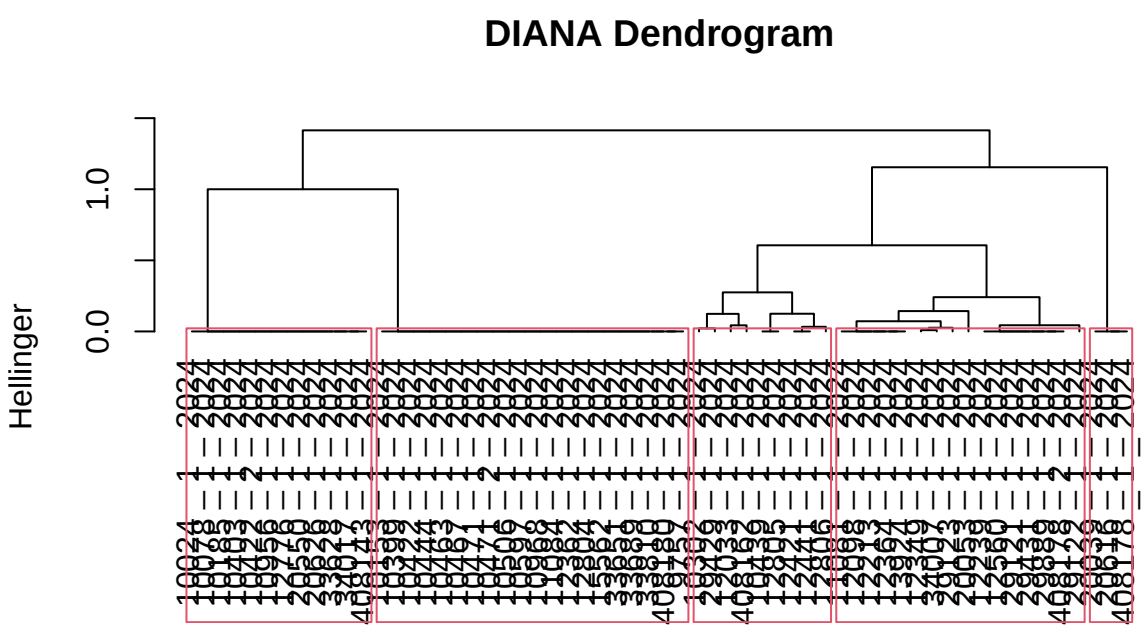
Divisive clustering is more suitable for well-separated groups but is computationally intensive. Agglomerative clustering is more computationally efficient but is sensitive to outliers.

This approach does look a little more organized than the ones above.

```
diana_sals <- diana(as.dist(sal.hel))
hclus.scree(diana_sals)
```



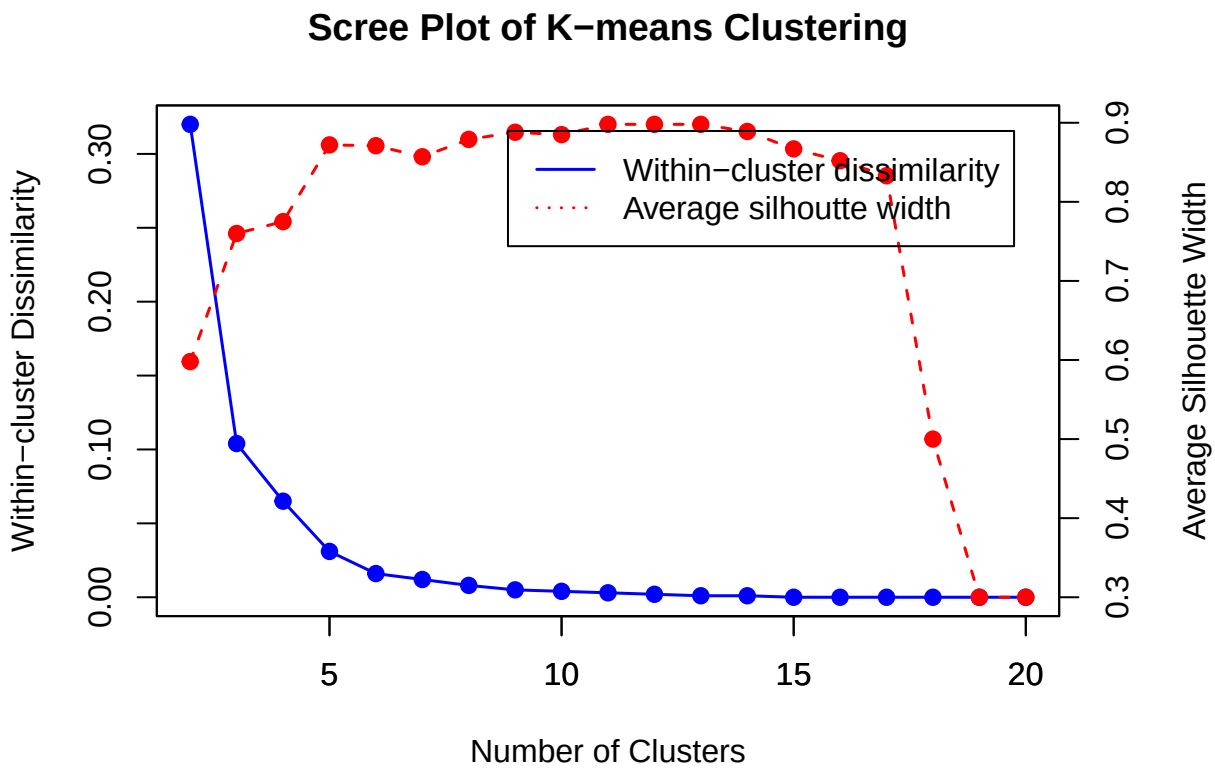

```
plot(diana_sals, which=2, main="DIANA Dendrogram",
             xlab="Sites",
             ylab="Hellinger",
             hang=-1)
             rect.hclust(diana_sals, k=5)
```



###Question 6) Explain the k-means clustering algorithm and list key steps. When you run a k-means partitioning analysis using your data, how do the results compare to the hierarchical clustering output(s)? Based on the sum of squares error, do you think the resulting solution makes sense?

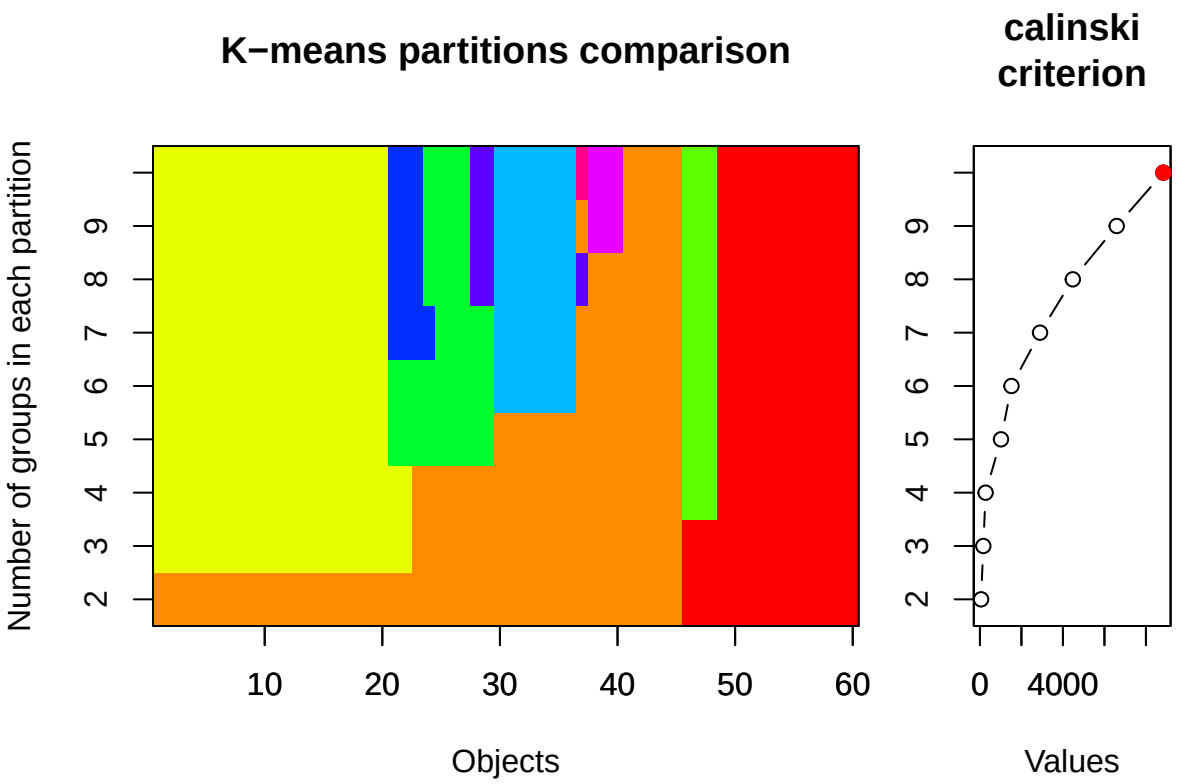
K-means partitioning is meant to reduce the variation within clusters, but requires that you set the number of clusters beforehand. First is initialization, then assigning points to clusters, assigning a new centroid from the new mean, and repeat.

```
nhclus.scree(sal.hel, max.k = 20)
```




```
##    no. clusters sum within-clus diss ave silhouette width
## 1             2                0.320                0.598
## 2             3                0.104                0.760
## 3             4                0.065                0.775
## 4             5                0.031                0.872
## 5             6                0.016                0.871
## 6             7                0.012                0.857
## 7             8                0.008                0.879
## 8             9                0.005                0.888
## 9            10                0.004                0.885
## 10           11                0.003                0.898
## 11           12                0.002                0.898
## 12           13                0.001                0.898
## 13           14                0.001                0.889
## 14           15                0.000                0.867
## 15           16                0.000                0.852
## 16           17                0.000                0.833
## 17           18                0.000                0.500
## 18           19                0.000                0.300
## 19           20                0.000                0.300
```

```
salcl.kmeans.cas <- cascadeKM(sal.hel, inf.gr=2, sup.gr=10, iter=100)
plot(salcl.kmeans.cas, sortg=T)
```

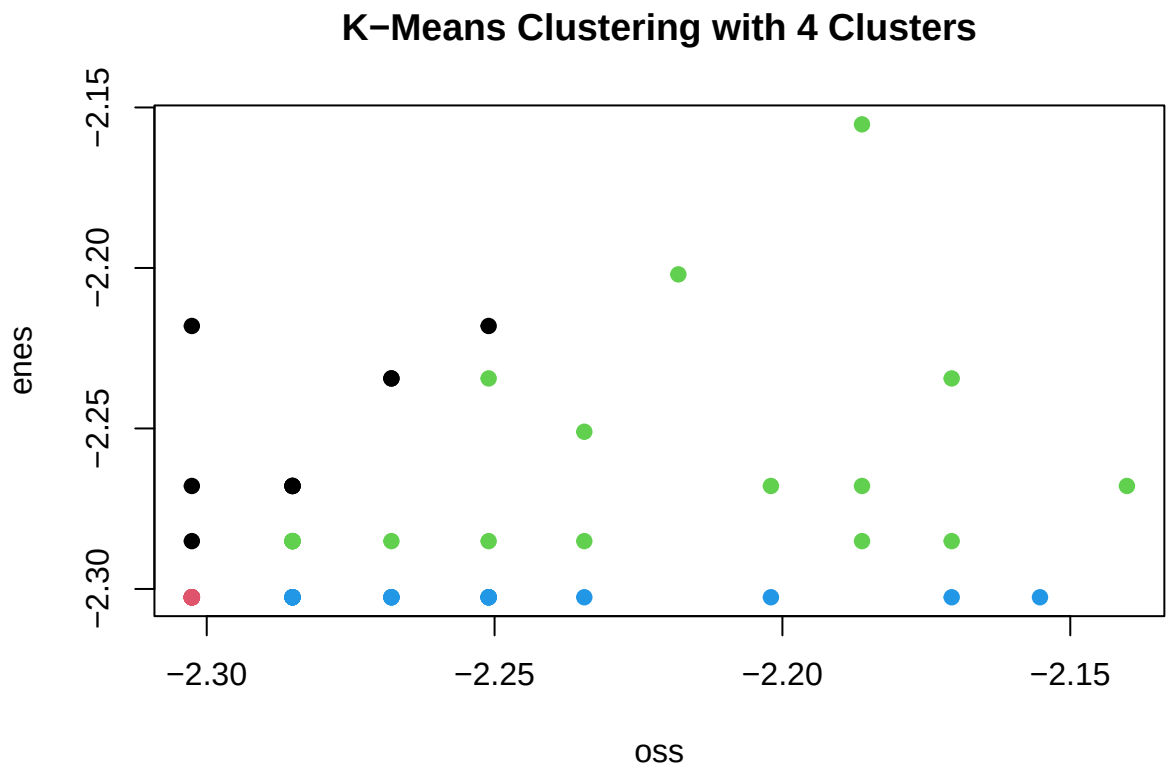



```
salcl.kmeans.cas$results
```

```
##           2 groups  3 groups  4 groups    5 groups    6 groups   7 groups
## SSE      288.63951  84.37072  36.10154    7.551995    3.986908    1.71702
## calinski  55.49125 162.28457 273.36587 1014.576784 1519.151812 2896.79016
##              8 groups     9 groups    10 groups
## SSE         0.9356806    0.5456649    0.3548135
## calinski 4476.5946657 6592.1121800 8837.8086944
```

```
sal.kmeans <- kmeans(sal.hel, centers=4, iter.max=10000, nstart=10)

plot(log_sal_dens$oss, log_sal_dens$enes,
            col = sal.kmeans$cluster,
            pch = 19,
            main = "K-Means Clustering with 4 Clusters",
            xlab="oss",
            ylab="enes")
```



###Question 7) Discuss the role of initialization in k-means clustering. How do different initialization methods impact the convergence and final results of the algorithm?

Initialization uses the number of clusters you chose and creates that many initial centroids from which to start creating clusters. There are several methods and they impact the quality of the final clusters and can make the analysis slow. Most of them are a trade off between computationaly simple/fast and more in depth analysis.

###Question 8) Explain the concept of indicator species in ecological analysis. How are indicator species identified and used to interpret ecological data? Do there appear to be any indicator species in your data, and if so, for what site groups and are the indicator values significant?

Indicator species are sensitive to environmental change and are therefor important for monitoring environmental health and stability. They are identified using indicator species analysis and other methods. My main study species is known to be an indicator species, but that isnt showing up in my data (i have theories about why)...

This shows that ENES is associated with BU (burned, unharvested) and UU (unburned, unharvested) sites, and ODF and BLM land; these are considered our most pristine or least changed sites. ENES are much more mobile than OSS and I think they just moved out of the sites where they didnt like the conditions, whereas OSS has to stick around and are persisting so far.

```
site_num <- env$trt
        ind_sp <- multipatt(log_sal_dens, site_num, func = "IndVal.g", control = how(nperm=999))
        summary(ind_sp)
```

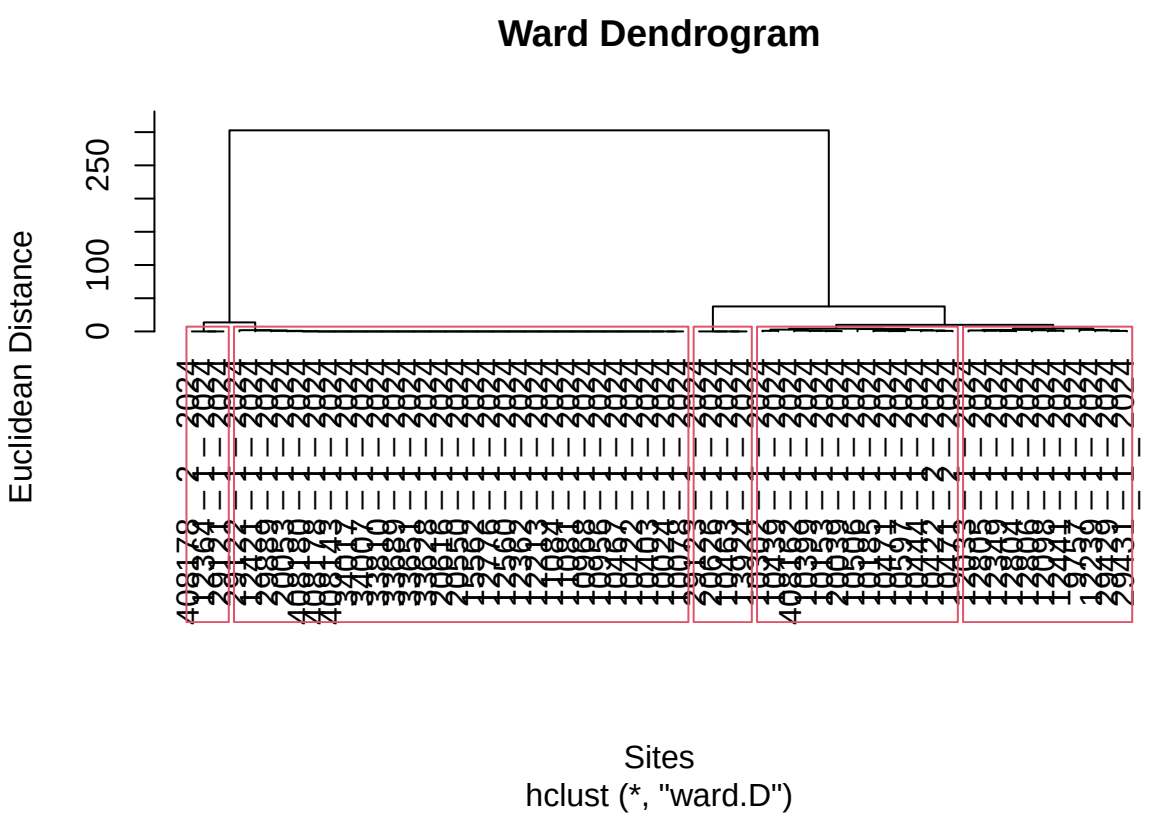
```
##
##  Multilevel pattern analysis
```

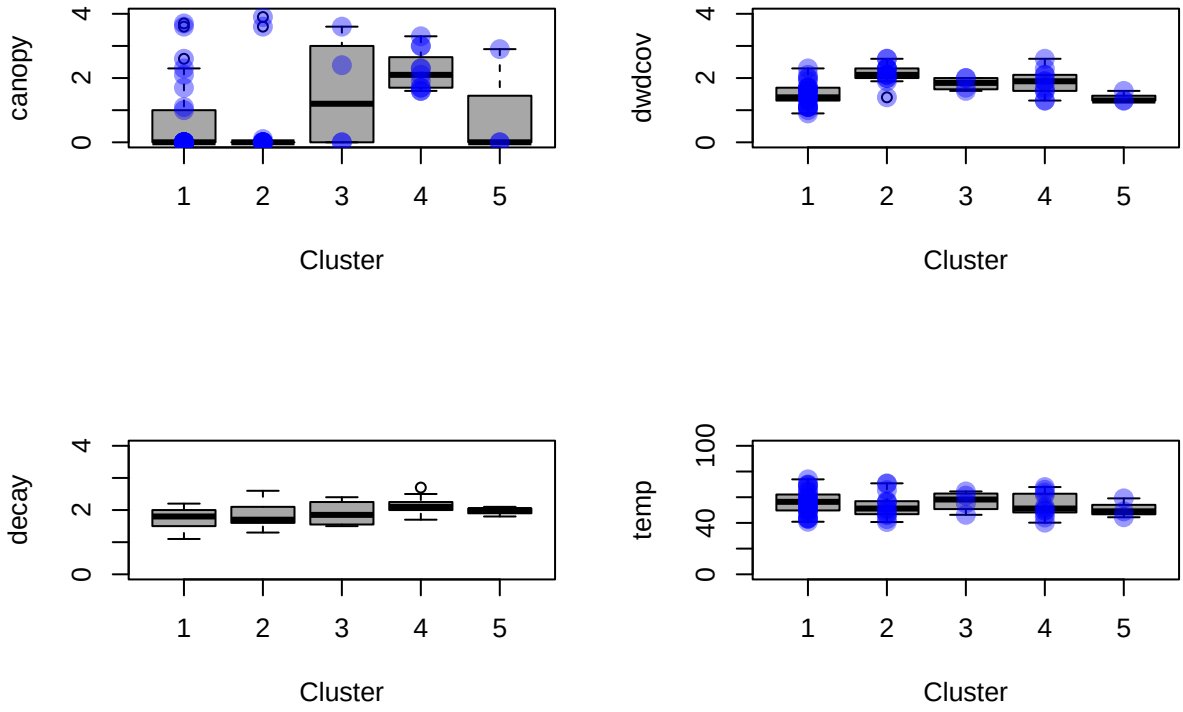
```
##  ---------------------------
##
##  Association function: IndVal.g
##  Significance level (alpha): 0.05
##
##  Total number of species: 2
##  Selected number of species: 0
##  Number of species associated to 1 group: 0
##  Number of species associated to 2 groups: 0
##  Number of species associated to 3 groups: 0
##  Number of species associated to 4 groups: 0
##
##  List of species associated to each combination:
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
site_num <- dat2$landowner
        ind_sp <- multipatt(log_sal_dens, site_num, func = "IndVal.g", control = how(nperm=999))
        summary(ind_sp)
```

```
##
##  Multilevel pattern analysis
##  ---------------------------
##
##  Association function: IndVal.g
##  Significance level (alpha): 0.05
##
##  Total number of species: 2
##  Selected number of species: 0
##  Number of species associated to 1 group: 0
##  Number of species associated to 2 groups: 0
##  Number of species associated to 3 groups: 0
##
##  List of species associated to each combination:
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here OSS is associated with clusters 2, 3, 4 #2: no canopy, high dwd, med decay (no canopy means harvested so this is weird!) #3: high canopy, med dwd,high decay #4: no canopy, med decay, med dwdcov


Ward Dendrogram
Euclidean Distance
0
100
250
Sites
hclust (*, "ward.D")




```
##
##  Multilevel pattern analysis
##  ---------------------------
##
##  Association function: IndVal.g
##  Significance level (alpha): 0.05
##
##  Total number of species: 2
##  Selected number of species: 0
##  Number of species associated to 1 group: 0
##  Number of species associated to 2 groups: 0
##  Number of species associated to 3 groups: 0
##  Number of species associated to 4 groups: 0
##
##  List of species associated to each combination:
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```