Unsupervised Learning Analysis

William Norfolk

11/18/2019

This script uses unsupervised modeling to determine similaraity characteristics between individual sampling sites, as well as uses cluster analysis to measure similarity of year to year measures.

Start by loading required libraries

library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## -- Attaching packages ------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v readr 1.3.1  
## v tibble 2.1.3 v purrr 0.3.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v ggplot2 3.2.1 v forcats 0.4.0

## -- Conflicts ---------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(forcats)  
library(ggthemes)  
library(knitr)  
library(naniar)  
library(broom)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(ggpubr)

## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(visdat)  
library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(dendextend)

##   
## ---------------------  
## Welcome to dendextend version 1.12.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------

##   
## Attaching package: 'dendextend'

## The following object is masked from 'package:ggpubr':  
##   
## rotate

## The following object is masked from 'package:stats':  
##   
## cutree

library(cluster)  
library(NbClust)  
library(mclust)

## Package 'mclust' version 5.4.5  
## Type 'citation("mclust")' for citing this R package in publications.

##   
## Attaching package: 'mclust'

## The following object is masked from 'package:purrr':  
##   
## map

#Note it will be very important to call functions using the :: library notation due to the frequency of function masking by our libraries.

Next load the data a take a look.

WQ\_clean\_data <- readRDS("../../data/processed\_data/processeddata.rds")  
  
glimpse(WQ\_clean\_data)

## Observations: 522  
## Variables: 15  
## $ Month <chr> "01", "02", "02", "02", "02", "02", "02", "02...  
## $ Day <chr> "08", "08", "08", "08", "09", "09", "09", "12...  
## $ Year <chr> "16", "16", "16", "16", "16", "16", "16", "16...  
## $ military\_time <dbl> 1415, 1515, 1550, 1555, 1001, 1015, 1022, 103...  
## $ location <chr> "Boat Ramp", "Grecian Dry Rocks", "Grecian Dr...  
## $ instructor\_name <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, Tomm...  
## $ group\_name <chr> "NA", "McLean High School", "McLean High Scho...  
## $ ph <dbl> 8.0, 8.4, 8.2, 8.4, 8.0, 8.0, 8.0, 8.0, 8.4, ...  
## $ ammonia <dbl> 0.00, 0.00, 0.00, 0.00, 0.25, 0.00, 0.00, 0.0...  
## $ dissolved\_oxygen <dbl> 5.0, 4.0, 4.0, 6.0, 8.0, 4.0, 5.0, 6.0, 6.0, ...  
## $ water\_temp <dbl> NA, 23.5, 21.0, 36.0, 18.0, 18.0, 18.0, 18.3,...  
## $ salinity <dbl> 36, 40, 44, 35, 33, 30, 33, 35, 40, 30, 35, 2...  
## $ equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "ki...  
## $ island\_side <chr> "ocean", "ocean", "ocean", "ocean", NA, "bay"...  
## $ site\_type <chr> "Seagrass/Mangrove", "Coral Reef", "Coral Ree...

## Unsupervised Cluster Analysis

Since our data is not specifically targeted at particular outcome (i.e. occurance of disease, algal bloom, etc.) and focuses more as a monitoring program; it may be best to use unsupervised modeling techniques to to cluster our data. We will use agglomerative clustering to explore the data for various cluster patterns. Agglomerative clustering works in a bottom-up fashion, each individual object within a data set is considered a distinct cluster (leaves) initially and is grouped to the nearest cluster (via the input variables) at each step of the algorithm until all data is in a single big cluster (root) thus producing a dendrogram. The inverse of agglomerative clustering is called divisive cluster analysis and works opposite from a root to leaf manner. Divisive clustering is acomplished through the function diana() and is used substantially less in analyses than agnes(). We will compare a diana() cluster to agnes() and hclut() below.

### Questions for this Analysis

*How do sampling sites cluster based on individual water quality measurements* *Are there any site-wise, annual, and/or seasonal patterns that cna be identified based on cluster analysis?*

First we must remove any NAs from our data as cluster functions cannot handle them. Additionally, we need to select our branch outcome (in this case location) as well as our five numeric water quality parameters of interest.

cluster\_trim <- WQ\_clean\_data %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
cluster\_trim <- na.omit(cluster\_trim)

We will cluster by site means to produce a readable dendrogram.

With a retrospective view, there are four locations which will greatly skew our data when producing a dendrogram. All four locations have a single measure comprising the data for site. When allowed to remain in our sample set these data points force our dendrogram to subdivide further grouping themselves into indivual clusters. Though this information is interesting, considering the n of each of these locations and the large impact they play on the final dendrogram, we will remove them here to produce a generalized view of the clusters.

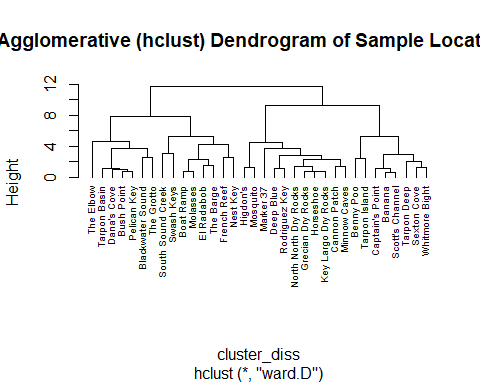
In addition to pruning, all data must be scaled so it can be accurately compared across multiple variables. Each of our water quality parameters have a different scale of measure (i.e. temperature is relatively variable from approximately 10 degrees to 40 degrees, where as ammonia is very narrow from 0 to 1 mg/L). The variability among the scales of the variables would skew the results to to very strongly express those with greater differences (likely temperature and salinity). To ensure all variables are accounted for correctly we will use the scale function which subtracts the mean from each observation then divides by the standard deviation. This allows variables which are measured on different scales to be compareable for clustering.

location\_average <- aggregate(. ~ location, cluster\_trim, mean)  
  
drop\_buttonwood <- location\_average[!location\_average$location == "Buttonwood", ]  
drop\_orc <- drop\_buttonwood[!drop\_buttonwood$location == "Ocean Reef Club", ]  
drop\_lagoon <- drop\_orc[!drop\_orc$location == "Lagoon", ]  
drop\_hidden <- drop\_lagoon[!drop\_lagoon$location == "Hidden Lake", ]  
drop\_NA\_site <- drop\_hidden[!drop\_hidden$location == "NA", ]#There is also a site listed as NA we must remove  
  
#set rownames to the locations names, then the location variable must be dropped to eliminate any non-numeric variables in the dataframe.   
rownames(drop\_NA\_site) <- drop\_NA\_site$location  
  
drop\_location <- drop\_NA\_site %>% dplyr::select(-location)  
  
#Scale the data so all variables are compareable, after using the scale function you must convert the resulting object back to a data fram for further analysis  
cluster\_ready <- scale(drop\_location, center = TRUE, scale = TRUE)  
  
cluster\_ready <- as.data.frame(cluster\_ready)

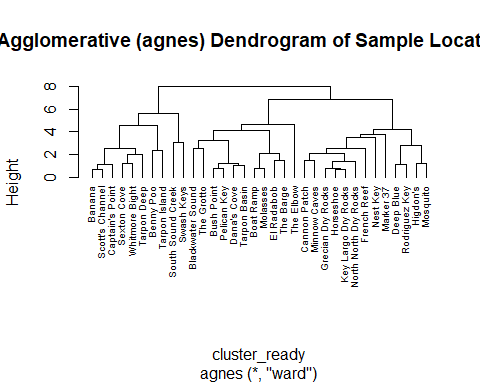
Next we produce the dendrograms with the code below. We will use three different cluster methods hclust, agnes, and diana to compare the results initially. hclust and agnes will preform agglomerative cluster analysis and diana will preform divisive cluster analysis.

Note prior preforming agglomerative clustering we must first compute a dissimilarity martix to determine how different individual observations are from one another, this is acomplished with the dist() function for the hclust function and is built into the functions agnes and diana.

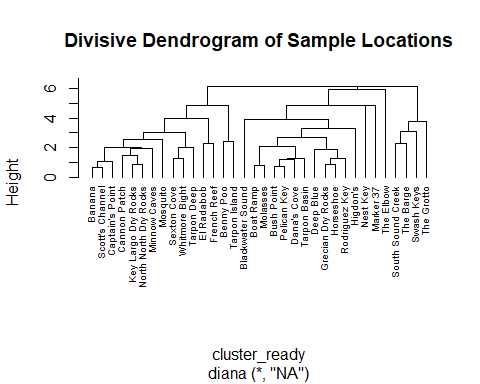
set.seed(1111) #Ensure reproducible  
  
#rownames(cluster\_ready) <- cluster\_ready$location #Assign the location names to rownames to produce dendrograms that show site information.  
  
cluster\_diss <- dist(cluster\_ready, method = "euclidean") #For the hclust function we must first produce a dissimilarity matrix before runing cluster analysis  
  
  
cluster\_site\_hclust <- hclust(cluster\_diss, method = "ward.D") #Hierarchial clustering with Ward's minimum variance  
cluster\_site\_agnes <- agnes(cluster\_ready, method = "ward") #Agglomerative clustering with Ward's minimum variance  
cluster\_site\_diana <- diana(cluster\_ready) #Divisive clustering is similar to the agnes function, however no method is needed to provide to this function  
  
#Plot the trees. Note agnes can use pltree or plot but hclust must use plot  
good\_tree\_hclust <- plot(cluster\_site\_hclust, cex = 0.6, hang = -1, main = "Agglomerative (hclust) Dendrogram of Sample Locations")



good\_tree\_agnes <- pltree(cluster\_site\_agnes, cex = 0.6, hang = -1, main = "Agglomerative (agnes) Dendrogram of Sample Locations")



good\_tree\_diana <- pltree(cluster\_site\_diana, cex = 0.6, hang = -1, main = "Divisive Dendrogram of Sample Locations")



So it looks like all three dendrograms look fairly different. There are no very distinct patterns that jump out immediately but we will need to do a bit more analysis to learn our optimal cluster number (k) before we can really start to interpret the results.

Our primary question is targeted at determining how different individual sampling site locations are from one another. While our sites represent an array of different habitat types, we are interested in clustering locations by the specific abiotic conditions of the locale. Since our question is looking at these small specific differences we will use agglomerative clustering for the duration of the analysis as this method is better suited for identifying small clusters. Agglomerative clustering uses a more exhaustive algorithm (pairs at each iteration and upon completion) to make comparisons so it is generally better with small differences in data however, this is a trade-off for computational ability. Since our dataset is fairly small and computationally easy this will not be an issue. For larger datasets divisive clustering would be the more efficent choice.

Before moving on we will check our agglomerative method. We started with Ward’s minimum variance method however there are others we could use (single, complete, and average). We will calculate the agglomerative coefficent for each below. A coefficent closer to 1 indicates a stronger clustering.

cluster\_site\_agnes$ac

## [1] 0.7893823

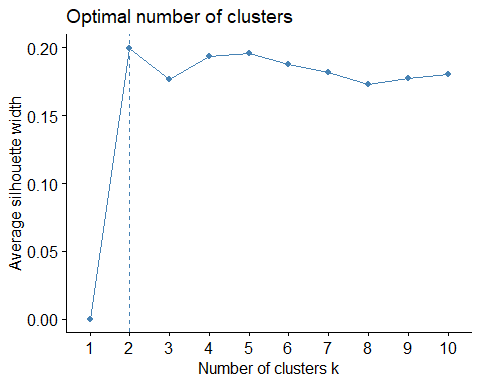
methods <- c( "average", "single", "complete", "ward")  
names(methods) <- c( "average", "single", "complete", "ward")  
  
#Note we must use agnes to get the agglomerative coefficent as hclust objects do not permit this function  
coeff <- function(x) {  
 agnes(cluster\_ready, method = x)$ac  
}  
  
map\_dbl(methods, coeff)

## average single complete ward   
## 0.6050743 0.5175979 0.7225867 0.7893823

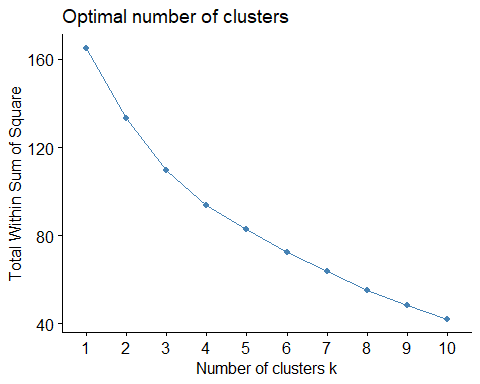
It looks like Ward’s method is the best option so we will keep that for all future clusters to be consistent.

Next we will check our optimal number of clusters. There are two available methods to check for optimal clusters wss and silhouette. These functions work with cluster algorithms to determine the ideal k value for samples. The wss (within-cluster sum of squares) or “Elbow Method” measures the compactness of clustering with the within cluster sum of squares, a smaller wss equals the optimal cluster numbers. The wss plot produced by the function shows a curve where the Elbow in the the curve is the optimal clusters. The silhouette method uses the average silhouette to determine how well an object lies in a cluster. There greater avaerage silhouette the more optimal the cluster level is.

set.seed(1111)  
  
fviz\_nbclust(cluster\_ready, FUN = hcut, method = "silhouette", print.summary = TRUE)



fviz\_nbclust(cluster\_ready, FUN = hcut, method = "wss", print.summary = TRUE)

 It looks like five clusters is the optimal number for both the wss and silhouette methods so we will use this k value to produce our good dendrograms for the manuscript.

Before moving to the next code set. Just a quick code bit to count the number of values in each cluster.This is more useful if there is a very high number of branches.

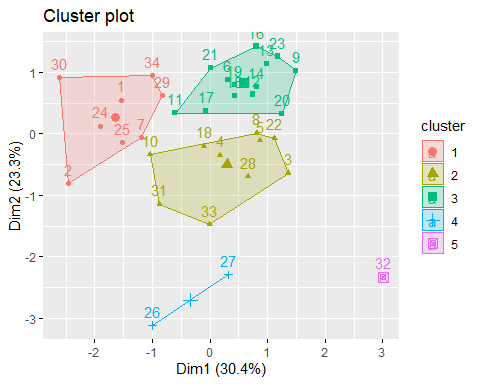
sub\_group\_dendro\_1 <- cutree(cluster\_site\_agnes, k = 5)  
  
table(sub\_group\_dendro\_1)

## sub\_group\_dendro\_1  
## 1 2 3 4 5   
## 8 10 13 2 1

Looks like one of our clusters is substantially larger than the rest of the of the clusters. Ideally we can find a common factor between these location points which we cna interpret.

Next for fun we will visualize our data as a scatter cluster plot.

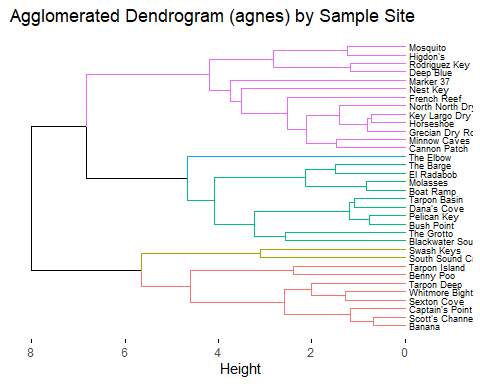
add\_cluster\_groups <- cluster\_ready %>% dplyr::mutate(cluster = sub\_group\_dendro\_1)  
  
  
  
cluster\_site <- agnes(add\_cluster\_groups, method = "average")  
  
cluster\_data <- add\_cluster\_groups[, c(2:6)]  
  
#Uncomment to add names to plot (the look is confusing in this situation)  
rownames(add\_cluster\_groups) <- cluster\_ready$location  
  
  
  
fviz\_cluster(list(data = cluster\_data, cluster = sub\_group\_dendro\_1))



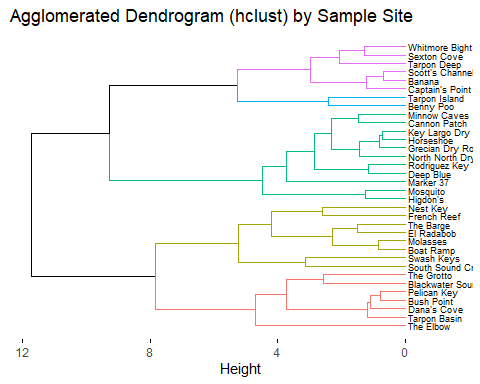
Looks like our clusters are relatively distinct with a little overlap between 1 and 2. Five may be a cluster to keep an eye on as it is very small, but it is possible that cluster contains two sites which are very different from others in the data set.

Next lets make a nice dendrogram.

sample\_site\_dendro\_agnes <- fviz\_dend(cluster\_site\_agnes, k = 5, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Agglomerated Dendrogram (agnes) by Sample Site")  
  
sample\_site\_dendro\_agnes



sample\_site\_dendro\_hclust <- fviz\_dend(cluster\_site\_hclust, k = 5, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Agglomerated Dendrogram (hclust) by Sample Site")  
  
sample\_site\_dendro\_hclust

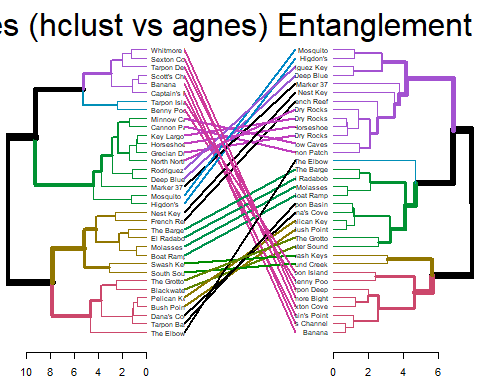


Now that we can better visualize differences it looks like our two agglomerative dendrograms are very similar (which is good news!) the hclust function and agnes functions are slightly different in practice so it is not out of the question to see small differences in the results.

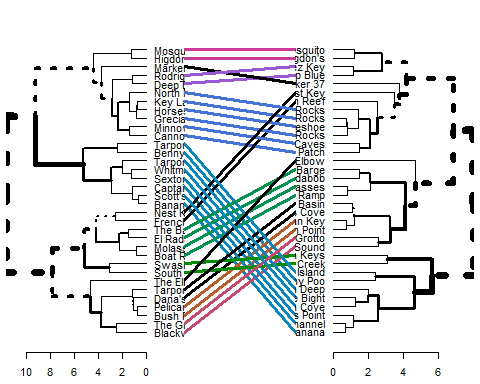
Based on this dendrogram it looks like the most similar sites visited by MarineLab in terms of water quality data are: Key Largo Dry Rocks & Grecian Dry Rocks and Scott’s Channel & Banana. The former pair are two sites which are adjacent to one another on the reef line so this is expected. The latter pair are highly different sites both geograpically and biologically so this is an interesting match. Looking at differences there are no individual sites which stand out a distinctly different from others within the data set. The two major divisions of the dendrogram (blue/purple and red/yellow/green) show a little favoritism between Seagrass/Mangrove and not. The majority of Seagrass/Mangrove sites are found in the red/yellow/green branch whereas the majority of Coral reefs are in the blue/purple division however, this situation is not consistent for all sites within each type. Patch Reef/Hardbottom sites appear sporadically in the clusters.

We can check the similarity of these two dendrograms by using a tanglegram. A tanglegram compares the alignment of two dendrograms (the must have the same information and number of samples) to show how entangled they are. The greater the entanglemnt (closer to 1) the less similar the two dendrograms are. Matching nodes in each dendrogram are connected by a colored line which corresponds to the cluster location, unmatched nodes are black, these represent disagreement between the two dendrograms.

p <- as.dendrogram(cluster\_site\_agnes)  
q <- as.dendrogram(cluster\_site\_hclust)  
  
all\_site\_dend <- dendlist(q, p)  
  
  
xx <- tanglegram(cluster\_site\_hclust, cluster\_site\_agnes, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = FALSE, lwd = 2, k\_branches = 5, lab.cex = 0.7, main = paste("All Sites (hclust vs agnes) Entanglement =", round(entanglement(all\_site\_dend), 2)))



all\_site\_dend %>% untangle("step1side") %>% tanglegram()

 From the tanglegram it looks like most of our sites are in agreement between the two methods. Most of the disagreement appears to be within the 5 major clusters (i.e. disagreement about where in the cluster exaclty the site belongs) which are relatively minor. The only large difference is the Elbow site which jumps clustering with the different alogrithms, it will be important to keep an eye on this location.

### Unsupervised Modeling Discussion

The unsupervised method of modeling allows us to see data clusters that occur as a result of the input variables. This type of analysis is very useful for datasets like ours where no defined outcome is a componet. From this type of analysis we can determine which sites (or other assigned criteria) are most similar across the dataset due to changes in the input variables.

However, this analysis is not without limiation. While the clustering provides important information, clustering does not explain why the data is grouped in a particular pattern. Water quality measuares feed into the model which in turn influence the decisions of the clustering however, this type of model cannot account for any external factors (unless previously controlled) which influence the variables (such as seasonality and annual variance). In the absence of a visible pattern of clustering attributable to some known characteristic of the sites (i.e. type, depth, season, distance from shore, etc) the cluster groups are relatively uninfomative beyond showing which are the most similar. This is an issue we have seen with our analysis above. Small patterns of site data can be seen within individual branches however, no overall pattern of clustering can be seen which applies to all sites.

This result is not entirely uninformative, this suggests that site clustering may be strongly affected by additional factors (which we will explore below) and/or water quality parameters are simply not powerful enough to distinguish these sites either entirely or through the usage of the tools employed by MarineLab.

To explore some more depths of our dataset we will subset the cleaned data to target specific aspects of the analyis in effort to visualize additional patterns which cannot be seen using the total data. These results are shown in the sections below.

## Site Type Clustering

Now that we have the overall dendrogram for each site lets compare our three site types individuall and see how different each of them are from one another. For this analysis we will use average values for each parameter as we did above and compare clustering for Seagrass/Mangrove, Patch Reef/Hardbottom, and Coral Reef sites respectively.

### Questions for this section

*Which of the three site types (Seagrass/Mangrove, Patch Reef/Hardbottom, or Coral Reef) are the most homogeneous and heterogeneous?*

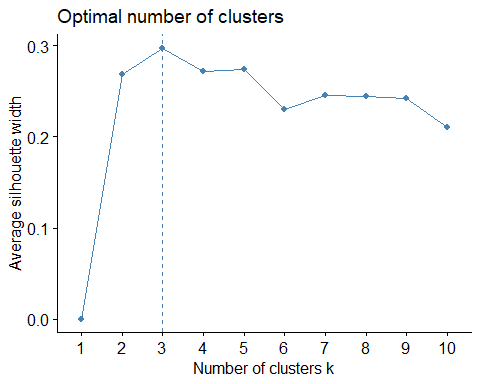
*Are there any notable site patterns which might explain why the clusters are distributed in this manner?*

First we need to subset data from the clean data, select our predictors, remove any NAs in the predictors, and finally aggregate the means of each site.

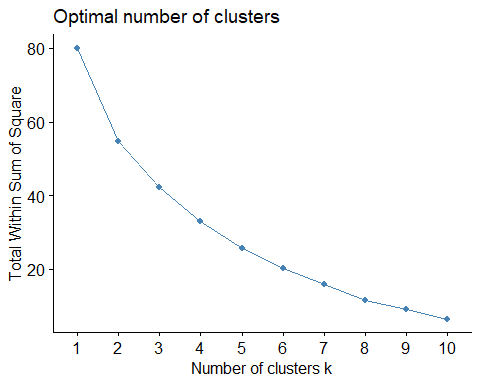
#Subset the sites  
sg\_mg\_subset <- subset(WQ\_clean\_data, site\_type == "Seagrass/Mangrove")  
#We will drop the same observations again as we did in the above cluster analysis  
sg\_mg\_redrop1 <- sg\_mg\_subset[!sg\_mg\_subset$location == "Buttonwood", ]  
sg\_mg\_redrop2 <- sg\_mg\_redrop1[!sg\_mg\_redrop1$location == "Lagoon", ]  
sg\_mg\_redrop3 <- sg\_mg\_redrop2[!sg\_mg\_redrop2$location == "Hidden Lake", ]  
  
pr\_h\_subset <- subset(WQ\_clean\_data, site\_type == "Patch Reef/Hardbottom")  
cr\_subset <- subset(WQ\_clean\_data, site\_type == "Coral Reef")  
  
#Reduce the data to the predictors of interest  
sg\_mg\_reduce <- sg\_mg\_redrop3 %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
pr\_h\_reduce <- pr\_h\_subset %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
cr\_reduce <- cr\_subset %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
  
#Drop any NAs  
sg\_mg\_ready <- na.omit(sg\_mg\_reduce)  
pr\_h\_ready <- na.omit(pr\_h\_reduce)  
cr\_ready <- na.omit(cr\_reduce)  
  
#Aggregate to get the mean value for each variable at each location  
sg\_mg\_ready <- aggregate(. ~ location, sg\_mg\_ready, mean)  
pr\_h\_ready <- aggregate(. ~ location, pr\_h\_ready, mean)  
cr\_ready <- aggregate(. ~ location, cr\_ready, mean)  
  
#Assign rownames as the site names  
rownames(sg\_mg\_ready) <- sg\_mg\_ready$location  
rownames(pr\_h\_ready) <- pr\_h\_ready$location  
rownames(cr\_ready) <- cr\_ready$location  
  
#Drop the location variable then scale the data to normalize and finally convert back to a data frame  
sg\_drop\_location <- sg\_mg\_ready %>% dplyr::select(-location)  
sg\_scale <- scale(sg\_drop\_location)  
sg\_mg\_ready <- as.data.frame(sg\_scale)  
  
pr\_h\_drop\_location <- pr\_h\_ready %>% dplyr::select(-location)  
pr\_h\_scale <- scale(pr\_h\_drop\_location)  
pr\_h\_ready <- as.data.frame(pr\_h\_scale)  
  
cr\_drop\_location <- cr\_ready %>% dplyr::select(-location)  
cr\_scale <- scale(cr\_drop\_location)  
cr\_ready <- as.data.frame(cr\_scale)

Now we can make our dendrograms, we will start with Seagrass/Mangrove. First we will determine our optimal number of clusters for analysis.

fviz\_nbclust(sg\_mg\_ready, FUN = hcut, method = "silhouette")

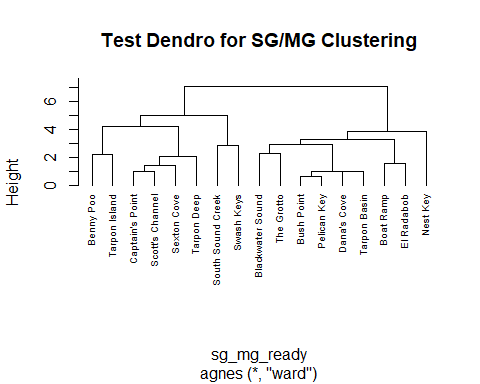


fviz\_nbclust(sg\_mg\_ready, FUN = hcut, method = "wss")

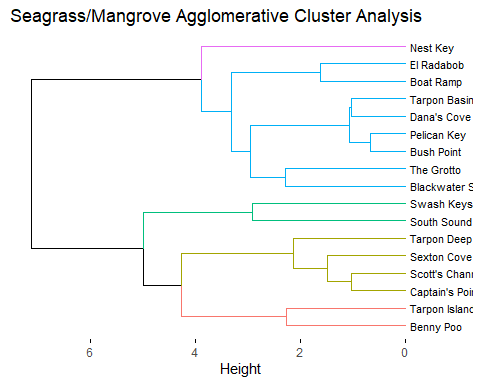


Five clusters it is! Now we can make our good dendrogram!

set.seed(1111)  
  
sg\_mg\_clust <- agnes(sg\_mg\_ready, method = "ward")  
  
#Note this code produces a raw dendrogram where as the fviz\_dend function allows for tuning and modification for publication/manuscript usage  
sg\_mg\_dend <- pltree(sg\_mg\_clust, cex = 0.6, hang = -1, main = "Test Dendro for SG/MG Clustering")



sg\_mg\_good\_dend <- fviz\_dend(sg\_mg\_clust, k = 5, show\_labels = TRUE, cex = 0.6, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Seagrass/Mangrove Agglomerative Cluster Analysis")  
  
sg\_mg\_good\_dend

 It looks like our Seagrass/Mangrove sites are reasonably distinct! It is unsprising that Nest Key is given a distinct cluster due to the fact that this site is very geographically different from all other sampling locations in the data set. Beyond this seperation, there are not noticable distinct patterns in the data. The lack of notable correlation may be due to the fact that Seagrass/Mangrove sites tend to be the most variable in terms of conditions and seasonality may be a stronger influencer of water quality parameters.

Lets take a look at patch reefs and hardbottom sites.

Cluster evaluation with the silhouette and wss method (as we have done for all samples) is not possible for the Patch Reef/Hardbottom data due to the fact that the number of observations is too low to use the function. We will use the NbClust function to compute the silhouette index and determine the best number of clusters. Note this method will not produce a silhouette curve but will still provide the ideal cluster number in the output.

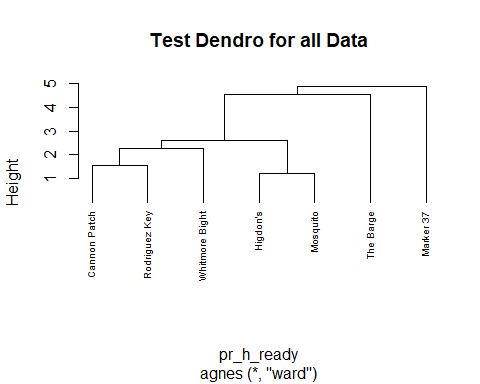
#A min and max number of clusters must be supplied in this function. We will use a 2 min to ensure we get some informaiton and a max of 6 as the highest max must be n-1  
pr\_h\_diss <- NbClust(pr\_h\_ready, distance = "euclidean", min.nc = 2, max.nc = 6, method = "ward.D2", index = "silhouette")  
  
#The code below provides the best number of columns, silhouette index, and partition of observations into clusters  
pr\_h\_diss$Best.nc

## Number\_clusters Value\_Index   
## 6.0000 0.7876

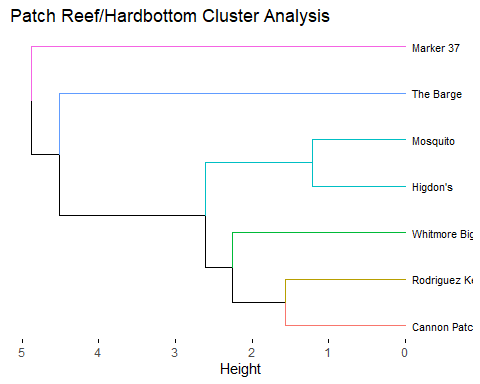
pr\_h\_diss$Best.partition

## Cannon Patch Higdon's Marker 37 Mosquito Rodriguez Key   
## 1 2 3 2 4   
## The Barge Whitmore Bight   
## 5 6

set.seed(1111)  
  
pr\_h\_clust <- agnes(pr\_h\_ready, method = "ward")  
  
pr\_h\_dend <- pltree(pr\_h\_clust, cex = 0.6, hang = -1, main = "Test Dendro for all Data", labels = pr\_h\_ready$location)



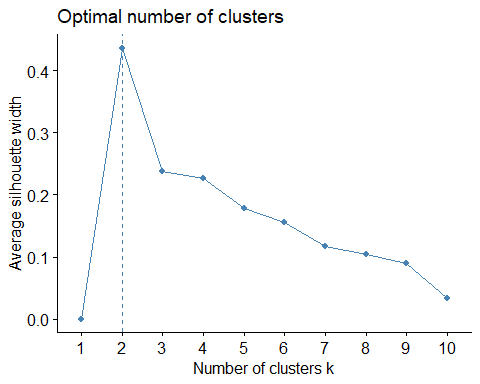
pr\_h\_good\_dend <- fviz\_dend(pr\_h\_clust, k = 6, show\_labels = TRUE, cex = 0.6, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Patch Reef/Hardbottom Cluster Analysis")  
  
pr\_h\_good\_dend



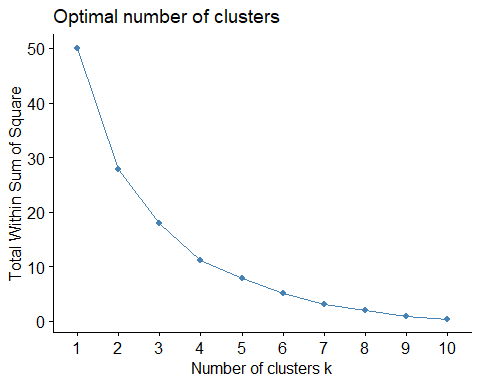
It looks like our Patch Reef/Hardbottom sites are very distinct across the majority of samples. This is likely due to the fact that all of these sites are distant from each other, this also my be sue to sampling bias as these sites are visited the lest frequently of all the three site types due to weather inaccessibility.

Lastly lets investigate Coral Reefs.

fviz\_nbclust(cr\_ready, FUN = hcut, method = "silhouette")

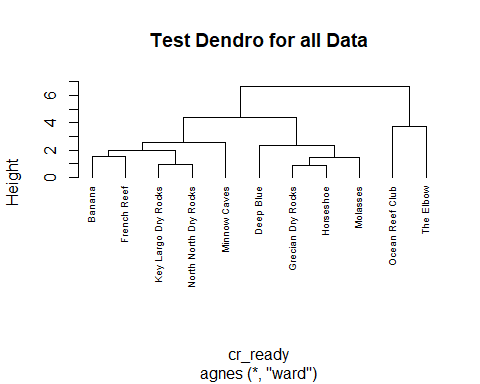


fviz\_nbclust(cr\_ready, FUN = hcut, method = "wss")

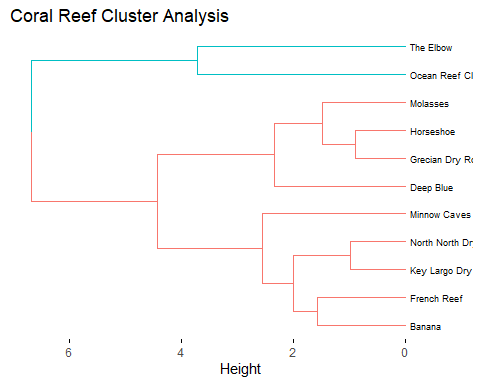


Looks like ideal clustering is a bit less obvious this time around. We will use two for our analysis.

set.seed(1111)  
  
cr\_clust <- agnes(cr\_ready, method = "ward")  
  
cr\_dend <- pltree(cr\_clust, cex = 0.6, hang = -1, main = "Test Dendro for all Data", labels = cr\_ready$location)



cr\_good\_dend <- fviz\_dend(cr\_clust, k = 2, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Coral Reef Cluster Analysis")  
  
cr\_good\_dend



Coral Reef clustering shows two distinct clusters, the The Elbow and Ocean Reef Club represent the two northernmost sites in our data set so it is interesting to see them seperated from the remained of the reef. All other sites are fairly variable. There is so clustering that appears to be geographically inclied (Key Kargo Dry Rocks & North North Dry Rocks as well as French Reef & Banana). Additionally, the clustering of Horseshoe and Grecian rocks may be due to depth similarity.

Overall, our site clustering did not show a consistent pattern of grouping across all samples of site types. Notably, the most similar/dissimilar sites in the early clusters often can be linked by a similar characteristic such as: geographical location, similar biota, water depth, etc.

For Seagrass/Mangrove our most similar site pair was Pelican Key and Bush Point. These two sites are quite dissimilar from each other interms of location and biota, however they are the most favorable sites for winter wind direction. It is possible this clustering is due to a bias for winter sampling. The most distinct site was Nest Key which is geographically the furthest Seagras/Mangrove site from any of the other sites in the data set.

For Patch Reef/Hardbottom there was a wide range of dissimilarirty. The most similar sites were Higdon’s and Mosquito. These two sites are almost exactly the same distance from shore and as a result have very similar coral biota as the primary reef structure. In terms of distance, Marker 37 is the most distinct site. Marker 37 is geographically distant from many of the other sites (and is probably closer to a Coral Reef location than other Patch Reefs). It should be noted that many sites were fairly distinct in this sampling set which shows that these locations are the most variable. This may be in part due to substantially lower sampling values for these sites due to inaccessibility.

For Coral Reefs the most similar site pairs are Horseshoe & Grecian Rocks as well as North North Dry Rocks (NNDR) and Key Largo Dry Rocks (KLDR). NNDR and KLDR are two reef sites which are adjacent to each other on the reef line so this result is unsuprising. Horseshoe and Grecian are not as close (Grecian is actually closest to KLDR) however these two sites are highly similar in depth (shallower) which may explain this clustering. The most distinct sites in this subset are Ocean Reef Club and The Elbow both of which are the northernmost sites for all Coral Reef locations.

## Year-to-Year Comparison

### Questions for this section

*How does our cluster analysis compare across sampling years?*

We will make a comparison across years using a tanglegram to see how clustering of invividual site locations compares across the ~4 years of study.

First we must subset our data into years and normalize our locations so we only have repedative locations for each dendrogram.

#Subset into individual years  
year\_16 <- subset(WQ\_clean\_data, Year == "16")  
year\_17 <- subset(WQ\_clean\_data, Year == "17")  
year\_18 <- subset(WQ\_clean\_data, Year == "18")  
year\_19 <- subset(WQ\_clean\_data, Year == "19")  
  
#Remove NAs  
year\_16 <- na.omit(year\_16)  
year\_17 <- na.omit(year\_17)  
year\_18 <- na.omit(year\_18)  
year\_19 <- na.omit(year\_19)  
  
#Select predictors  
year\_16 <- year\_16 %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
year\_17 <- year\_17 %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
year\_18 <- year\_18 %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
year\_19 <- year\_19 %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
#Aggregate the sample sites by mean of each water quality variable  
year\_16 <- aggregate(. ~ location, year\_16, mean)  
year\_17 <- aggregate(. ~ location, year\_17, mean)  
year\_18 <- aggregate(. ~ location, year\_18, mean)  
year\_19 <- aggregate(. ~ location, year\_19, mean)  
  
#Assign the location to the rowname  
rownames(year\_16) <- year\_16$location  
rownames(year\_17) <- year\_17$location  
rownames(year\_18) <- year\_18$location  
rownames(year\_19) <- year\_19$location

#Remove all rows that are not present in all four years of study  
year\_16 <- year\_16[!year\_16$location == "Scott's Channel", ]  
year\_16 <- year\_16[!year\_16$location == "Sexton Cove", ]  
year\_16 <- year\_16[!year\_16$location == "Ocean Reef Club", ]  
year\_16 <- year\_16[!year\_16$location == "North North Dry Rocks", ]  
year\_16 <- year\_16[!year\_16$location == "NA", ]  
year\_16 <- year\_16[!year\_16$location == "Higdon's", ]  
year\_16 <- year\_16[!year\_16$location == "The Elbow", ]  
year\_16 <- year\_16[!year\_16$location == "The Barge", ]  
year\_16 <- year\_16[!year\_16$location == "Bayside", ]  
year\_16 <- year\_16[!year\_16$location == "Blackwater Sound", ]  
year\_16 <- year\_16[!year\_16$location == "Benny Poo", ]  
year\_16 <- year\_16[!year\_16$location == "French Reef", ]  
year\_16 <- year\_16[!year\_16$location == "Deep Blue", ]  
year\_16 <- year\_16[!year\_16$location == "Marker 37", ]  
year\_16 <- year\_16[!year\_16$location == "Lagoon", ]  
year\_16 <- year\_16[!year\_16$location == "Tarpon Deep", ]  
year\_16 <- year\_16[!year\_16$location == "Mosquito", ]  
  
#Remove the location variable, scale the data to normalize, and return back to a data frame  
year\_16 <- year\_16 %>% dplyr::select(-location)  
year\_16 <- scale(year\_16)  
year\_16 <- as.data.frame(year\_16)

year\_17 <- year\_17[!year\_17$location == "Scott's Channel", ]  
year\_17 <- year\_17[!year\_17$location == "Sexton Cove", ]  
year\_17 <- year\_17[!year\_17$location == "Ocean Reef Club", ]  
year\_17 <- year\_17[!year\_17$location == "North North Dry Rocks", ]  
year\_17 <- year\_17[!year\_17$location == "NA", ]  
year\_17 <- year\_17[!year\_17$location == "Higdon's", ]  
year\_17 <- year\_17[!year\_17$location == "The Elbow", ]  
year\_17 <- year\_17[!year\_17$location == "The Barge", ]  
year\_17 <- year\_17[!year\_17$location == "Bayside", ]  
year\_17 <- year\_17[!year\_17$location == "Blackwater Sound", ]  
year\_17 <- year\_17[!year\_17$location == "Benny Poo", ]  
year\_17 <- year\_17[!year\_17$location == "French Reef", ]  
year\_17 <- year\_17[!year\_17$location == "Deep Blue", ]  
year\_17 <- year\_17[!year\_17$location == "Marker 37", ]  
year\_17 <- year\_17[!year\_17$location == "Lagoon", ]  
year\_17 <- year\_17[!year\_17$location == "Tarpon Deep", ]  
year\_17 <- year\_17[!year\_17$location == "Mosquito", ]  
year\_17 <- year\_17[!year\_17$location == "Swash Keys", ]  
year\_17 <- year\_17[!year\_17$location == "Captain's Point", ]  
  
year\_17 <- year\_17 %>% dplyr::select(-location)  
year\_17 <- scale(year\_17)  
year\_17 <- as.data.frame(year\_17)

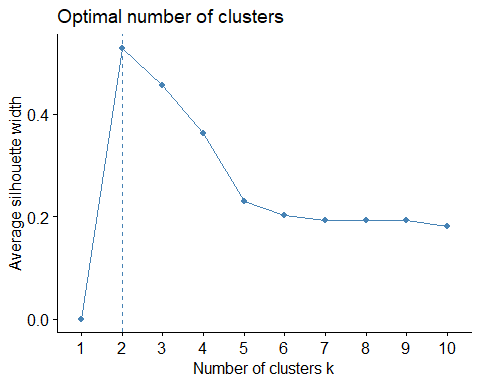
year\_18 <- year\_18[!year\_18$location == "Scott's Channel", ]  
year\_18 <- year\_18[!year\_18$location == "Sexton Cove", ]  
year\_18 <- year\_18[!year\_18$location == "Ocean Reef Club", ]  
year\_18 <- year\_18[!year\_18$location == "North North Dry Rocks", ]  
year\_18 <- year\_18[!year\_18$location == "NA", ]  
year\_18 <- year\_18[!year\_18$location == "Higdon's", ]  
year\_18 <- year\_18[!year\_18$location == "The Elbow", ]  
year\_18 <- year\_18[!year\_18$location == "The Barge", ]  
year\_18 <- year\_18[!year\_18$location == "Bayside", ]  
year\_18 <- year\_18[!year\_18$location == "Blackwater Sound", ]  
year\_18 <- year\_18[!year\_18$location == "Benny Poo", ]  
year\_18 <- year\_18[!year\_18$location == "French Reef", ]  
year\_18 <- year\_18[!year\_18$location == "Deep Blue", ]  
year\_18 <- year\_18[!year\_18$location == "Marker 37", ]  
year\_18 <- year\_18[!year\_18$location == "Lagoon", ]  
year\_18 <- year\_18[!year\_18$location == "Tarpon Deep", ]  
year\_18 <- year\_18[!year\_18$location == "Mosquito", ]  
year\_18 <- year\_18[!year\_18$location == "Swash Keys", ]  
year\_18 <- year\_18[!year\_18$location == "Buttonwood", ]  
year\_18 <- year\_18[!year\_18$location == "Captain's Point", ]  
  
year\_18 <- year\_18 %>% dplyr::select(-location)  
year\_18 <- scale(year\_18)  
year\_18 <- as.data.frame(year\_18)

year\_19 <- year\_19[!year\_19$location == "Scott's Channel", ]  
year\_19 <- year\_19[!year\_19$location == "Sexton Cove", ]  
year\_19 <- year\_19[!year\_19$location == "Ocean Reef Club", ]  
year\_19 <- year\_19[!year\_19$location == "North North Dry Rocks", ]  
year\_19 <- year\_19[!year\_19$location == "NA", ]  
year\_19 <- year\_19[!year\_19$location == "Higdon's", ]  
year\_19 <- year\_19[!year\_19$location == "The Elbow", ]  
year\_19 <- year\_19[!year\_19$location == "The Barge", ]  
year\_19 <- year\_19[!year\_19$location == "Bayside", ]  
year\_19 <- year\_19[!year\_19$location == "Blackwater Sound", ]  
year\_19 <- year\_19[!year\_19$location == "Benny Poo", ]  
year\_19 <- year\_19[!year\_19$location == "French Reef", ]  
year\_19 <- year\_19[!year\_19$location == "Deep Blue", ]  
year\_19 <- year\_19[!year\_19$location == "Marker 37", ]  
year\_19 <- year\_19[!year\_19$location == "Lagoon", ]  
year\_19 <- year\_19[!year\_19$location == "Tarpon Deep", ]  
year\_19 <- year\_19[!year\_19$location == "Mosquito", ]  
year\_19 <- year\_19[!year\_19$location == "Swash Keys", ]  
year\_19 <- year\_19[!year\_19$location == "Buttonwood", ]  
year\_19 <- year\_19[!year\_19$location == "Tarpon Island", ]  
year\_19 <- year\_19[!year\_19$location == "South Sound Creek", ]  
year\_19 <- year\_19[!year\_19$location == "Hidden Lake", ]  
year\_19 <- year\_19[!year\_19$location == "Captain's Point", ]  
  
year\_19 <- year\_19 %>% dplyr::select(-location)  
year\_19 <- scale(year\_19)  
year\_19 <- as.data.frame(year\_19)

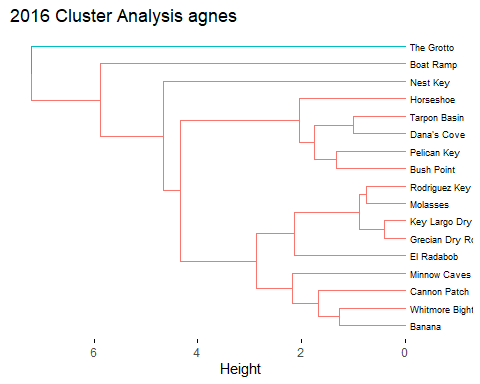
Phew lots of sorting, now we can start on our dendros.

We will make 2016 first.

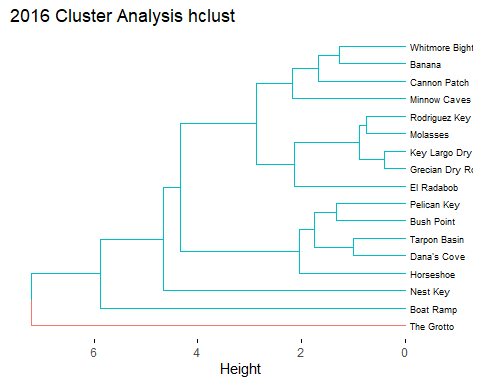
set.seed(1111)  
  
fviz\_nbclust(year\_16, FUN = hcut, method = "silhouette")



clust\_16 <- agnes(year\_16, method = "ward")  
  
good\_adend\_16 <- fviz\_dend(clust\_16, k = 2, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2016 Cluster Analysis agnes")  
  
good\_adend\_16

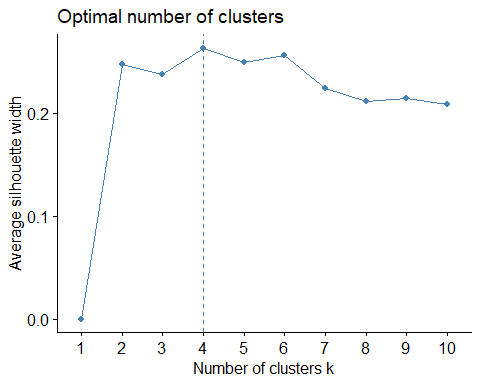


dist\_16 <- dist(year\_16, method = "euclidean")  
hclust\_16 <- hclust(dist\_16, method = "ward.D2")  
dend\_ready\_16 <- as.dendrogram(hclust\_16)  
  
good\_hdend\_16 <- fviz\_dend(hclust\_16, k = 2, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2016 Cluster Analysis hclust")  
good\_hdend\_16

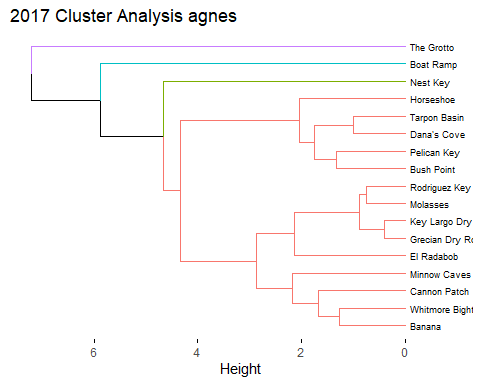
 Looks like the Grotto is very different from all other measures this year. Similarly, the boat ramp and Nest Key are quite distinct also.

Now for 2017!

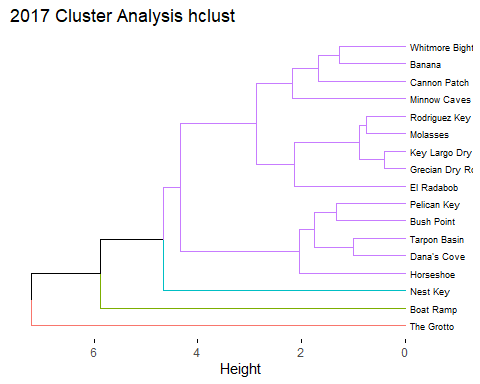
set.seed(1111)  
  
fviz\_nbclust(year\_17, FUN = hcut, method = "silhouette")



clust\_17 <- agnes(year\_17, method = "ward")  
  
good\_adend\_17 <- fviz\_dend(clust\_16, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2017 Cluster Analysis agnes")  
  
good\_adend\_17

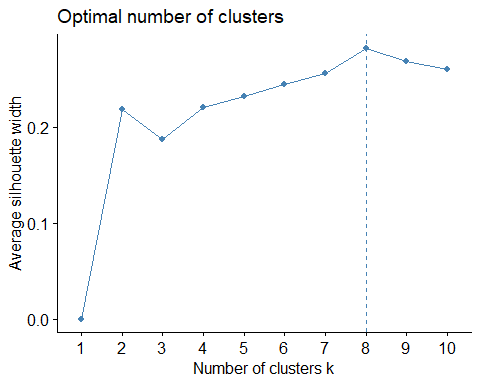


dist\_17 <- dist(year\_17, method = "euclidean")  
hclust\_17 <- hclust(dist\_17, method = "ward.D2")  
dend\_ready\_17 <- as.dendrogram(hclust\_17)  
  
good\_hdend\_17 <- fviz\_dend(hclust\_16, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2017 Cluster Analysis hclust")  
  
good\_hdend\_17

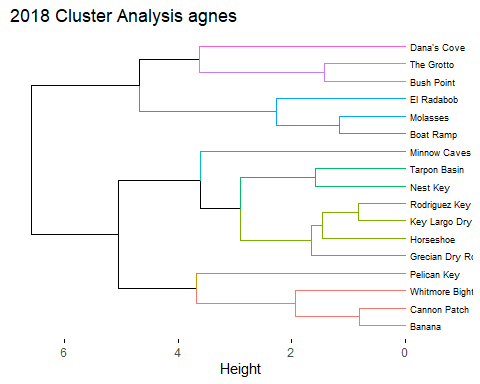
 2017 looks very similar to 2016, structurally with The Grotto, Boat Ramp, and Nest Key showing as the most distant from all other locations.

Next 2018

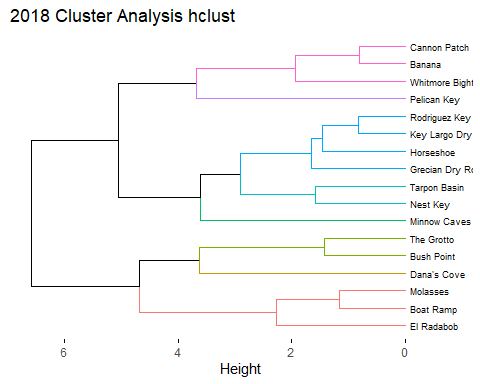
set.seed(1111)  
  
fviz\_nbclust(year\_18, FUN = hcut, method = "silhouette")



clust\_18 <- agnes(year\_18, method = "ward")  
  
good\_adend\_18 <- fviz\_dend(clust\_18, k = 8, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2018 Cluster Analysis agnes")  
  
good\_adend\_18



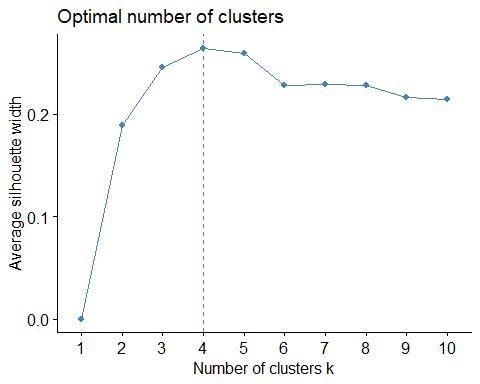
dist\_18 <- dist(year\_18, method = "euclidean")  
hclust\_18 <- hclust(dist\_18, method = "ward.D2")  
dend\_ready\_18 <- as.dendrogram(hclust\_18)  
  
good\_hdend\_18 <- fviz\_dend(hclust\_18, k = 8, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2018 Cluster Analysis hclust")  
  
good\_hdend\_18



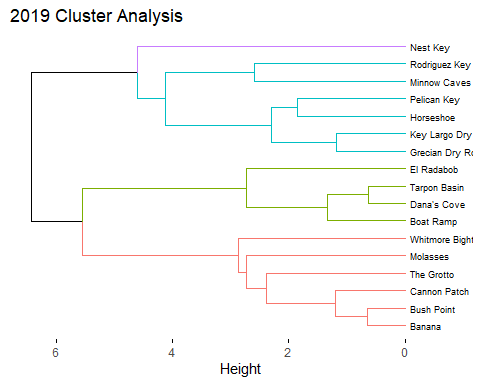
2018 is the most unique so far however, no distinct patterns are notable.

And finally 2019

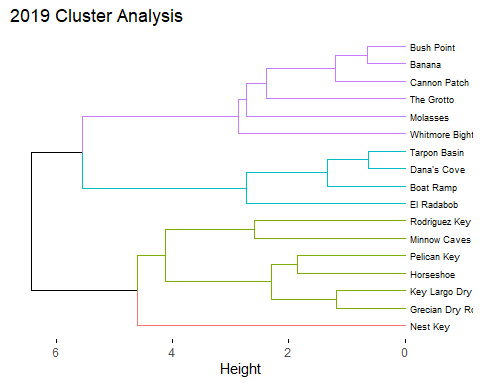
set.seed(1111)  
  
fviz\_nbclust(year\_19, FUN = hcut, method = "silhouette")



clust\_19 <- agnes(year\_19, method = "ward")  
  
good\_adend\_19 <- fviz\_dend(clust\_19, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2019 Cluster Analysis")  
  
good\_adend\_19



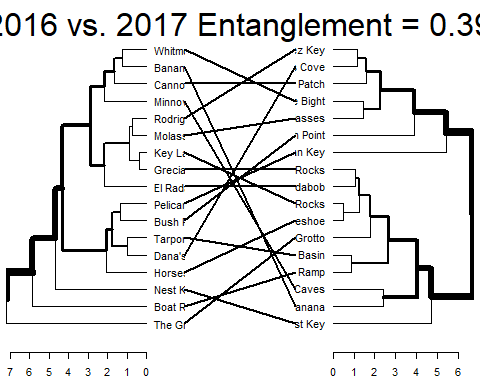
dist\_19 <- dist(year\_19, method = "euclidean")  
hclust\_19 <- hclust(dist\_19, method = "ward.D2")  
dend\_ready\_19 <- as.dendrogram(hclust\_19)  
  
good\_hdend\_19 <- fviz\_dend(hclust\_19, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "2019 Cluster Analysis")  
  
good\_hdend\_19



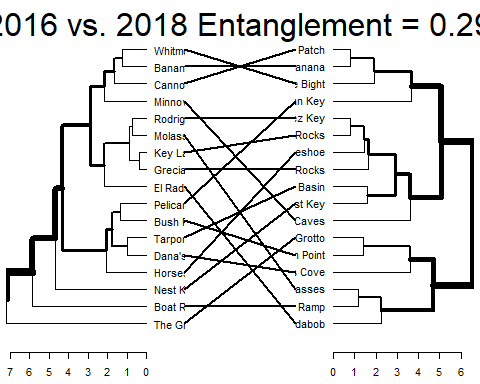
2019 looks the most consistent with geographical location as the previous years however, this pattern is not seen in all sites.

Now we can build our tanglegrams. Tanglegrams compare the alignment of two dendrograms. The entanglement coefficent represents the degree by which the two samples are entangled (or improperly aligned). A low entanglemnt shows a good alignment whereas a high entanglement shows disalignment.

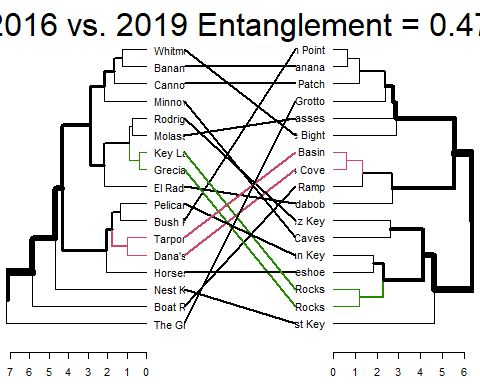
entang\_16\_17 <- dendlist(dend\_ready\_16, dend\_ready\_17)  
  
tangle\_16\_17 <- tanglegram(dend\_ready\_16, dend\_ready\_17, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2016 vs. 2017 Entanglement =", round(entanglement(entang\_16\_17), 2)))



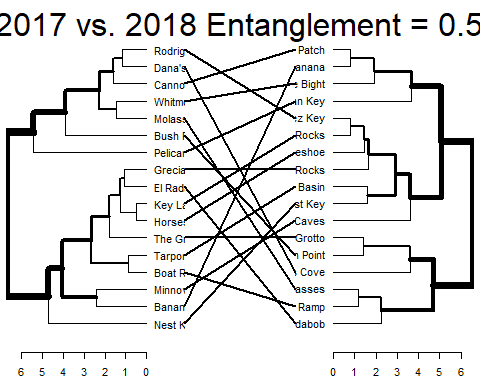
entang\_16\_18 <- dendlist(dend\_ready\_16, dend\_ready\_18)  
  
tangle\_16\_18 <- tanglegram(dend\_ready\_16, dend\_ready\_18, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2016 vs. 2018 Entanglement =", round(entanglement(entang\_16\_18), 2)))



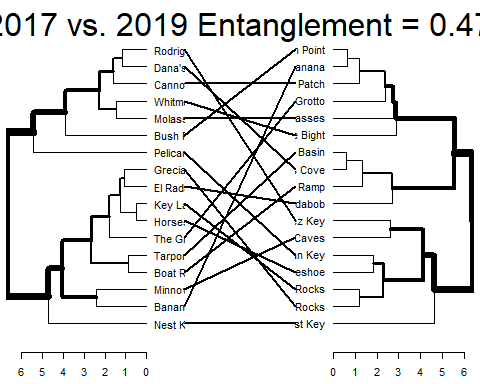
entang\_16\_19 <- dendlist(dend\_ready\_16, dend\_ready\_19)  
  
tangle\_16\_19 <- tanglegram(dend\_ready\_16, dend\_ready\_19, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2016 vs. 2019 Entanglement =", round(entanglement(entang\_16\_19), 2)))



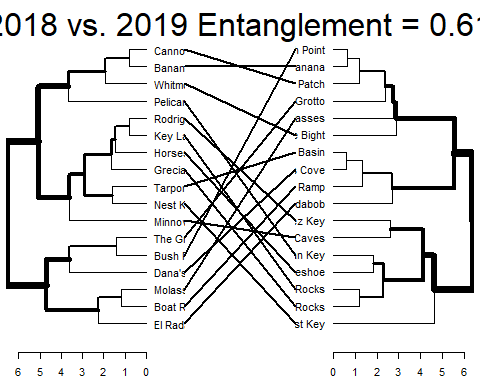
entang\_17\_18 <- dendlist(dend\_ready\_17, dend\_ready\_18)  
  
tangle\_17\_18 <- tanglegram(dend\_ready\_17, dend\_ready\_18, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2017 vs. 2018 Entanglement =", round(entanglement(entang\_17\_18), 2)))



entang\_17\_19 <- dendlist(dend\_ready\_17, dend\_ready\_19)  
  
tangle\_17\_19 <- tanglegram(dend\_ready\_17, dend\_ready\_19, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2017 vs. 2019 Entanglement =", round(entanglement(entang\_17\_19), 2)))



entang\_18\_19 <- dendlist(dend\_ready\_18, dend\_ready\_19)  
  
tangle\_18\_19 <- tanglegram(dend\_ready\_18, dend\_ready\_19, highlight\_distinct\_edges = FALSE, common\_subtrees\_color\_branches = TRUE, lwd = 2, lab.cex = 1, main = paste("2018 vs. 2019 Entanglement =", round(entanglement(entang\_18\_19), 2)))



Based on the tanglegrams and entanglement measures it seems like MarineLab is not very consistent with measurement across multiple years. Only one of the tanglegrams (2016 vs. 2019) shows any alignment of nodes. Notably some of the greatest entanglement is seen in tanglegrams containing 2019, this my be the result of the fact that this data set ended before the fall of this year.

While consistency may be a goal of sampling, it is important to note that this analysis does not account for any seasonal variation between sites and could show skewed values if a particualr site was only visited during a certian part of the year for one or more of the data subsets. Additionally, inconsistancy may also be a result of improving methods, training, or equipment over time.

## Seasonal Clustering

### Questions for this section

*Which of the three representitive sites is more strongly influenced by seasonal changes*

Based on the poor entanglement measures of year-to-year comparison it may be more prudent to cluster data based on month collected rather than simply measuring specific sites across years due to annual and seasonal fluctuation.

For this analysis we will use our most frequently visted sites for comparison of each of the three site types. Seagrass/Mangrove will be represented by Tarpon Basin, Patch Reef, Hardbottom will be represented by Rodriguez Key and Coral Reefs will be represented by Grecian Dry Rocks.

First we need to make our subsets.

tarpon\_only <- subset(WQ\_clean\_data, location == "Tarpon Basin")  
grecian\_only <- subset(WQ\_clean\_data, location == "Grecian Dry Rocks")  
roddy\_only <- subset(WQ\_clean\_data, location == "Rodriguez Key")

Next as above, we must clean our data, aggregate by means of Month, assin rownames, scale, and prep the dataframe.

tarpon\_reduce <- tarpon\_only %>% dplyr::select(Month, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
tarpon\_mean <- aggregate(. ~ Month, tarpon\_reduce, mean)  
  
rownames(tarpon\_mean) <- tarpon\_mean$Month  
  
tarpon\_drop\_month <- tarpon\_mean %>% dplyr::select(-Month)  
  
tarpon\_scale <- scale(tarpon\_drop\_month)  
  
tarpon\_ready <- as.data.frame(tarpon\_scale)

Since these dendrograms will have less than 10 total bins we will need to modify our silhouete code as above using the NbClust. The fviz\_nbclust function cannot handle datasets with less than 10 observations in the format as above, additionall we cannot produce the silhouette curves however, we can calculate the values.

library(NbClust)  
#In this function we must specify a min and mx number of clusters, we will use a min of two to get some information and a max of four to see if we can get seasonal clustering  
tarpon\_diss <-NbClust(tarpon\_ready, distance = "euclidean", min.nc = 2, max.nc = 4,   
 method = "ward.D2", index = "silhouette")  
  
#The best number of clusters calcuated by the value index  
tarpon\_diss$Best.nc

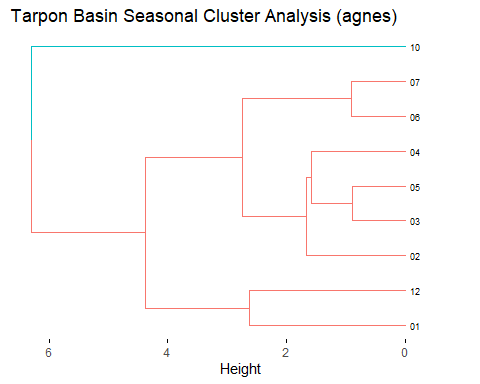
## Number\_clusters Value\_Index   
## 2.0000 0.5904

#the partition of observations into their respective clusters  
tarpon\_diss$Best.partition

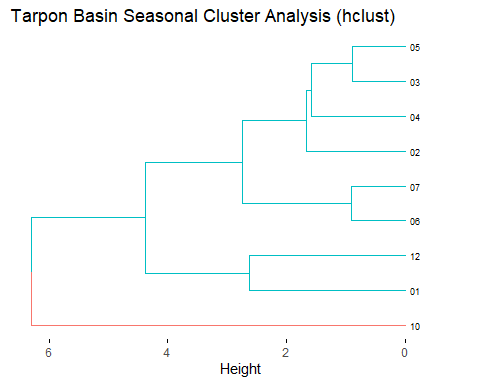
## 01 02 03 04 05 06 07 10 12   
## 1 1 1 1 1 1 1 2 1

Next we make the dendgrograms as above.

set.seed(1111)  
  
  
clust\_tarpon\_agnes <- agnes(tarpon\_ready, method = "ward")  
  
good\_adend\_tarpon <- fviz\_dend(clust\_tarpon\_agnes, k = 2, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Tarpon Basin Seasonal Cluster Analysis (agnes)")  
  
good\_adend\_tarpon



dist\_tarpon <- dist(tarpon\_ready, method = "euclidean")  
hclust\_tarpon <- hclust(dist\_tarpon, method = "ward.D2")  
dend\_ready\_tarpon <- as.dendrogram(hclust\_tarpon)  
  
good\_hdend\_tarpon <- fviz\_dend(hclust\_tarpon, k = 2, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Tarpon Basin Seasonal Cluster Analysis (hclust)")  
  
good\_hdend\_tarpon

 It looks like our data groups relatively close to a seasonal timeline. The closest clusters are often adjacent or near months. It appears October is a very different month from the rest of the samples, this may be due to the fact that September and November observations are absent so it is our only representitive of fall.

Next we will look at Rodriguez Key

roddy\_reduce <- roddy\_only %>% dplyr::select(Month, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
roddy\_mean <- aggregate(. ~ Month, roddy\_reduce, mean)  
  
rownames(roddy\_mean) <- roddy\_mean$Month  
  
roddy\_drop\_month <- roddy\_mean %>% dplyr::select(-Month)  
  
roddy\_scale <- scale(roddy\_drop\_month)  
  
roddy\_ready <- as.data.frame(roddy\_scale)

roddy\_diss <- NbClust(roddy\_ready, distance = "euclidean", min.nc = 2, max.nc = 4,   
 method = "ward.D2", index = "silhouette")  
  
roddy\_diss$Best.nc

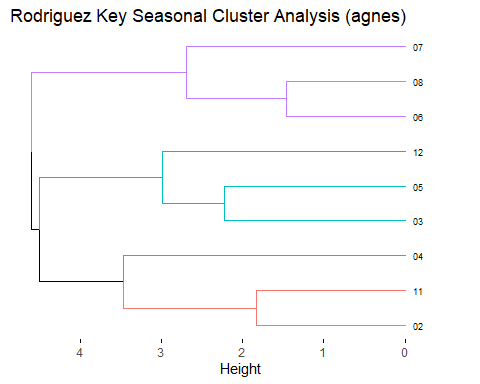
## Number\_clusters Value\_Index   
## 4.0000 0.2923

roddy\_diss$Best.partition

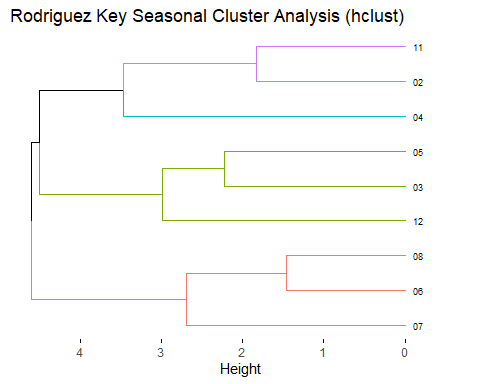
## 02 03 04 05 06 07 08 11 12   
## 1 2 3 2 4 4 4 1 2

Lookd likr four clusters is best here!

set.seed(1111)  
  
  
clust\_roddy\_agnes <- agnes(roddy\_ready, method = "ward")  
  
good\_adend\_roddy <- fviz\_dend(clust\_roddy\_agnes, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Rodriguez Key Seasonal Cluster Analysis (agnes)")  
  
good\_adend\_roddy



dist\_roddy <- dist(roddy\_ready, method = "euclidean")  
hclust\_roddy <- hclust(dist\_roddy, method = "ward.D2")  
dend\_ready\_roddy <- as.dendrogram(hclust\_roddy)  
  
good\_hdend\_roddy <- fviz\_dend(hclust\_roddy, k = 4, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Rodriguez Key Seasonal Cluster Analysis (hclust)")  
  
good\_hdend\_roddy

 So it looks like our summer clustering is fairly strong, however the remainder of the months are not grouped in any notable pattern.

Last we will look at Grecian Dry Rocks

grecian\_reduce <- grecian\_only %>% dplyr::select(Month, water\_temp, salinity, dissolved\_oxygen, ammonia, ph)  
  
grecian\_mean <- aggregate(. ~ Month, grecian\_reduce, mean)  
  
rownames(grecian\_mean) <- grecian\_mean$Month  
  
grecian\_drop\_month <- grecian\_mean %>% dplyr::select(-Month)  
  
grecian\_scale <- scale(grecian\_drop\_month)  
  
grecian\_ready <- as.data.frame(grecian\_scale)

grecian\_diss <- NbClust(grecian\_ready, distance = "euclidean", min.nc = 2, max.nc = 4, method = "ward.D2", index = "silhouette")  
  
grecian\_diss$Best.nc

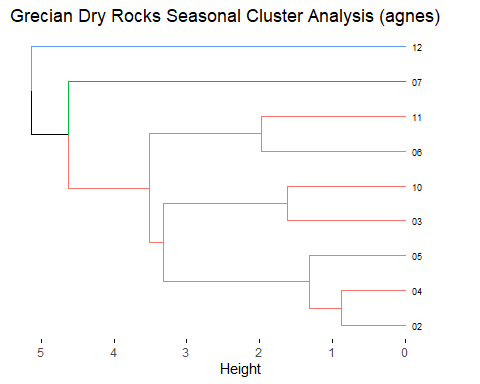
## Number\_clusters Value\_Index   
## 3.0000 0.5197

grecian\_diss$Best.partition

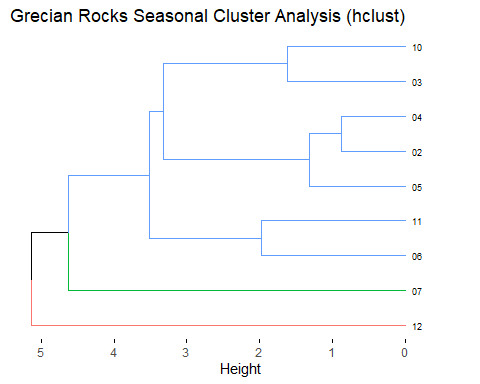
## 02 03 04 05 06 07 10 11 12   
## 1 1 1 1 1 2 1 1 3

Three clusters for this subset.

set.seed(1111)  
  
clust\_grecian\_agnes <- agnes(grecian\_ready, method = "ward")  
  
good\_adend\_grecian <- fviz\_dend(clust\_grecian\_agnes, k = 3, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Grecian Dry Rocks Seasonal Cluster Analysis (agnes)")  
  
good\_adend\_grecian



dist\_grecian <- dist(grecian\_ready, method = "euclidean")  
hclust\_grecian <- hclust(dist\_grecian, method = "ward.D2")  
dend\_ready\_grecian <- as.dendrogram(hclust\_grecian)  
  
good\_hdend\_grecian <- fviz\_dend(hclust\_grecian, k = 3, show\_labels = TRUE, cex = 0.5, rect = FALSE, color\_labels\_by\_k = FALSE, horiz = TRUE, rect\_lty = 5, rect\_border = 8, main = "Grecian Rocks Seasonal Cluster Analysis (hclust)")  
  
good\_hdend\_grecian



The clustering in this instance does not appear to show seasonal grouping. It should be noted that the two individual clusters July and December are the two most extreme seasonal months so it is possible their seperation may be due to seasonal extremes. However, the majority of months do not show a distinct pattern.

### Further Cluster Analysis Discussion

Through the use of hiearacical clustering and data subsets were were able to identify how closely related observations are to one another. Using site type cluster analysis we explored the distrubution of sampling sites under each of the three site types in the study. For Seagrass/Mangrove sites our data is fairly heterogeneous. These locaitons divided into 6 distinct clusters. No specific common attribute of the sites can explain the division pattern seen however, some notable differences and similarities appear to be correlated with geographic location (i.e. very different clusters contain sites which are far from all others whereas similar clusters are close in proximity). This pattern was not observed across all site locations. Path Reef/Hardbottom locations showed a very distinct spread across all of the sites with 6 clusters as ideal for a total of 7 sites. This suggests that these sites are the most distinct from one another. The two similar sites in the group Higdon’s and Mosquito have very similar coral biota and depth ranges compared to all others. Coral reef analysis shows the least clustering of all of the site types, primairly divinding into the northernmost reefs (The Elbow and Ocean Reef Club) and all of the remaining southern reefs. This suggests that coral reefs are the least heterogenous of the three site types which is consistant previous analysis which show the narrowest range of abiotic conditions.

*Which of the three site types (Seagrass/Mangrove, Patch Reef/Hardbottom, or Coral Reef) are the most homogeneous and heterogeneous?* In terms of water quality parameters the most heterogeneous site types were Seagrass/Mangrove and Patch Reef/Hardbottom. The most homogeneous are Coral Reefs. This observation is consistent with previous analyses which show a wider range of abiotic conditions in Seagrass/Mangrove sites and a narrower range in Coral Reefs. *Are there any notable site patterns which might explain why the clusters are distributed in this manner?* There is no discernable pattern that applies to all clustering situations that would explain the distribution. Large differences and samll differences may likely be attributable to site characteristics (geographic proximity, depth similarity, etc.) however, no one pattern explains all of the clustering.

Annual clustering was used to determine the similarity between subsequent years of sampling. Dendrograms of sampling sites which were measured all four years were subsetted and compared using a tanglegram to calculate the level of entanglement. Greater entanglement shows a greater level of dissimilarity between years of sampling. This analysis suggests that year-to-year measures of water quality are not very consistentusing the MarineLab sampling techniques. Though the analysis suggests this, it is important to understand the limiations of this technique. Entanglement is dependant entirely on the structure of the respective dendrograms which are in turn dependant on the water quality parameters which feed into the model. This method is limited in its ability to assess this type of question because it cannot account for differences in seasonal flucutation, annual fluctuation, equipment optimization, training optimization, or other confounding factors. As a result we will determine the quesitons of this subsection to be inconclusive due to model limitations. The usage of a tanglegram/entanglment would be an applicable method to compare different samling teams who visited the same sites at the same time of year (inter-rater relatibility).

*How does our cluster analysis compare across sampling years?* The results of cluster comparison across multiple years are inconclusive due to model restrictions. Upon deeper investigation of the entanglemnt calculations and model abilities, data is not comparable across multiple years due to the presence of confounding factors which cannot be eliminated using this modeling technique. The methodology would work as a comparison of inter-rater relatability for observers using the same methods.

Lastly, seasonal clustering investigated the grouping of Month groups when observed at the most frequently visited sampling sites of each of the three site types: Tarpon Basin, Grecian Dry Rocks, and Rodriguez Key. The results showed a seasonally organized clustering for Tarpon Basin which suggests that Seagrass/Mangrove sites are more strongly influenced by seasonal factors, this observation is consistent with previous analyses. Rodriguez Key showed a summer month cluster however the remaining clusters were not consistent with a seasonal pattern. Grecian Dry Rocks showed the least seasonally grouped clustering pattern of all three sites which suggests that coral reefs are least influenced by seasonal fluctuation.

*Which of the three representitive sites is more strongly influenced by seasonal changes* Based on month-wise clustering, Tarpon Basin is the site which is most strongly influenced by seasonal flucutation in water quality. This result suggests that Seagrass/MAngrove sites are strongly influenced by seasonal change. Patch Reefs/Hardbottom sites appear to be moderately influenced and Coral Reefs are influenced the least.

### Revisiting Analysis Questions

*How do sampling sites cluster based on individual water quality measurements?* Overall sampling sites do not cluster with any discernable pattern that is applicable for the entire data set. Individual subsets of the data can show specific clustering patterns that may be attributable to season, year, geographical distance, or local biota however; these atrributes are inferences based on individual dendrograms and do not apply to the entire data set.

*Are there any site-wise, annual, and/or seasonal patterns that can be identified based on cluster analysis?* Based on cluster analysis, it was observed that Seagrass/Mangrove sites are the most heterogeneous sites in terms of water quality. Patch Reefs/Hardbottom were moderate and Coral Reefs were the most homogenous. Additionally, Seagrass/Mangrove sites are the most influenced by seasonal fluctuation, Patch Reefs/Hardbottom are moderately influenced, and Coral Reefs are the least influenced.

### Pros and Cons of Hierarchical Clustering

*Pros* 1. Hierarchical cluster analysis is applicable to datasets without a predetermined outcome. This is ideal for exploratory data, monitoring, or taxanomic grouping research. 2. There are multiple options for analysis algorithms and methods which can be optimized for individual data sets. For this data set we experimented with a number of methods but found our optimal model approach used agglomerative clustering with Ward’s minimum variance. Optimal cluster number was determined using the silhouette method (though the Elbow method was equivlant in all cases as well). 3. Hierarchical clustering allows for a dynamic analysis approach to modeling. A large dendrogram can be made to visualize gross patterns in the data which can then be subsestted to investigate individual categories/parameters/timeframes of interest. *Cons* 1. Hierarchical clustering will only show the resulting clusters that observations have been assigned to, it does not provide any evidence as to why these clusters were desginated. 2. Dendrograms are not comparable between one another unless they conatin data which are replicates of the data in another dendrogram (i.e. they cannot control for confounding factors when comparing). 3. Cluster analysis (in general) is relatively slow compared to other modeling methods especially when larger datasets are in use. 4. Since there is no specific outcome for hierarchical clustering, if the resulting analysis does not show any informative information (through clustering patterns) it is likely the results of the analysis will not be useful to the research. 5. Hierarchical clustering cannot handle any predictor variables which are categorical. The cluster results can be categorical (as done here) but must be assigned to the rownames and removed from the data frame.