*Figure **Error! No text of specified style in document.**-1 This is our results by making*

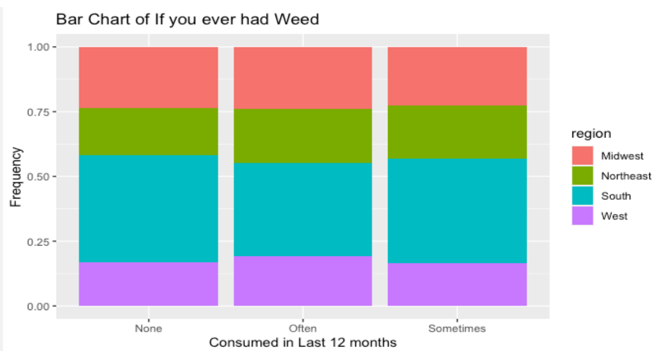
Clustering on MFT Study

By Aretha Kassegnin

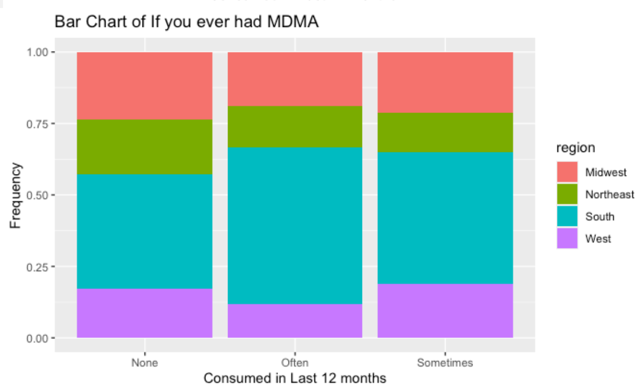
1) For our data we have no numerical variables which means that we cannot use K-Means clustering therefore we must use another method of clustering. We used a method called hierarchical clustering because it works with categorical data. The goal is still the same we are trying to find similarities in the dataset and grouping them together as a cluster. Because we are not able to use the pairs command in R, we are not able to see patterns in our dataset. But we can use our EDA to see where most of the data will be clustered. We predict that most of the data points will be grouped in in South.

Below is our Exploratory Data Analysis (EDA)

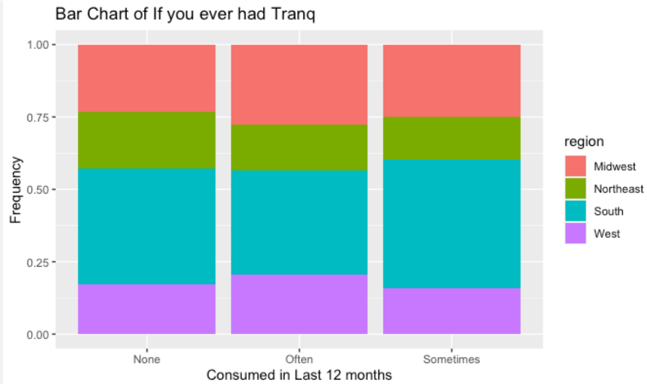
As we can see from this overlay bar graph is that for "None" most students where from the South and this also shows that most students in the Northeast consumed alcohol is the last 12 months is often.



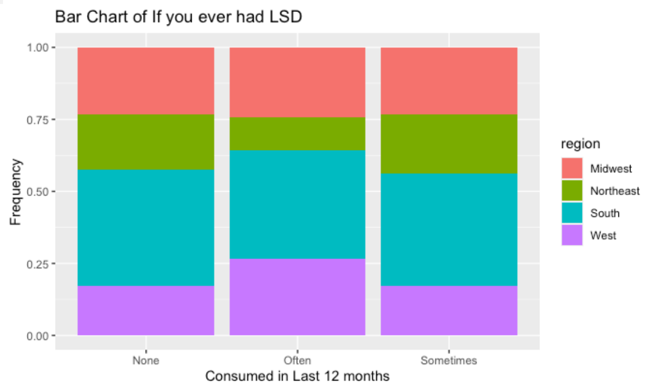
This shows Consumed region vs Weed in last 12 months, we can see most students in the South are the majority in this bar graph.



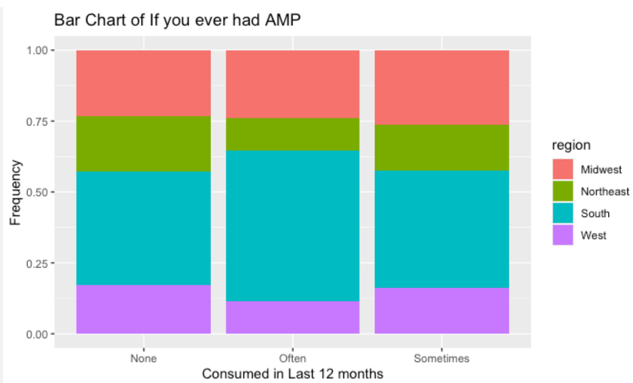
This bar graph is of Students that consumed MDMA in the last 12 months. It's obvious that the majority of students in the South often consume MDMA.



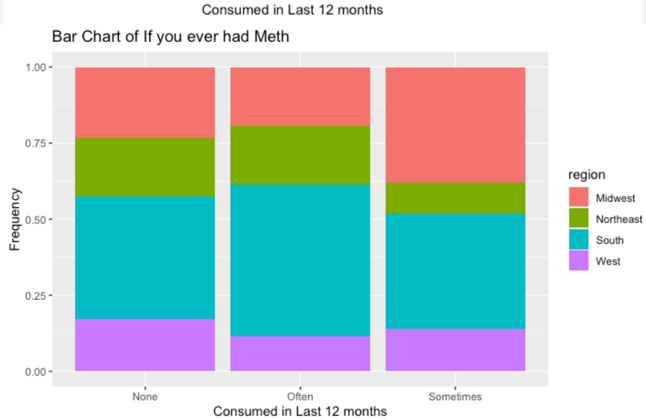
Next, we have a bar graph that shows the relationship between Tranq in the last 12 months and Region and as can see like the others South has majority in all categories with Midwest being the 2nd with the region that often uses this drug.



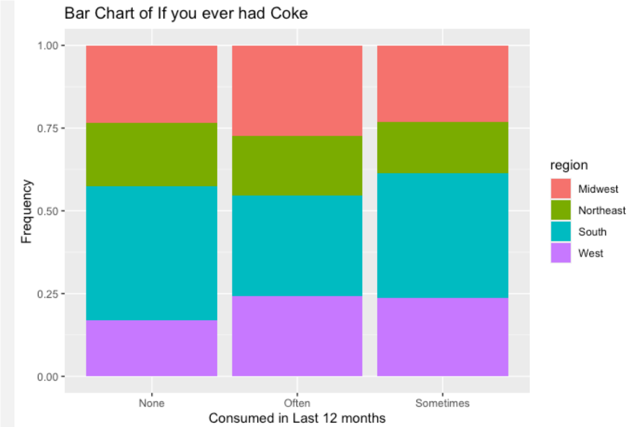
Another graph shows Region VS LSD usage in the last 12 months. It's the same as the other variable the South is dominate in all categories, but we can see in "Often" the South and West and close in the amount of students.



This bar graph shows a AMP consumption in the last 12 months and as we can see again the South is majority in all sections but it's also clear that the northeast consumes the least amount of a AMP "often" in the last 12 months.



This bar graph is showing the relationship between Meth use in the last 12 months and Region. As we can see again the South is really majority of all 3 choices but most students in the South picked "often".



Lasty, this is the bar graph that shows the relationship between Coke in the last 12 months and Region and again most categories are majority but this time it looks like most students in the South picked "Sometimes".

2) When using hierarchical clustering different methods give you different results and tells us how we can compare them using with different k values which is the amount of cluster would work best for our data using Dunn's metric to show us by plotting below.

3) Using the ward.D method didn't give use anything about the data because the graph we created using this method is constant. Which means there isn't that variation in the clusters.

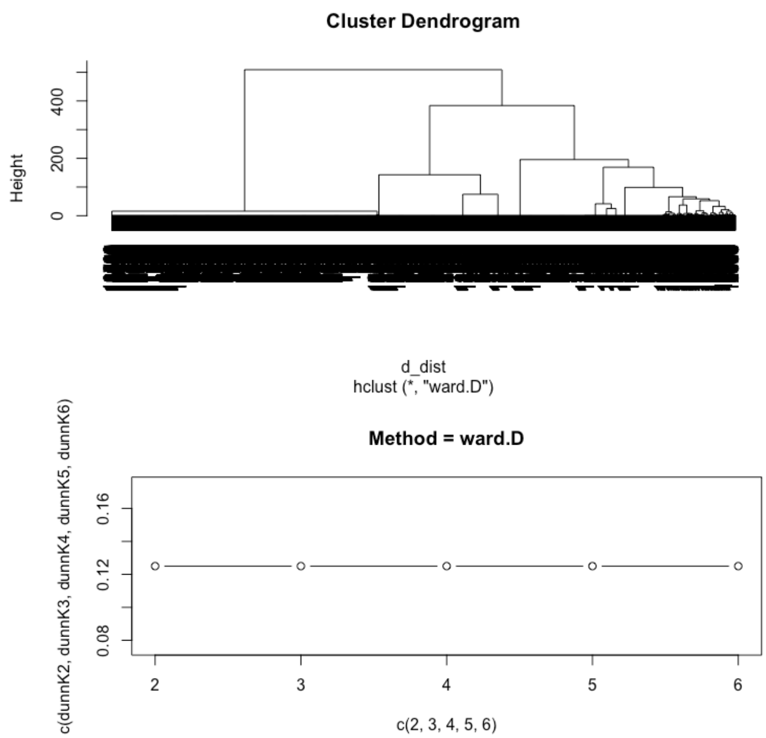


Figure 1-1 Plot using ward.D method

Using the average method for the hierarchical cluster didn't work as well as you can see from the plot below we got a constant line as well which means the these clusters will not work.

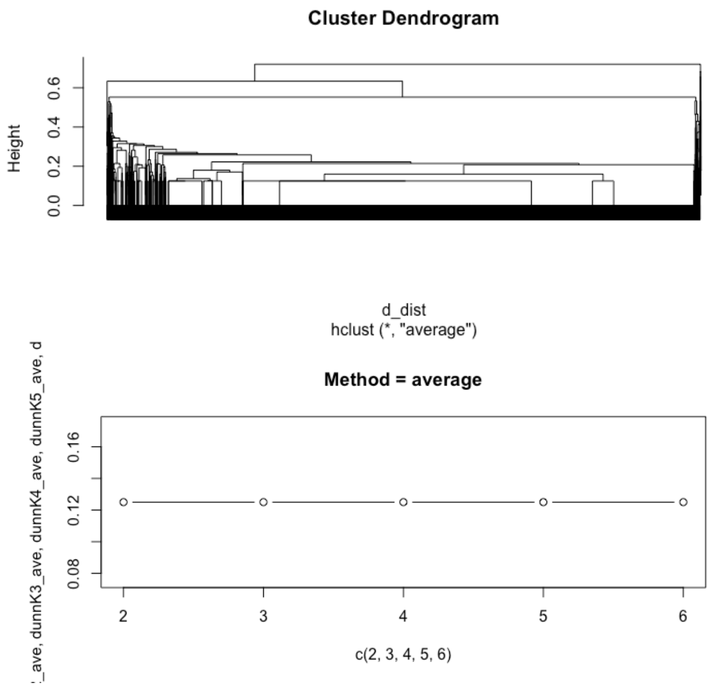


Figure 1-2 plot using average method

Finally using the compete method and when you look at the plot below we see that there is some variation in the plot as you can from the x axis the number is 5 which means that K=5

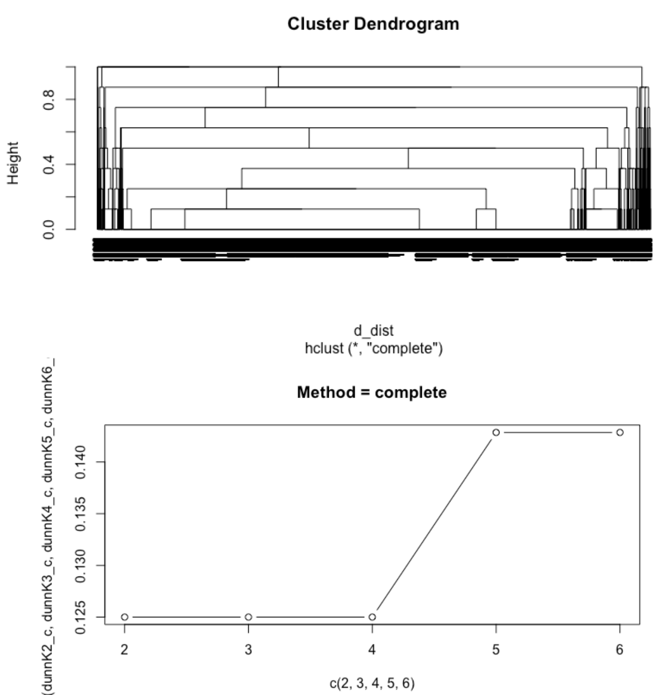


*Figure 1-**Error! No text of specified style in document.**-2 Plot using complete method*

4)

	1	2	3	4	5
None	8805	171	65	4	1
Often	995	56	16	11	3
Sometimes	3476	37	65	8	0

This table details the alcohol variable and the clustering using k=5 as you can see the 1st cluster has most of the alcohol data and we can see that None has the most data points in the cluster. This tells us that that the data in cluster 1 has similarities different from the others

	1	2	3	4	5
None	8587	15	28	19	0
Often	1997	236	67	2	4
Sometimes	2692	13	51	2	0

This table shows weed variable vs the clusters. We can see that Cluster 1 has the most datapoints with a mix of None, Often and Sometimes. We can also note that the other clusters don't have that many data points.

	1	2	3	4	5
None	12999	64	122	14	4
Often	6	35	4	0	0
Sometimes	271	165	20	9	0

This shows LSD variable vs the clusters and it's the same thing as the other two tables above. It is important to note that the other clusters don't have as many data points for LSD .

	1	2	3	4	5
None	13168	215	0	19	2
Often	27	11	0	4	0
Sometimes	81	38	146	0	2

We see the same thing when looking at the variable for MDMA vs clusters we can see that the 1st cluster has a lot of none data points and really nothing else compares.

	1	2	3	4	5
None	13117	183	119	15	1
Often	41	16	3	3	3
Sometimes	118	65	24	5	0

This table shows coke vs clusters it's the same thing with the clusters we see that for cluster 1 None has most datapoints out of any other option.

	1	2	3	4	5
None	12795	90	127	21	2
Often	74	58	6	2	1
Sometimes	407	116	13	0	1

This table shows the AMP vs clusters and it is the same thing where we see cluster one with most of the data points are in the 1st cluster which none.

<table><tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr><tr><td>None</td><td>12998</td><td>100</td><td>115</td><td>23</td><td>0</td></tr><tr><td>Often</td><td>34</td><td>41</td><td>7</td><td>0</td><td>1</td></tr><tr><td>Sometimes</td><td>244</td><td>123</td><td>24</td><td>0</td><td>3</td></tr></table>	This table shows TRANQ vs Clusters and we can see like the rest the pattern in cluster 1
<table><tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr><tr><td>None</td><td>13220</td><td>261</td><td>140</td><td>8</td><td>0</td></tr><tr><td>Often</td><td>15</td><td>2</td><td>0</td><td>9</td><td>0</td></tr><tr><td>Sometimes</td><td>41</td><td>1</td><td>6</td><td>6</td><td>4</td></tr></table>	The last table shows METH vs Clusters and its again the same same pattern with most the data points being in the 1st cluster

5)

	1	2	3	4	5
Midwest	3106	68	32	5	1
Northeast	2552	42	21	5	1
South	5333	107	67	11	2
West	2285	47	26	2	0

*Figure 5-**Error! No text of specified style in document.**-3 Target variable vs the Clusters*

As we expected most of the data points are in the South for the 1st cluster just like we saw on the EDA and most of our data is in the 1st cluster which is expected when looking at the table. It might have been better to do K=4 because the 5th cluster is really insignificant with only a few data points in that cluster. But we am not sure what the other data points are because it could be anything more analysis on the data can help us figure out the patterns for the clusters

Appenix

```
library(datasets) # contains iris dataset
library(cluster)  # clustering algorithms
library(factoextra) # visualization
library(purrr)
library(ClusterR)
library(cluster)
library(clValid)
#remove target variable
data2 <- data[, -1]
df<-data2
# calculate distance
d_dist<-daisy(df, metric = "gower", weights =c(1))

# hierarchical clustering using differeny methods
hc <- hclust(d_dist, method = 'ward.D')
hc1<-hclust(d_dist, method = 'average')
hc2<-hclust(d_dist, method = "complete")

plot(hc, labels= FALSE)
plot(hc1, labels= FALSE)
plot(hc2, labels= FALSE)

#seeing if method =ward.D has a good clusters

cutK2 <- cutree(tree = hc, k = 2)
cutK3 <- cutree(tree = hc, k = 3)
cutK4<- cutree(tree = hc, k = 4)
cutK5 <- cutree(tree = hc, k = 5)
cutK6 <- cutree(tree = hc, k = 6)

dunnK2 <- dunn(distance = d_dist, clusters = cutK2)
dunnK3 <- dunn(distance = d_dist, clusters = cutK3)
dunnK4 <- dunn(distance = d_dist, clusters = cutK4)
dunnK5<- dunn(distance = d_dist, clusters = cutK5)
dunnK6<- dunn(distance = d_dist, clusters = cutK6)

plot(x = c(2,3,4,5,6),y = c(dunnK2, dunnK3, dunnK4, dunnK5, dunnK6),
    type = "b",
    main ="Method = ward.D")
#Since the graph is constant this isnt a good method to use for our dataset
#Next we will use another method = average
```

```
cutK2_ave <- cutree(tree = hc1, k = 2)
cutK3_ave <- cutree(tree = hc1, k = 3)
cutK4_ave<- cutree(tree = hc1, k = 4)
cutK5_ave <- cutree(tree = hc1, k = 5)
cutK6_ave <- cutree(tree = hc1, k = 6)

dunnK2_ave <- dunn(distance = d_dist, clusters = cutK2_ave)
dunnK3_ave <- dunn(distance = d_dist, clusters = cutK3_ave)
dunnK4_ave <- dunn(distance = d_dist, clusters = cutK4_ave)
dunnK5_ave <- dunn(distance = d_dist, clusters = cutK5_ave)
dunnK6_ave <- dunn(distance = d_dist, clusters = cutK6_ave)

plot(x = c(2,3,4,5,6),y = c(dunnK2_ave, dunnK3_ave, dunnK4_ave, dunnK5_ave, dunnK6_ave),
    type = "b",
    main ="Method = average")
#using the complete method
cutK2_c <- cutree(tree = hc2, k = 2)
cutK3_c <- cutree(tree = hc2, k = 3)
cutK4_c<- cutree(tree = hc2, k = 4)
cutK5_c <- cutree(tree = hc2, k = 5)
cutK6_c <- cutree(tree = hc2, k = 6)

dunnK2_c <- dunn(distance = d_dist, clusters = cutK2_c)
dunnK3_c <- dunn(distance = d_dist, clusters = cutK3_c)
dunnK4_c <- dunn(distance = d_dist, clusters = cutK4_c)
dunnK5_c <- dunn(distance = d_dist, clusters = cutK5_c)
dunnK6_c <- dunn(distance = d_dist, clusters = cutK6_c)

plot(x = c(2,3,4,5,6),y = c(dunnK2_c, dunnK3_c, dunnK4_c, dunnK5_c, dunnK6_c),
    type = "b",
    main ="Method = complete")
# the Dunns method tells us the K=5 is the best
k5 <- cutK5_c

table(data$alcohol_12, as.factor(k5))
table(data$weed_12, as.factor(k5))
table(data2$LSD_12, as.factor(k5))
table(data$MDMA_12, as.factor(k5))
table(data$coke_12, as.factor(k5))
table(data$amp_12, as.factor(k5))
table(data$tranq_12, as.factor(k5))
table(data$meth_12, as.factor(k5))

#5
table(data$region, as.factor(k5))
```