Preprocess your dataset to run with the kMeans, Hierarchical Clusterer, and DB Scan algorithms and run the algorithms in R or Python or h2o (kMeans). Optimize the configurations as much as you can, then create a visual of the clusters.  Then answer the following questions on one single-spaced page:

1. How do the cluster distributions that the different algorithms generate differ from one another?

Student Alcohol Consumption is our dataset , which is regarding student alcohol consumption, and where which could interrupt the education with grade of education , what factors would interrupt the education would interrupt the education and grade of the student. In the dataset, and for the dataset, we would like to have a different cluster to maintain what does a machine try to learn when I give the dataset.

Number of clusters which was created with the dataset by the machine by using different algorithms as given below, and it was 5, and different values are been provided by different clusters.   
  
student4<- student3

student3$Alc <- NULL

(kmeans.result <- kmeans(student3, 5))

kmeans.result

table(student4$Alc, kmeans.result$cluster)

plot(student3[c("absences", "G2")], col = kmeans.result$cluster)

points(kmeans.result$centers[c("age", "G2")], col = 1:5, pch = 8, cex=2)

> table(student4$Alc, kmeans.result$cluster)

1 2 3 4 5

1 120 73 17 101 43

2 54 51 7 37 29

3 24 21 9 4 15

4 7 7 7 2 6

5 3 7 3 1 1

Cluster means:

age Medu Fedu traveltime studytime failures famrel freetime goout health

1 16.40385 2.403846 2.250000 1.576923 2.014423 0.07211538 4.086538 3.129808 3.072115 3.658654

2 16.78616 2.465409 2.264151 1.603774 1.798742 0.19496855 3.767296 3.094340 3.295597 3.496855

3 17.37209 2.348837 2.395349 1.534884 1.651163 0.53488372 3.813953 3.116279 3.441860 3.604651

4 16.73103 3.041379 2.662069 1.400000 2.234483 0.01379310 4.013793 3.158621 3.055172 3.351724

5 17.15957 2.106383 1.914894 1.765957 1.627660 0.77659574 3.787234 3.500000 3.329787 3.585106

absences G1 G2 G3

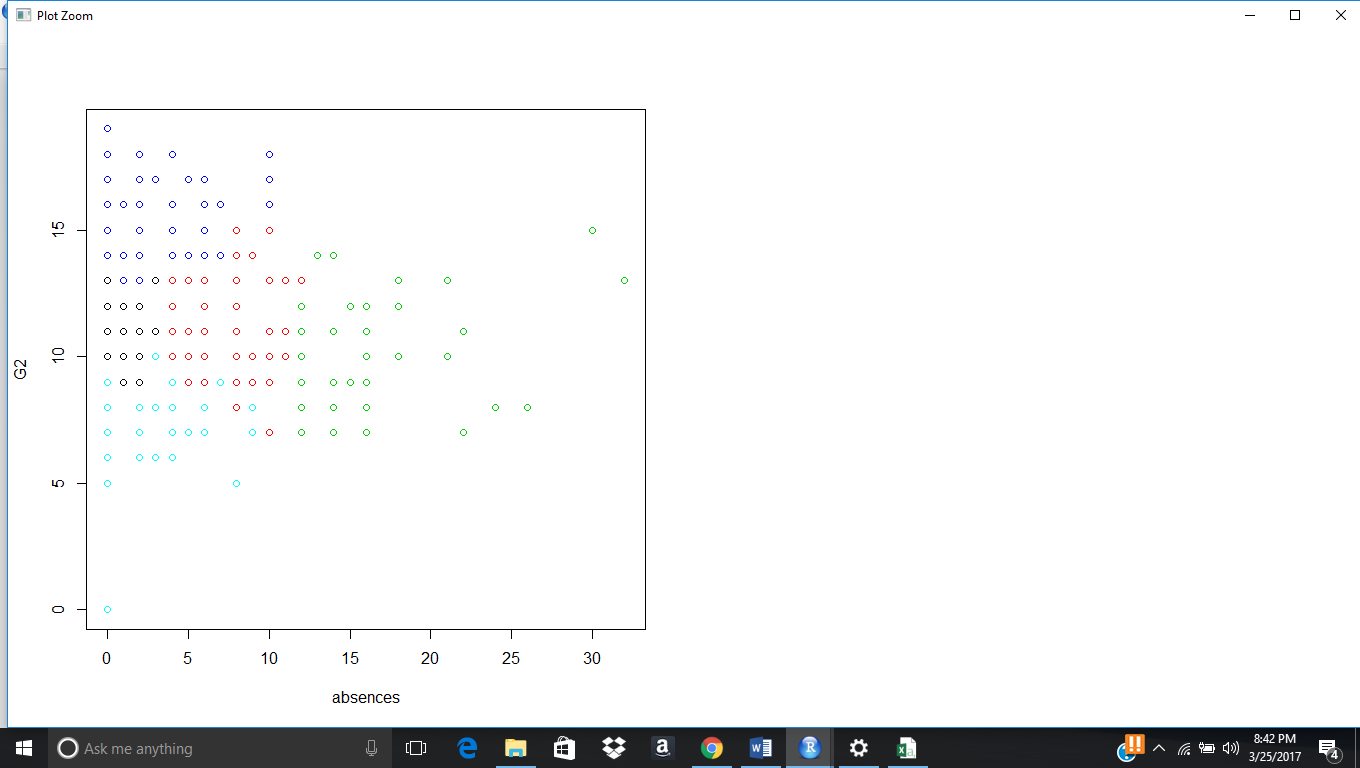
1 0.6730769 11.384615 11.46154 11.923077

2 6.4591195 10.723270 10.98113 11.477987

3 16.2558140 10.232558 10.25581 10.511628

4 2.0758621 14.944828 15.35862 15.786207

5 2.2127660 7.638298 7.56383 7.244681



###################### k Medoids ######################

library(fpc)

pamk.result <- pamk(student3,5)

pamk.result$nc ## number of clusters

## Let's check the clusters against the actual species

table(pamk.result$pamobject$clustering, student4$Alc)

layout(matrix(c(1,2),1,2)) ## 2 graphs per page

plot(pamk.result$pamobject)

layout(matrix(1)) ## change back to one graph per page

> pamk.result$nc ## number of clusters

[1] 5

> table(pamk.result$pamobject$clustering, student4$Alc)

1 2 3 4 5

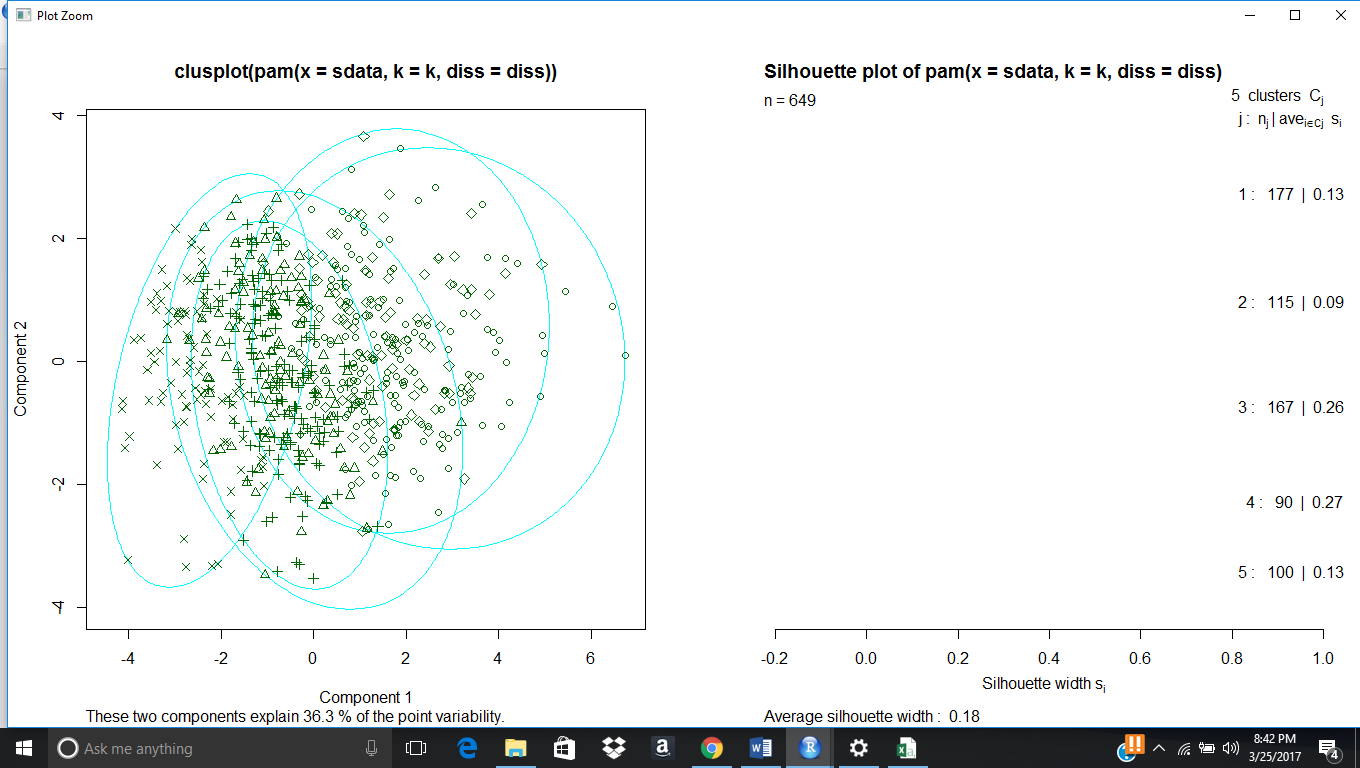
1 87 45 29 13 3

2 71 35 7 1 1

3 100 47 14 4 2

4 61 26 1 1 1

5 35 25 22 10 8



idx <- sample(1:dim(student4)[1], 40)

studentSample <- student4[idx,]

studentSample$Alc <- NULL

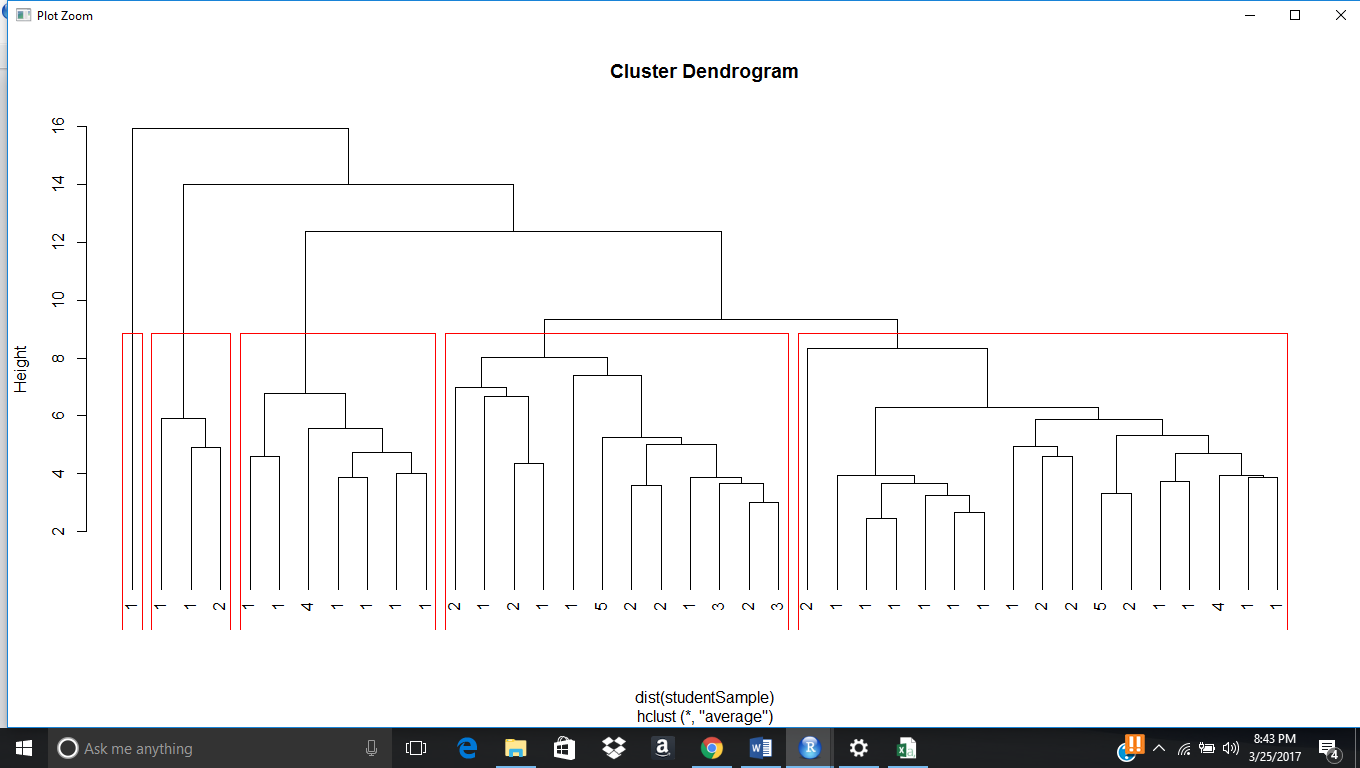
hc <- hclust(dist(studentSample), method="ave")

plot(hc, hang = -1, labels=student4$Alc[idx])

## Let's cut the tree into 3 clusters

rect.hclust(hc, k=5)

groups <- cutree(hc, k=5)



1. Which algorithm produces the best results?

Generally density algorithm would generally best results for dataset, and then in our results, we find hierarchical density is providing most effective results, and then the results in terms of graph, it would be a effective results for our dataset,

idx <- sample(1:dim(student4)[1], 40)

studentSample <- student4[idx,]

studentSample$Alc <- NULL

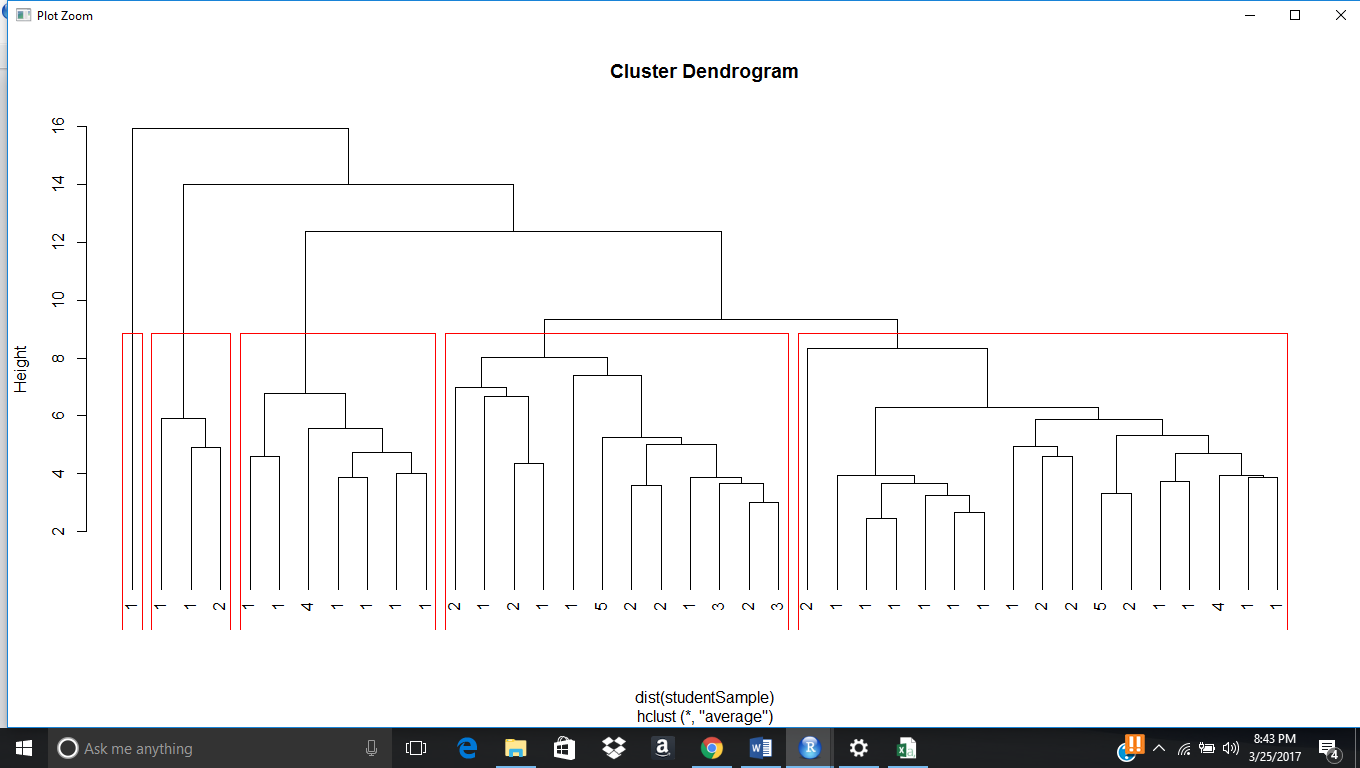
hc <- hclust(dist(studentSample), method="ave")

plot(hc, hang = -1, labels=student4$Alc[idx])

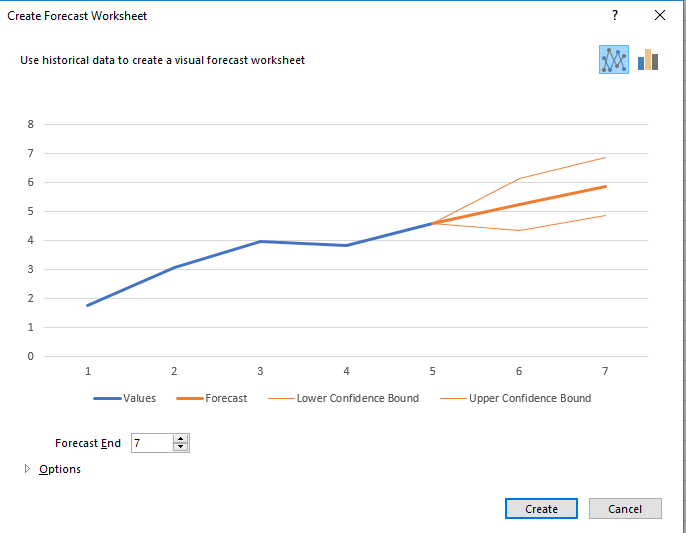
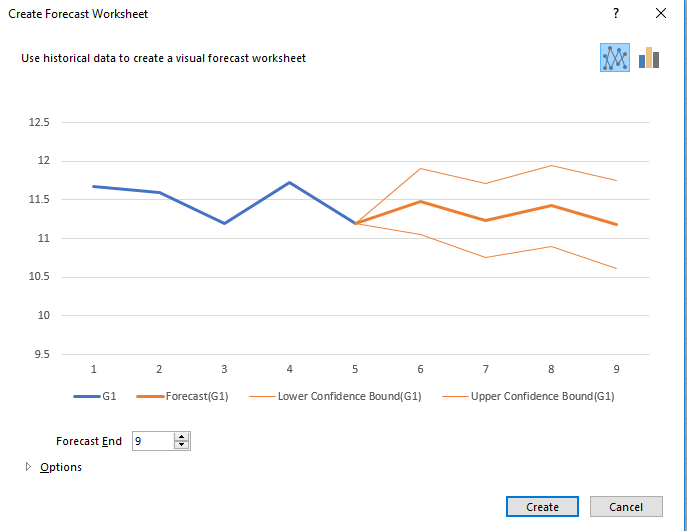
## Let's cut the tree into 3 clusters

rect.hclust(hc, k=5)

groups <- cutree(hc, k=5)



1. Given the nature of your dataset and its attribute values, why would that algorithm produce the best results?

Our Dataset was regarding with student alcohol where how much each factor creates impact for student grades by taking drink, and in our dataset we are having 32 attributes, and by having drinking by weekdays and weekends how much famrel and health would impact he student education in terms of grade, so in order to get a forecast of what attributes best relates to the ideology of the dataset, and then taking attributes in consideration, we are having,   
  
school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup paid activities nursery higher internet romantic famrel freetime goout Dalc Walc health absences G1 G2 G3  
  
these are the attributes of our dataset, and for this we made Dalc and Walc are the basic class attribute which we thought and for we made the close analysis what best can lead and best decision for the dataset to provide best results, and for this we made a realistic analysis using forecast analysis to best replicate with clustering, and association analysis, and for this we got results in this way.   
  
  
  
  
  
  
Yes by this and our algorithm, every data points get effected down and reach at a single points in a cluster in a near by attributes where correlation of each attrbiutes creates impact, but specificed attributes like Dalc, Walc, health, Sex, to impact the grade of the student and this got proved in earlier with earlier results like with coorelation, and kmeans, chisq test , Euclidean and manhattan distance with data points are near by with the respective attributes provided and with grade. Here its get proved.

1. What do the results tell you about the data?  What business, political, or medical decisions could management make based on the results of your analysis?

Our results decide that we would get good understanding that machine learning helped us in identifying what best clusters takes in consideration ,and how much data points will have best and near distance, where it creates a impact in terms of making student education, and where a government can take a effective decision in terms of analysis which it brought. Our results relies what our forecast we made like what made that sex, health Dalc and Walc are created with best impact with student grade and it has got provided with the latest algorithm test which we made to identify and best interpret what factors impact student grade. By using this attributes and results a political can be made like   
  
  
  
First, student grade is impact people are drinking more at weekends, so government can have them stop selling drinks in the weekend.

Clusters which has given that student health is also impacting the student grade in terms of student grade.

People with good famrel is having good education and number of datapoints which is there in there third clusters best implies how much relation is given.

There are considered to be many more relations, but we find many more points what best reactions can be taking government and political, but as per business we have like if we stop drinking that would impact government and people investing in drinking and liquor factory get away, so we can have one decision as none of the drinking shop should be near by school, so that atleast this can stop people coming to shop to by drinks