BUDA535Module1

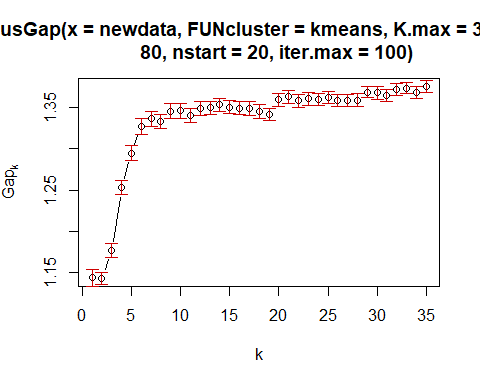
## Problem 1 :

Using the College Data in the ISLR Package we want to better understand if there are groups of colleges. Pick a subset of numeric variables (don’t use Private), you must use at least 4, that you believe would segment colleges into some meaningful groups. From there use both k-means and variations of hierarchical clustering to see what groups you can possibly identify. Please comment on how you selected your clusters, what meaning your clusters have (if any), and if scaling your data had any effect on the results.

##   
## Attaching package: 'dplyr'

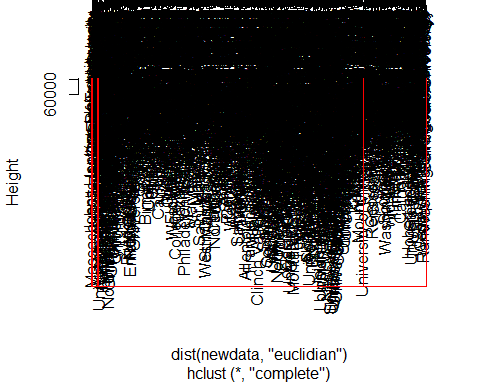
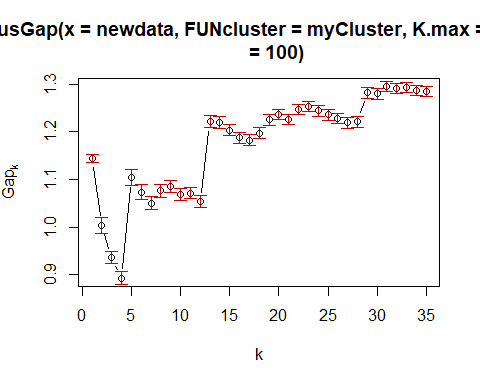
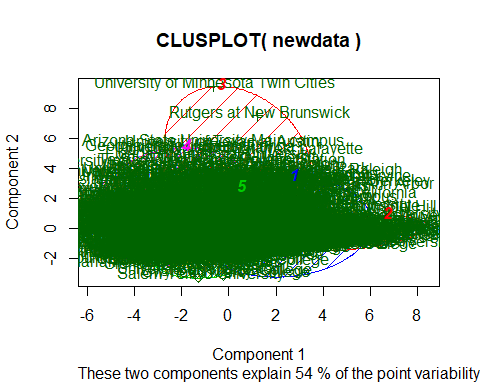
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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



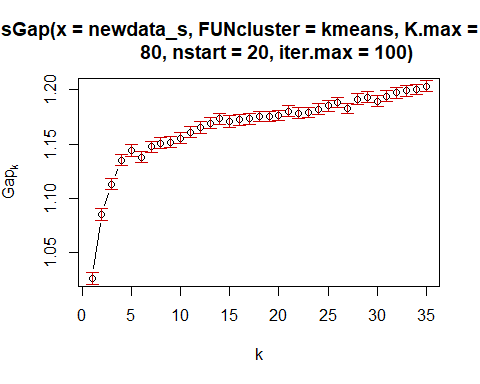
## Accept Top10perc P.Undergrad Outstate Room.Board Books Personal  
## 1 2076.576 45.84746 257.3983 16741.559 5386.297 570.5593 997.339  
## 2 3234.438 77.06250 312.5000 18084.625 5965.750 580.1250 1300.562  
## 3 8127.689 32.59016 3428.0492 8409.607 4252.525 607.3934 1923.262  
## 4 1578.261 18.61071 986.8071 6716.875 3582.129 538.1643 1481.218  
## 5 1106.364 25.06954 476.0828 11436.526 4610.474 538.1589 1228.887  
## PhD Terminal S.F.Ratio Expend Grad.Rate  
## 1 86.55085 92.62712 10.79153 15363.220 78.77119  
## 2 93.81250 96.68750 6.36875 34815.938 90.87500  
## 3 85.81967 91.03279 15.88033 10011.574 61.26230  
## 4 66.43214 73.28929 16.51464 6416.450 55.52857  
## 5 69.22848 77.41060 13.17748 9035.517 68.97682

## [1] 118 16 61 280 302



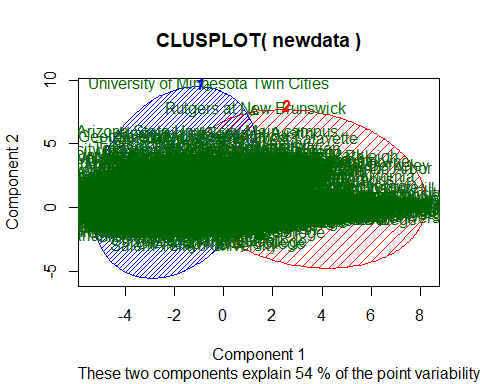
## v1  
## 1 2 3 4 5   
## 614 147 12 1 3

## Group.1 Accept Top10perc P.Undergrad Outstate Room.Board Books Personal  
## 1 1 1005.0 20.0 466 8840 4000.0 500 1300  
## 2 2 1598.0 42.0 81 16260 5450.0 525 900  
## 3 3 2332.5 82.5 69 18607 5987.5 575 1252  
## 4 4 3446.0 75.0 1569 18800 6740.0 500 1040  
## 5 5 18744.0 29.0 4065 9556 3990.0 570 1060  
## PhD Terminal S.F.Ratio Expend Grad.Rate  
## 1 71.0 78.0 14.40 7714.5 62.0  
## 2 88.0 94.0 11.10 14213.0 79.0  
## 3 95.5 97.5 5.55 35197.5 96.5  
## 4 96.0 97.0 3.30 56233.0 90.0  
## 5 90.0 95.0 18.20 10474.0 71.0



## Accept Top10perc P.Undergrad Outstate Room.Board Books  
## 1 -0.07697502 -0.4467637 0.1334696 -0.5580183 -0.4393584 -0.04085915  
## 2 0.13362214 0.7755440 -0.2316919 0.9686726 0.7626891 0.07092802  
## Personal PhD Terminal S.F.Ratio Expend Grad.Rate  
## 1 0.1879572 -0.4276507 -0.4282695 0.3323019 -0.4424866 -0.3958639  
## 2 -0.3262777 0.7423655 0.7434396 -0.5768480 0.7681194 0.6871862

## [1] 493 284



### Interpretation :

**Data :** Removed Apps, Enroll and F.Undergraducate variables as those are inter-related/dependent with Accept Variable. Removed perc.alumni because that did not seem a lot of significant as its donation from partners for alumni.

**Clustering with kmeans :** Using gap clustering method and based on it plot, 5 clusters can be created. Based on the clusters groups, Cluster 1 has colleges with the least part time graduate students having less personal expenses with average graduation rate of 78.7 %. Cluster 2 have colleges with higher PhD and Terminal faculties have top 10% HSC students enrolled with high expenses of room.board, outstate and Expend along with lowest S.F.Ratio (students faculty ratio) have highest graduation rate. Cluster 3 have colleges with high applications accepted having maximum number of Part time Undergrad enrolled with high Expend cost. Cluster 4 have colleges with least graduation rate having least top10% of HSC student with least PHD and Terminal degree faculties and least expenses. Cluster 5 is not too meaningfull except that it floats in medium range for all variables except colleges with least accepted applications. Clustering plot shows a lot of overlapping of these clusters. To add, using nstart = 100/60/30 gives almost similar clusters results.

**Hierarchical clustering :** After doing cuttree and looking at the median in hierarchical clustering, cluster groups have similar meaning as in kmeans. Looking at table of cutree, clusters 4 and 5 have 1 and 3 records respectively which are outliers.

**Overall Conclusion as per my knowledge:** Colleges with highest number of PhD and Terminal degree faculties have Top 10% of HSC students have high graduation rates along with high expenses of Outstate, room.Board and Expend. These variables are likely proporational to each other in all clusters such as percentage of expenses and PhD/Terminal faculties decreases, top 10% of HSC and grad rate decreases. Part time students are more where acceptance rate of colleges are more.

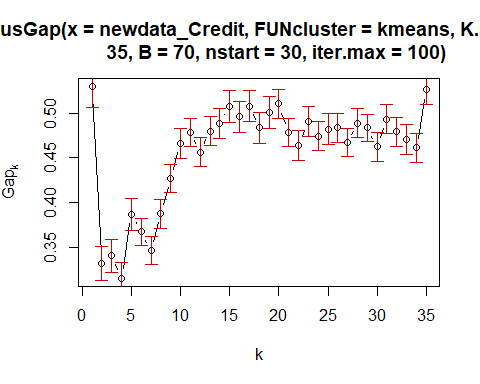
**Results on scaled data :** Scaled the data and used Kmeans with 2 clusters. Based on plot, 5 clusters can be possible but thinking as maximum gap statistics, 2 clusters seems best. Cluster 1 have colleges with top 10% of HSC students with high expenses of outstate, room.board, books and Expend. These colleges have highest number of PhD and Terminal degree faculties and least Part time graducate students and less S.F. Ratio. Cluster 2 have least Graduation rate with less number of top 10% of HSC students and less number of PhD and terminal degree faculties and high part time graducate students. Students of these colleges have less expenses except Personal expenses.

**Comparision :** Results of scaled data have aggregated the groupings in 2 with reasonable size clusters. While on raw data, grouping are widened and have 1-2 clusters with outliers. But results of scaled and raw data varies depending on dataset. Graph of scaled data shows overlapping clusters but quite clear. I tried different size of clusters such as 3 and 4 on raw data with kmeans and hclust after cutting it to 5 but somehow 5 makes more sense though it has few outliers. Varified the results for 3 but clusters with 5 gave me more meaningful subgroups. Please suggest if my understanding is not correct.

## Problem 2 :

The Credit data in the ISLR package describes 400 customers. Using only the numeric variables, investigate if the customers break out into any subgroups (clusters). If so provide insights on those subgroups. Hint: Sometimes it’s useful to compare the groups to baselines such as underlying factors i.e. distribution of education, student. It can also be useful to just summarize the group based on the cluster itself such as cluster mean and median.

library(ISLR)  
data("Credit")  
  
newdata\_Credit <- Credit[,-c(1,8,9,10,11)]  
#head(newdata\_Credit)  
  
Credit\_sel <- clusGap(newdata\_Credit,FUN = kmeans, nstart = 30, K.max = 35, B=70, iter.max = 100 )  
plot(Credit\_sel)



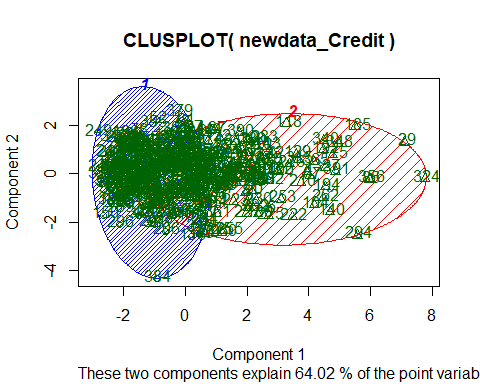
#Kmeans method  
set.seed(6591)  
m1<- kmeans(newdata\_Credit,2)  
m1$centers

## Income Limit Rating Cards Age Education Balance  
## 1 30.78704 3462.034 269.4662 2.943609 54.81203 13.59398 272.4211  
## 2 73.86718 7263.724 524.6119 2.985075 57.36567 13.16418 1011.5075

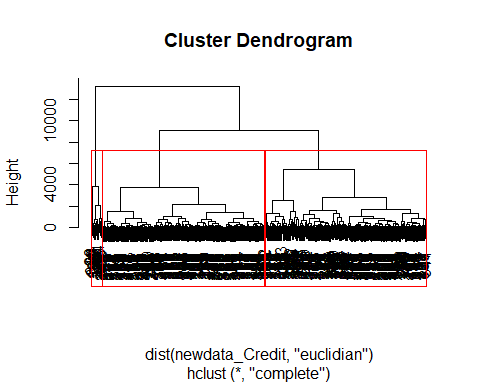
m1$size

## [1] 266 134

clusplot(newdata\_Credit, m1$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)



#hclust method  
h1<-hclust(dist(newdata\_Credit,"euclidian"), method="complete")  
plot(h1)  
rect.hclust(h1,k=3,border="red")



c1<-cutree(h1,k=3)  
table(c1)

## c1  
## 1 2 3   
## 194 193 13

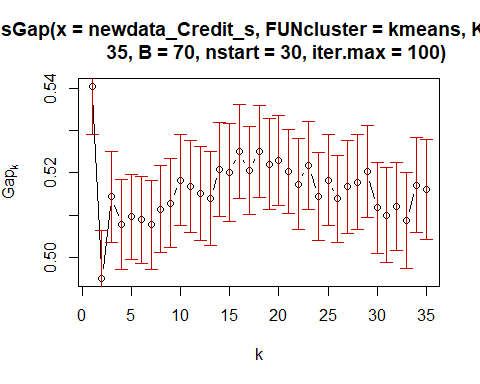
aggregate(newdata\_Credit,list(c1), mean)

## Group.1 Income Limit Rating Cards Age Education  
## 1 1 26.92624 2903.093 232.7938 2.958763 54.63402 13.54124  
## 2 2 56.64213 6134.264 448.0000 2.937824 56.11399 13.31606  
## 3 3 148.61015 11317.462 796.1538 3.230769 64.46154 14.07692  
## Balance  
## 1 160.0773  
## 2 814.4301  
## 3 1520.4615

aggregate(newdata\_Credit,list(c1), median)

## Group.1 Income Limit Rating Cards Age Education Balance  
## 1 1 24.5015 3031.5 238.5 3 56 14 60  
## 2 2 49.5700 5673.0 427.0 3 55 14 822  
## 3 3 149.3160 10748.0 754.0 3 62 16 1448

#Scaled Data  
newdata\_Credit\_s = scale(newdata\_Credit)  
  
Credit\_sel\_s <- clusGap(newdata\_Credit\_s,FUN = kmeans, nstart = 30, K.max = 35, B=70, iter.max = 100 )  
plot(Credit\_sel\_s)



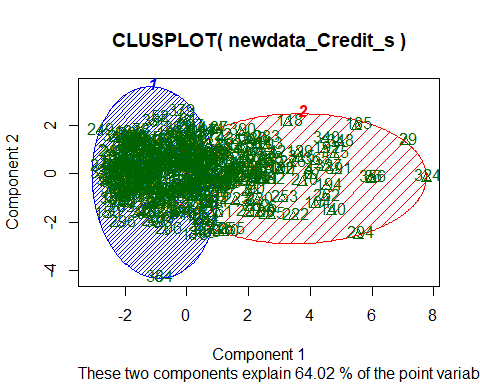
#Kmeans method  
set.seed(6571)  
ms<- kmeans(newdata\_Credit\_s,2)  
ms$centers

## Income Limit Rating Cards Age Education  
## 1 -0.4143337 -0.4777333 -0.4788188 0.0008172611 -0.07187987 0.07006433  
## 2 1.0923342 1.2594787 1.2623405 -0.0021545976 0.18950146 -0.18471505  
## Balance  
## 1 -0.4284581  
## 2 1.1295713

ms$size

## [1] 290 110

clusplot(newdata\_Credit\_s, ms$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)



### Interpretation :

**Kmeans Method :** Based on graph, created 2 clusters which are of size 233 and 167 observations respectively. Looking at the centers of clusters, Cluster 1 makes group of customers having low income with less limit and rating along with less credit card balance. Also, age of these customers are less compared to cluster2. Cluster 2 have customers with high income having high limit with good rating and credit card balance.

**Hierarchical clustering :** Based on dendogram, created 3 clusters which have 194,193 and 13 observations respectively. Looking at median and mean of the clusters, Cluster 2 and 3 provides almost same insights : high income with good credit limit and rating along with good credit card balance. Cluster 1 have customers with less incomes, low limit and rating with less credit card balance. If we look at the size of groups closely, group 3 have just 13 records which can be considered as kind of outliers but surely they are imporant data which is 3% of the 400 records.

**Scaled Data :** Scaled data provides more insight for the clusters. Cluster 1 have customers having low income with low credit limit and low credit rating along with less credit card balance with less age and high education. Cluster 2 have customers with high income having high limit and rating and good credit card balance. These customers age are higher.

**Overall Conclusion :** Comparing both the methods, credit data provides 2 clusters with meaningfull insights. One with good income, limit, rating and balance and other with less income, limit, rating and balance. Looking at these 2 clusters, variation of age and education is less. If we consider variables like student, married and gender, that might have provided more insights to these clusters. Scaled data for this dataset provides few more meaningful insights with age and education.