Final Project Group 5 Multivariate

Multivariate Group 5

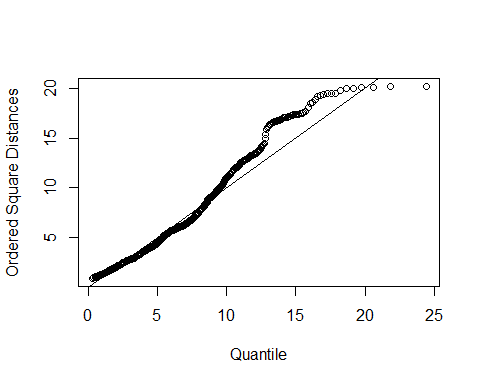
2023-12-01

AreaDrought=read.csv("areaDrought.csv",row.names = 1)

Data Cleaning

chi-square plot of the Mahalanobis distances

AreaDroughtNew=AreaDrought  
  
#Get column means + cov  
xbar=colMeans(AreaDroughtNew)  
S=cov(AreaDroughtNew)  
  
#using the mahalanobis function  
d2=mahalanobis(AreaDroughtNew,xbar,S)  
#finding the quantiles  
quantiles=qchisq((1:nrow(AreaDroughtNew)-1/2)/nrow(AreaDroughtNew),df=ncol(AreaDroughtNew))  
sd2=sort(d2)  
#plotting quantiles versus ordered squared distances  
plot(quantiles,sd2,xlab='Quantile',ylab='Ordered Square Distances')  
abline(a=0,b=1)



#text(quantiles,sd2,labels=row.names(AreaDroughtNew),pos = 1,cex=0.4)

#explain this garbage  
outliers <- which(d2> qchisq(0.975, ncol(S)))  
outliers

## 1 Jan 2013 2 Jan 2013 1 Jan 2020 2 Jan 2020 3 Jan 2020 4 Jan 2020 1 Feb 2020   
## 679 680 1045 1046 1047 1048 1049   
## 2 Feb 2020 3 Feb 2020 4 Feb 2020 1 Mar 2020 2 Mar 2020 3 Mar 2020 4 Mar 2020   
## 1050 1051 1052 1053 1054 1055 1056   
## 5 Mar 2020 1 Apr 2020 2 Apr 2020 3 Apr 2020 4 Apr 2020 1 May 2020 2 May 2020   
## 1057 1058 1059 1060 1061 1062 1063   
## 3 May 2020 4 May 2020 1 Jun 2020 2 Jun 2020 3 Jun 2020 4 Jun 2020 5 Jun 2020   
## 1064 1065 1066 1067 1068 1069 1070   
## 1 Jul 2020 2 Jul 2020 3 Jul 2020 4 Jul 2020 1 Aug 2020 2 Aug 2020 3 Aug 2020   
## 1071 1072 1073 1074 1075 1076 1077   
## 4 Aug 2020 1 Sep 2020 2 Sep 2020 3 Sep 2020 4 Sep 2020 5 Sep 2020 1 Oct 2020   
## 1078 1079 1080 1081 1082 1083 1084   
## 2 Oct 2020 3 Oct 2020 4 Oct 2020 1 Nov 2020 2 Nov 2020 3 Nov 2020 4 Nov 2020   
## 1085 1086 1087 1088 1089 1090 1091   
## 1 Dec 2020 2 Dec 2020 3 Dec 2020 4 Dec 2020 5 Dec 2020   
## 1092 1093 1094 1095 1096

#lots of outliers need to go to each index and delete the row #INTERESTING!!!!!! Most outliers are during covid year 2020

AreaDroughtClean=AreaDrought[-outliers,]  
  
round(cor(AreaDrought),2)

## PAD1 FCET AGSBillions YIR PPIF PPD1  
## PAD1 1.00 -0.01 -0.05 0.20 -0.10 0.70  
## FCET -0.01 1.00 -0.12 0.02 -0.23 -0.05  
## AGSBillions -0.05 -0.12 1.00 0.19 0.03 -0.25  
## YIR 0.20 0.02 0.19 1.00 -0.20 -0.04  
## PPIF -0.10 -0.23 0.03 -0.20 1.00 0.04  
## PPD1 0.70 -0.05 -0.25 -0.04 0.04 1.00

round(cor(AreaDroughtClean),2)

## PAD1 FCET AGSBillions YIR PPIF PPD1  
## PAD1 1.00 -0.02 -0.03 0.21 -0.10 0.69  
## FCET -0.02 1.00 -0.10 0.01 -0.22 -0.06  
## AGSBillions -0.03 -0.10 1.00 0.55 -0.22 -0.23  
## YIR 0.21 0.01 0.55 1.00 -0.17 -0.06  
## PPIF -0.10 -0.22 -0.22 -0.17 1.00 0.07  
## PPD1 0.69 -0.06 -0.23 -0.06 0.07 1.00

PCA Analysis

areaDroughtPCA=princomp(AreaDroughtClean,cor=T)  
  
  
print(summary(areaDroughtPCA,loading=T),cut=0.4)

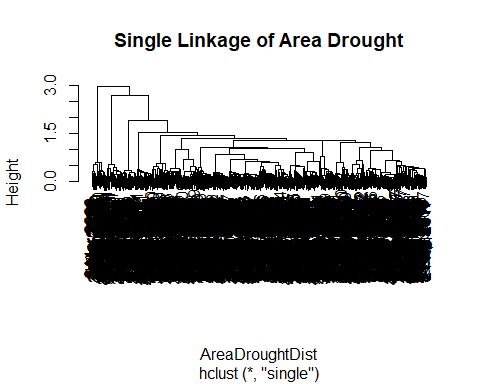
## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 1.3269512 1.2926833 1.0880288 0.8523803 0.63635984  
## Proportion of Variance 0.2934666 0.2785050 0.1973011 0.1210920 0.06749231  
## Cumulative Proportion 0.2934666 0.5719716 0.7692727 0.8903647 0.95785703  
## Comp.6  
## Standard deviation 0.50284970  
## Proportion of Variance 0.04214297  
## Cumulative Proportion 1.00000000  
##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  
## PAD1 0.448 0.559 0.695  
## FCET 0.787 0.572   
## AGSBillions -0.501 0.404 -0.721   
## YIR 0.549 0.578   
## PPIF -0.540 0.721   
## PPD1 0.610 -0.654

AreaDroughtScore=areaDroughtPCA$scores[,1:3]  
#round(AreaDroughtScore,2)  
  
round(cor(AreaDroughtScore),2)

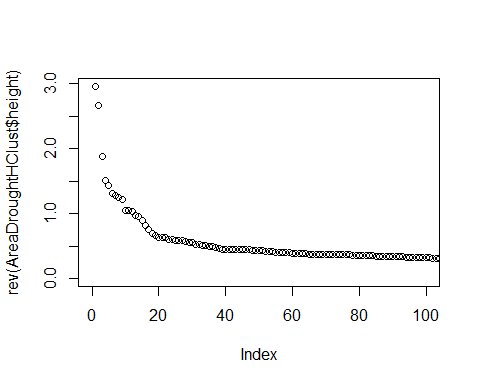
## Comp.1 Comp.2 Comp.3  
## Comp.1 1 0 0  
## Comp.2 0 1 0  
## Comp.3 0 0 1

Cluster Analysis

AreaDroughtNew=AreaDrought  
  
AreaDroughtScale=scale(AreaDroughtNew)  
AreaDroughtDist=dist(AreaDroughtScale)  
  
AreaDroughtHClust=hclust(AreaDroughtDist,method='single')  
plot(AreaDroughtHClust,main="Single Linkage of Area Drought")



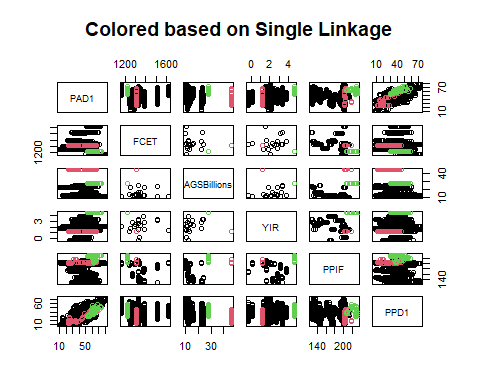
plot(rev(AreaDroughtHClust$height),xlim = c(0,100))



ct=cutree(AreaDroughtHClust,k=3)  
table(ct)

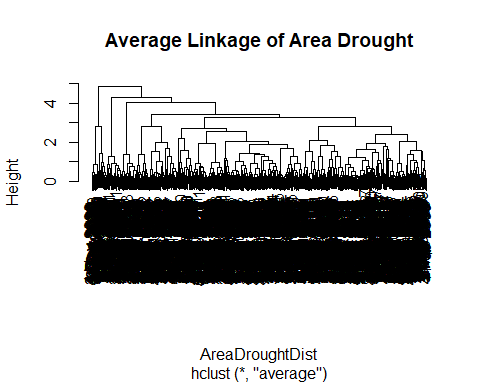
## ct  
## 1 2 3   
## 1044 52 52

plot(AreaDroughtNew,col=ct,main="Colored based on Single Linkage")

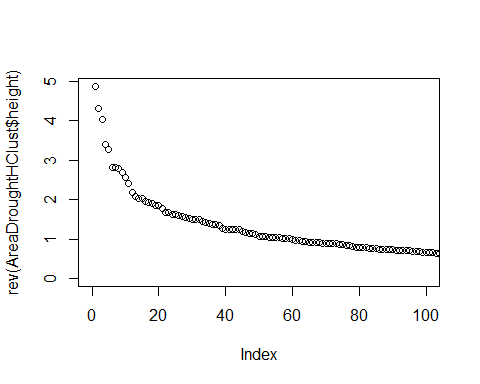


Average Hclust

AreaDroughtHClust=hclust(AreaDroughtDist,method='average')  
plot(AreaDroughtHClust,main="Average Linkage of Area Drought")



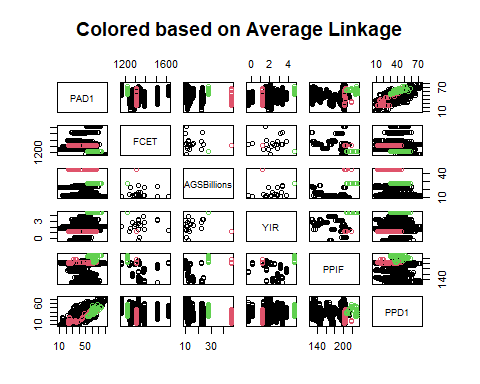
plot(rev(AreaDroughtHClust$height),xlim = c(0,100))



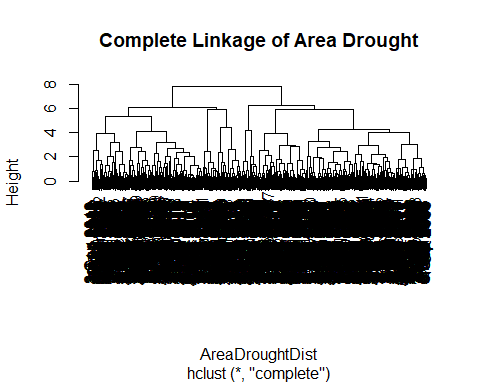
ct=cutree(AreaDroughtHClust,k=3)  
table(ct)

## ct  
## 1 2 3   
## 1044 52 52

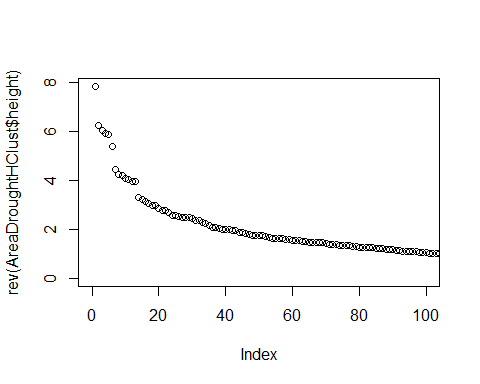
plot(AreaDroughtNew,col=ct,main="Colored based on Average Linkage")



AreaDroughtHClust=hclust(AreaDroughtDist,method='complete')  
plot(AreaDroughtHClust,main="Complete Linkage of Area Drought")



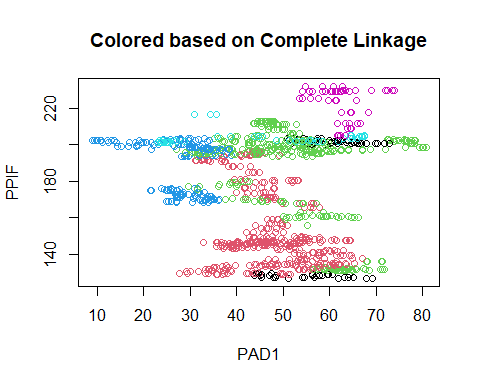
plot(rev(AreaDroughtHClust$height),xlim = c(0,100))



ct=cutree(AreaDroughtHClust,k=6)  
table(ct)

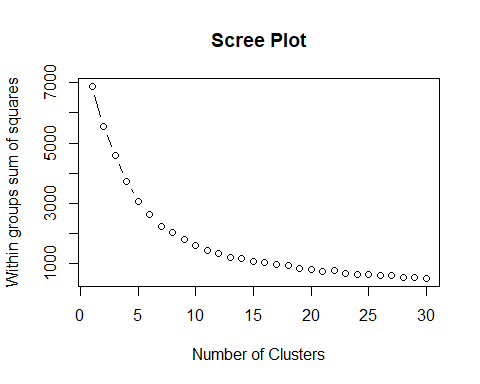
## ct  
## 1 2 3 4 5 6   
## 84 412 381 167 52 52

plot(AreaDroughtNew[,c(1,5)],col=ct,main="Colored based on Complete Linkage")

 K-means

#Scree plot thingy for k-means  
plot.wgss = function(mydata, maxc = nrow(mydata) - 1) {  
   
wgss = numeric(maxc)  
   
for (i in 1:maxc) {  
   
km = kmeans(mydata, centers = i, nstart = 10)  
   
wgss[i] = km$tot.withinss  
   
}  
   
plot(1:maxc, wgss, type = "b", xlab = "Number of Clusters", ylab =  
"Within groups sum of squares",  
   
main = "Scree Plot")  
}

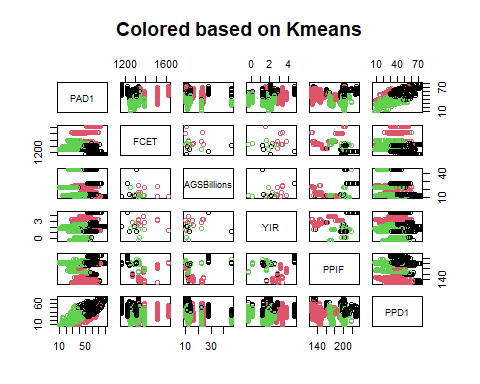
plot.wgss(AreaDroughtScale,maxc=30)



AreaDroughtKMean=kmeans(AreaDroughtScale,centers=3,nstart=10)  
table(AreaDroughtKMean$cluster)

##   
## 1 2 3   
## 227 540 381

plot(AreaDroughtNew,col=AreaDroughtKMean$cluster,main="Colored based on Kmeans")

 Column Means for Each Cluster

clusterMeans=AreaDroughtKMean$centers  
rownames(clusterMeans)=c("Clust1","Clust2","Clust3")  
library(knitr)  
  
kable(clusterMeans,caption="Column Means for Each Cluster")

Column Means for Each Cluster

|  | PAD1 | FCET | AGSBillions | YIR | PPIF | PPD1 |
| --- | --- | --- | --- | --- | --- | --- |
| Clust1 | 0.9796945 | -0.9644277 | 0.3643360 | 0.2583599 | 0.7260874 | 0.9668258 |
| Clust2 | 0.2101822 | 0.6028553 | -0.0661409 | 0.4880050 | -0.7390956 | -0.0573961 |
| Clust3 | -0.8815986 | -0.2798341 | -0.1233286 | -0.8455916 | 0.6149338 | -0.4946865 |

#model based clustering

library(mclust)

## Warning: package 'mclust' was built under R version 4.2.3

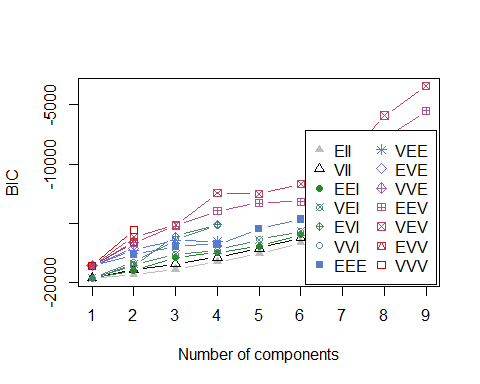
## Package 'mclust' version 6.0.1  
## Type 'citation("mclust")' for citing this R package in publications.

mc=Mclust(AreaDroughtScale)  
table(mc$classification)

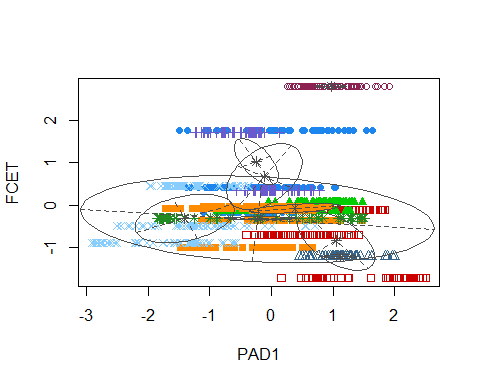
##   
## 1 2 3 4 5 6 7 8 9   
## 156 157 156 105 260 157 53 52 52

#plot(mc, what="BIC")

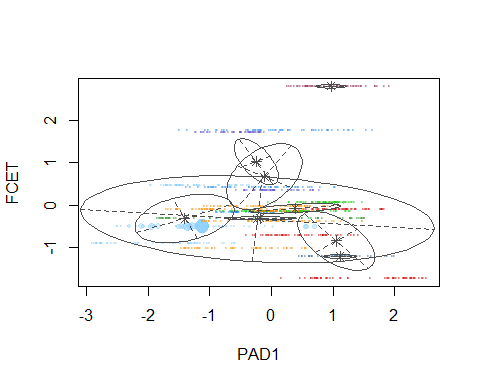
plot(mc, what="BIC")



plot(mc,what="classification",c(1,2))



plot(mc,what="uncertainty",c(1,2))



Exploratory Factory Analysis

AreaDroughtFactorAnalysis=factanal(AreaDroughtScale, factors=3)  
  
AreaDroughtFactorAnalysis

##   
## Call:  
## factanal(x = AreaDroughtScale, factors = 3)  
##   
## Uniquenesses:  
## PAD1 FCET AGSBillions YIR PPIF PPD1   
## 0.365 0.780 0.747 0.649 0.686 0.005   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## PAD1 0.776 0.164   
## FCET -0.106 -0.455   
## AGSBillions -0.149 0.451 0.167   
## YIR 0.127 0.542 -0.202   
## PPIF -0.154 0.537   
## PPD1 0.959 -0.261   
##   
## Factor1 Factor2 Factor3  
## SS loadings 1.564 0.628 0.577  
## Proportion Var 0.261 0.105 0.096  
## Cumulative Var 0.261 0.365 0.461  
##   
## The degrees of freedom for the model is 0 and the fit was 4e-04

print(AreaDroughtFactorAnalysis$loadings,cut=.4)

##   
## Loadings:  
## Factor1 Factor2 Factor3  
## PAD1 0.776   
## FCET -0.455   
## AGSBillions 0.451   
## YIR 0.542   
## PPIF 0.537   
## PPD1 0.959   
##   
## Factor1 Factor2 Factor3  
## SS loadings 1.564 0.628 0.577  
## Proportion Var 0.261 0.105 0.096  
## Cumulative Var 0.261 0.365 0.461

f.loading=AreaDroughtFactorAnalysis$loadings[,1:2]  
  
corHat=f.loading%\*%t(f.loading)+diag(AreaDroughtFactorAnalysis$uniquenesses)  
corr=cor(AreaDroughtScale)  
  
  
print('This is the Root mean Square Error')

## [1] "This is the Root mean Square Error"

rmse=sqrt(mean((corHat-corr)^2))  
rmse

## [1] 0.09609365

Confirmatory Factor Analysis

library(sem)

## Warning: package 'sem' was built under R version 4.2.3

modelCFA=specifyModel(text="  
 Drought->PAD1,lambda1,NA  
 Drought->PPD1,lambda2,NA  
 Inflation->AGSBillions,lambda3,NA  
 Inflation->YIR,lambda4,NA  
 Food\_Supply->FCET,lambda5,NA  
 Food\_Supply->PPIF,lambda6,NA  
   
 Drought<->Inflation,rho,NA  
 Inflation<->Food\_Supply,rho2,NA  
 Food\_Supply<->Drought,rho3,NA  
   
   
 PAD1<->PAD1,theta1,NA  
 FCET<->FCET,theta2,NA  
 AGSBillions<->AGSBillions,theta3,NA  
 YIR<->YIR,theta4,NA  
 PPIF<->PPIF,theta5,NA  
 PPD1<->PPD1,theta6,NA  
 Drought<->Drought,NA,1  
 Inflation<->Inflation,NA,1  
 Food\_Supply<->Food\_Supply,NA,1")

## NOTE: it is generally simpler to use specifyEquations() or cfa()  
## see ?specifyEquations

covthing=cov(AreaDrought)  
covthing

## PAD1 FCET AGSBillions YIR PPIF  
## PAD1 176.014336 -7.651136 -5.557211 3.0974076 -38.438705  
## FCET -7.651136 9978.210340 -98.128446 2.5668003 -663.643710  
## AGSBillions -5.557211 -98.128446 68.600001 1.7930724 6.698724  
## YIR 3.097408 2.566800 1.793072 1.3264388 -6.837508  
## PPIF -38.438705 -663.643710 6.698724 -6.8375084 862.587367  
## PPD1 121.331798 -59.780697 -26.874265 -0.5577278 15.227115  
## PPD1  
## PAD1 121.3317981  
## FCET -59.7806974  
## AGSBillions -26.8742649  
## YIR -0.5577278  
## PPIF 15.2271153  
## PPD1 173.1013373

ability\_sem=sem(modelCFA,covthing,nrow(AreaDrought))  
summary(ability\_sem)