

**聚类分析实验报告**

院 系 数学与统计学院

专 业 应用统计学

姓名-学号 庞浩哲 201813870023

曹行健 201813870028

杨钦 201813870052

龙福渊

指导教师 来 鹏

**二Ｏ 20 年 4 月 11 日**

第三工业数据普查结果分析

1. 摘要

1995年4月3号, 第三次全国工业普查全国召开，为了更好地发展国家工业，今早步入世界一流工业强国，亟需对普查结果进行数据分析，本文通过多种聚类分析方法，并针对各个聚类方法的特点，给出分析结果及发展建议

1. 数据初准备分析

data<-read.table("D:/MRDATA/CI.csv", header=T, sep=",")  
library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

summary( data[,-1] )

## 工业企业和生产单位数 工业总产值 从业人员年末人数  
## Min. : 250 Min. : 5.93 Min. : 1.76   
## 1st Qu.:2938 1st Qu.: 491.63 1st Qu.: 86.80   
## Median :3658 Median : 814.47 Median :152.90   
## Mean :3933 Mean : 2458.59 Mean :155.07   
## 3rd Qu.:5566 3rd Qu.: 1387.91 3rd Qu.:223.50   
## Max. :8040 Max. :47391.00 Max. :337.26

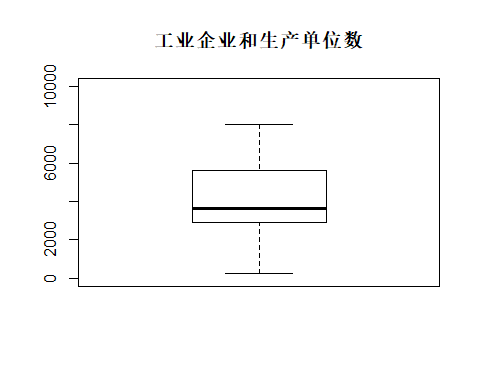
library(Rmisc)

## Warning: package 'Rmisc' was built under R version 3.6.3

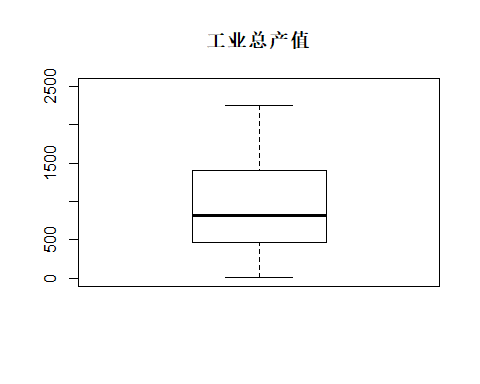
## Loading required package: lattice

## Loading required package: plyr

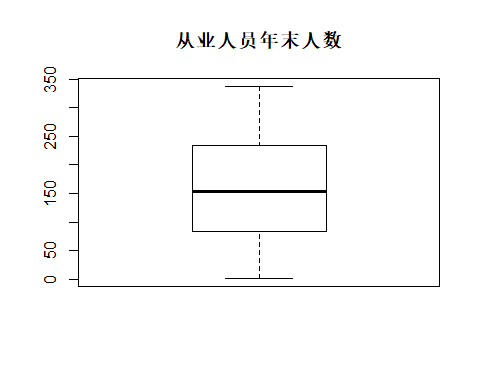
boxplot(data$工业企业和生产单位数,  
 main = "工业企业和生产单位数",  
 ylim = c(0,10000))



boxplot(data$工业总产值,  
 main = "工业总产值",  
 ylim = c(0,2500))



boxplot(data$从业人员年末人数,  
 main = "从业人员年末人数")



data<-read.table("D:/MRDATA/PI.csv", header=T, sep=",")  
library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

summary( data[,-1] )

## 工业企业和生产单位数 工业总产值 从业人员年末人数  
## Min. : 6019 Min. : 0.51 Min. : 0.93   
## 1st Qu.: 46533 1st Qu.: 40.94 1st Qu.: 17.25   
## Median :141980 Median : 132.67 Median : 60.27   
## Mean :180121 Mean : 321.08 Mean : 85.88   
## 3rd Qu.:276942 3rd Qu.: 518.53 3rd Qu.:138.54   
## Max. :564135 Max. :1711.87 Max. :297.19

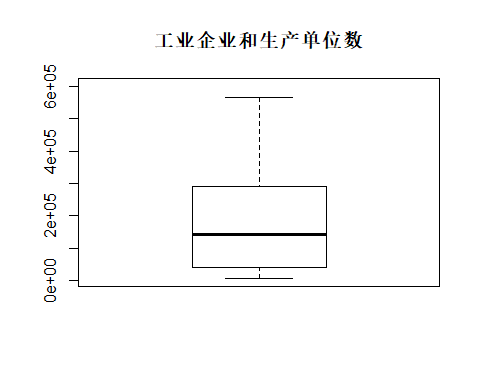
library(Rmisc)

## Warning: package 'Rmisc' was built under R version 3.6.3

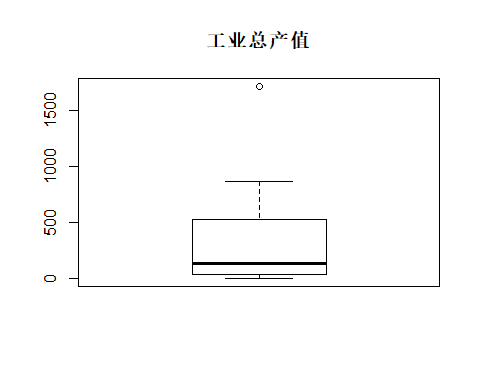
## Loading required package: lattice

## Loading required package: plyr

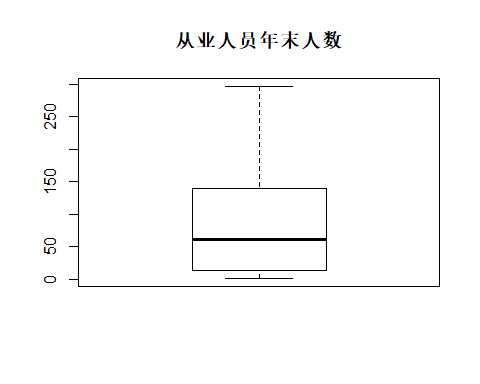
boxplot(data$工业企业和生产单位数,  
 main = "工业企业和生产单位数",  
 ylim = c(6000,600000))



boxplot(data$工业总产值,  
 main = "工业总产值")



boxplot(data$从业人员年末人数,  
 main = "从业人员年末人数")



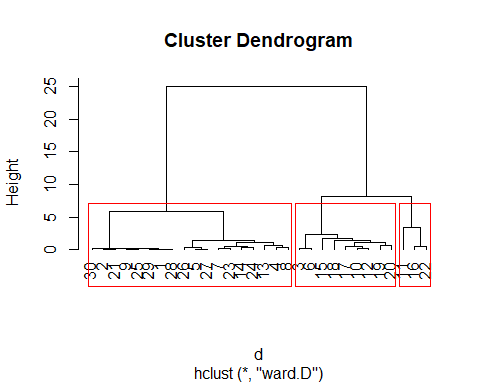
**通过箱线图可以很直观地看出我国无论是个体经济、国有经济的工业总产值平均值都略低于中位数，但是企业数量和从业人数却基本持平。结合我国的国情可以很容易解释这一现象，我国的经济发展采取先富带动后富，逐步改革开放，这对后面的聚类分析具有一定参考价值。**

1. 利用ward法分析

**（1）个体经济**

d<-dist(scale(data[,-1]), method="euclidean", diag=T, upper=F, p=2)   
hc<-hclust(d, "ward.D")   
cbind(hc$merge, round(hc$height,2))

plot(hc, hang=-1) #聚类树形图，hang指定标签在图形中所处的高度（负值时挂在0下面）  
rect.hclust(hc, k=3) #将聚成的三类用边框界定



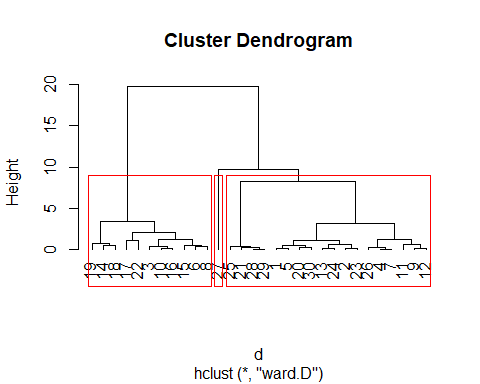
（2）国有经济

d<-dist(scale(data[,-1]), method="euclidean", diag=T, upper=F, p=2)  
d<-dist(scale(data[,-1]), method="euclidean", diag=T, upper=F, p=2)   
hc<-hclust(d, "ward.D")

cbind(hc$merge, round(hc$height,2))

plot(hc, hang=-1)

rect.hclust(hc, k=3)

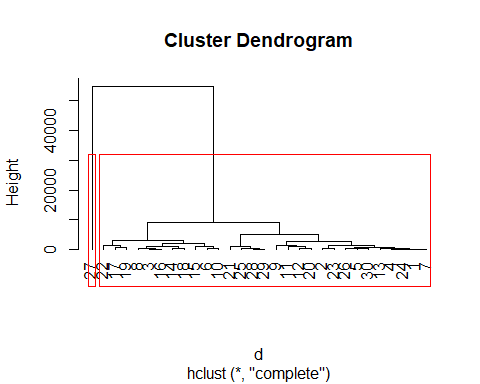


**结合实际情况来看，个体经济将青海与天津划分为一类有些不合理的，结合ward法的特点可知，ward法无法很好的判断噪点，以此保留聚类分析的结果。**

1. 最长距离法

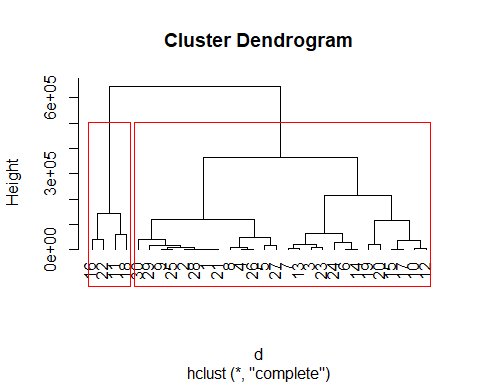
**（1）个体经济**

hc<-hclust(d, "complete") #最长距离法  
plot(hc, hang=-1) #树形图  
rect.hclust(hc, k=2) #将聚成的两类用边框界定



**（2）国有经济**

hc<-hclust(d, "complete") #最长距离法  
plot(hc, hang=-1) #树形图  
rect.hclust(hc, k=2) #将聚成的两类用边框界定



**同样作为层次聚类法的另一种分析方法，最大距离法直观来说也没能很好地处理数据中的噪点，故下面我们使用其他聚类方法进行分析。**

1. 改进k-means及DBSCAN聚类方法

**（1）个体经济**

# data3<-read.table("D:/MRDATA/PI.csv", header=T, row.names="region", sep=",") #读取文本文件  
km<-kmeans(scale(data[,-1]), 3) #k均值法，聚成3类  
sort(km$cluster) #对聚类结果进行排序

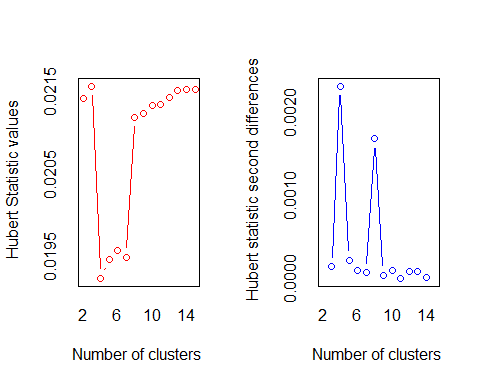
## [1] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3

KM<-kmeans(data2[,-1],3,nstart=20,algorithm="Hartigan-Wong") #聚类的个数为4, 随机集合#的个数为20, 算法为"Hartigan-Wong", 其他备选算法为"Lloyd","Forgy", "MacQueen" 算法不同，得到的聚类结果也可能不同  
sort(KM$cluster) #对分类结果进行排序并查看分类情况

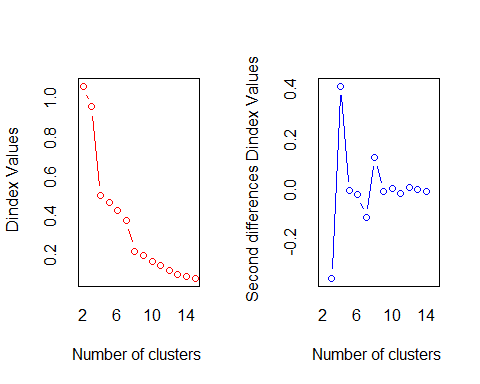
## [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3

library(NbClust)  
# Selecting the number of clusters  
  
data1.scaled <- scale(data[,-1])  
nc <- NbClust(data1.scaled, distance="euclidean",   
 min.nc=2, max.nc=15, method="average")

## Warning in pf(beale, pp, df2): 产生了NaNs  
  
## Warning in pf(beale, pp, df2): 产生了NaNs



## \*\*\* : The Hubert index is a graphical method of determining the number of clusters.  
## In the plot of Hubert index, we seek a significant knee that corresponds to a   
## significant increase of the value of the measure i.e the significant peak in Hubert  
## index second differences plot.   
##

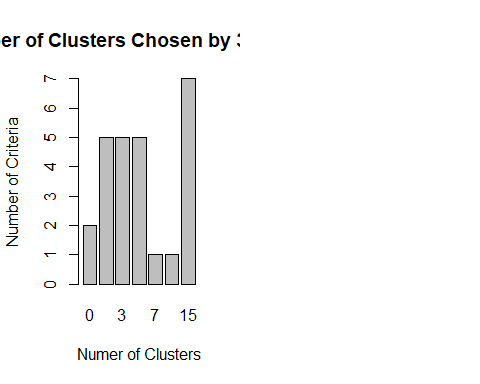


## \*\*\* : The D index is a graphical method of determining the number of clusters.   
## In the plot of D index, we seek a significant knee (the significant peak in Dindex  
## second differences plot) that corresponds to a significant increase of the value of  
## the measure.   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## \* Among all indices:   
## \* 5 proposed 2 as the best number of clusters   
## \* 5 proposed 3 as the best number of clusters   
## \* 5 proposed 4 as the best number of clusters   
## \* 1 proposed 7 as the best number of clusters   
## \* 1 proposed 12 as the best number of clusters   
## \* 7 proposed 15 as the best number of clusters   
##   
## \*\*\*\*\* Conclusion \*\*\*\*\*   
##   
## \* According to the majority rule, the best number of clusters is 15   
##   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

table(nc$Best.n[1,])

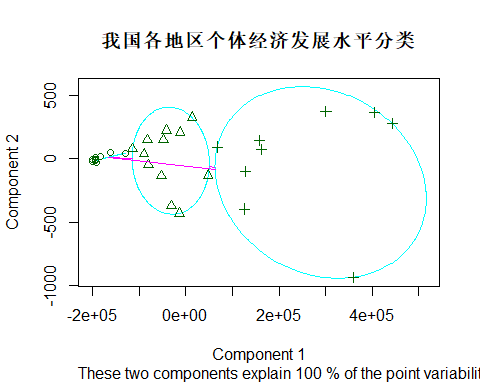
##   
## 0 2 3 4 7 12 15   
## 2 5 5 5 1 1 7

barplot(table(nc$Best.n[1,]),   
 xlab="Numer of Clusters", ylab="Number of Criteria",  
 main="Number of Clusters Chosen by 30 Criteria")  
##查询结果  
#################################################################

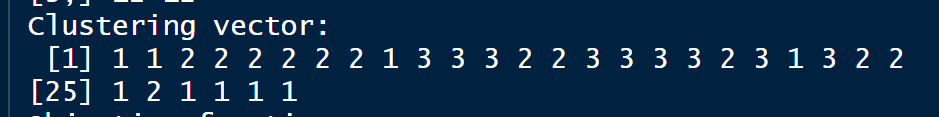


由这个图象显示聚类成3类左右比较合理

cc<-pam(d,3,stand=T)  
clusplot(cc,main = "我国各地区个体经济发展水平分类")



**最终利用DBSCAN显示结果也可显示聚类结果为三类**

**最终结果为**

**第一类：北京、天津、上海、广东、甘肃、青海、宁夏、新疆、海南**

**第二类：河北、山西、内蒙古、辽宁、吉林、黑龙江、福建、江西、陕西、云南、贵州**

**第三类：江苏、浙江、安徽、山东、河南、湖北、湖南、广西、四川**

**分类结果来看个体经济分类呈现出地域聚集的特点**

**（2）国有经济**

km<-kmeans(scale(data[,-1]), 3) #k均值法，聚成3类  
sort(km$cluster) #对聚类结果进行排序并查看分类情况

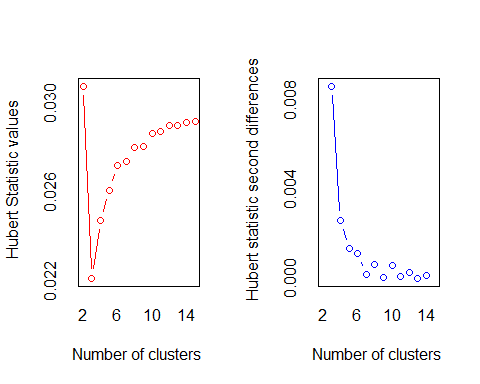
## [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3

KM<-kmeans(data2[,-1],3,nstart=20,algorithm="Hartigan-Wong")   
sort(KM$cluster)

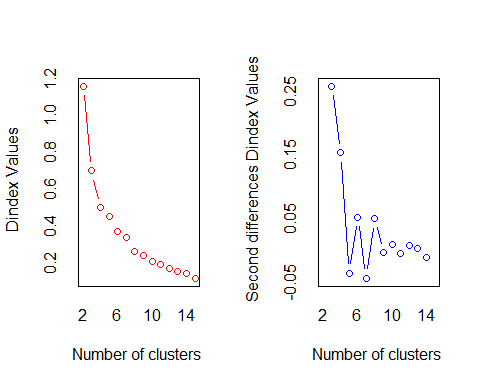
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3

library(NbClust)  
# Selecting the number of clusters  
  
data1.scaled <- scale(data[,-1])  
nc <- NbClust(data1.scaled, distance="euclidean",   
 min.nc=2, max.nc=15, method="average")

## Warning in pf(beale, pp, df2): 产生了NaNs



## \*\*\* : The Hubert index is a graphical method of determining the number of clusters.  
## In the plot of Hubert index, we seek a significant knee that corresponds to a   
## significant increase of the value of the measure i.e the significant peak in Hubert  
## index second differences plot.   
##

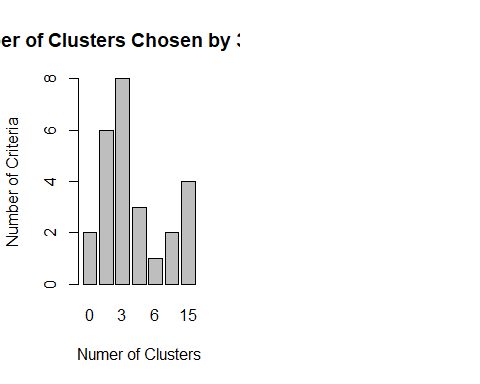


## \*\*\* : The D index is a graphical method of determining the number of clusters.   
## In the plot of D index, we seek a significant knee (the significant peak in Dindex  
## second differences plot) that corresponds to a significant increase of the value of  
## the measure.   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## \* Among all indices:   
## \* 6 proposed 2 as the best number of clusters   
## \* 8 proposed 3 as the best number of clusters   
## \* 3 proposed 4 as the best number of clusters   
## \* 1 proposed 6 as the best number of clusters   
## \* 2 proposed 12 as the best number of clusters   
## \* 4 proposed 15 as the best number of clusters   
##   
## \*\*\*\*\* Conclusion \*\*\*\*\*   
##   
## \* According to the majority rule, the best number of clusters is 3   
##   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

table(nc$Best.n[1,])

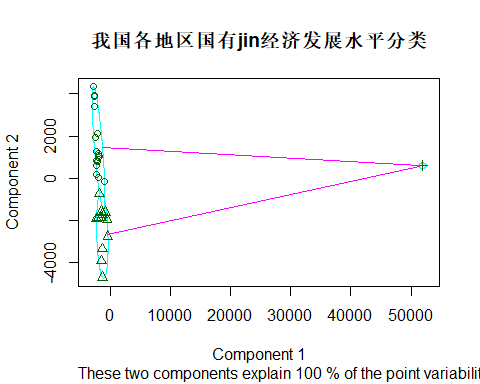
##   
## 0 2 3 4 6 12 15   
## 2 6 8 3 1 2 4

barplot(table(nc$Best.n[1,]),   
 xlab="Numer of Clusters", ylab="Number of Criteria",  
 main="Number of Clusters Chosen by 30 Criteria")  
  
#################################################################

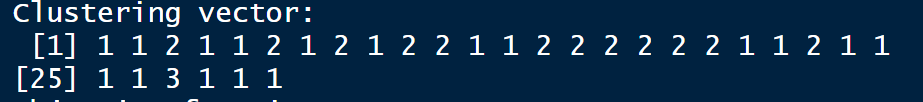


**根据示意图显示划分为三类比较适宜**

cc<-pam(d,3,stand=T)  
clusplot(cc,main = "我国各地区国有jin经济发展水平分类")



**由图像可以很明显的看出有一个离群点**



**最终结果为**

**第一类：天津、北京 山西 内蒙古 吉林 上海 安徽 福建 广西 海南 贵州云南 西藏 陕西 青海 宁夏 新疆**

**第二类：河北 辽宁 黑龙江 江苏 浙江 江西 山东 河南 湖北 湖南 广东 四川**

**第三类：甘肃**

1. 总结

**（1）经济发展水平深受历史、政策、资源等多因素的影响，1995年正值改革开放初期，国家各省整体发展水平虽然差别不大，但可以看到已经出现了一定差距，结合当今社会现状来看，分类现象愈发明显，但长久来看，为了发展成为现代化强国，向世界其他强国看齐，我们仍有很远的路要走。**

**（2）借助计算机的帮助，我们能够大幅度减少运算量，同时也能利用计算机的绘图功能很直观得看到每个点的分布情况以及类簇的内部情况，也在一定程度上反应了不同聚类分析的方法优劣之处**