Homework 5

张子乐,3230104237

第一题

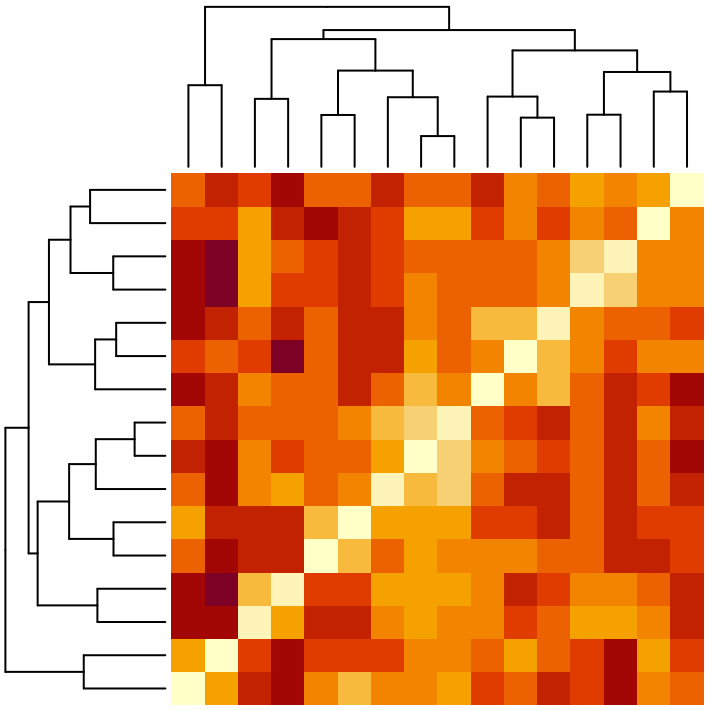
```
df <- read_excel("ex4.2.xls")
```

```
## New names:
## * `` -> `...1`
```

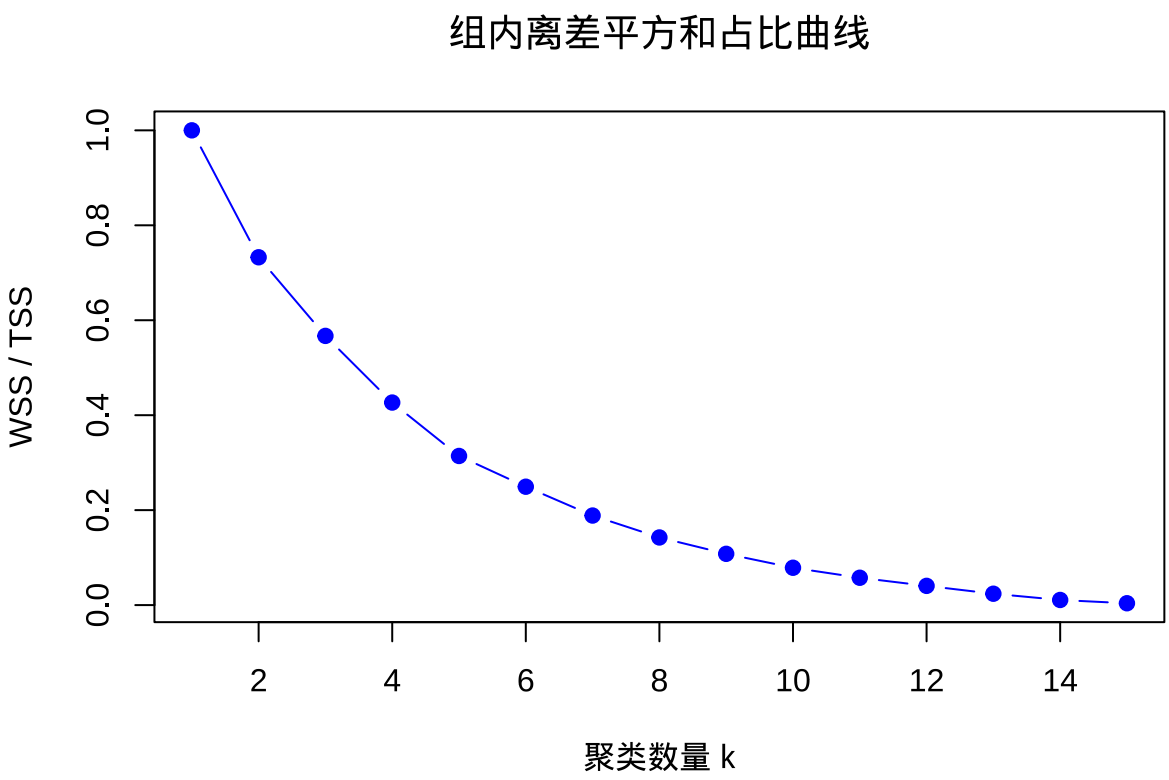
```
X <- df[,2:5] # 要聚类的变量

clust <- function(X,rows,num_k)
{
  X <- scale(X, center=TRUE, scale=TRUE)# 按列规范化，均值为 0，方差为 1
  d <- dist(X, method="euclidean", diag=F, upper=F)

  heatmap(as.matrix(d),labRow = F, labCol = F) # 热力图

  center_total <- colMeans(X)
  TSS <- sum((scale(X, center = center_total, scale = FALSE))^2)  # 总离差平方和
  k_max <- rows-1
  wss <- numeric(k_max)
  diff <- 0
  idx <- 1
  for (k in 1:k_max) {
    km <- kmeans(X, centers = k, nstart = 25, iter.max = 100)
    wss[k] <- km$tot.withinss / TSS # 计算组内离差平方和/总离差平方和
    if(k > 2){
      diff_cur <- wss[k-1] - wss[k]
      # 找变化最剧烈的拐点
      if(diff_cur > diff){
        diff <- diff_cur
```

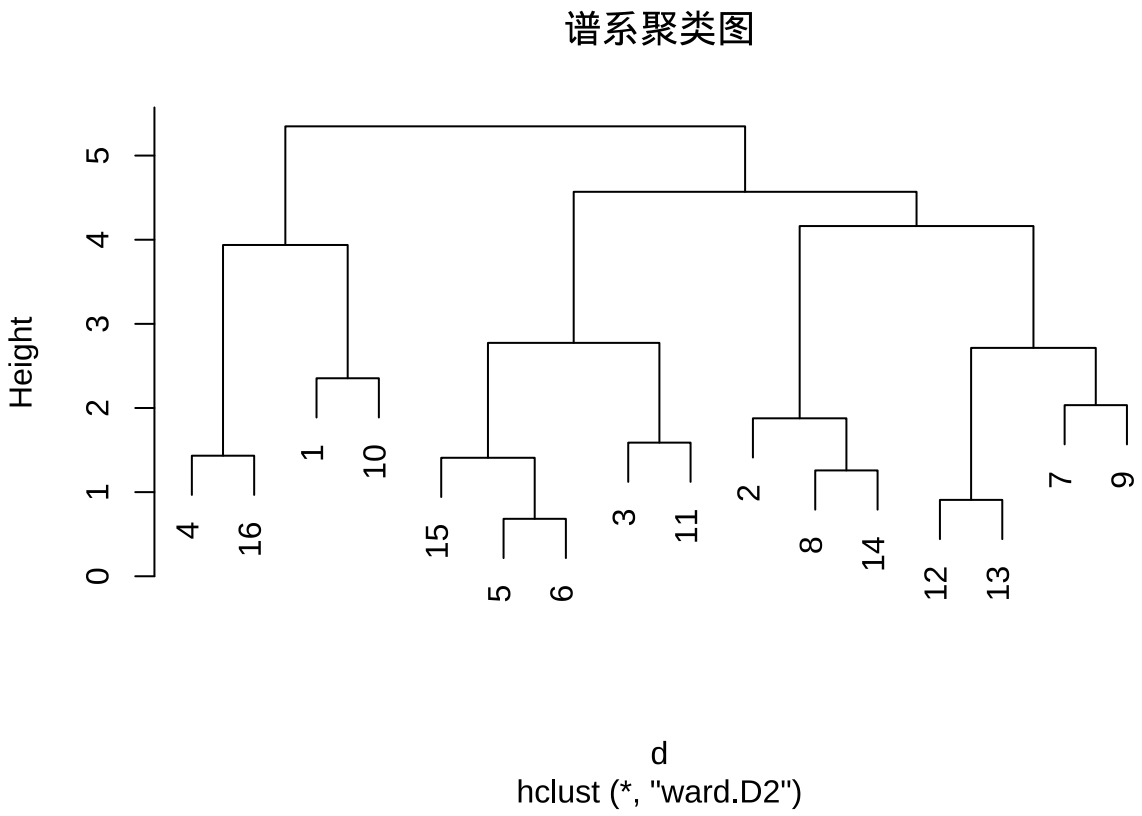
```
      idx <- k
    }
  }
}
cat(" 组内离差平方和/总离差平方和: ")
print(wss)
cat(" 拐点的索引: ")
print(idx)

# 绘制碎石图
plot(1:k_max, wss,
     type = "b", pch = 19, col = "blue",
     xlab = " 聚类数量 k",
     ylab = "WSS / TSS",
     main = " 组内离差平方和占比曲线")

model <- hclust(d,method = "ward.D2")
result=cutree(model,k=num_k)
plot(model,main = " 谱系聚类图")

mds=cmdscale(d,k=2,eig=T)
x = mds$points[,1]
y = mds$points[,2]
df_p <- data.frame(
  x = mds$points[, 1],
  y = mds$points[, 2],
  cluster = factor(result)
)
p <- ggplot(df_p, aes(x = x, y = y, color = cluster)) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_manual(
    values = c("#E41A1C", "#377EB8", "#4DAF4A"),
    name = "Cluster"
  ) +
  theme_minimal()
print(p)

# 用 kmeans 聚类
cat(" 用 k-means 方法聚类: ")
```

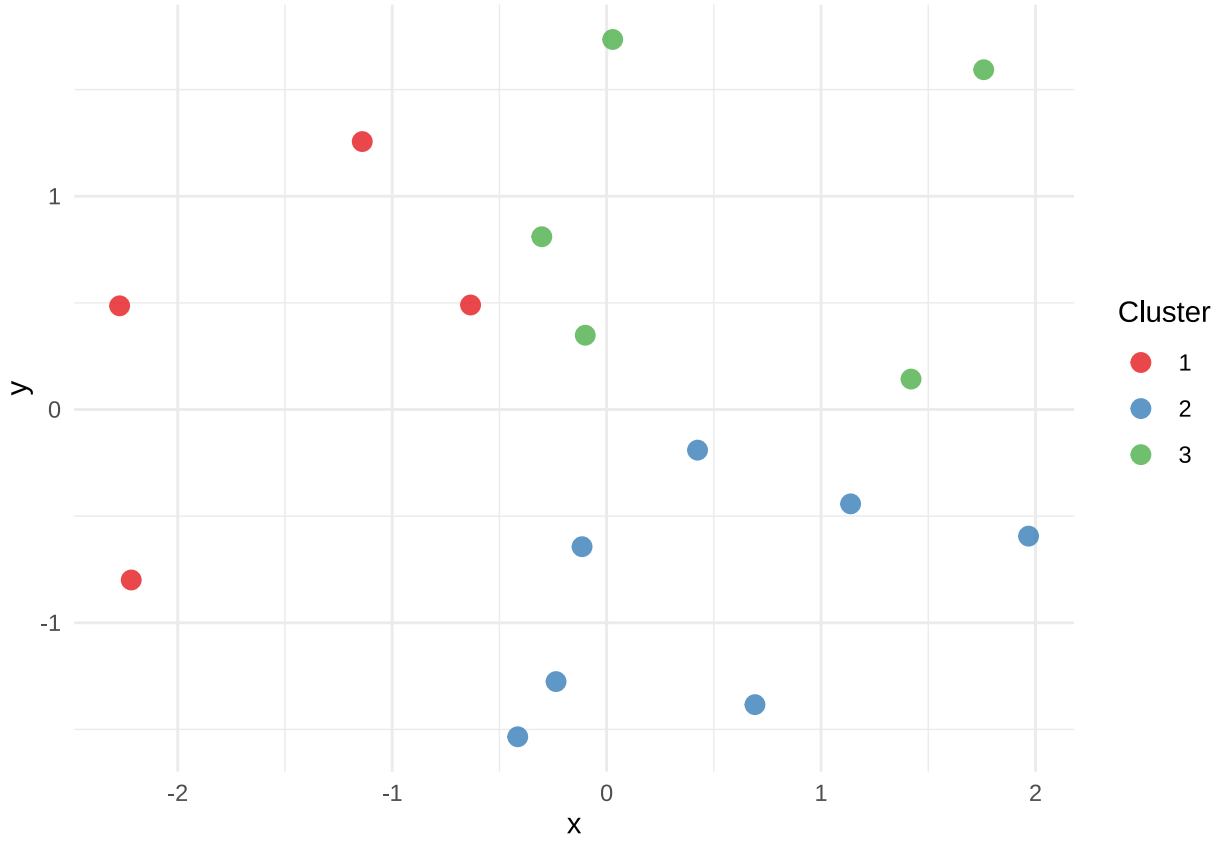
```
  model2=kmeans(X,centers=num_k,nstart=10)
  print(model2$cluster)
}

clust(X,16,3)
```



```
## 组内离差平方和/总离差平方和： [1] 1.00000000 0.73264994 0.56701966 0.42663191 0.31411620 0.2493
##  [7] 0.18860097 0.14244737 0.10797747 0.07860496 0.05758566 0.04047010
## [13] 0.02394163 0.01075135 0.00388392
## 拐点的索引: [1] 3
```


组内离差平方和占比曲线
WSS / TSS
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
14
聚类数量 k


谱系聚类图
Height
0
1
2
3
4
5
4
16
1
10
15
5
6
3
11
2
8
14
12
13
7
9
d
hclust (*, "ward.D2")




```
## 用k-means方法聚类： [1] 3 1 2 1 1 1 2 2 2 3 1 2 2 2 1 1
```

首先画出了热力图，颜色越深代表样本间距离越近。接着用 R^2 统计量确定分类个数，找到使组内离差平方和/总离差平方和变化最剧烈的分类个数，并画出了碎石图寻找拐点，最终确定分类个数为 3。

用系统聚类法中的 ward 法将样本分为三类，从谱系聚类图可以看出，{1,4,10,16} 为一类，{3,5,6,11,15} 为一类，{2,7,8,9,12,13,14} 为一类。降维后用散点图将其可视化，发现分类效果较好。

用 K-means 将样本分为三类，其中 {1,10} 为一类，{3,7,8,9,12,13,14} 为一类，{2,4,5,6,11,15,16} 为一类。两种分类方法仅有少量的区别。

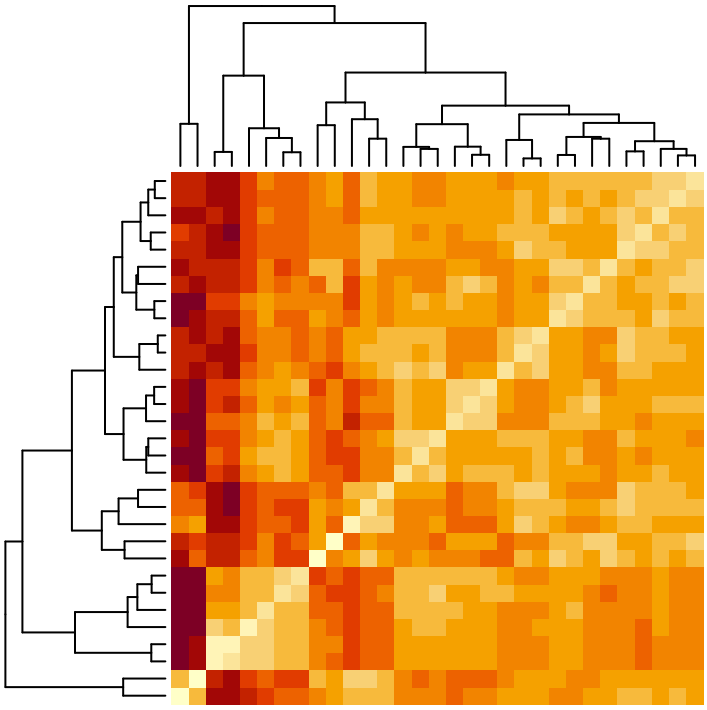
第二题

```
df2 <- read_excel("ex4.3.xls")
```

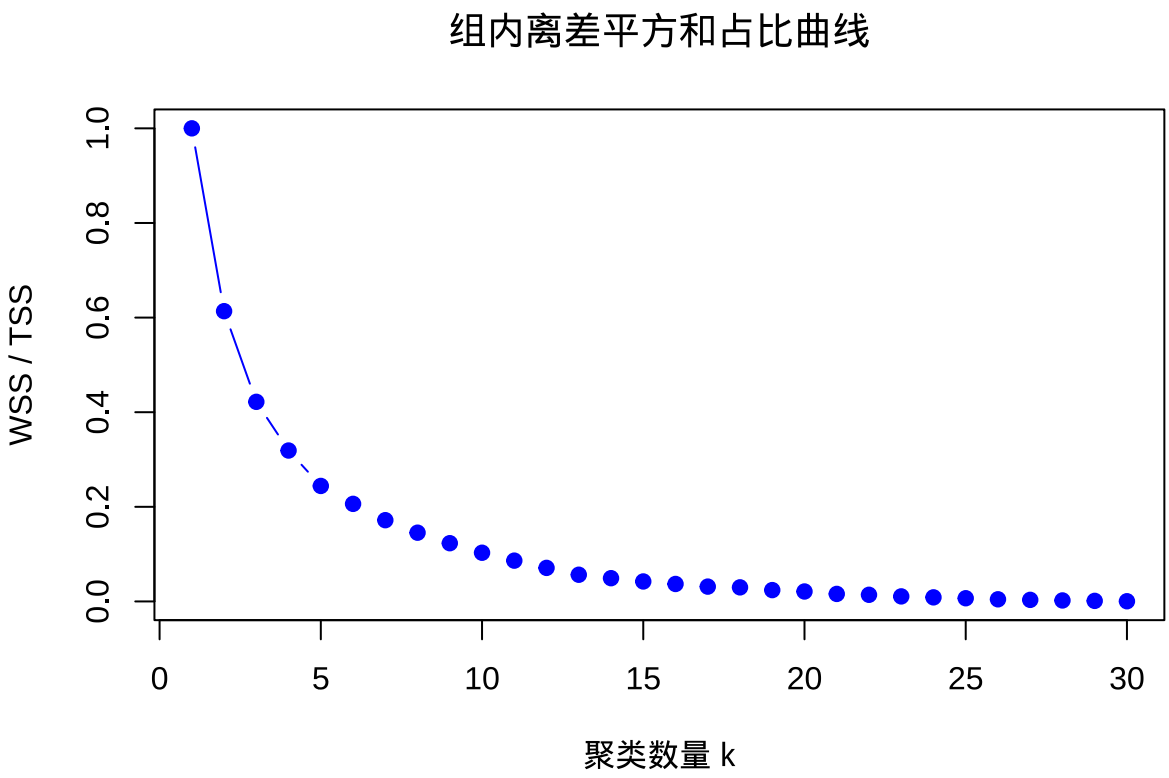
```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
```

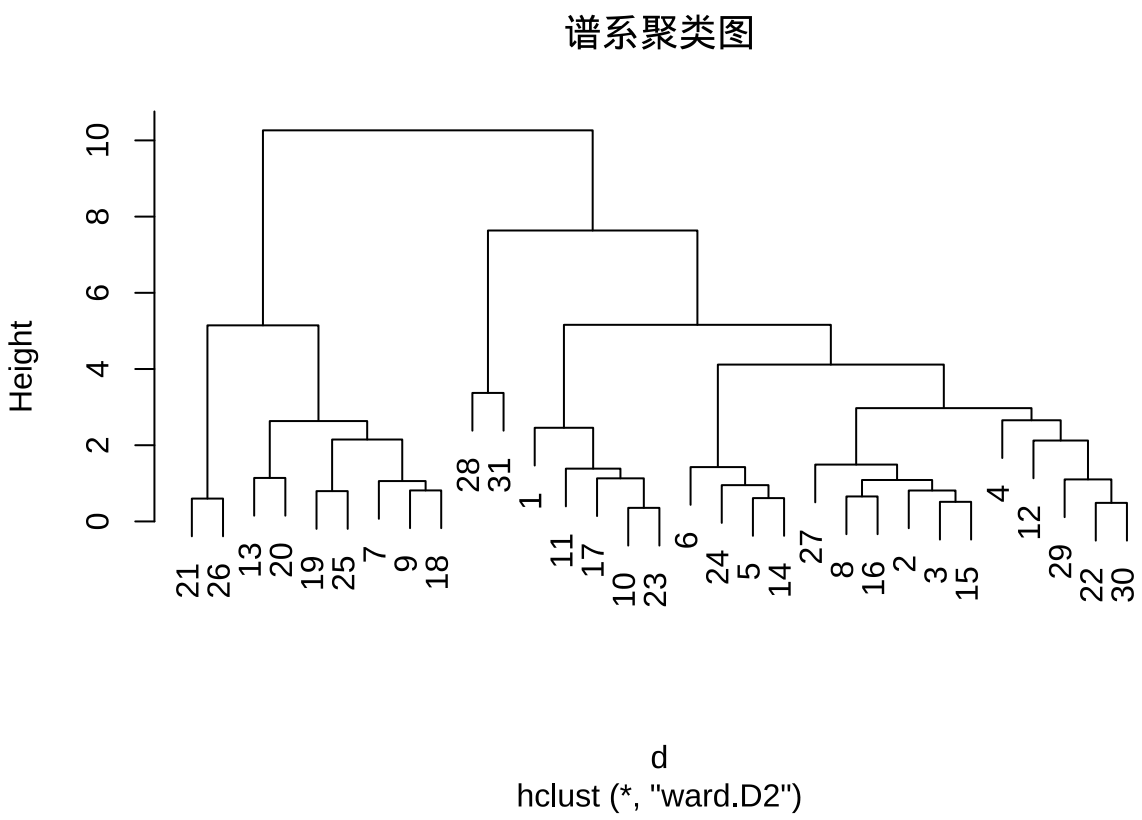
```
## * `` -> `...5`
## * `` -> `...6`
```

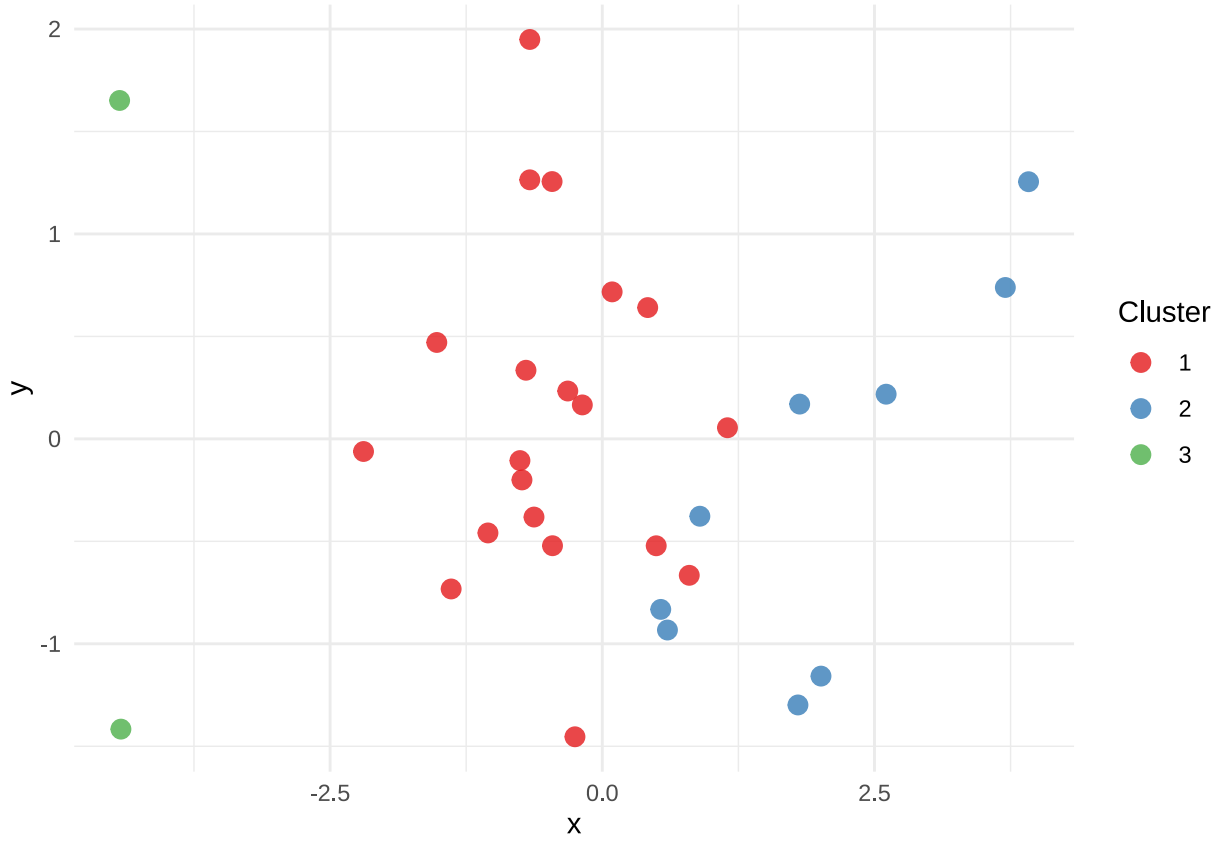
```
X2 <- df2[,2:6] # 要聚类的变量
names <- X2[1, ] %>% as.character()
data <- X2[-1, ]
colnames(data) <- names
X2 <- data %>%
  mutate(across(everything(), ~ parse_number(.x)))
clust(X2,31,3)
```



```
## 组内离差平方和/总离差平方和： [1] 1.0000000000 0.6135576207 0.4219312018 0.3190118600 0.2441521
##  [6] 0.2062438221 0.1717364496 0.1453428117 0.1231118149 0.1029531307
## [11] 0.0862793006 0.0709356912 0.0564477112 0.0491681055 0.0422515522
## [16] 0.0369476727 0.0313322787 0.0297514386 0.0239749019 0.0210518775
## [21] 0.0158918513 0.0140328556 0.0106648057 0.0086527364 0.0067491153
## [26] 0.0045431533 0.0033437071 0.0020939423 0.0012134763 0.0004232711
## 拐点的索引： [1] 3
```


组内离差平方和占比曲线
WSS / TSS
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
聚类数量 k


谱系聚类图
Height
0
2
4
6
8
10
21
26
13
20
19
25
7
9
18
28
31
1
11
17
10
23
6
24
5
14
27
8
16
2
3
15
4
12
29
22
30
d
hclust (*, "ward.D2")




```
## 用k-means方法聚类： [1] 3 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 1 2 2 2 1 1 2 3 2 2 3
```

与第一题同理，先画出热力图，颜色越深代表样本间距离越近。接着根据 R^2 统计量确定分类个数，仍然分为 3 类。

用系统聚类法中的 ward 法将样本分为三类,从谱系聚类图可以看出,{28,31} 为一类,{7,9,13,18,19,20,21,25,26} 为一类，剩下的样本为一类。降维后用散点图将其可视化，发现分类效果较好。

用 K-means 将样本分为三类，其中 {1,28,31} 为一类，{13,19,20,21,25,26} 为一类，剩下的样本为一类。两种分类方法仅有少量的区别。