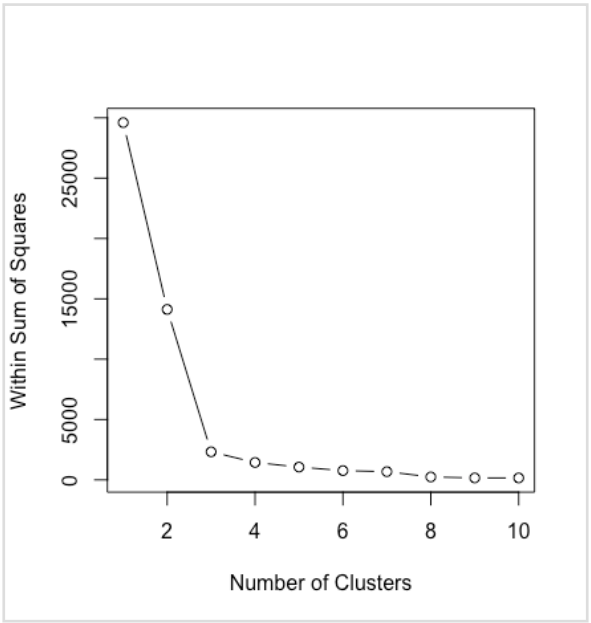
1. **已知在cluster.csv中保存着某时期某地区传染病发病及死亡情况的统计信息。使用K均值聚类方法对传染病进行聚类分析，要求完成以下任务：**

1. **尝试找出一个较为优化的聚类个数K（K最大为10），并说明理由；**
2. **使用前面找出的K值，对给的数据进行聚类分析，并说明聚类结果中各个簇包含了哪些传染病；**
3. **以发病率为横轴，病死率为纵轴，聚类簇号为颜色，绘制散点图，并为每个点标记疾病名称，结合图形及聚类数据，解释聚类结果的各簇传染病，各自的典型特征是什么。**

答：

1. library(cluster)
data = as.data.frame(read.csv("/Users/mayuheng/Desktop/ex/ex5/cluster.csv",fileEncoding = "GBK"))
rownames(data)=data[,1]
data=data[,-1]
wss <- numeric(10)
for (k in 1:10)
  wss[k] <- sum(kmeans(data, centers=k, nstart=k)$withinss)
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within Sum of Squares”)

绘图如下




所以选取k=3为聚类个数。
2.
result=kmeans(data,3,nstart = 3)
data$cluster = factor(result$cluster)
输出result如下

K-means clustering with 3 clusters of sizes 2, 23, 2

Cluster means:

| | incidence | deathrate | fatalityrate |
|---|---|---|---|
| 1 | 94.16125 | 0.179400000 | 0.198700 |
| 2 | 3.40063 | 0.008826087 | 2.343691 |
| 3 | 0.12585 | 0.123250000 | 82.357450 |

Clustering vector:

| 鼠疫 | 霍乱 | 病毒性肝炎 | 痢疾 |
|---|---|---|---|
| 2 | 2 | 1 | 2 |
| 伤寒副伤寒 | 艾滋病 | 淋病 | 梅毒 |
| 2 | 2 | 2 | 2 |
| 脊髓灰质炎 | 麻疹 | 百日咳 | 白喉 |
| 2 | 2 | 2 | 2 |
| 流脑 | 猩红热 | 出血热 | 狂犬病 |
| 2 | 2 | 2 | 3 |
| 钩端螺旋体病 | 布氏杆菌病 | 炭疽 | 乙脑 |
| 2 | 2 | 2 | 2 |
| 血吸虫 | 疟疾 | 登革热 | 新生儿破伤风 |
| 2 | 2 | 2 | 2 |
| 肺结核 | 传染性非典型肺炎 | 人禽流感 | |
| 1 | 2 | 3 | |

Within cluster sum of squares by cluster:
[1] 125.6909 1705.7287 492.4606

(between_SS / total_SS = 92.1 %)

Available components:

[1] "cluster" "centers" "totss" "withinss"
[5] "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"

输出data如下

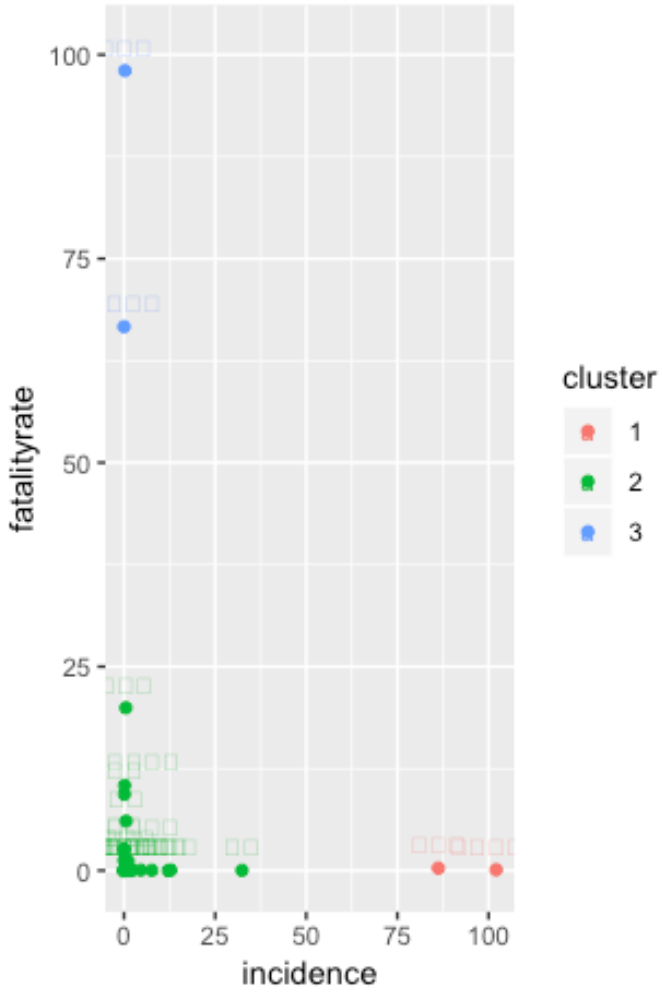
| | incidence | deathrate | fatalityrate | cluster |
|---|---|---|---|---|
| 鼠疫 | 0.0001 | 0.0000 | 0.0000 | 2 |
| 霍乱 | 0.0122 | 0.0002 | 1.2579 | 2 |
| 病毒性肝炎 | 102.0878 | 0.1034 | 0.1013 | 1 |
| 痢疾 | 32.3604 | 0.0085 | 0.0262 | 2 |
| 伤寒副伤寒 | 1.9874 | 0.0013 | 0.0654 | 2 |
| 艾滋病 | 0.5102 | 0.1018 | 19.9520 | 2 |
| 淋病 | 12.1444 | 0.0002 | 0.0019 | 2 |
| 梅毒 | 12.8002 | 0.0066 | 0.0514 | 2 |
| 脊髓灰质炎 | 0.0000 | 0.0000 | 0.0000 | 2 |
| 麻疹 | 7.6174 | 0.0027 | 0.0351 | 2 |
| 百日咳 | 0.1948 | 0.0003 | 0.1570 | 2 |
| 白喉 | 0.0001 | 0.0000 | 0.0000 | 2 |

| | | | | |
|---|---|---|---|---|
| 流脑 | 0.1276 | 0.0119 | 9.3469 | 2 |
| 猩红热 | 2.1123 | 0.0000 | 0.0000 | 2 |
| 出血热 | 1.1547 | 0.0132 | 1.1458 | 2 |
| 狂犬病 | 0.2508 | 0.2459 | 98.0482 | 3 |
| 钩端螺旋体病 | 0.0480 | 0.0012 | 2.5518 | 2 |
| 布氏杆菌病 | 1.4541 | 0.0000 | 0.0000 | 2 |
| 炭疽 | 0.0345 | 0.0009 | 2.6608 | 2 |
| 乙脑 | 0.5845 | 0.0354 | 6.0578 | 2 |
| 血吸虫 | 0.2349 | 0.0002 | 0.0977 | 2 |
| 疟疾 | 4.6035 | 0.0026 | 0.0565 | 2 |
| 登革热 | 0.0798 | 0.0000 | 0.0000 | 2 |
| 新生儿破伤风 | 0.1534 | 0.0160 | 10.4407 | 2 |
| 肺结核 | 86.2347 | 0.2554 | 0.2961 | 1 |
| 传染性非典型肺炎 | 0.0000 | 0.0000 | 0.0000 | 2 |
| 人禽流感 | 0.0009 | 0.0006 | 66.6667 | 3 |

3.

```
data$disease=row.names(data)
ggplot(data=data, aes(x=incidence, y=fatalityrate, color=cluster )) + geom_point() + geom_text(aes(label=disease), size=3,nudge_y=3)
```

绘图如下



所以三类疾病特点分别是少发少死，少发多死和多发少死。

2. 已知某交易数据库如下，如果支持度为0.5，置信度为0.6，请给出所有频繁项集，已经所有的关联规则，要求给出计算过程。

| Transaction ID | Items Bought |
|---|---|
| 2000 | A,B,C |
| 1000 | A,C |
| 4000 | A,D |
| 5000 | B,E,F |

答:

A 3 B 2 C 2 D 1 E 1 F 1
AB 1 BC 1 AC 2
所以频繁项集有{A}{B}{C}{AC}
关联规则有A-C，C-A