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：**
2. **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；**
3. **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；**
4. **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. 解决方法是尝试按照K从1个簇到最多10个簇分别进行聚类分析，然后对于每个K值，计算其内平方和WSS的大小，格局WSS的改善程度来进行选择。相关R语言代码如下：

library(cluster)

disease\_input = as.data.frame(read.csv("f:/tmp/cluster.csv"))

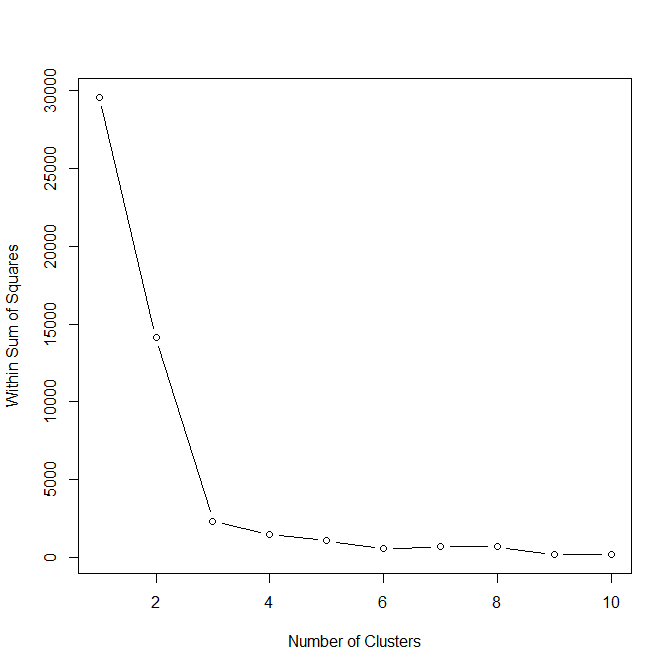
dis\_data = as.matrix(disease\_input[,c("incidence","deathrate","fatalityrate")])

wss <- numeric(10)

for (k in 1:10)

wss[k] <- sum(kmeans(dis\_data, centers=k, nstart=k)$withinss)

plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within Sum of Squares")



根据绘制的WSS和K值的散点图，可见从K为3到4之后WSS改善有限，而之前K的增加都会使得WSS显著改善。因此选择K=4作为优化的簇数。（注：这里选择K=3也是可以的，因为相比较而言，K从3到4的改善不如K从2到3的改善。）

1. 使用K等于4进行K均值聚类分析，R语言代码如下：

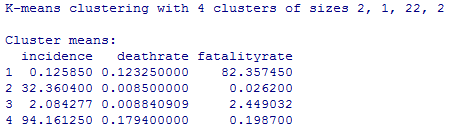
km = kmeans(dis\_data,4, nstart=4)

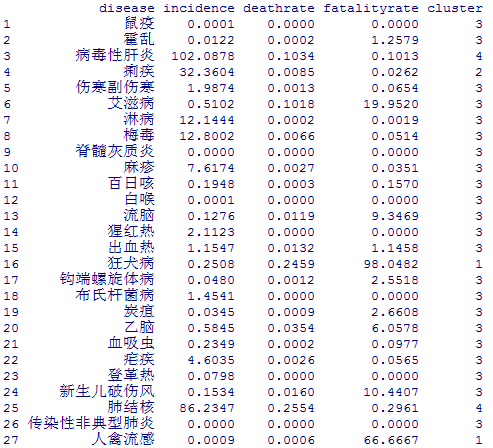
km

disease\_input$cluster = factor(km$cluster)

disease\_input

根据输出





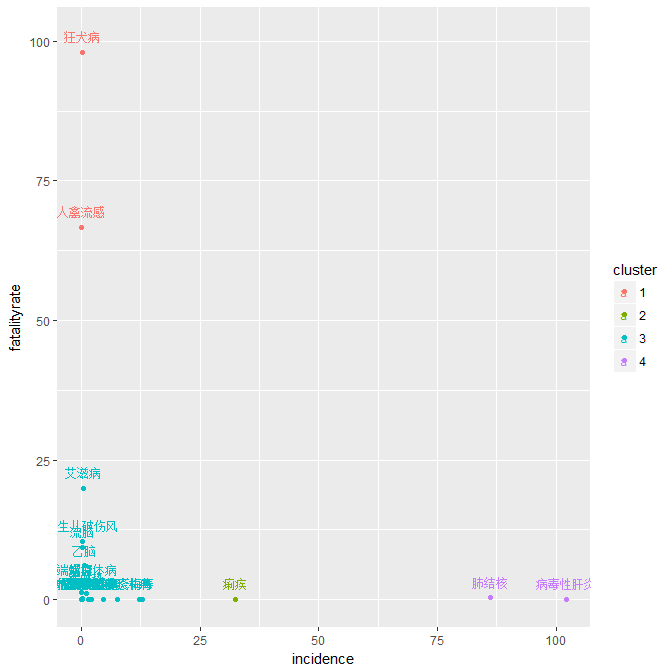
可知，所有传染病被分成四个簇，第1簇有2种传染病（狂犬病、人禽流感），第2簇有1种传染病（痢疾），第4簇有2种传染病（病毒性肝炎、肺结核），第3簇有22种传染病（其他传染病）。

注：这里可能聚类结果中每簇的传染病数为20,3,2,2等。

1. 绘制散点图的R语言代码为：

ggplot(data=disease\_input, aes(x=incidence, y=fatalityrate, color=cluster )) +

geom\_point() + geom\_text(aes(label=disease), size=3,nudge\_y=3)



结合散点图及上面聚类中心数据，可以发现，第1簇有2种传染病（狂犬病、人禽流感）都是发病率很低，但是病死率很高的，第2簇有1种传染病（痢疾）属于发病率较高但是病死率很低，第4簇有2种传染病（病毒性肝炎、肺结核）属于发病率很高但是病死率很低的，第3簇有22种传染病（其他传染病）属于发病率和病死率都很低。

1. 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。



答：参见课程PPT。