美国各州犯罪率数据的聚类分析

数据描述

该数据集(USArrests)包含的统计数据是，1973年美国50个州中的每10万居民因袭击，谋杀和强奸而被捕。此外，还给出了居住在城市地区的人口百分比。

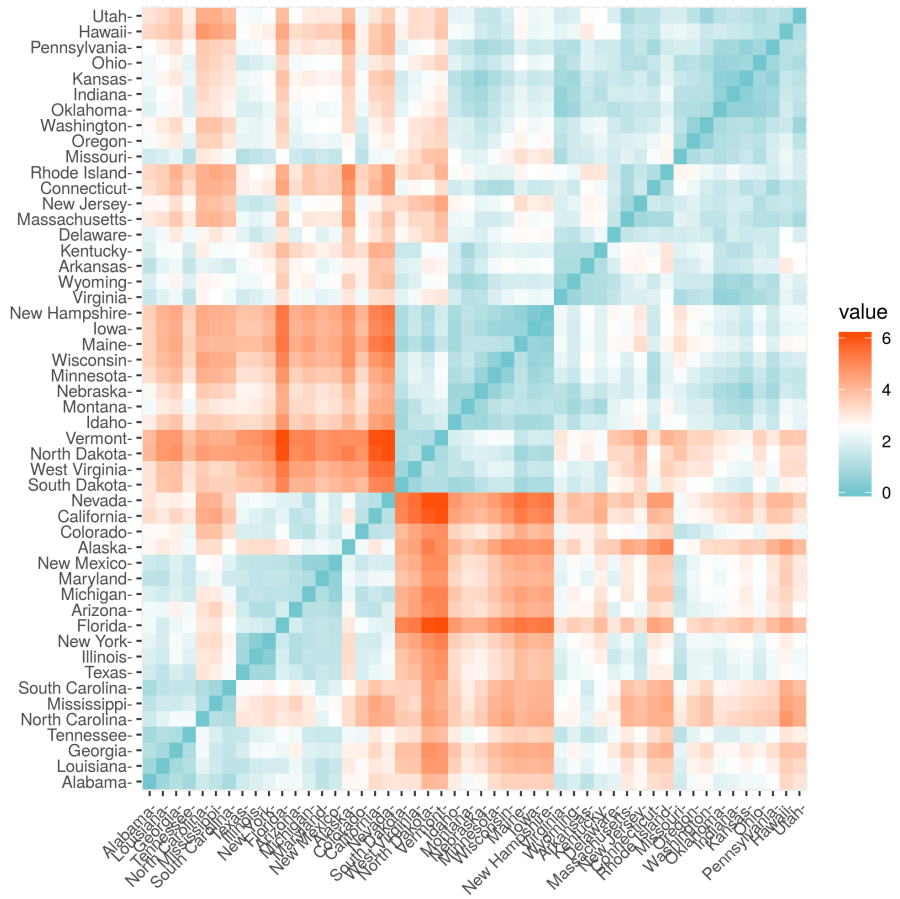
即每个州包含四个字段，分别是`Murder`(谋杀)(人数/每10万人), `Assault`(袭击), `UrbanPop`(城镇人口该州比例),`Rape`(抢劫).

预处理与EDA

数据质量良好，无重复值和缺失值。

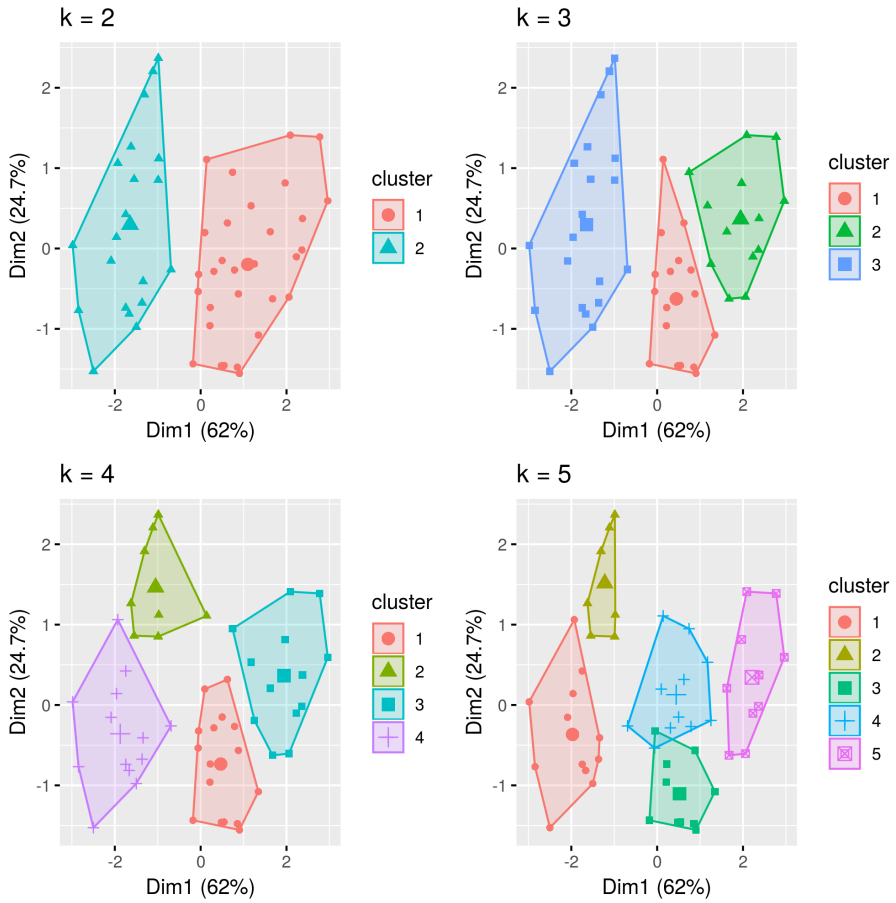
因为聚类涉及距离的度量，所以我们对数据进行标准化。

随后得到的距离矩阵如下。

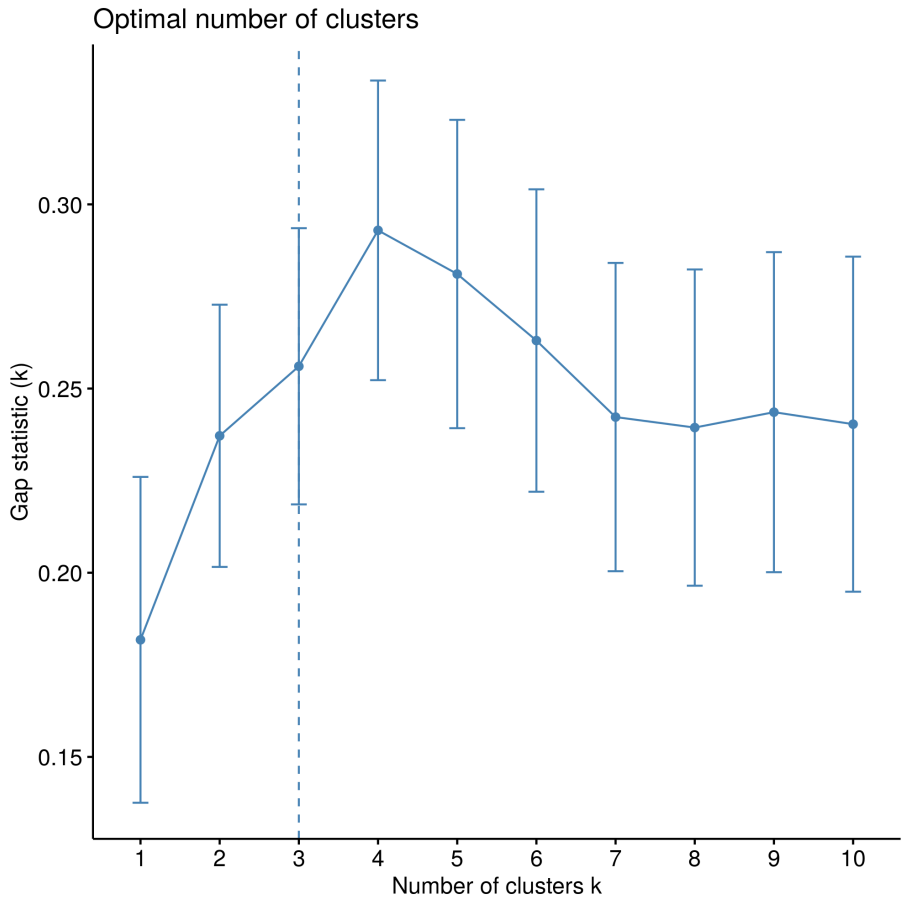


聚类算法

这里我们采用了`KMeans`算法，首先尝试选取不同个数的聚类中心来聚类。



之后采用`Gap Statistic Method`方法定出最优的类中心个数为4.



然后执行最终的聚类程序。最后给出每个聚类中，各种犯罪的平均比率。最后的聚类结果如下。

结论

CLUSTER	MURDER	ASSAULT	URBANPOP	RAPE
1	13.93750	243.62500	53.75000	21.41250
2	3.60000	78.53846	52.07692	12.17692
3	5.65625	138.87500	73.87500	18.78125
4	10.81538	257.38462	76.00000	33.19231

聚类1中各州的城镇人口比率较低，同时各项犯罪均属于中上水平，属于经济较为落后且犯罪率较高的区域；

聚类2中的虽然城镇人口比率最低，但是各项犯罪指标均是出于最低的水平，属于治安工作做的较好的区域；

聚类3中虽然城市人口比率多，经济比较好，各项犯罪指标处于中游，治安工作出于一般水平；

聚类4是平均城市人口比率最多的一些州，但是其各项犯罪都普遍多于其他区域，尤其是袭击和抢劫等犯罪率最高，属于明显的疏于治理。