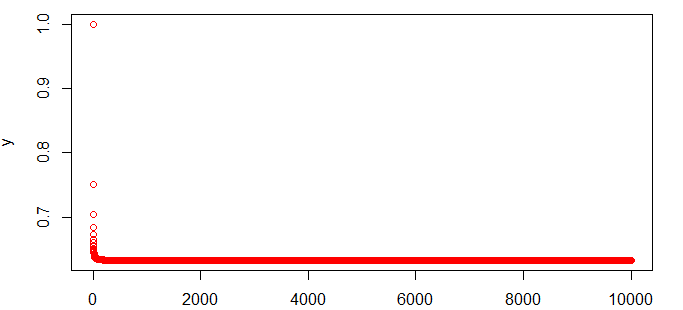
】

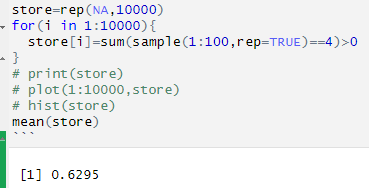
A、1-1/n

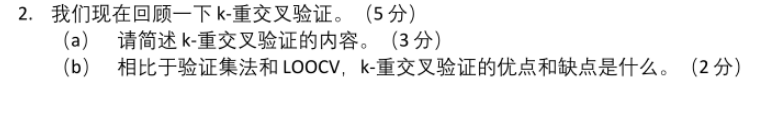
B、

C、（1-1/n）^n

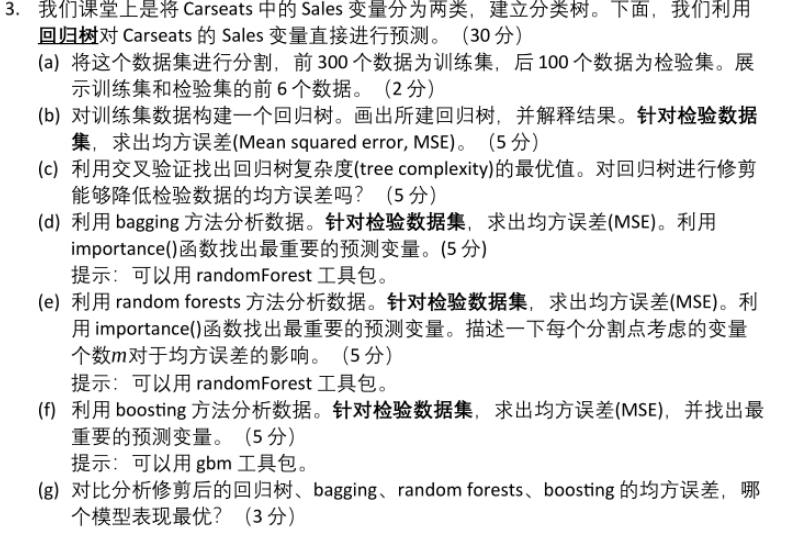
D、每个样本被选中的概率为1/n，bootstarp数据集中的概率为1-（1-1/n）^n，当n越大，概率越低，最后无限趋近0



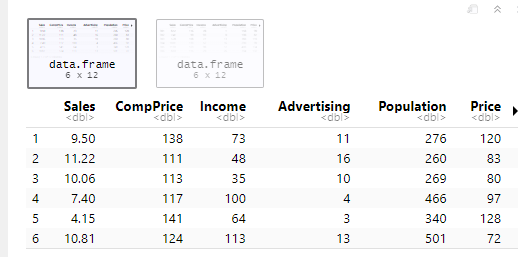
E、包含第四个数据的概率与估计包含第四个数据的概率存在误差



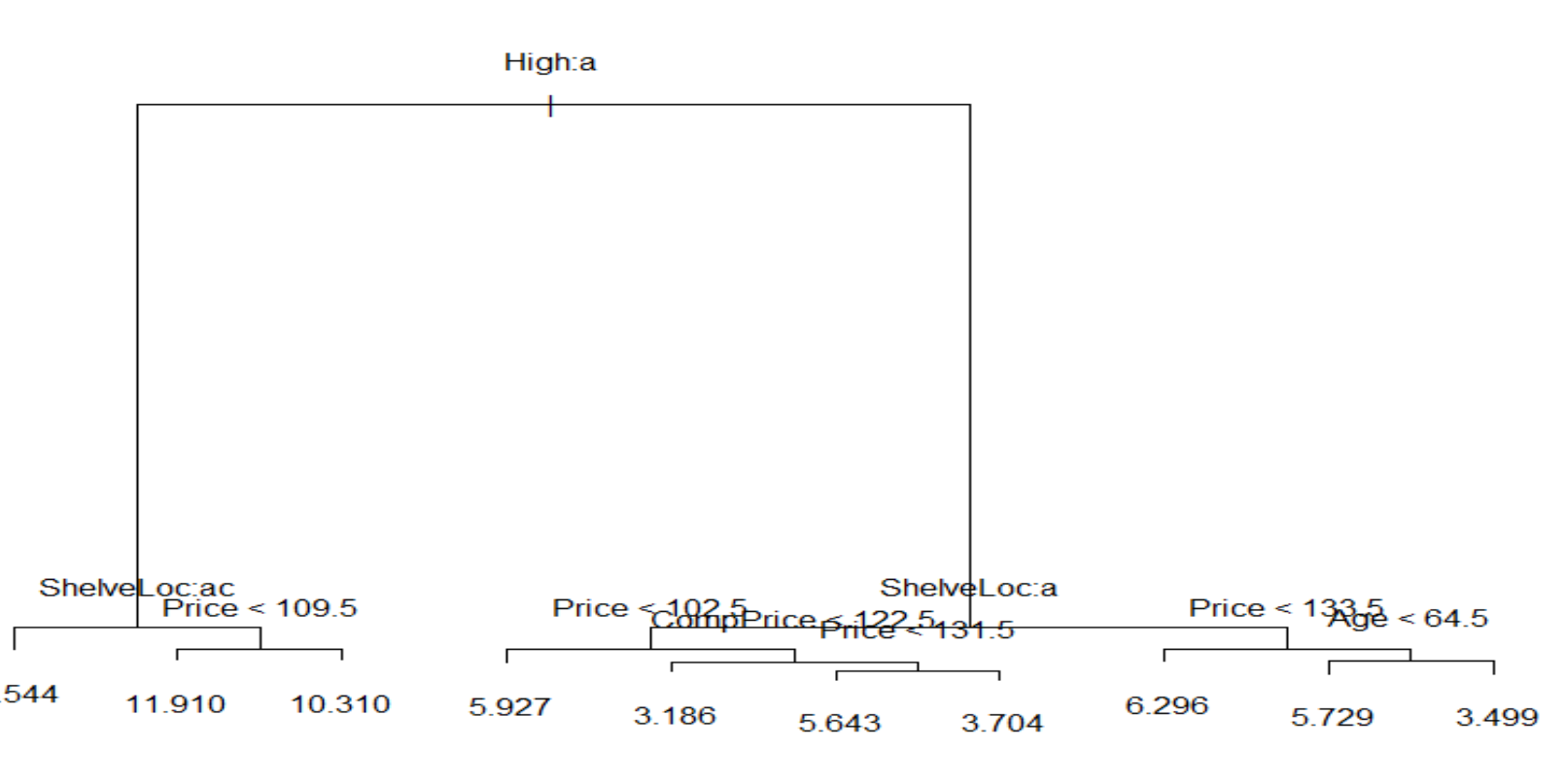
1. 将数据集平均分成k组，k重交叉验证每次选择1组作为验证集，采用平方差作为评价指标
2. 相比于验证集法和loocv，k重交叉验证的优点是以k组作为训练集，大大减少了计算量，缩短运行时间，也不会对模型结构造成太大影响，loocv是k重交叉验证的特殊情况，k等于数据集大小n，缺点：数据都只被用溢出，没有充分利用。



a

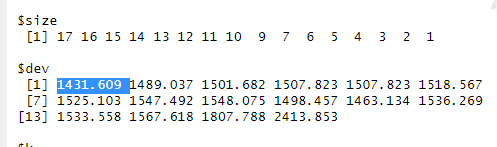


B



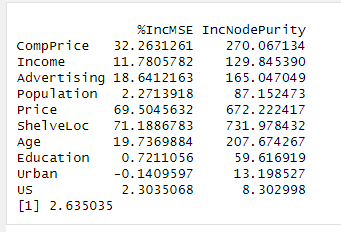
回归树MSE为5.223056

C、

最优值是10，

交叉验证剪枝后MSE是4.941963

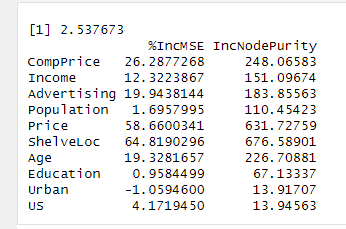
D、



RandomForest-Mse=2.537673

Bagging-MSE: 2.635035

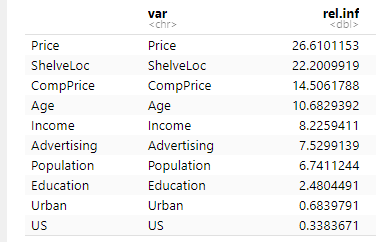
E



最重要的预测变量是ComPrice

RandomForest-Mse=2.537673

F、

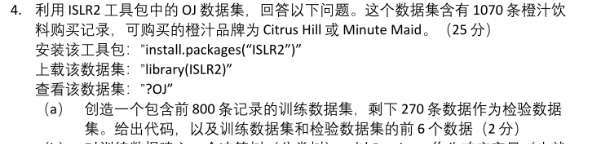


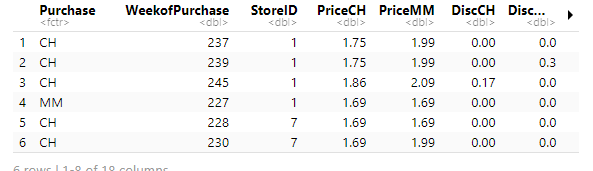
最重要的预测变量是Price

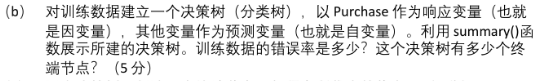
Boost-MSE : 2.221459

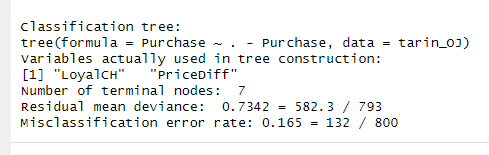
G、

根据上述求得的MSE，boosting求得的最小，所以Boosting模型表现最优



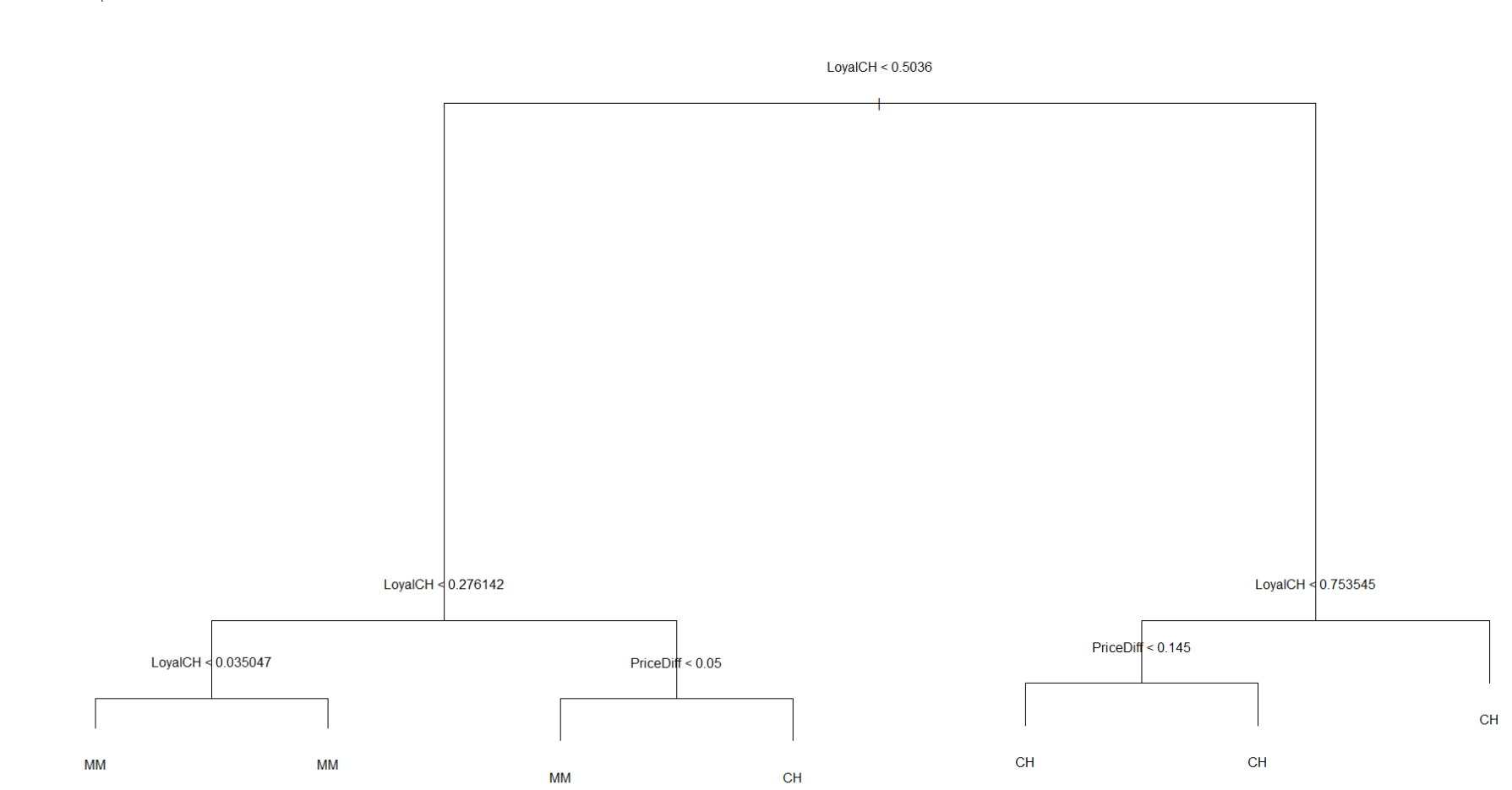






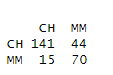
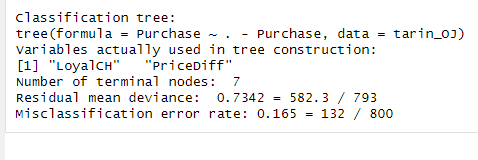
准确率为0.7342，所以错误率为1-0.7342，有7个终端节点





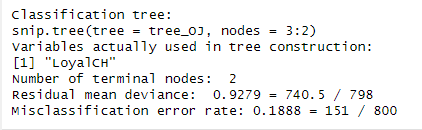
我选择最左边结果为MM的终端节点，当LoyalCH小于0.276142，且大于0.035047时，Purhcase将预测为MM



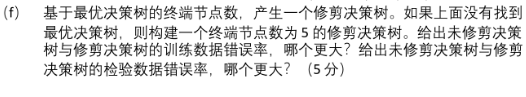
 

错误率为1-（0.5222+0.2593）





Tree size为2



训练数据

修剪后：Misclassification error rate: 0.1888 = 151 / 800

修剪前：Misclassification error rate: 0.165 = 132 / 800

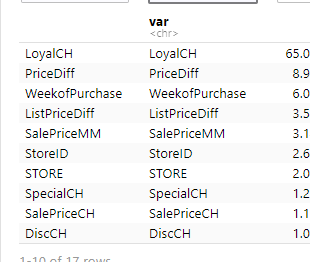
检验数据错误率

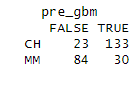
修剪后：63/270

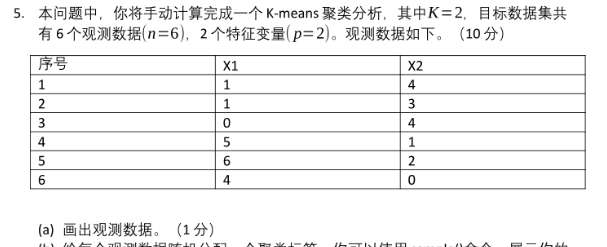
修剪前：59/270

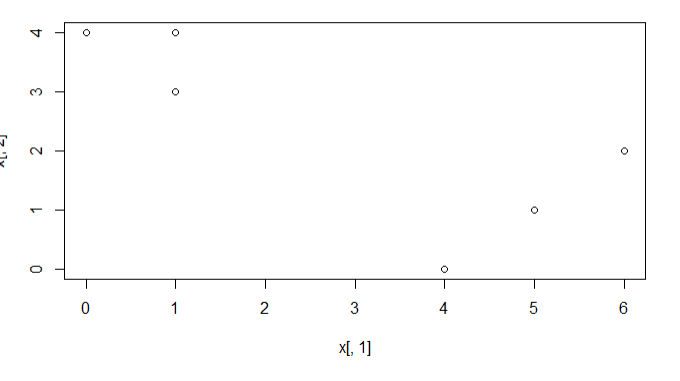
修建后的错误率更大



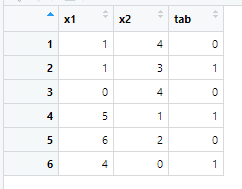
LoyalCH、PreciDiff,

检验数据错误率为53/270

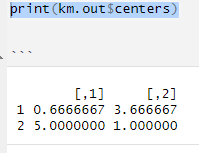






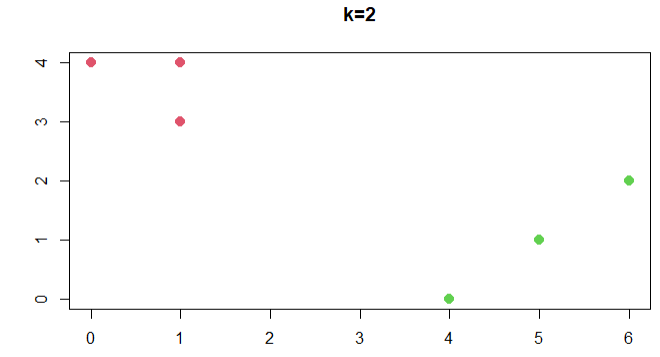


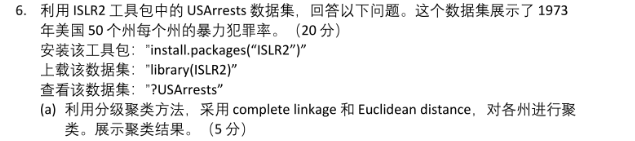


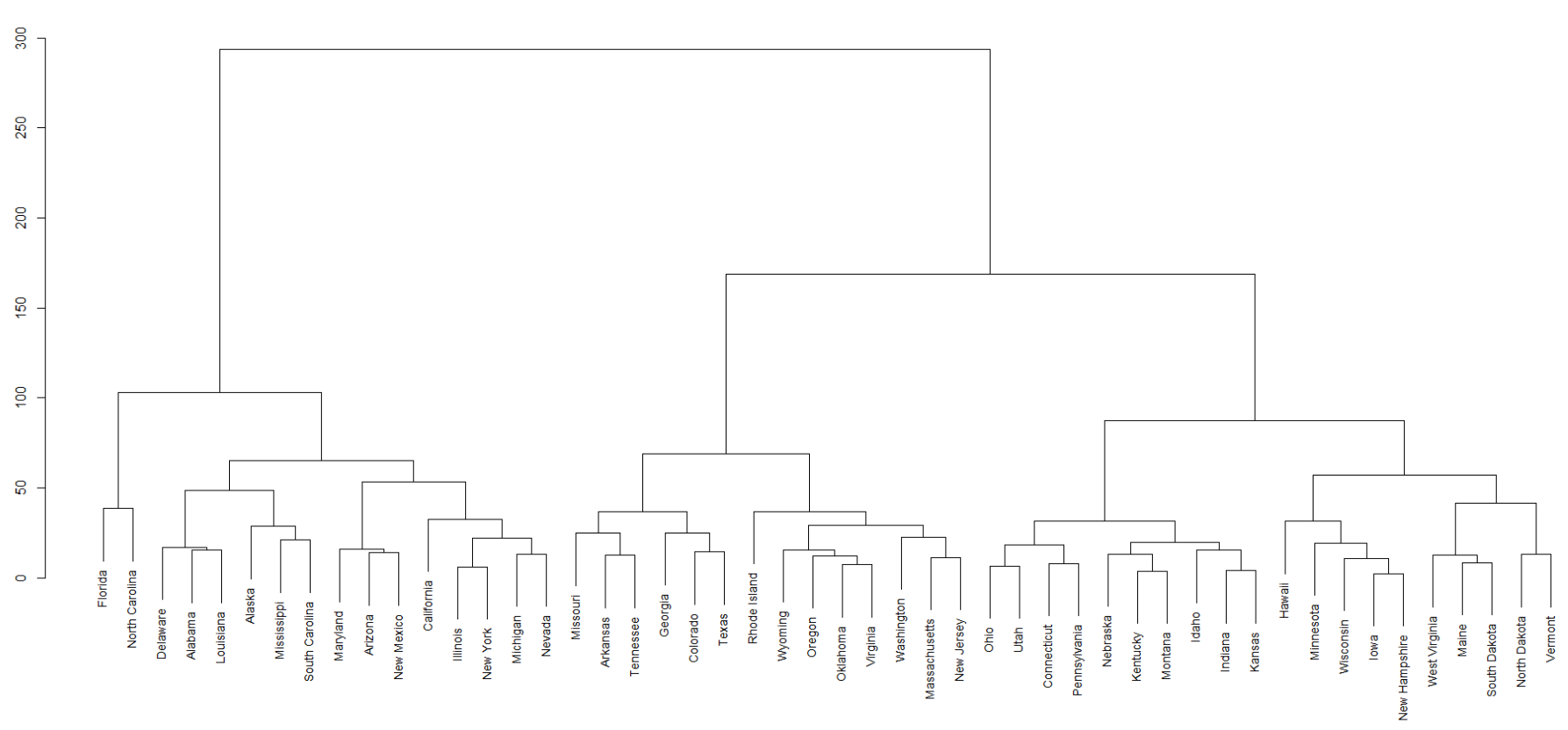




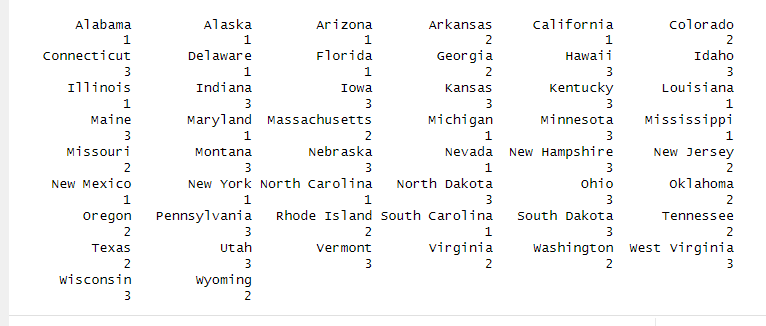




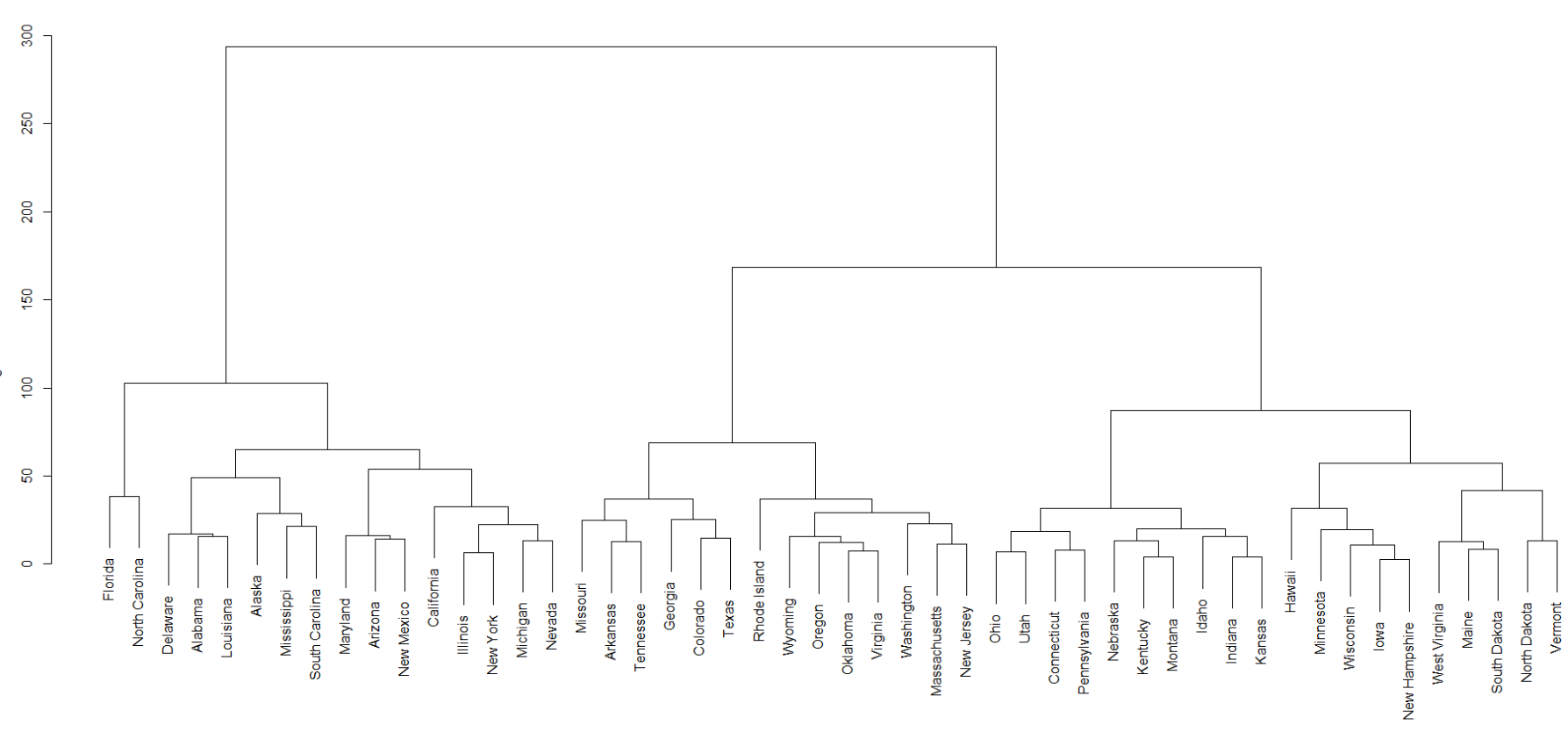










****



对变量进行归一化处理可以在一定程度上影响分级聚类的结果，因为如果某个变量的变化范围比其他变量大很多，那么在未归一化时，这个变量会对聚类结果产生更大的影响。在归一化之后，所有变量的重要性都变得相同，避免了某个变量对聚类结果的过度影响。因此，进行归一化处理可以使聚类结果更加准确可靠。

一般来说，在进行聚类分析之前，建议对数据进行归一化处理，以确保不同变量之间的权重是平等的，聚类结果具有更高的可靠性。常用的归一化方法包括最大-最小值归一化、z-score标准化、分位数归一化等。具体选择哪种方法取决于数据集的特点和分析目的。