Data Science II: Homework 4

Name: Jasmin Martinez (JRM2319) Date: 04/20/25

QUESTION 1: In this exercise, we will build tree-based models using the College data (see "College.csv" in Homework 2). The response variable is the out-of-state tuition (Outstate). Partition the dataset into two parts: training data (80%) and test data (20%).

```
# initial data steps--importing and partitioning
College = read.csv("College.csv")
head(College)
```

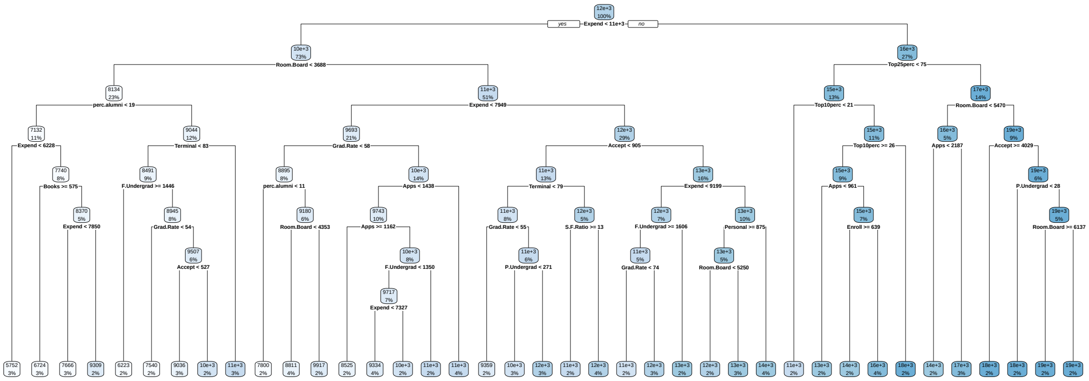
```
##                        College Apps Accept Enroll Top10perc Top25perc
## 1 Abilene Christian University 1660   1232    721        23        52
## 2           Adelphi University 2186   1924    512        16        29
## 3               Adrian College 1428   1097    336        22        50
## 4          Agnes Scott College  417    349    137        60        89
## 5    Alaska Pacific University  193    146     55        16        44
## 6            Albertson College  587    479    158        38        62
##   F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1        2885         537     7440       3300   450     2200  70       78
## 2        2683        1227    12280       6450   750     1500  29       30
## 3        1036          99    11250       3750   400     1165  53       66
## 4         510          63    12960       5450   450      875  92       97
## 5         249         869     7560       4120   800     1500  76       72
## 6         678          41    13500       3335   500      675  67       73
##   S.F.Ratio perc.alumni Expend Grad.Rate
## 1      18.1          12   7041        60
## 2      12.2          16  10527        56
## 3      12.9          30   8735        54
## 4       7.7          37  19016        59
## 5      11.9           2  10922        15
## 6       9.4          11   9727        55
```

```
datSplit = initial_split(data = College, prop = 0.8)
trainData = training(datSplit)
testData = testing(datSplit)
head(trainData)
```

```
##                      College Apps Accept Enroll Top10perc Top25perc F.Undergrad
## 1               Hope College 1712   1483    624        37        69        2505
## 2           Hamilton College 3140   1783    454        40        82        1646
## 3          Notre Dame College  379    324    107        15        37         500
## 4         Assumption College 2135   1700    491        23        59        1708
## 5          Blackburn College  500    336    156        25        55         421
## 6 St. Thomas Aquinas College  861    609    215         10         27         1117
##   P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1         208    12275       4341   465     1100  72       81      12.5
## 2          24    19700       5050   300      800  91       96       9.6
## 3         311     9990       4900   400      600  44       47      12.1
## 4         689    12000       5920   500      500  93       93      13.8
## 5          27     6500       2700   500     1000  76       76      14.3
## 6         815     8650       5700   500     1750  69       73      16.1
##   perc.alumni Expend Grad.Rate
## 1          40   9284        72
## 2          60  17761        91
## 3          26   4948        33
## 4          30   7100        88
## 5          53   8377        51
## 6          13   6534        67
```

```r
set.seed(1)
tree1 = rpart(formula = Outstate ~ . - College,
              data = trainData,
              control = rpart.control(cp=0))
rpart.plot(tree1) #this gives the full tree, but we want a more complex and smaller tree
```

1.A: Build a regression tree on the training data to predict the response (10pts). Create a plot

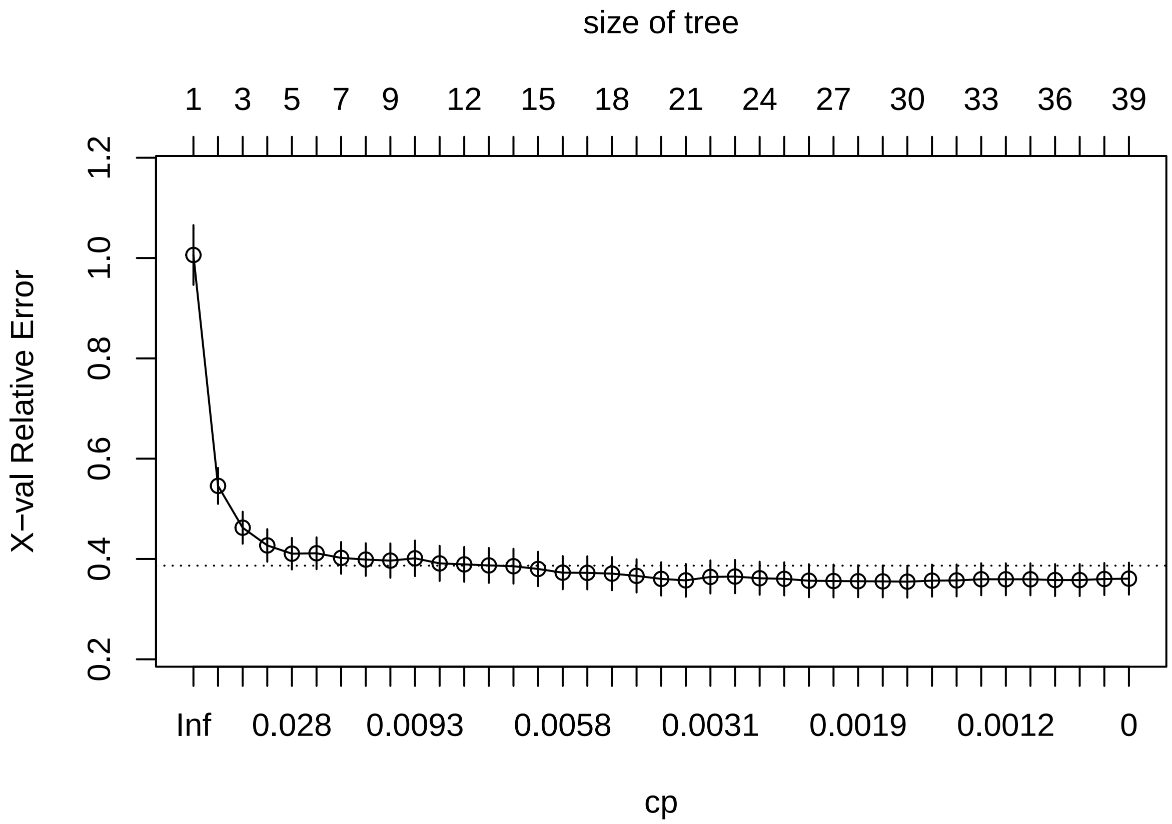


of the tree (10pts).

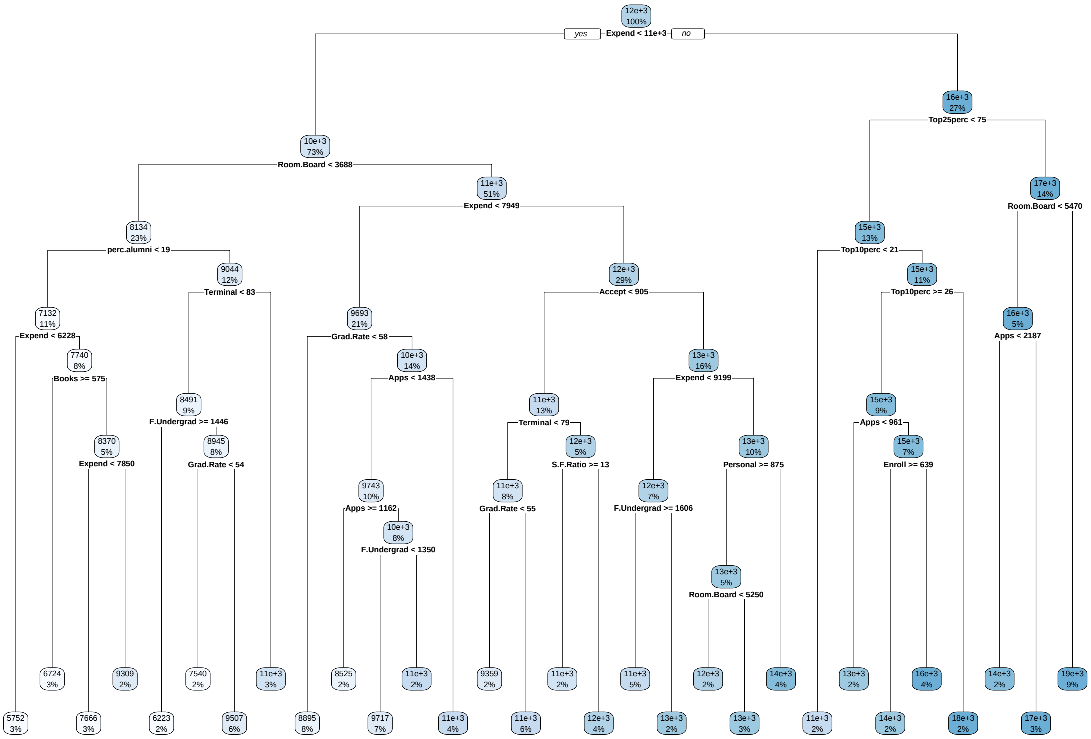
```r
printcp(tree1)
```

```
##
## Regression tree:
## rpart(formula = Outstate ~ . - College, data = trainData, control = rpart.control(cp = 0))
##
## Variables actually used in tree construction:
##  [1] Accept      Apps        Books       Enroll      Expend      F.Undergrad
##  [7] Grad.Rate   P.Undergrad perc.alumni Personal    Room.Board  S.F.Ratio
## [13] Terminal    Top10perc   Top25perc
##
## Root node error: 6220836924/452 = 13762914
##
## n= 452
##
##            CP nsplit rel error  xerror     xstd
## 1  0.50478790      0   1.00000 1.00612 0.059671
## 2  0.09802256      1   0.49521 0.54577 0.035959
## 3  0.05086808      2   0.39719 0.46234 0.031965
## 4  0.03836392      3   0.34632 0.42700 0.032558
## 5  0.02006144      4   0.30796 0.41052 0.031474
## 6  0.01745407      5   0.28790 0.41138 0.031811
## 7  0.01510399      6   0.27044 0.40214 0.031793
## 8  0.01475475      7   0.25534 0.39877 0.032594
## 9  0.00941134      8   0.24058 0.39672 0.034347
## 10 0.00928328      9   0.23117 0.40126 0.035514
## 11 0.00696378     10   0.22189 0.39118 0.035065
## 12 0.00694585     11   0.21492 0.38921 0.034855
## 13 0.00661551     12   0.20798 0.38715 0.034807
## 14 0.00590749     13   0.20136 0.38555 0.034818
## 15 0.00581922     14   0.19546 0.38006 0.034358
## 16 0.00572648     15   0.18964 0.37269 0.033111
## 17 0.00568659     16   0.18391 0.37233 0.033130
## 18 0.00534971     17   0.17822 0.37075 0.033066
## 19 0.00444544     18   0.17287 0.36629 0.033273
## 20 0.00395809     19   0.16843 0.36025 0.033239
## 21 0.00349895     20   0.16447 0.35732 0.032738
## 22 0.00281525     21   0.16097 0.36415 0.033097
## 23 0.00258565     22   0.15816 0.36491 0.033220
## 24 0.00242768     23   0.15557 0.36154 0.033036
## 25 0.00230820     24   0.15314 0.36034 0.033008
## 26 0.00223183     25   0.15083 0.35663 0.032938
## 27 0.00198894     26   0.14860 0.35603 0.032806
## 28 0.00190646     27   0.14661 0.35563 0.031900
## 29 0.00182891     28   0.14471 0.35527 0.031902
## 30 0.00170134     29   0.14288 0.35476 0.031904
## 31 0.00146860     30   0.14118 0.35673 0.031914
## 32 0.00130244     31   0.13971 0.35720 0.031905
## 33 0.00121780     32   0.13841 0.35954 0.031953
## 34 0.00119379     33   0.13719 0.35950 0.031953
## 35 0.00117946     34   0.13599 0.35950 0.031953
## 36 0.00113651     35   0.13482 0.35796 0.031890
## 37 0.00103429     36   0.13368 0.35796 0.031890
## 38 0.00024057     37   0.13264 0.36001 0.031931
## 39 0.00000000     38   0.13240 0.36072 0.031920
```

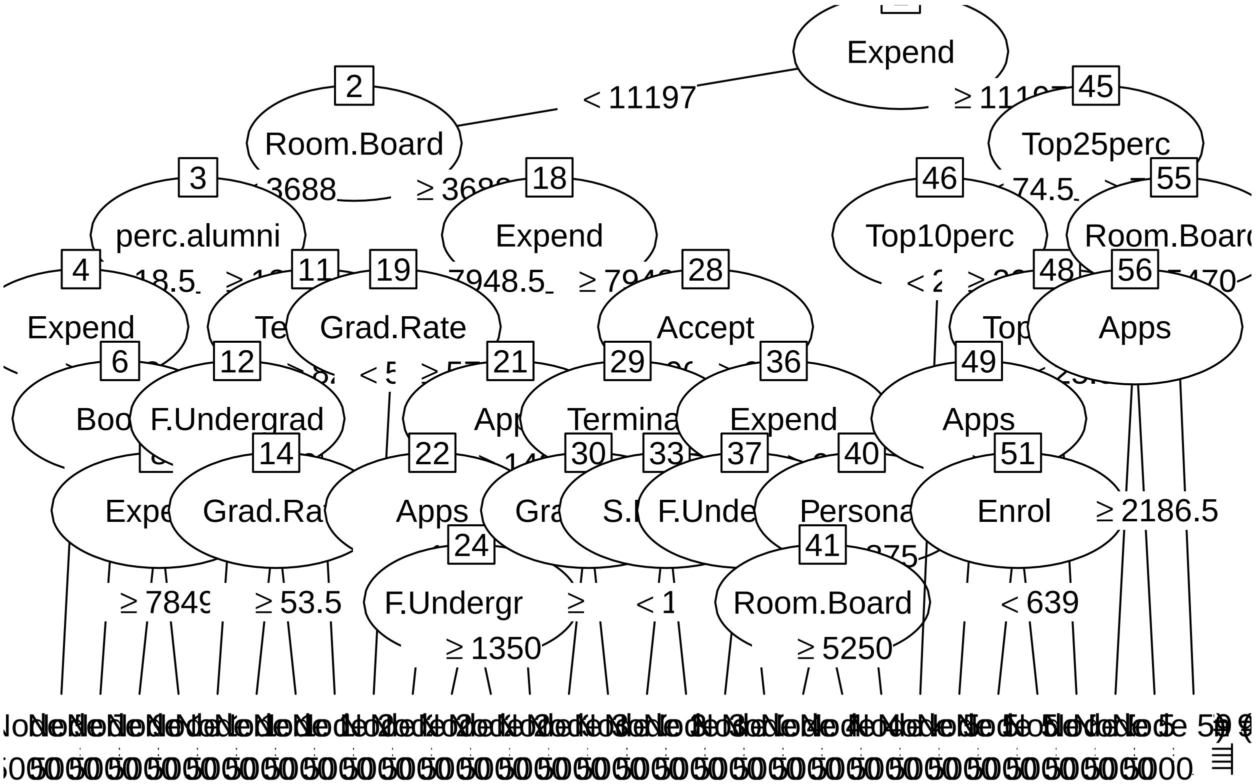
```
cpTable = tree1$cptable
plotcp(tree1)
```



```
# Picking the cp that yields the minimum cross-validation error
minErr = which.min(cpTable[,4])
tree3 = rpart::prune(tree1, cp = cpTable[minErr,1])
rpart.plot(tree3)
```



```r
plot(as.party(tree3)) #another visual
```



```r
summary(tree3) # summary of Tree3 (the final condensed version of the regression tree)
```

```
## Call:
## rpart(formula = Outstate ~ . - College, data = trainData, control = rpart.control(cp = 0))
##   n= 452
##
##             CP nsplit rel error    xerror       xstd
## 1  0.504787902      0 1.0000000 1.0061173 0.05967051
## 2  0.098022562      1 0.4952121 0.5457670 0.03595857
## 3  0.050868077      2 0.3971895 0.4623360 0.03196542
## 4  0.038363918      3 0.3463215 0.4270040 0.03255821
## 5  0.020061436      4 0.3079575 0.4105223 0.03147442
## 6  0.017454066      5 0.2878961 0.4113816 0.03181078
## 7  0.015103992      6 0.2704420 0.4021428 0.03179258
## 8  0.014754751      7 0.2553380 0.3987677 0.03259436
## 9  0.009411337      8 0.2405833 0.3967205 0.03434670
## 10 0.009283285      9 0.2311720 0.4012642 0.03551407
## 11 0.006963780     10 0.2218887 0.3911805 0.03506468
## 12 0.006945855     11 0.2149249 0.3892067 0.03485469
## 13 0.006615514     12 0.2079790 0.3871485 0.03480680
## 14 0.005907488     13 0.2013635 0.3855483 0.03481777
## 15 0.005819222     14 0.1954560 0.3800623 0.03435814
## 16 0.005726478     15 0.1896368 0.3726909 0.03311053
## 17 0.005686589     16 0.1839103 0.3723264 0.03313035
## 18 0.005349708     17 0.1782237 0.3707487 0.03306639
## 19 0.004445437     18 0.1728740 0.3662866 0.03327278
## 20 0.003958094     19 0.1684286 0.3602498 0.03323851
## 21 0.003498953     20 0.1644705 0.3573228 0.03273770
## 22 0.002815248     21 0.1609716 0.3641505 0.03309730
## 23 0.002585651     22 0.1581563 0.3649121 0.03322033
## 24 0.002427676     23 0.1555707 0.3615358 0.03303571
## 25 0.002308201     24 0.1531430 0.3603369 0.03300782
## 26 0.002231825     25 0.1508348 0.3566334 0.03293800
## 27 0.001988937     26 0.1486030 0.3560312 0.03280611
## 28 0.001906456     27 0.1466140 0.3556269 0.03189967
## 29 0.001828907     28 0.1447076 0.3552668 0.03190171
## 30 0.001701345     29 0.1428787 0.3547615 0.03190371
##
## Variable importance
##      Expend   Top10perc    Terminal         PhD   Top25perc   S.F.Ratio
##          26          13          12          11          11           8
##  Room.Board        Apps   Grad.Rate F.Undergrad      Enroll      Accept
##           5           3           3           2           2           2
## perc.alumni P.Undergrad
##           1           1
##
## Node number 1: 452 observations,    complexity param=0.5047879
##   mean=11739.82, MSE=1.376291e+07
##   left son=2 (332 obs) right son=3 (120 obs)
##   Primary splits:
##       Expend     < 11197   to the left,  improve=0.5047879, (0 missing)
##       Terminal   < 89.5    to the left,  improve=0.3759708, (0 missing)
##       PhD        < 78.5    to the left,  improve=0.3706877, (0 missing)
##       Top10perc  < 35.5    to the left,  improve=0.3099237, (0 missing)
##       Room.Board < 4053    to the left,  improve=0.3006840, (0 missing)
##   Surrogate splits:
##       Terminal  < 93.5    to the left,  agree=0.856, adj=0.458, (0 split)
##       Top10perc < 39.5    to the left,  agree=0.852, adj=0.442, (0 split)
##       PhD       < 89.5    to the left,  agree=0.841, adj=0.400, (0 split)
##       Top25perc < 74.5    to the left,  agree=0.834, adj=0.375, (0 split)
##       S.F.Ratio < 10.35   to the right, agree=0.814, adj=0.300, (0 split)
##
## Node number 2: 332 observations,    complexity param=0.09802256
##   mean=10155.18, MSE=6382441
##   left son=4 (103 obs) right son=5 (229 obs)
##   Primary splits:
##       Room.Board < 3688    to the left,  improve=0.2877730, (0 missing)
##       Expend     < 8545.5  to the left,  improve=0.2778285, (0 missing)
##       Terminal   < 77.5    to the left,  improve=0.1757753, (0 missing)
##       Accept     < 1169    to the left,  improve=0.1631625, (0 missing)
##       Grad.Rate  < 59.5    to the left,  improve=0.1606487, (0 missing)
##   Surrogate splits:
##       P.Undergrad < 50.5    to the left,  agree=0.720, adj=0.097, (0 split)
##       Expend      < 5641.5  to the left,  agree=0.714, adj=0.078, (0 split)
##       Grad.Rate   < 38.5    to the left,  agree=0.702, adj=0.039, (0 split)
##       F.Undergrad < 479     to the left,  agree=0.699, adj=0.029, (0 split)
##       PhD         < 27.5    to the left,  agree=0.699, adj=0.029, (0 split)
##
## Node number 3: 120 observations,    complexity param=0.03836392
##   mean=16124, MSE=8013862
##   left son=6 (57 obs) right son=7 (63 obs)
##   Primary splits:
##       Top25perc  < 74.5    to the left,  improve=0.2481697, (0 missing)
##       Expend     < 14711.5 to the left,  improve=0.2463772, (0 missing)
##       Grad.Rate  < 67.5    to the left,  improve=0.2336111, (0 missing)
##       Room.Board < 5470    to the left,  improve=0.2240147, (0 missing)
##       Top10perc  < 20.5    to the left,  improve=0.2206538, (0 missing)
##   Surrogate splits:
##       Top10perc < 43      to the left,  agree=0.942, adj=0.877, (0 split)
##       Grad.Rate < 73.5    to the left,  agree=0.767, adj=0.509, (0 split)
##       PhD       < 87.5    to the left,  agree=0.733, adj=0.439, (0 split)
##       Apps      < 1753    to the left,  agree=0.725, adj=0.421, (0 split)
##       Expend    < 14649   to the left,  agree=0.725, adj=0.421, (0 split)
##
## Node number 4: 103 observations,    complexity param=0.01510399
##   mean=8134.408, MSE=4878635
##   left son=8 (49 obs) right son=9 (54 obs)
##   Primary splits:
##       perc.alumni < 18.5    to the left,  improve=0.1869842, (0 missing)
##       Expend      < 6215.5  to the left,  improve=0.1343670, (0 missing)
##       S.F.Ratio   < 15.45   to the right, improve=0.1257252, (0 missing)
##       Books       < 507.5   to the right, improve=0.1070805, (0 missing)
##       Terminal    < 84      to the left,  improve=0.1061297, (0 missing)
##   Surrogate splits:
##       Terminal  < 67.5    to the left,  agree=0.660, adj=0.286, (0 split)
##       Grad.Rate < 62.5    to the left,  agree=0.660, adj=0.286, (0 split)
##       Top25perc < 40.5    to the left,  agree=0.641, adj=0.245, (0 split)
##       Books     < 507.5   to the right, agree=0.631, adj=0.224, (0 split)
##       PhD       < 64.5    to the left,  agree=0.631, adj=0.224, (0 split)
##
## Node number 5: 229 observations,    complexity param=0.05086808
##   mean=11064.09, MSE=4396020
##   left son=10 (97 obs) right son=11 (132 obs)
##   Primary splits:
##       Expend    < 7948.5  to the left,  improve=0.3143396, (0 missing)
##       Apps      < 1383.5  to the left,  improve=0.1944680, (0 missing)
##       Accept    < 1224    to the left,  improve=0.1900224, (0 missing)
##       Terminal  < 77.5    to the left,  improve=0.1824427, (0 missing)
##       Grad.Rate < 55.5    to the left,  improve=0.1782686, (0 missing)
##   Surrogate splits:
##       Terminal  < 73.5    to the left,  agree=0.707, adj=0.309, (0 split)
##       PhD       < 70.5    to the left,  agree=0.694, adj=0.278, (0 split)
##       S.F.Ratio < 14.05   to the right, agree=0.659, adj=0.196, (0 split)
##       Top10perc < 15.5    to the left,  agree=0.651, adj=0.175, (0 split)
##       Grad.Rate < 58.5    to the left,  agree=0.651, adj=0.175, (0 split)
##
## Node number 6: 57 observations,    complexity param=0.01745407
##   mean=14641.39, MSE=7559612
##   left son=12 (9 obs) right son=13 (48 obs)
##   Primary splits:
##       Top10perc < 20.5    to the left,  improve=0.2519829, (0 missing)
##       Top25perc < 44      to the left,  improve=0.2519829, (0 missing)
##       Personal  < 2026.5  to the right, improve=0.1780328, (0 missing)
##       PhD       < 85.5    to the left,  improve=0.1669229, (0 missing)
##       Grad.Rate < 54.5    to the left,  improve=0.1606177, (0 missing)
##   Surrogate splits:
##       Top25perc   < 44      to the left,  agree=1.000, adj=1.000, (0 split)
##       PhD         < 48      to the left,  agree=0.912, adj=0.444, (0 split)
##       Terminal    < 70      to the left,  agree=0.912, adj=0.444, (0 split)
##       perc.alumni < 9.5     to the left,  agree=0.912, adj=0.444, (0 split)
##       Books       < 1250    to the right, agree=0.877, adj=0.222, (0 split)
##
## Node number 7: 63 observations,    complexity param=0.02006144
##   mean=17465.41, MSE=4636664
##   left son=14 (24 obs) right son=15 (39 obs)
##   Primary splits:
##       Room.Board < 5470    to the left,  improve=0.4272328, (0 missing)
##       Expend     < 14711.5 to the left,  improve=0.2664701, (0 missing)
##       Grad.Rate  < 82.5    to the left,  improve=0.2397540, (0 missing)
##       Apps       < 2460.5  to the left,  improve=0.2246702, (0 missing)
##       Accept     < 1172.5  to the left,  improve=0.1535362, (0 missing)
##   Surrogate splits:
##       Expend      < 14749.5 to the left,  agree=0.810, adj=0.500, (0 split)
##       Enroll      < 510.5   to the left,  agree=0.762, adj=0.375, (0 split)
##       F.Undergrad < 2115.5  to the left,  agree=0.762, adj=0.375, (0 split)
##       Apps        < 4115    to the left,  agree=0.730, adj=0.292, (0 split)
##       Grad.Rate   < 82.5    to the left,  agree=0.730, adj=0.292, (0 split)
##
## Node number 8: 49 observations,    complexity param=0.006615514
##   mean=7131.755, MSE=3451055
##   left son=16 (15 obs) right son=17 (34 obs)
##   Primary splits:
##       Expend     < 6228    to the left,  improve=0.2433685, (0 missing)
##       Room.Board < 3056    to the left,  improve=0.1708790, (0 missing)
##       S.F.Ratio  < 13.1    to the right, improve=0.1660126, (0 missing)
##       Apps       < 1327    to the right, improve=0.1301367, (0 missing)
##       Enroll     < 211.5   to the right, improve=0.1247013, (0 missing)
##   Surrogate splits:
##       S.F.Ratio   < 14.95   to the right, agree=0.918, adj=0.733, (0 split)
##       Room.Board  < 3018    to the left,  agree=0.796, adj=0.333, (0 split)
##       P.Undergrad < 658     to the right, agree=0.755, adj=0.200, (0 split)
##       Apps        < 2379    to the right, agree=0.735, adj=0.133, (0 split)
##       perc.alumni < 5.5     to the left,  agree=0.735, adj=0.133, (0 split)
##
## Node number 9: 54 observations,    complexity param=0.009283285
##   mean=9044.222, MSE=4434042
##   left son=18 (42 obs) right son=19 (12 obs)
##   Primary splits:
##       Terminal    < 82.5    to the left,  improve=0.2411887, (0 missing)
##       S.F.Ratio   < 16.8    to the right, improve=0.2388366, (0 missing)
##       Grad.Rate   < 55.5    to the left,  improve=0.1843976, (0 missing)
##       P.Undergrad < 432.5   to the right, improve=0.1559643, (0 missing)
##       Books       < 440     to the right, improve=0.1271977, (0 missing)
##   Surrogate splits:
##       PhD         < 78      to the left,  agree=0.907, adj=0.583, (0 split)
##       Top10perc   < 29.5    to the left,  agree=0.889, adj=0.500, (0 split)
##       Top25perc   < 59.5    to the left,  agree=0.852, adj=0.333, (0 split)
##       P.Undergrad < 51      to the right, agree=0.852, adj=0.333, (0 split)
##       Expend      < 8410.5  to the left,  agree=0.833, adj=0.250, (0 split)
## 
## Node number 10: 97 observations,    complexity param=0.005349708
##   mean=9692.794, MSE=2498873 
##   left son=20 (34 obs) right son=21 (63 obs)
##   Primary splits:
##       Grad.Rate   < 57.5    to the left,  improve=0.1372976, (0 missing)
##       Expend      < 6331    to the left,  improve=0.1366361, (0 missing)
##       Apps        < 1438    to the left,  improve=0.1196904, (0 missing)
##       perc.alumni < 14.5    to the left,  improve=0.1036239, (0 missing)
##       F.Undergrad < 1102    to the left,  improve=0.1010706, (0 missing)
##   Surrogate splits:
##       Apps        < 480.5   to the left,  agree=0.722, adj=0.206, (0 split)
##       Room.Board  < 3830    to the left,  agree=0.722, adj=0.206, (0 split)
##       Enroll      < 174.5   to the left,  agree=0.711, adj=0.176, (0 split)
##       Expend      < 5864.5  to the left,  agree=0.711, adj=0.176, (0 split)
##       F.Undergrad < 934     to the left,  agree=0.701, adj=0.147, (0 split)
## 
## Node number 11: 132 observations,    complexity param=0.01475475
##   mean=12071.78, MSE=3392847 
##   left son=22 (58 obs) right son=23 (74 obs)
##   Primary splits:
##       Accept      < 905     to the left,  improve=0.2049475, (0 missing)
##       Apps        < 1108    to the left,  improve=0.1997585, (0 missing)
##       F.Undergrad < 1064.5  to the left,  improve=0.1930986, (0 missing)
##       Enroll      < 300.5   to the left,  improve=0.1920852, (0 missing)
##       Grad.Rate   < 55.5    to the left,  improve=0.1410561, (0 missing)
##   Surrogate splits:
##       Apps        < 1108    to the left,  agree=0.977, adj=0.948, (0 split)
##       Enroll      < 300.5   to the left,  agree=0.924, adj=0.828, (0 split)
##       F.Undergrad < 1362.5  to the left,  agree=0.864, adj=0.690, (0 split)
##       S.F.Ratio   < 12.45   to the left,  agree=0.742, adj=0.414, (0 split)
##       Terminal    < 79.5    to the left,  agree=0.689, adj=0.293, (0 split)
## 
## Node number 12: 9 observations
##   mean=11454, MSE=4554780 
## 
## Node number 13: 48 observations,    complexity param=0.009411337
##   mean=15239.02, MSE=5860957 
##   left son=26 (41 obs) right son=27 (7 obs)
##   Primary splits:
##       Top10perc   < 25.5    to the right, improve=0.2081088, (0 missing)
##       P.Undergrad < 31      to the right, improve=0.1738244, (0 missing)
##       Personal    < 1324    to the right, improve=0.1730307, (0 missing)
##       Room.Board  < 4913    to the left,  improve=0.1402596, (0 missing)
##       Top25perc   < 55.5    to the right, improve=0.1354151, (0 missing)
##   Surrogate splits:
##       Top25perc  < 53.5    to the right, agree=0.896, adj=0.286, (0 split)
##       Room.Board < 6538    to the left,  agree=0.875, adj=0.143, (0 split)
## 
## Node number 14: 24 observations,    complexity param=0.005686589
##   mean=15671.25, MSE=5587777 
##   left son=28 (9 obs) right son=29 (15 obs)
##   Primary splits:
##       Apps        < 2186.5  to the left,  improve=0.2637851, (0 missing)
##       F.Undergrad < 1362    to the left,  improve=0.1990182, (0 missing)
##       Grad.Rate   < 82.5    to the left,  improve=0.1682183, (0 missing)
##       Expend      < 13171   to the left,  improve=0.1518581, (0 missing)
##       Enroll      < 399.5   to the left,  improve=0.1427091, (0 missing)
##   Surrogate splits:
##       Enroll      < 361     to the left,  agree=0.875, adj=0.667, (0 split)
##       F.Undergrad < 1339    to the left,  agree=0.875, adj=0.667, (0 split)
##       Accept      < 952     to the left,  agree=0.833, adj=0.556, (0 split)
##       Expend      < 13306   to the left,  agree=0.792, adj=0.444, (0 split)
##       Room.Board  < 4299    to the left,  agree=0.708, adj=0.222, (0 split)
##
## Node number 15: 39 observations
##   mean=18569.51, MSE=851391.7
##
## Node number 16: 15 observations
##   mean=5752, MSE=813191.5
##
## Node number 17: 34 observations,    complexity param=0.003498953
##   mean=7740.471, MSE=3404406
##   left son=34 (13 obs) right son=35 (21 obs)
##   Primary splits:
##       Books       < 575     to the right, improve=0.18804710, (0 missing)
##       Expend      < 7653    to the left,  improve=0.10490960, (0 missing)
##       F.Undergrad < 651.5   to the left,  improve=0.09435098, (0 missing)
##       Enroll      < 211.5   to the right, improve=0.08667413, (0 missing)
##       Grad.Rate   < 40      to the left,  improve=0.08508979, (0 missing)
##   Surrogate splits:
##       Top25perc   < 28.5    to the left,  agree=0.735, adj=0.308, (0 split)
##       Enroll      < 503     to the right, agree=0.676, adj=0.154, (0 split)
##       F.Undergrad < 1973.5  to the right, agree=0.676, adj=0.154, (0 split)
##       P.Undergrad < 102.5   to the left,  agree=0.676, adj=0.154, (0 split)
##       Room.Board  < 3645    to the right, agree=0.676, adj=0.154, (0 split)
##
## Node number 18: 42 observations,    complexity param=0.006945855
##   mean=8491.452, MSE=3639141
##   left son=36 (7 obs) right son=37 (35 obs)
##   Primary splits:
##       F.Undergrad < 1446    to the right, improve=0.2827004, (0 missing)
##       PhD         < 68.5    to the right, improve=0.2796709, (0 missing)
##       S.F.Ratio   < 13.2    to the right, improve=0.2258541, (0 missing)
##       Enroll      < 371.5   to the right, improve=0.2127179, (0 missing)
##       Apps        < 938     to the right, improve=0.2117170, (0 missing)
##   Surrogate splits:
##       Apps      < 1458.5  to the right, agree=0.929, adj=0.571, (0 split)
##       Accept    < 1023    to the right, agree=0.929, adj=0.571, (0 split)
##       Enroll    < 578     to the right, agree=0.929, adj=0.571, (0 split)
##       Expend    < 4721    to the left,  agree=0.929, adj=0.571, (0 split)
##       S.F.Ratio < 16.95   to the right, agree=0.905, adj=0.429, (0 split)
##
## Node number 19: 12 observations
##   mean=10978.92, MSE=2403712
##
```

```
## Node number 20: 34 observations
##   mean=8895.471, MSE=1717528
##
## Node number 21: 63 observations,    complexity param=0.003958094
##   mean=10123.1, MSE=2392303
##   left son=42 (46 obs) right son=43 (17 obs)
##   Primary splits:
##       Apps       < 1438    to the left,  improve=0.1633722, (0 missing)
##       Terminal   < 86      to the left,  improve=0.1405912, (0 missing)
##       Expend     < 7079    to the left,  improve=0.1302950, (0 missing)
##       Personal   < 1273    to the right, improve=0.1161823, (0 missing)
##       Room.Board < 4922.5  to the left,  improve=0.1153151, (0 missing)
##   Surrogate splits:
##       Accept      < 1189    to the left,  agree=0.968, adj=0.882, (0 split)
##       Enroll      < 424     to the left,  agree=0.937, adj=0.765, (0 split)
##       F.Undergrad < 2227    to the left,  agree=0.825, adj=0.353, (0 split)
##       PhD         < 67.5    to the left,  agree=0.794, adj=0.235, (0 split)
##       Room.Board  < 5740    to the left,  agree=0.778, adj=0.176, (0 split)
##
## Node number 22: 58 observations,    complexity param=0.002815248
##   mean=11129.88, MSE=1849875
##   left son=44 (34 obs) right son=45 (24 obs)
##   Primary splits:
##       Terminal  < 78.5    to the left,  improve=0.1632281, (0 missing)
##       Grad.Rate < 55.5    to the left,  improve=0.1615433, (0 missing)
##       Expend    < 8556.5  to the left,  improve=0.1376245, (0 missing)
##       Top25perc < 51.5    to the left,  improve=0.1217833, (0 missing)
##       PhD       < 69      to the left,  improve=0.1180594, (0 missing)
##   Surrogate splits:
##       PhD         < 72.5    to the left,  agree=0.828, adj=0.583, (0 split)
##       F.Undergrad < 1057.5  to the left,  agree=0.707, adj=0.292, (0 split)
##       S.F.Ratio   < 11.95   to the left,  agree=0.707, adj=0.292, (0 split)
##       Apps        < 752.5   to the left,  agree=0.690, adj=0.250, (0 split)
##       Accept      < 561.5   to the left,  agree=0.690, adj=0.250, (0 split)
##
## Node number 23: 74 observations,    complexity param=0.00696378
##   mean=12810.03, MSE=3361839
##   left son=46 (31 obs) right son=47 (43 obs)
##   Primary splits:
##       Expend      < 9199    to the left,  improve=0.17413460, (0 missing)
##       perc.alumni < 21      to the left,  improve=0.15643560, (0 missing)
##       Grad.Rate   < 63.5    to the left,  improve=0.13023400, (0 missing)
##       Personal    < 875     to the right, improve=0.12697000, (0 missing)
##       S.F.Ratio   < 14.05   to the right, improve=0.08149241, (0 missing)
##   Surrogate splits:
##       Terminal  < 84.5    to the left,  agree=0.716, adj=0.323, (0 split)
##       PhD       < 75.5    to the left,  agree=0.703, adj=0.290, (0 split)
##       Apps      < 1701    to the left,  agree=0.676, adj=0.226, (0 split)
##       Top25perc < 56.5    to the left,  agree=0.676, adj=0.226, (0 split)
##       Grad.Rate < 83.5    to the right, agree=0.676, adj=0.226, (0 split)
##
## Node number 26: 41 observations,    complexity param=0.005907488
##   mean=14782.68, MSE=4663425
##   left son=52 (11 obs) right son=53 (30 obs)
##   Primary splits:
##       Apps       < 961     to the left,  improve=0.1922042, (0 missing)
##       Room.Board < 4760    to the left,  improve=0.1787202, (0 missing)
##       Grad.Rate  < 60.5    to the left,  improve=0.1702441, (0 missing)
##       Terminal   < 91.5    to the left,  improve=0.1683315, (0 missing)
##       PhD        < 85.5    to the left,  improve=0.1670680, (0 missing)
##   Surrogate splits:
##       Accept      < 597.5   to the left,  agree=0.902, adj=0.636, (0 split)
##       Enroll      < 217.5   to the left,  agree=0.902, adj=0.636, (0 split)
##       F.Undergrad < 845     to the left,  agree=0.902, adj=0.636, (0 split)
##       Grad.Rate   < 64.5    to the left,  agree=0.829, adj=0.364, (0 split)
##       Terminal    < 81      to the left,  agree=0.805, adj=0.273, (0 split)
##
## Node number 27: 7 observations
##   mean=17911.86, MSE=4511305
##
## Node number 28: 9 observations
##   mean=14103.89, MSE=2464935
##
## Node number 29: 15 observations
##   mean=16611.67, MSE=5103127
##
## Node number 34: 13 observations
##   mean=6723.538, MSE=2621027
##
## Node number 35: 21 observations,    complexity param=0.002231825
##   mean=8370, MSE=2852859
##   left son=70 (12 obs) right son=71 (9 obs)
##   Primary splits:
##       Expend      < 7849.5  to the left,  improve=0.23174450, (0 missing)
##       Personal    < 1512.5  to the right, improve=0.16657010, (0 missing)
##       F.Undergrad < 825.5   to the right, improve=0.11935210, (0 missing)
##       Grad.Rate   < 57.5    to the right, improve=0.11130180, (0 missing)
##       PhD         < 63.5    to the left,  improve=0.09049605, (0 missing)
##   Surrogate splits:
##       PhD         < 65.5    to the left,  agree=0.762, adj=0.444, (0 split)
##       Personal    < 1250    to the right, agree=0.714, adj=0.333, (0 split)
##       Terminal    < 68      to the left,  agree=0.714, adj=0.333, (0 split)
##       S.F.Ratio   < 14.45   to the left,  agree=0.714, adj=0.333, (0 split)
##       perc.alumni < 7.5     to the right, agree=0.714, adj=0.333, (0 split)
##
## Node number 36: 7 observations
##   mean=6223.429, MSE=3989025
##
## Node number 37: 35 observations,    complexity param=0.004445437
##   mean=8945.057, MSE=2334620
##   left son=74 (10 obs) right son=75 (25 obs)
##   Primary splits:
##       Grad.Rate   < 53.5    to the left,  improve=0.3384379, (0 missing)
##       Accept      < 708     to the left,  improve=0.1804239, (0 missing)
##       F.Undergrad < 989.5   to the left,  improve=0.1752059, (0 missing)
##       Room.Board  < 2920    to the left,  improve=0.1738597, (0 missing)
##       S.F.Ratio   < 13.2    to the right, improve=0.1318418, (0 missing)
##   Surrogate splits:
```

```
##       F.Undergrad < 443.5   to the left,  agree=0.800, adj=0.3, (0 split)
##       Enroll      < 96      to the left,  agree=0.771, adj=0.2, (0 split)
##       PhD         < 34      to the left,  agree=0.771, adj=0.2, (0 split)
##       S.F.Ratio   < 16.55   to the right, agree=0.771, adj=0.2, (0 split)
##       perc.alumni < 42.5    to the right, agree=0.771, adj=0.2, (0 split)
##
## Node number 42: 46 observations,    complexity param=0.002308201
##   mean=9743.043, MSE=1593312
##   left son=84 (8 obs) right son=85 (38 obs)
##   Primary splits:
##       Apps      < 1161.5  to the right, improve=0.1959132, (0 missing)
##       Expend    < 7312    to the left,  improve=0.1677488, (0 missing)
##       Enroll    < 354     to the right, improve=0.1382364, (0 missing)
##       Personal  < 1273    to the right, improve=0.1335470, (0 missing)
##       Top25perc < 50.5    to the left,  improve=0.1053780, (0 missing)
##   Surrogate splits:
##       Accept    < 899.5   to the right, agree=0.913, adj=0.500, (0 split)
##       Enroll    < 367.5   to the right, agree=0.891, adj=0.375, (0 split)
##       PhD       < 77      to the right, agree=0.870, adj=0.250, (0 split)
##       Top10perc < 32.5    to the right, agree=0.848, adj=0.125, (0 split)
##
## Node number 43: 17 observations
##   mean=11151.47, MSE=3105888
##
## Node number 44: 34 observations,    complexity param=0.002427676
##   mean=10668.21, MSE=1524362
##   left son=88 (7 obs) right son=89 (27 obs)
##   Primary splits:
##       Grad.Rate   < 54.5    to the left,  improve=0.2913886, (0 missing)
##       Room.Board  < 4745    to the left,  improve=0.2291231, (0 missing)
##       F.Undergrad < 527.5   to the left,  improve=0.1958720, (0 missing)
##       Top25perc   < 51.5    to the left,  improve=0.1641989, (0 missing)
##       Top10perc   < 22.5    to the left,  improve=0.1622200, (0 missing)
##   Surrogate splits:
##       Enroll      < 81.5    to the left,  agree=0.853, adj=0.286, (0 split)
##       F.Undergrad < 311.5   to the left,  agree=0.853, adj=0.286, (0 split)
##       P.Undergrad < 48      to the left,  agree=0.853, adj=0.286, (0 split)
##       Terminal    < 52      to the left,  agree=0.853, adj=0.286, (0 split)
##       S.F.Ratio   < 9.1     to the left,  agree=0.853, adj=0.286, (0 split)
##
## Node number 45: 24 observations,    complexity param=0.001988937
##   mean=11783.92, MSE=1581302
##   left son=90 (8 obs) right son=91 (16 obs)
##   Primary splits:
##       S.F.Ratio   < 13.25   to the right, improve=0.32601960, (0 missing)
##       Expend      < 8822    to the left,  improve=0.30259820, (0 missing)
##       Books       < 525     to the right, improve=0.17664230, (0 missing)
##       F.Undergrad < 789     to the right, improve=0.07887553, (0 missing)
##       Enroll      < 215     to the right, improve=0.07158222, (0 missing)
##   Surrogate splits:
##       Expend      < 8822    to the left,  agree=0.792, adj=0.375, (0 split)
##       Top10perc   < 29.5    to the right, agree=0.750, adj=0.250, (0 split)
##       Top25perc   < 36.5    to the left,  agree=0.750, adj=0.250, (0 split)
##       P.Undergrad < 91      to the left,  agree=0.750, adj=0.250, (0 split)
##       F.Undergrad < 897     to the right, agree=0.708, adj=0.125, (0 split)
##
## Node number 46: 31 observations,    complexity param=0.002585651
##   mean=11908.9, MSE=2274824
##   left son=92 (22 obs) right son=93 (9 obs)
##   Primary splits:
##       F.Undergrad < 1606    to the right, improve=0.2280916, (0 missing)
##       Grad.Rate   < 66.5    to the left,  improve=0.1860533, (0 missing)
##       Apps        < 1608    to the right, improve=0.1617918, (0 missing)
##       Books       < 525     to the right, improve=0.1368413, (0 missing)
##       perc.alumni < 21      to the left,  improve=0.1291313, (0 missing)
##   Surrogate splits:
##       Accept      < 1193.5  to the right, agree=0.935, adj=0.778, (0 split)
##       Apps        < 1436    to the right, agree=0.903, adj=0.667, (0 split)
##       Enroll      < 390.5   to the right, agree=0.903, adj=0.667, (0 split)
##       Top10perc   < 42      to the left,  agree=0.774, adj=0.222, (0 split)
##       P.Undergrad < 163     to the right, agree=0.774, adj=0.222, (0 split)
##
## Node number 47: 43 observations,    complexity param=0.005819222
##   mean=13459.67, MSE=3138048
##   left son=94 (24 obs) right son=95 (19 obs)
##   Primary splits:
##       Personal    < 875     to the right, improve=0.2682784, (0 missing)
##       perc.alumni < 19.5    to the left,  improve=0.2126111, (0 missing)
##       S.F.Ratio   < 15.25   to the right, improve=0.2088310, (0 missing)
##       Grad.Rate   < 80.5    to the left,  improve=0.1992718, (0 missing)
##       Expend      < 10712   to the left,  improve=0.1324686, (0 missing)
##   Surrogate splits:
##       Grad.Rate   < 80.5    to the left,  agree=0.791, adj=0.526, (0 split)
##       perc.alumni < 26.5    to the left,  agree=0.744, adj=0.421, (0 split)
##       F.Undergrad < 1892    to the right, agree=0.721, adj=0.368, (0 split)
##       Enroll      < 476     to the right, agree=0.698, adj=0.316, (0 split)
##       Accept      < 2457.5  to the right, agree=0.674, adj=0.263, (0 split)
##
## Node number 52: 11 observations
##   mean=13219.18, MSE=4462580
##
## Node number 53: 30 observations,    complexity param=0.005726478
##   mean=15355.97, MSE=3512084
##   left son=106 (11 obs) right son=107 (19 obs)
##   Primary splits:
##       Enroll      < 639     to the right, improve=0.3381040, (0 missing)
##       F.Undergrad < 2760    to the right, improve=0.3381040, (0 missing)
##       P.Undergrad < 346.5   to the right, improve=0.2533396, (0 missing)
##       Personal    < 1324    to the right, improve=0.2392008, (0 missing)
##       Room.Board  < 4913    to the left,  improve=0.1913443, (0 missing)
##   Surrogate splits:
##       F.Undergrad < 2760    to the right, agree=1.000, adj=1.000, (0 split)
##       Accept      < 2600.5  to the right, agree=0.933, adj=0.818, (0 split)
##       Apps        < 3154    to the right, agree=0.867, adj=0.636, (0 split)
##       Top10perc   < 37.5    to the right, agree=0.867, adj=0.636, (0 split)
##       P.Undergrad < 346.5   to the right, agree=0.867, adj=0.636, (0 split)
##
## Node number 70: 12 observations
##   mean=7665.833, MSE=1236177
##
## Node number 71: 9 observations
##   mean=9308.889, MSE=3465788
##
## Node number 74: 10 observations
##   mean=7539.6, MSE=2418300
##
## Node number 75: 25 observations
##   mean=9507.24, MSE=1194975
##
## Node number 84: 8 observations
##   mean=8525.375, MSE=641878.7
##
## Node number 85: 38 observations,    complexity param=0.001828907
##   mean=9999.395, MSE=1415747
##   left son=170 (30 obs) right son=171 (8 obs)
##   Primary splits:
##       F.Undergrad < 1350    to the left,  improve=0.2114810, (0 missing)
##       PhD         < 61      to the left,  improve=0.2085145, (0 missing)
##       Expend      < 7312    to the left,  improve=0.2061474, (0 missing)
##       Accept      < 683     to the left,  improve=0.2053201, (0 missing)
##       Top25perc   < 50.5    to the left,  improve=0.1824312, (0 missing)
##   Surrogate splits:
##       Apps   < 1100    to the left,  agree=0.842, adj=0.250, (0 split)
##       Accept < 931.5   to the left,  agree=0.842, adj=0.250, (0 split)
##       Enroll < 351     to the left,  agree=0.842, adj=0.250, (0 split)
##       Books  < 625     to the left,  agree=0.842, adj=0.250, (0 split)
##       Expend < 7659    to the left,  agree=0.816, adj=0.125, (0 split)
##
## Node number 88: 7 observations
##   mean=9359.286, MSE=1029803
##
## Node number 89: 27 observations
##   mean=11007.56, MSE=1093241
##
## Node number 90: 8 observations
##   mean=10768.5, MSE=1361606
##
## Node number 91: 16 observations
##   mean=12291.62, MSE=917846.9
##
## Node number 92: 22 observations
##   mean=11448.18, MSE=2017996
##
## Node number 93: 9 observations
##   mean=13035.11, MSE=1115412
##
## Node number 94: 24 observations,    complexity param=0.001906456
##   mean=12643.29, MSE=2222168
##   left son=188 (11 obs) right son=189 (13 obs)
##   Primary splits:
##       Room.Board < 5250    to the left,  improve=0.22237580, (0 missing)
##       Top25perc  < 56      to the right, improve=0.19688710, (0 missing)
##       S.F.Ratio  < 14.55   to the right, improve=0.15977350, (0 missing)
##       Top10perc  < 35.5    to the right, improve=0.09492025, (0 missing)
##       PhD        < 79      to the right, improve=0.09138872, (0 missing)
##   Surrogate splits:
##       Apps        < 2012.5  to the left,  agree=0.792, adj=0.545, (0 split)
##       F.Undergrad < 2550.5  to the left,  agree=0.792, adj=0.545, (0 split)
##       Accept      < 1487.5  to the left,  agree=0.750, adj=0.455, (0 split)
##       Enroll      < 495.5   to the left,  agree=0.750, adj=0.455, (0 split)
##       Top25perc   < 64.5    to the right, agree=0.750, adj=0.455, (0 split)
##
## Node number 95: 19 observations
##   mean=14490.89, MSE=2389663
##
## Node number 106: 11 observations
##   mean=13923.82, MSE=3887045
##
## Node number 107: 19 observations
##   mean=16185.11, MSE=1420081
##
## Node number 170: 30 observations
##   mean=9716.833, MSE=842589
##
## Node number 171: 8 observations
##   mean=11059, MSE=2142921
##
## Node number 188: 11 observations
##   mean=11879.09, MSE=2822642
##
## Node number 189: 13 observations
##   mean=13289.92, MSE=801785.6
```

```r
set.seed(1)
bagging = randomForest(Outstate ~ . - College,
                       data = trainData,
                       mtry = 16)

set.seed(1)
rf = randomForest(Outstate ~ . - College,
                       data = trainData,
                       mtry = 5)

set.seed(1)
rf2 = ranger(Outstate ~ . - College,
                       data = trainData,
                       mtry = 5)

pred.rf = predict(rf, newdata = testData)
pred.rf2 = predict(rf2, data = testData)$predictions

# Test Error:
RMSE(pred.rf, testData$Outstate)
```

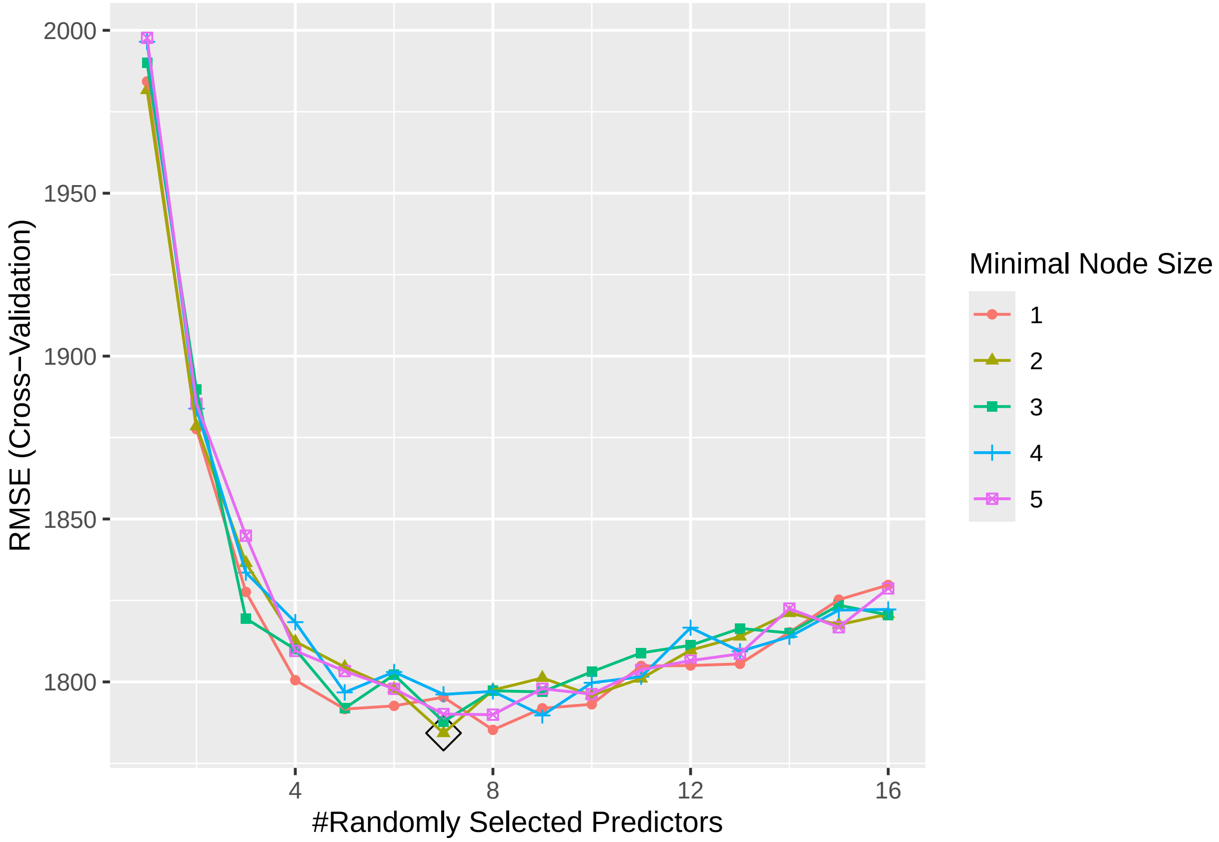
1.B: Perform random forest on the training data (10pts). Report the variable importance (5pts) and the test error (5pts).

```
## [1] 1704.415
```

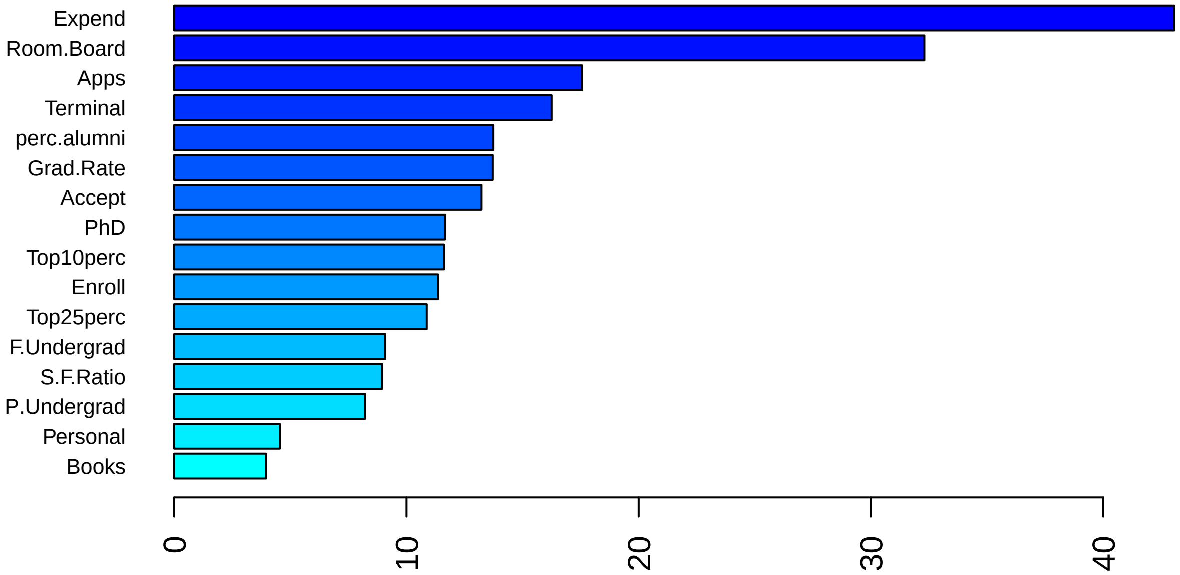
```
RMSE(pred.rf2, testData$Outstate)
```

```
## [1] 1683.584
```

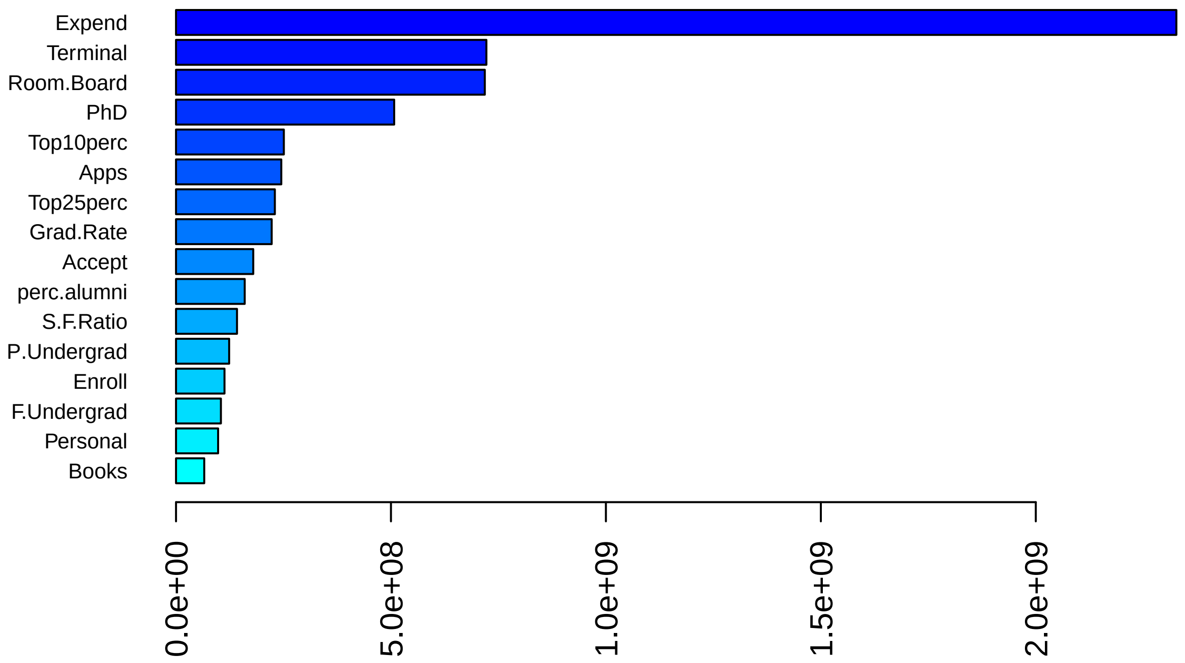
```
ctrl = trainControl(method = "cv")

rf.grid = expand.grid(mtry = 1:16,
                      splitrule = "variance",
                      min.node.size = 1:5)

set.seed(1)
rf.fit = train(Outstate ~ . - College,
               data = trainData,
               method = "ranger",
               tuneGrid = rf.grid,
               trControl = ctrl)

ggplot(rf.fit, highlight = TRUE)
```



```
#Extracting the variable importance from permutting
set.seed(1)
rf2.final.per = ranger(Outstate ~ . - College,
                       data = trainData,
                       mtry = rf.fit$bestTune[[1]],
                       splitrule = "variance",
                       min.node.size = rf.fit$bestTune[[3]],
                       importance = "permutation",
                       scale.permutation.importance = TRUE)

barplot(sort(ranger::importance(rf2.final.per), decreasing = FALSE),
        las = 2, horiz = TRUE, cex.names = 0.7,
        col = colorRampPalette(colors=c("cyan", "blue"))(16))
```

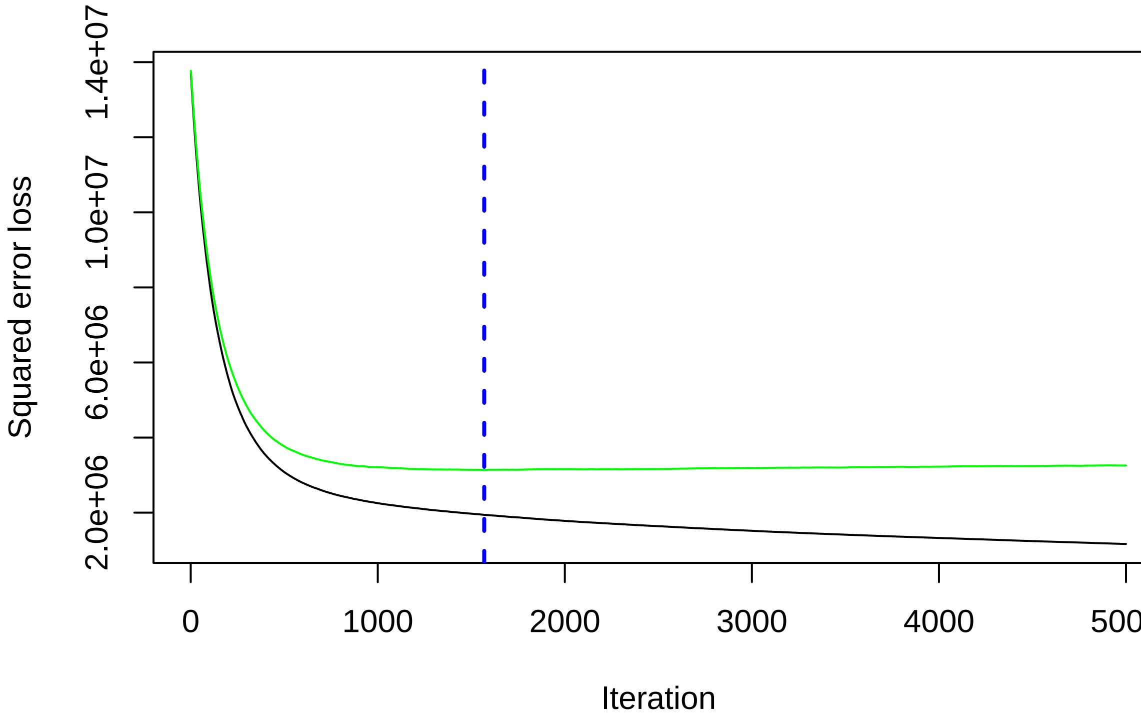


```
#Extracting the variable importance from node impurities

set.seed(1)
  rf2.final.imp = ranger(Outstate ~ . - College,
                         data = trainData,
                         mtry = rf.fit$bestTune[[1]],
                         splitrule = "variance",
                         min.node.size = rf.fit$bestTune[[3]],
                         importance = "impurity")

barplot(sort(ranger::importance(rf2.final.imp), decreasing = FALSE),
        las = 2, horiz = TRUE, cex.names = 0.7,
        col = colorRampPalette(colors = c("cyan","blue"))(16))
```



```
set.seed(1)
bst = gbm(Outstate ~ . - College,
          data = trainData,
          distribution = "gaussian",
          n.trees = 5000,
          interaction.depth = 2,
          shrinkage = 0.005,
          cv.folds =10)

gbm.perf(bst, method = "cv")
```

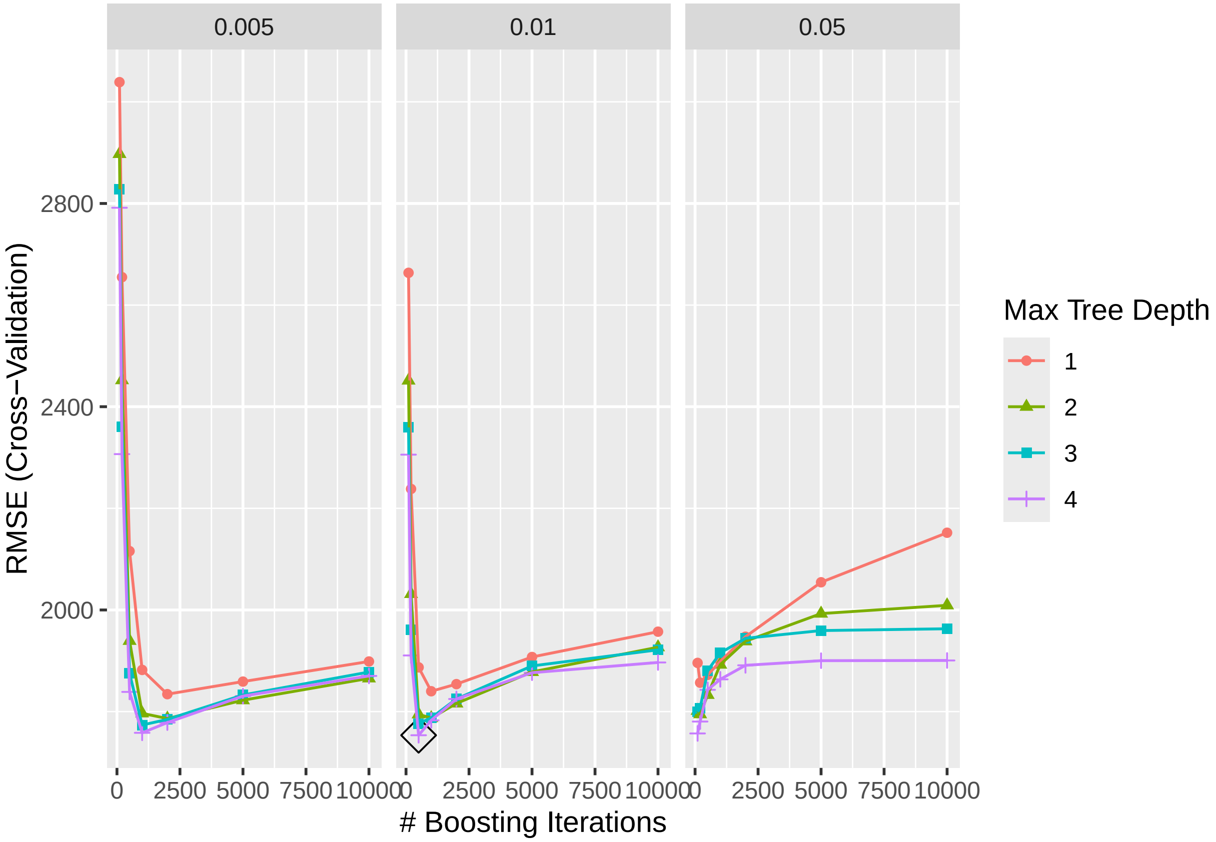
1.C: Perform boosting on the training data (10pts). Report the variable importance (5pts) and



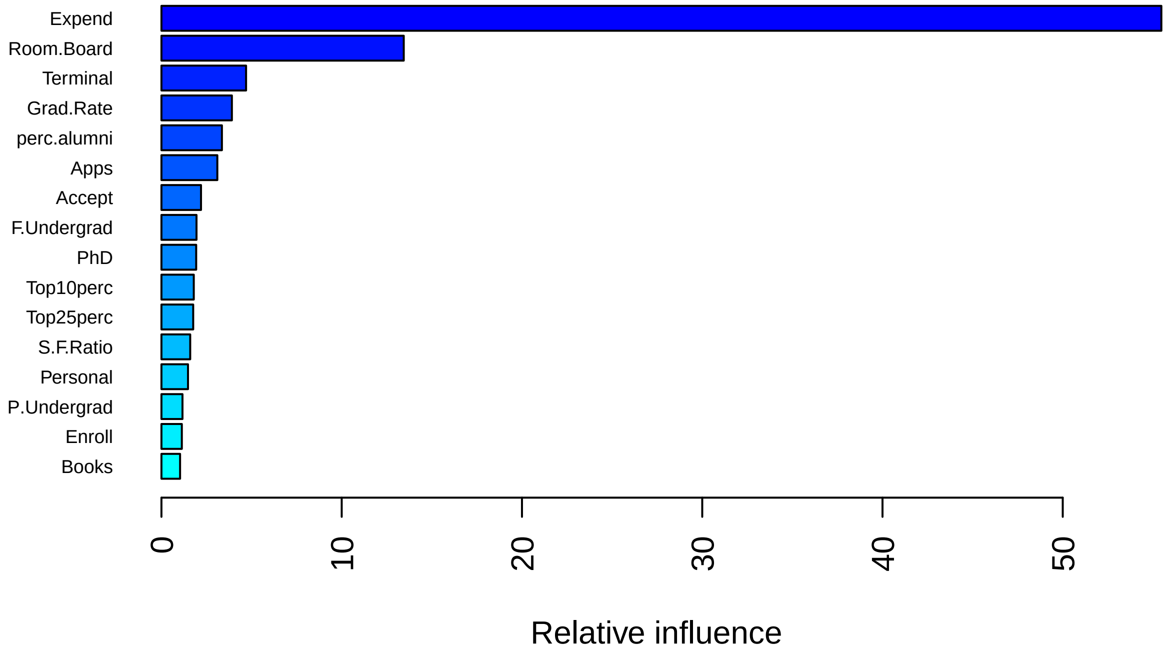
the test error (5pts).

```
## [1] 1569
```

```
ctrl = trainControl(method = "cv")

gbm.grid = expand.grid(n.trees = c(100,200,500,1000,2000,5000,10000),
                        interaction.depth = 1:4,
                        shrinkage = c(0.005,0.01,0.05),
                        n.minobsinnode = c(10))

set.seed(1)
gbm.fit = train(Outstate ~ . - College,
                data = trainData,
                method = "gbm",
                tuneGrid = gbm.grid,
                trControl = ctrl,
                verbose = FALSE
                )

ggplot(gbm.fit, highlight = TRUE)
```



```r
summary(gbm.fit$finalModel, las = 2, cBars = 19, cex.names = 0.6)
```



```
##                     var   rel.inf
## Expend           Expend 55.469914
## Room.Board   Room.Board 13.442969
## Terminal       Terminal  4.695168
## Grad.Rate     Grad.Rate  3.908733
## perc.alumni perc.alumni  3.355253
## Apps               Apps  3.100460
## Accept           Accept  2.195802
## F.Undergrad F.Undergrad  1.945905
## PhD                 PhD  1.927205
## Top10perc     Top10perc  1.794344
## Top25perc     Top25perc  1.761419
## S.F.Ratio     S.F.Ratio  1.596315
## Personal       Personal  1.473870
## P.Undergrad P.Undergrad  1.166772
## Enroll           Enroll  1.132208
## Books             Books  1.033661
```

QUESTION 2: This problem is based on the data "auto.csv" in Homework 3. Split the dataset into two parts: training data (70%) and test data (30%).

```
# initial data steps--importing and partitioning
auto = read.csv("auto.csv")
head(auto)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         8          307        130   3504         12.0   70      1     low
## 2         8          350        165   3693         11.5   70      1     low
## 3         8          318        150   3436         11.0   70      1     low
## 4         8          304        150   3433         12.0   70      1     low
## 5         8          302        140   3449         10.5   70      1     low
## 6         8          429        198   4341         10.0   70      1     low
```

```r
datSplit = initial_split(data = auto, prop = 0.7)
trainData_auto = training(datSplit)
testData_auto = testing(datSplit)
head(trainData_auto)
```

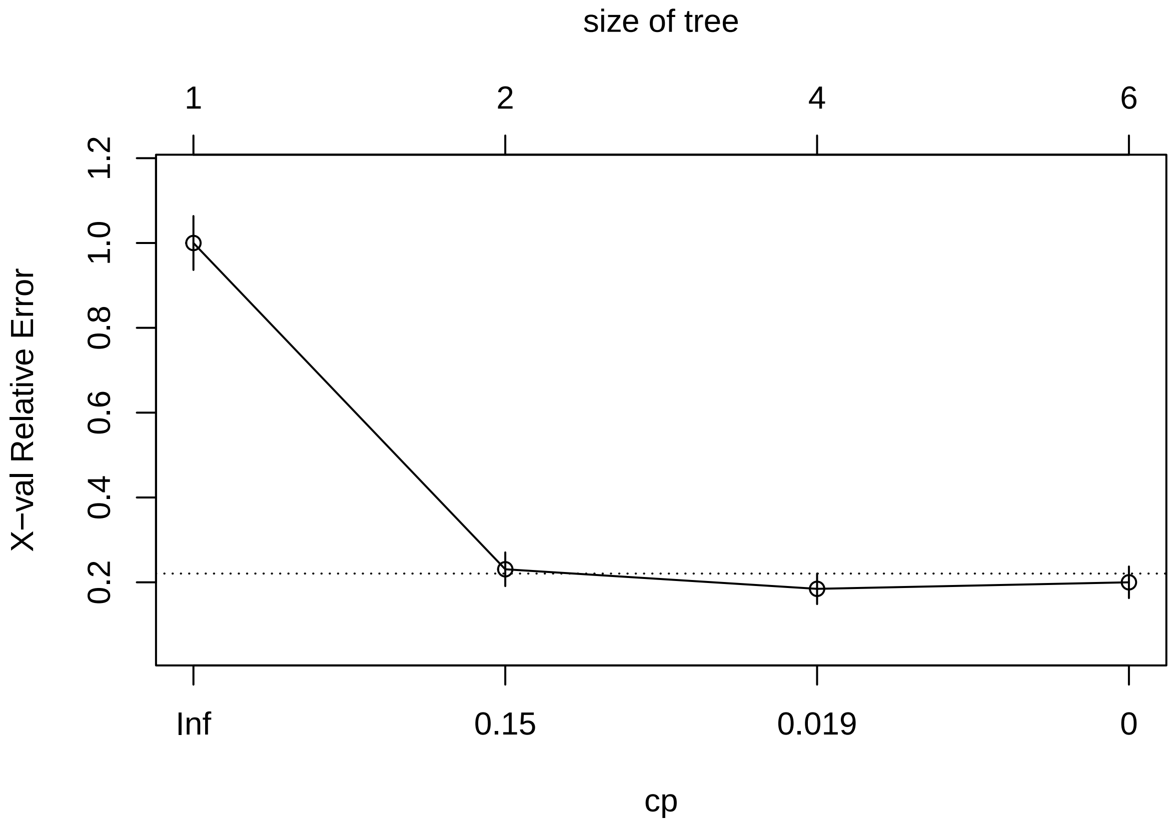
```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         4          140         90   2264         15.5   71      1    high
## 2         6          198         95   2904         16.0   73      1    high
## 3         6          231        110   3415         15.8   81      1     low
## 4         4           89         71   1990         14.9   78      2    high
## 5         4          120         87   2979         19.5   72      2     low
## 6         8          455        225   4425         10.0   70      1     low
```

```r
set.seed(1)
mpg1 = rpart(formula = mpg_cat ~ .,
              data = trainData_auto,
              control = rpart.control(cp=0))

cpTable = printcp(mpg1)
```

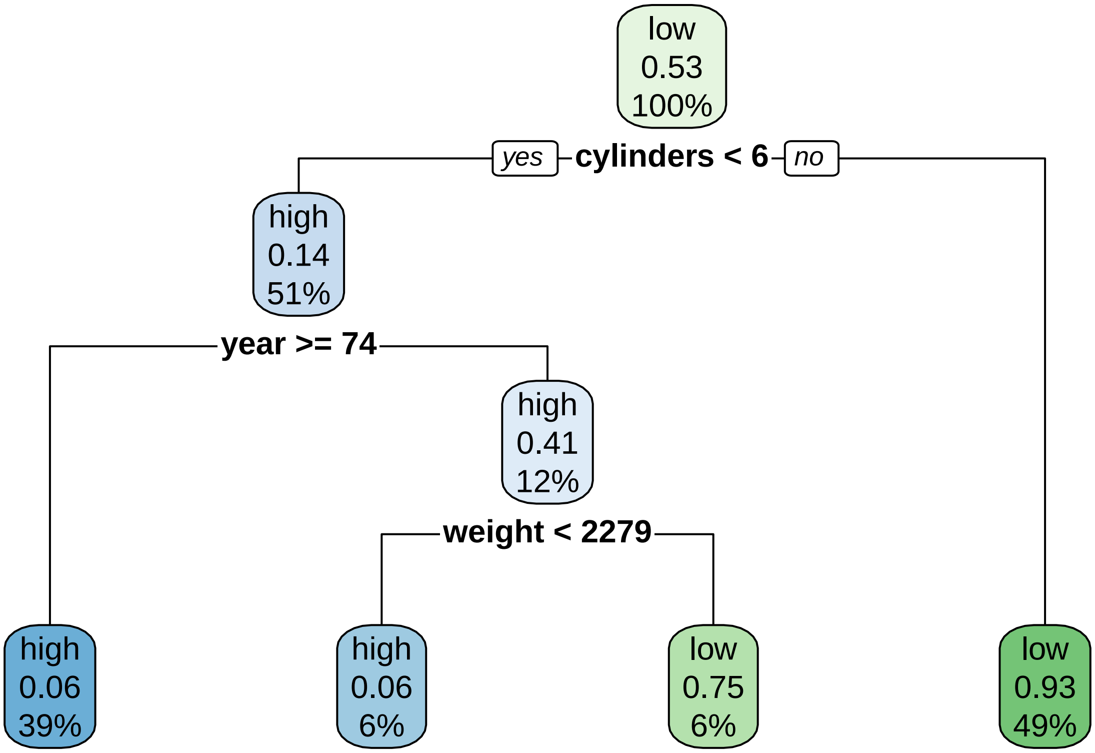
2.A: Build a classification tree using the training data, with mpg cat as the response (10pts). Which mpg category corresponds to the lowest cross-validation error? Is this the same as the tree size obtained using the 1 SE rule (10pts)?

```
##
## Classification tree:
## rpart(formula = mpg_cat ~ ., data = trainData_auto, control = rpart.control(cp = 0))
##
## Variables actually used in tree construction:
## [1] cylinders weight    year
##
## Root node error: 130/274 = 0.47445
##
## n= 274
##
##          CP nsplit rel error  xerror     xstd
## 1 0.776923      0   1.00000 1.00000 0.063582
## 2 0.030769      1   0.22308 0.23077 0.039759
## 3 0.011538      3   0.16154 0.18462 0.035996
## 4 0.000000      5   0.13846 0.20000 0.037316
```

```
plotcp(mpg1)
```

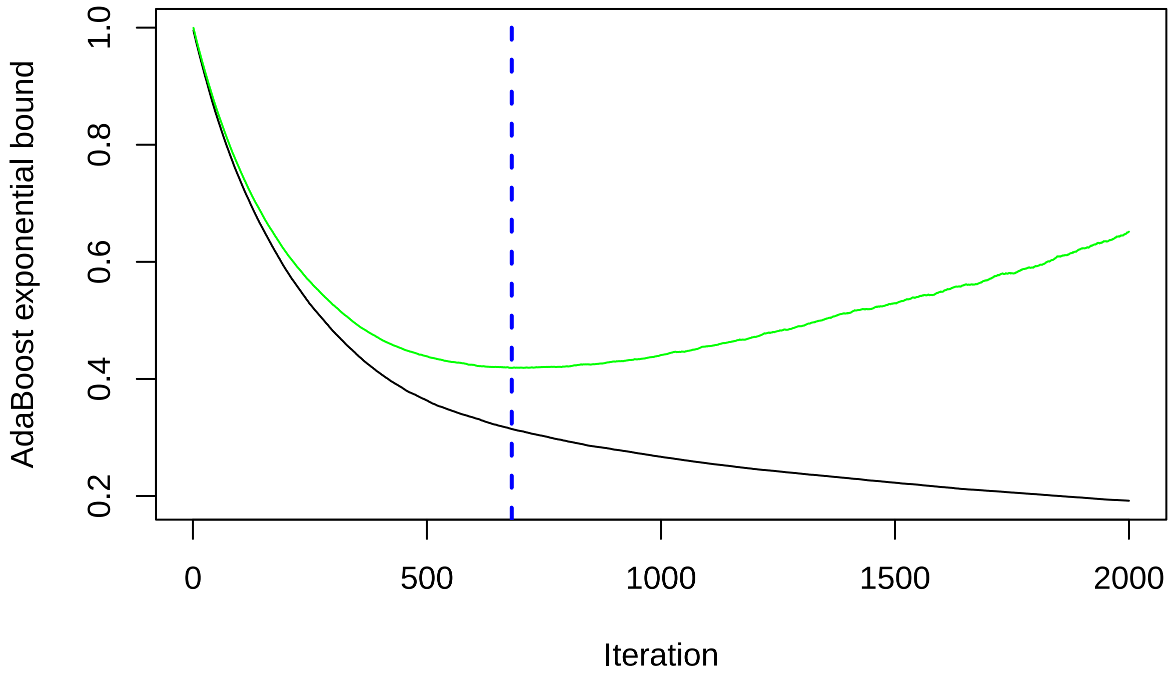


```
minErr = which.min(cpTable[,4])
mpg2 = rpart::prune(mpg1, cp = cpTable[minErr,1])
rpart.plot(mpg2)
```



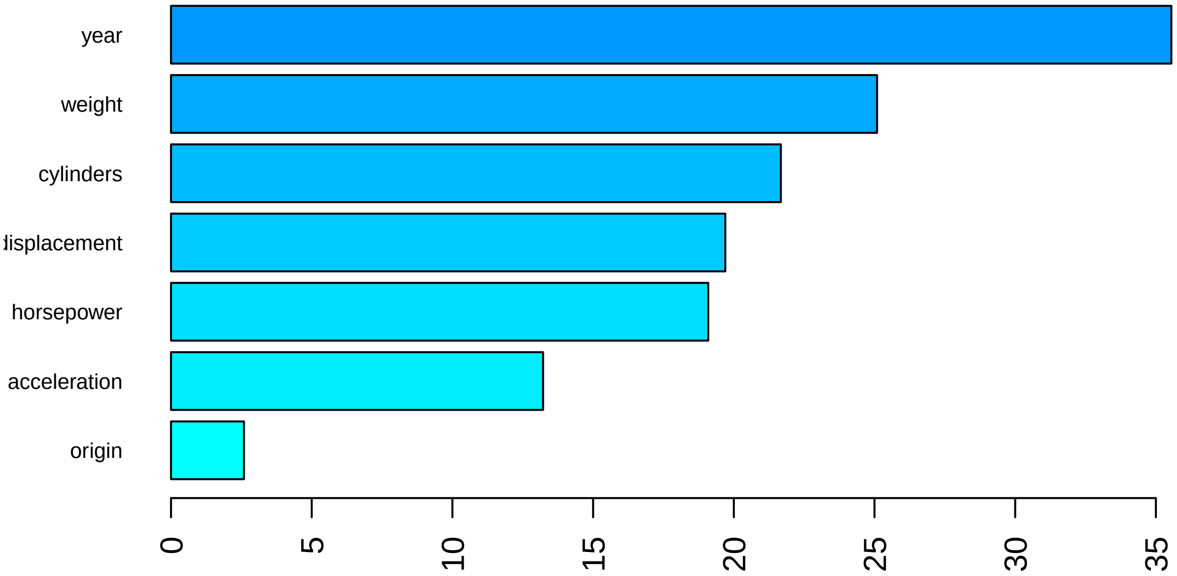
#### 2.B: Perform boosting on the training data and report the variable importance (10pts). Report the test data performance (10pts).

```
# Boosting
trainData_auto$mpg_bin <- ifelse(trainData_auto$mpg_cat == "low", 0, 1)

set.seed(1)
bst = gbm(mpg_bin ~ .,
          data = trainData_auto[, !names(trainData_auto) %in% "mpg_cat"],
          distribution = "adaboost",
          n.trees = 2000,
          interaction.depth = 2,
          shrinkage = 0.005,
          cv.folds = 10)

gbm.perf(bst, method = "cv")
```



```
## [1] 681
```

```
# Convert mpg_bin to factor (required for classification in ranger)
trainData_auto$mpg_bin <- factor(ifelse(trainData_auto$mpg_cat == "low", 0, 1))

# Optional: drop mpg_cat to prevent leakage
trainData_rf <- trainData_auto[, !names(trainData_auto) %in% "mpg_cat"]

# Fit final random forest
set.seed(1)
rf2.final.per <- ranger(mpg_bin ~ .,
                        data = trainData_rf,
                        mtry = 7,
                        splitrule = "gini",  # Classification rule
                        min.node.size = rf.fit$bestTune[[3]],
                        importance = "permutation",
                        scale.permutation.importance = TRUE)

# Plot variable importance
barplot(sort(ranger::importance(rf2.final.per), decreasing = FALSE),
        las = 2, horiz = TRUE, cex.names = 0.7,
        col = colorRampPalette(colors = c("cyan", "blue"))(16))
```


year
weight
cylinders
lisplacement
horsepower
acceleration
origin
0
5
10
15
20
25
30
35