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        Bethel College  502    384    104        11        28         347
## 2     Alfred University 1732   1425    472        37        75        1830
## 3      Caldwell College 1011    604    213        17        42         693
## 4  Tuskegee University 2267   1827    611        20        59        2825
## 5       Carroll College 1160    991    352        19        55        1357
```

```
## 6  St. Paul's College  651    581    243         8       17       617
##    P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1          74     6200       2900   600      800  63       63     11.7
## 2         110    16548       5406   500      600  82       88     11.3
## 3         868     8900       4600   425     1000  87       96     13.9
## 4         144     6735       3395   600     1425  70       74     12.2
## 5         737    12200       3880   480      930  74       81     17.8
## 6          34     5000       3650   600      600  45       45     14.0
##    perc.alumni Expend Grad.Rate
## 1          13   7623        35
## 2          31  10932        73
## 3          25   7922        55
## 4           7  10872        65
## 5          25   7666        79
## 6           8   8426        45
```

```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
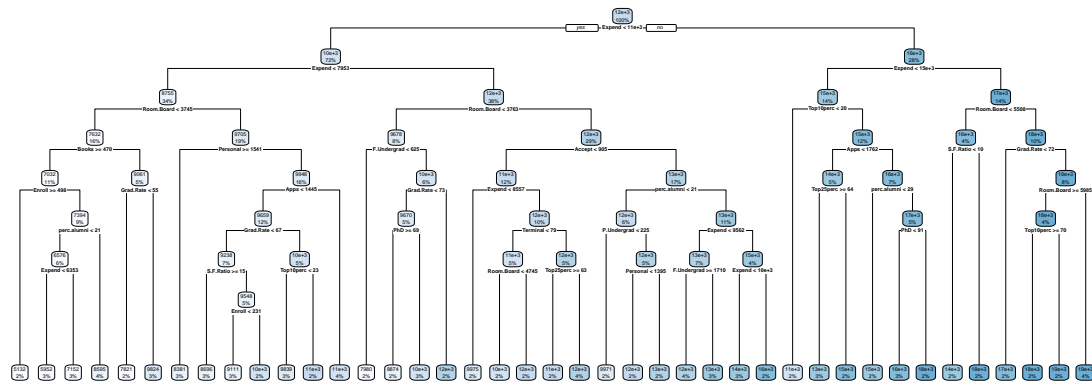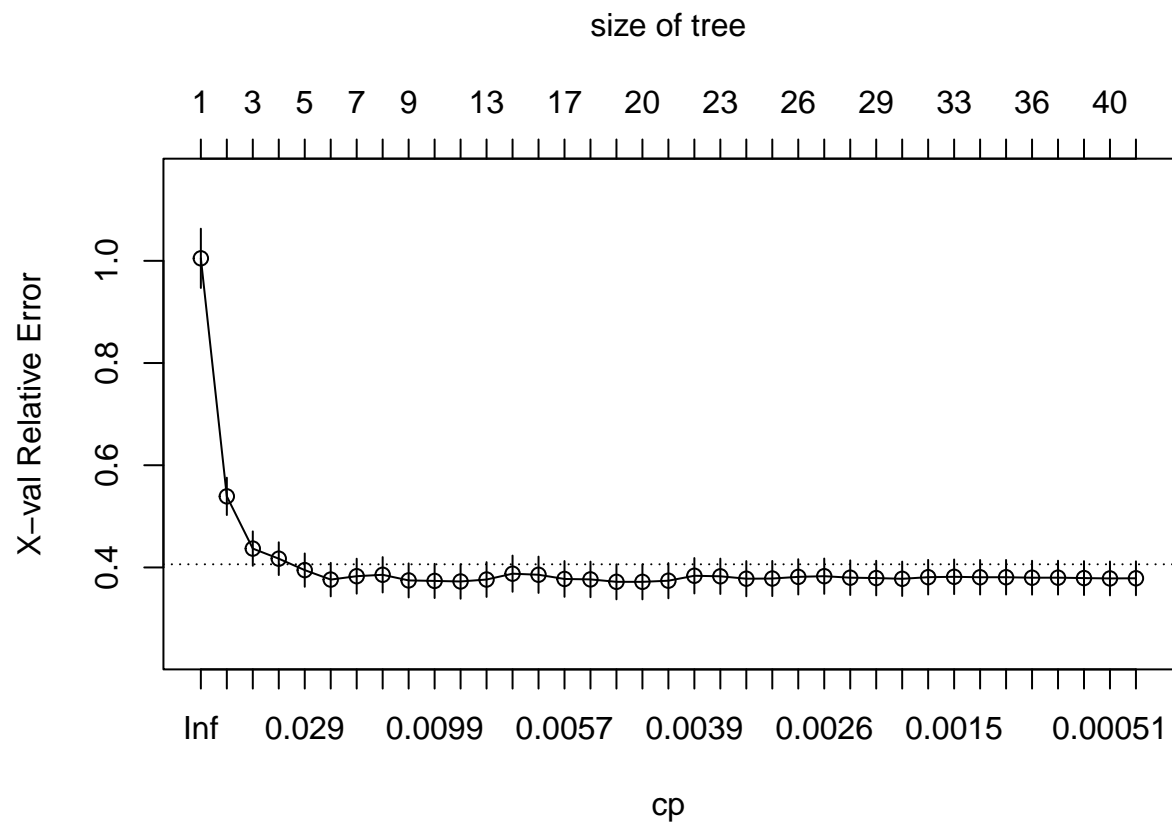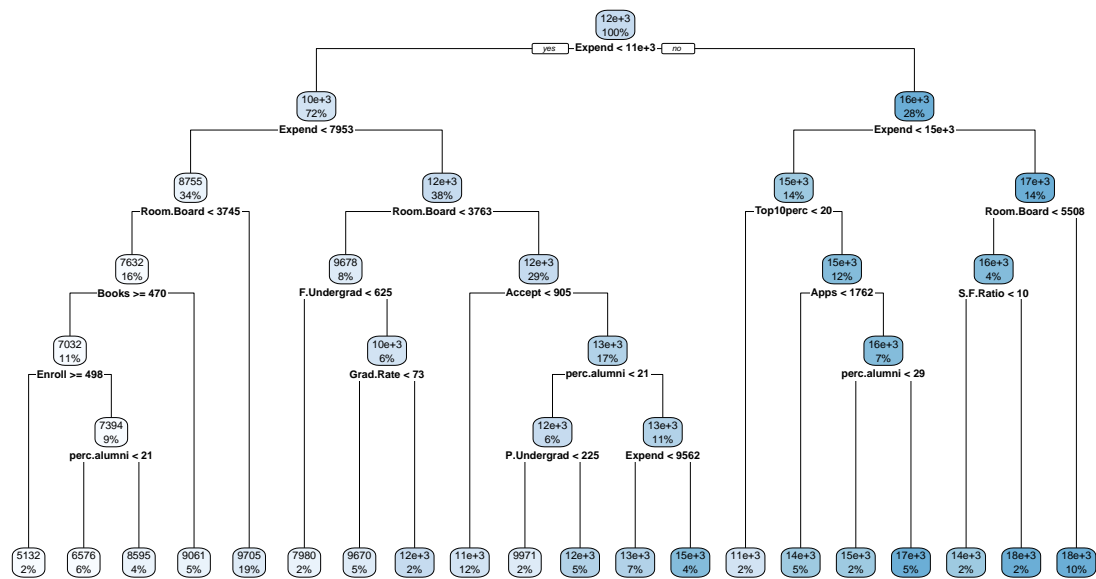

of the tree (10pts).

```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    PhD         Room.Board
## [13] S.F.Ratio   Terminal    Top10perc   Top25perc
##
## Root node error: 6174889736/452 = 13661260
##
## n= 452
##
##            CP nsplit rel error  xerror     xstd
## 1  0.49930434      0   1.00000 1.00490 0.057968
## 2  0.10919061      1   0.50070 0.53911 0.036529
## 3  0.04097275      2   0.39151 0.43691 0.033685
## 4  0.03048800      3   0.35053 0.41720 0.032085
## 5  0.02678798      4   0.32004 0.39478 0.032604
## 6  0.01654616      5   0.29326 0.37625 0.032798
## 7  0.01444453      6   0.27671 0.38291 0.034188
## 8  0.01328950      7   0.26227 0.38564 0.034623
## 9  0.00990755      8   0.24898 0.37462 0.033338
## 10 0.00986166      9   0.23907 0.37364 0.033476
## 11 0.00876820     10   0.22921 0.37254 0.033737
## 12 0.00875857     12   0.21167 0.37636 0.033999
## 13 0.00633400     13   0.20291 0.38782 0.035398
## 14 0.00612392     14   0.19658 0.38568 0.035515
## 15 0.00527293     16   0.18433 0.37748 0.035004
## 16 0.00512488     17   0.17906 0.37661 0.034847
## 17 0.00444872     18   0.17393 0.37187 0.034302
## 18 0.00436886     19   0.16948 0.37185 0.034318
## 19 0.00393330     20   0.16511 0.37409 0.034412
## 20 0.00389777     21   0.16118 0.38373 0.034665
## 21 0.00329589     22   0.15728 0.38268 0.034610
## 22 0.00321592     23   0.15399 0.37800 0.034403
## 23 0.00281595     24   0.15077 0.37831 0.034388
## 24 0.00270739     25   0.14796 0.38159 0.034543
## 25 0.00241497     26   0.14525 0.38290 0.034538
## 26 0.00236934     27   0.14283 0.37997 0.033893
## 27 0.00229161     28   0.14046 0.37924 0.033867
## 28 0.00185541     30   0.13588 0.37779 0.033617
## 29 0.00161652     31   0.13403 0.38094 0.033880
## 30 0.00145553     32   0.13241 0.38170 0.033876
## 31 0.00142533     33   0.13095 0.38078 0.033815
## 32 0.00127258     34   0.12953 0.38088 0.033813
## 33 0.00109761     35   0.12826 0.37992 0.033155
## 34 0.00081594     37   0.12606 0.38003 0.033157
## 35 0.00071260     38   0.12524 0.37916 0.033160
## 36 0.00036526     39   0.12453 0.37848 0.033162
## 37 0.00000000     40   0.12417 0.37879 0.033158
```
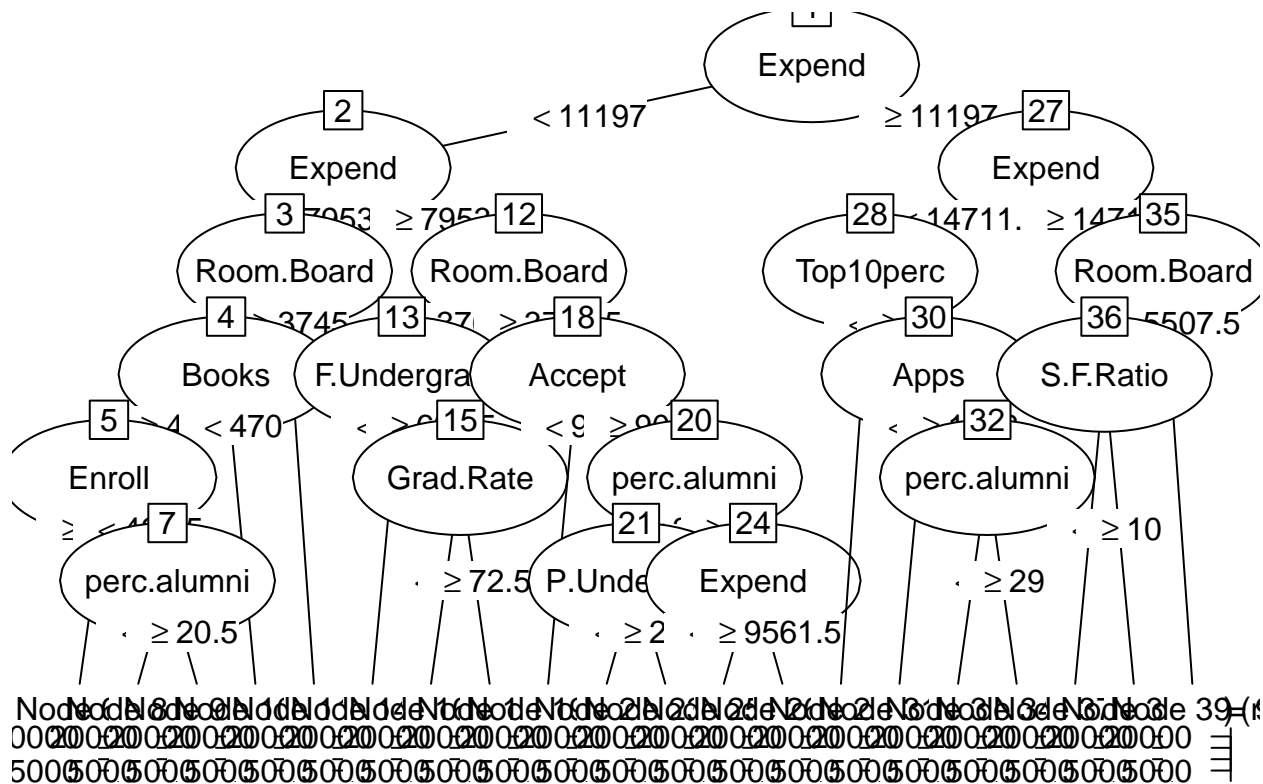
```
cpTable = tree1$cptable
plotcp(tree1)
```

size of tree


```
# 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)
```

Tree diagram:

- 12e+3 / 100% — Expend < 11e+3 (yes / no)
  - yes → 10e+3 / 72% — Expend < 7953
    - 8755 / 34% — Room.Board < 3745
      - 7632 / 16% — Books >= 470
        - 7032 / 11% — Enroll >= 498
          - 5132 / 2%
          - 6576 / 6%
        - 7394 / 9% — perc.alumni < 21
          - 8595 / 4%
          - 9061 / 5%
      - 9705 / 19%
    - 12e+3 / 38% — Room.Board < 3763
      - 9678 / 8% — F.Undergrad < 625
        - 7980 / 2%
        - 10e+3 / 6% — Grad.Rate < 73
          - 9670 / 5%
          - 12e+3 / 2%
      - 12e+3 / 29% — Accept < 905
        - 11e+3 / 12%
        - 13e+3 / 17% — perc.alumni < 21
          - 12e+3 / 6% — P.Undergrad < 225
            - 9971 / 2%
            - 12e+3 / 5%
          - 13e+3 / 11% — Expend < 9562
            - 13e+3 / 7%
            - 15e+3 / 4%
  - no → 16e+3 / 28% — Expend < 15e+3
    - 15e+3 / 14% — Top10perc < 20
      - 11e+3 / 2%
      - 15e+3 / 12% — Apps < 1762
        - 14e+3 / 5%
        - 16e+3 / 7% — perc.alumni < 29
          - 15e+3 / 2%
          - 17e+3 / 5%
    - 17e+3 / 14% — Room.Board < 5508
      - 16e+3 / 4% — S.F.Ratio < 10
        - 14e+3 / 2%
        - 18e+3 / 2%
      - 18e+3 / 10%

```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  0.499304343      0 1.0000000 1.0049030 0.05796832
## 2  0.109190611      1 0.5006957 0.5391143 0.03652891
## 3  0.040972754      2 0.3915050 0.4369093 0.03368548
## 4  0.030488001      3 0.3505323 0.4171952 0.03208462
## 5  0.026787982      4 0.3200443 0.3947814 0.03260383
## 6  0.016546163      5 0.2932563 0.3762456 0.03279813
## 7  0.014444532      6 0.2767101 0.3829088 0.03418845
## 8  0.013289499      7 0.2622656 0.3856422 0.03462300
## 9  0.009907546      8 0.2489761 0.3746178 0.03333792
## 10 0.009861665      9 0.2390686 0.3736388 0.03347560
## 11 0.008768197     10 0.2292069 0.3725418 0.03373734
## 12 0.008758567     12 0.2116705 0.3763635 0.03399948
## 13 0.006333997     13 0.2029119 0.3878160 0.03539827
## 14 0.006123922     14 0.1965779 0.3856843 0.03551511
## 15 0.005272927     16 0.1843301 0.3774793 0.03500447
## 16 0.005124879     17 0.1790572 0.3766075 0.03484688
## 17 0.004448721     18 0.1739323 0.3718712 0.03430230
## 18 0.004368860     19 0.1694836 0.3718464 0.03431808
```

```
## 
## Variable importance
##       Expend    Terminal    Top10perc         PhD    Top25perc    S.F.Ratio
##           29          12          11          10           9           9
## Room.Board         Apps   Grad.Rate       Enroll P.Undergrad      Accept
##            6           2           2           2           2           2
## F.Undergrad perc.alumni    Personal       Books
##            1           1           1           1
## 
## Node number 1: 452 observations,    complexity param=0.4993043
##   mean=11896.55, MSE=1.366126e+07
##   left son=2 (325 obs) right son=3 (127 obs)
##   Primary splits:
##       Expend     < 11197   to the left,  improve=0.4993043, (0 missing)
##       PhD        < 78.5    to the left,  improve=0.3585854, (0 missing)
##       Terminal   < 83.5    to the left,  improve=0.3473682, (0 missing)
##       Room.Board < 3953.5  to the left,  improve=0.3144625, (0 missing)
##       Top10perc  < 34.5    to the left,  improve=0.2716013, (0 missing)
##   Surrogate splits:
##       Terminal   < 92.5    to the left,  agree=0.843, adj=0.441, (0 split)
##       Top10perc  < 39.5    to the left,  agree=0.838, adj=0.425, (0 split)
##       Top25perc  < 76.5    to the left,  agree=0.821, adj=0.362, (0 split)
##       PhD        < 87.5    to the left,  agree=0.821, adj=0.362, (0 split)
##       S.F.Ratio  < 10.55   to the right, agree=0.801, adj=0.291, (0 split)
## 
## Node number 2: 325 observations,    complexity param=0.1091906
##   mean=10263.92, MSE=6567163
##   left son=4 (155 obs) right son=5 (170 obs)
##   Primary splits:
##       Expend     < 7953    to the left,  improve=0.3159027, (0 missing)
##       Room.Board < 3772.5  to the left,  improve=0.2849725, (0 missing)
##       Grad.Rate  < 62.5    to the left,  improve=0.1846756, (0 missing)
##       Terminal   < 77.5    to the left,  improve=0.1819224, (0 missing)
##       PhD        < 71.5    to the left,  improve=0.1283426, (0 missing)
##   Surrogate splits:
##       PhD        < 70.5    to the left,  agree=0.702, adj=0.374, (0 split)
##       Terminal   < 73.5    to the left,  agree=0.702, adj=0.374, (0 split)
##       Room.Board < 3898.5  to the left,  agree=0.662, adj=0.290, (0 split)
##       S.F.Ratio  < 14.05   to the right, agree=0.652, adj=0.271, (0 split)
##       Grad.Rate  < 62.5    to the left,  agree=0.640, adj=0.245, (0 split)
## 
## Node number 3: 127 observations,    complexity param=0.04097275
##   mean=16074.54, MSE=7538680
##   left son=6 (63 obs) right son=7 (64 obs)
##   Primary splits:
##       Expend     < 14711.5 to the left,  improve=0.2642563, (0 missing)
##       Room.Board < 5557.5  to the left,  improve=0.2556123, (0 missing)
##       Apps       < 1025.5  to the left,  improve=0.2017792, (0 missing)
##       Terminal   < 90.5    to the left,  improve=0.1959893, (0 missing)
##       Top10perc  < 20      to the left,  improve=0.1924720, (0 missing)
##   Surrogate splits:
##       S.F.Ratio  < 10.75   to the right, agree=0.756, adj=0.508, (0 split)
##       Apps       < 2691    to the left,  agree=0.693, adj=0.381, (0 split)
##       Top10perc  < 44.5    to the left,  agree=0.693, adj=0.381, (0 split)
```

```
##        Top25perc   < 77.5     to the left,   agree=0.693, adj=0.381, (0 split)
##        Room.Board  < 5585     to the left,   agree=0.693, adj=0.381, (0 split)
##
## Node number 4: 155 observations,    complexity param=0.02678798
##    mean=8755.49, MSE=4127556
##    left son=8 (71 obs) right son=9 (84 obs)
##    Primary splits:
##        Room.Board  < 3745     to the left,   improve=0.2585500, (0 missing)
##        Grad.Rate   < 64.5     to the left,   improve=0.1903534, (0 missing)
##        perc.alumni < 13.5     to the left,   improve=0.1291383, (0 missing)
##        Expend      < 7115.5   to the left,   improve=0.1247768, (0 missing)
##        Personal    < 1523.5   to the right,  improve=0.1128661, (0 missing)
##    Surrogate splits:
##        P.Undergrad < 168      to the left,   agree=0.671, adj=0.282, (0 split)
##        Expend      < 6872.5   to the left,   agree=0.671, adj=0.282, (0 split)
##        Terminal    < 63.5     to the left,   agree=0.639, adj=0.211, (0 split)
##        Grad.Rate   < 61.5     to the left,   agree=0.632, adj=0.197, (0 split)
##        Personal    < 1523.5   to the right,  agree=0.626, adj=0.183, (0 split)
##
## Node number 5: 170 observations,    complexity param=0.030488
##    mean=11639.25, MSE=4825394
##    left son=10 (38 obs) right son=11 (132 obs)
##    Primary splits:
##        Room.Board  < 3762.5   to the left,   improve=0.2294967, (0 missing)
##        Accept      < 915      to the left,   improve=0.2218655, (0 missing)
##        Apps        < 1108     to the left,   improve=0.2013545, (0 missing)
##        F.Undergrad < 1068     to the left,   improve=0.1833686, (0 missing)
##        Enroll      < 302.5    to the left,   improve=0.1487546, (0 missing)
##    Surrogate splits:
##        P.Undergrad < 50.5     to the left,   agree=0.818, adj=0.184, (0 split)
##        PhD         < 54.5     to the left,   agree=0.800, adj=0.105, (0 split)
##        Terminal    < 59.5     to the left,   agree=0.800, adj=0.105, (0 split)
##        Apps        < 159.5    to the left,   agree=0.788, adj=0.053, (0 split)
##        Accept      < 129      to the left,   agree=0.788, adj=0.053, (0 split)
##
## Node number 6: 63 observations,    complexity param=0.01654616
##    mean=14651.95, MSE=6275273
##    left son=12 (7 obs) right son=13 (56 obs)
##    Primary splits:
##        Top10perc   < 20       to the left,   improve=0.2584362, (0 missing)
##        Grad.Rate   < 61.5     to the left,   improve=0.2305391, (0 missing)
##        Apps        < 953.5    to the left,   improve=0.2291761, (0 missing)
##        Accept      < 716.5    to the left,   improve=0.2232408, (0 missing)
##        F.Undergrad < 955      to the left,   improve=0.2179193, (0 missing)
##    Surrogate splits:
##        Top25perc   < 44       to the left,   agree=0.984, adj=0.857, (0 split)
##        PhD         < 48       to the left,   agree=0.952, adj=0.571, (0 split)
##        Terminal    < 67.5     to the left,   agree=0.952, adj=0.571, (0 split)
##        Personal    < 2654     to the right,  agree=0.937, adj=0.429, (0 split)
##        perc.alumni < 5        to the left,   agree=0.937, adj=0.429, (0 split)
##
## Node number 7: 64 observations,    complexity param=0.01444453
##    mean=17474.91, MSE=4829186
##    left son=14 (20 obs) right son=15 (44 obs)
```

```
##     Primary splits:
##         Room.Board  < 5507.5  to the left,   improve=0.2885883, (0 missing)
##         Grad.Rate   < 67.5    to the left,   improve=0.2150844, (0 missing)
##         Apps        < 889.5   to the left,   improve=0.1583357, (0 missing)
##         F.Undergrad < 844     to the left,   improve=0.1318119, (0 missing)
##         Accept      < 714     to the left,   improve=0.1300508, (0 missing)
##     Surrogate splits:
##         P.Undergrad < 24.5    to the left,   agree=0.781, adj=0.30, (0 split)
##         Apps        < 1056    to the left,   agree=0.766, adj=0.25, (0 split)
##         Grad.Rate   < 67.5    to the left,   agree=0.766, adj=0.25, (0 split)
##         Accept      < 1177    to the left,   agree=0.750, adj=0.20, (0 split)
##         Enroll      < 141.5   to the left,   agree=0.750, adj=0.20, (0 split)
##
## Node number 8: 71 observations,    complexity param=0.009861665
##   mean=7631.845, MSE=3919718
##   left son=16 (50 obs) right son=17 (21 obs)
##     Primary splits:
##         Books       < 470     to the right,  improve=0.2188095, (0 missing)
##         Grad.Rate   < 65.5    to the left,   improve=0.2071713, (0 missing)
##         perc.alumni < 14.5    to the left,   improve=0.1659754, (0 missing)
##         Apps        < 1373    to the right,  improve=0.1337428, (0 missing)
##         Enroll      < 497.5   to the right,  improve=0.1294546, (0 missing)
##     Surrogate splits:
##         Room.Board  < 3690    to the left,   agree=0.761, adj=0.190, (0 split)
##         F.Undergrad < 3296    to the left,   agree=0.718, adj=0.048, (0 split)
##         PhD         < 82.5    to the left,   agree=0.718, adj=0.048, (0 split)
##         Terminal    < 85      to the left,   agree=0.718, adj=0.048, (0 split)
##         S.F.Ratio   < 10.2    to the right,  agree=0.718, adj=0.048, (0 split)
##
## Node number 9: 84 observations
##   mean=9705.238, MSE=2334028
##
## Node number 10: 38 observations,    complexity param=0.006333997
##   mean=9677.921, MSE=4457169
##   left son=20 (10 obs) right son=21 (28 obs)
##     Primary splits:
##         F.Undergrad < 624.5   to the left,   improve=0.2309215, (0 missing)
##         Top10perc   < 27.5    to the left,   improve=0.1926464, (0 missing)
##         Books       < 561.5   to the right,  improve=0.1808821, (0 missing)
##         Grad.Rate   < 73.5    to the left,   improve=0.1588082, (0 missing)
##         Top25perc   < 55.5    to the left,   improve=0.1466846, (0 missing)
##     Surrogate splits:
##         Enroll      < 134     to the left,   agree=0.895, adj=0.6, (0 split)
##         Accept      < 343.5   to the left,   agree=0.868, adj=0.5, (0 split)
##         Apps        < 274     to the left,   agree=0.816, adj=0.3, (0 split)
##         P.Undergrad < 24      to the left,   agree=0.816, adj=0.3, (0 split)
##         Room.Board  < 3071    to the left,   agree=0.816, adj=0.3, (0 split)
##
## Node number 11: 132 observations,    complexity param=0.0132895
##   mean=12203.87, MSE=3505185
##   left son=22 (53 obs) right son=23 (79 obs)
##     Primary splits:
##         Accept      < 905     to the left,   improve=0.1773589, (0 missing)
##         Apps        < 1108    to the left,   improve=0.1578164, (0 missing)
```

```
##         Expend      < 9464.5  to the left,  improve=0.1327338, (0 missing)
##         F.Undergrad < 1064.5  to the left,  improve=0.1231907, (0 missing)
##         Enroll      < 299     to the left,  improve=0.1211312, (0 missing)
##    Surrogate splits:
##         Apps        < 1108    to the left,  agree=0.970, adj=0.925, (0 split)
##         Enroll      < 318.5   to the left,  agree=0.947, adj=0.868, (0 split)
##         F.Undergrad < 1231.5  to the left,  agree=0.864, adj=0.660, (0 split)
##         S.F.Ratio   < 11.95   to the left,  agree=0.720, adj=0.302, (0 split)
##         Grad.Rate   < 59.5    to the left,  agree=0.667, adj=0.170, (0 split)
##
## Node number 12: 7 observations
##   mean=11050, MSE=3895984
##
## Node number 13: 56 observations,    complexity param=0.009907546
##   mean=15102.2, MSE=4748207
##   left son=26 (24 obs) right son=27 (32 obs)
##   Primary splits:
##         Apps        < 1762    to the left,  improve=0.2300793, (0 missing)
##         F.Undergrad < 955     to the left,  improve=0.1718941, (0 missing)
##         Room.Board  < 4371    to the left,  improve=0.1585929, (0 missing)
##         Accept      < 933.5   to the left,  improve=0.1551638, (0 missing)
##         Grad.Rate   < 74.5    to the left,  improve=0.1340003, (0 missing)
##    Surrogate splits:
##         Accept      < 1159    to the left,  agree=0.911, adj=0.792, (0 split)
##         Enroll      < 355.5   to the left,  agree=0.911, adj=0.792, (0 split)
##         F.Undergrad < 1198.5  to the left,  agree=0.875, adj=0.708, (0 split)
##         Room.Board  < 4682.5  to the left,  agree=0.696, adj=0.292, (0 split)
##         Grad.Rate   < 61.5    to the left,  agree=0.696, adj=0.292, (0 split)
##
## Node number 14: 20 observations,    complexity param=0.008758567
##   mean=15723.9, MSE=7890089
##   left son=28 (11 obs) right son=29 (9 obs)
##   Primary splits:
##         S.F.Ratio   < 10      to the left,  improve=0.3427286, (0 missing)
##         P.Undergrad < 40.5    to the right, improve=0.1931171, (0 missing)
##         Personal    < 830     to the right, improve=0.1685762, (0 missing)
##         Room.Board  < 4344    to the right, improve=0.1643041, (0 missing)
##         Expend      < 18163.5 to the right, improve=0.1388171, (0 missing)
##    Surrogate splits:
##         Room.Board  < 4147.5  to the right, agree=0.85, adj=0.667, (0 split)
##         P.Undergrad < 54.5    to the right, agree=0.75, adj=0.444, (0 split)
##         Apps        < 2830.5  to the right, agree=0.70, adj=0.333, (0 split)
##         Expend      < 18163.5 to the right, agree=0.70, adj=0.333, (0 split)
##         Enroll      < 394     to the right, agree=0.65, adj=0.222, (0 split)
##
## Node number 15: 44 observations
##   mean=18270.82, MSE=1410744
##
## Node number 16: 50 observations,    complexity param=0.006123922
##   mean=7031.66, MSE=3515178
##   left son=32 (8 obs) right son=33 (42 obs)
##   Primary splits:
##         Enroll      < 497.5   to the right, improve=0.1955700, (0 missing)
##         S.F.Ratio   < 11.65   to the right, improve=0.1751577, (0 missing)
```

```
##         Apps       < 1337.5  to the right, improve=0.1710200, (0 missing)
##         Grad.Rate  < 62.5    to the left,  improve=0.1597601, (0 missing)
##         perc.alumni < 13.5   to the left,  improve=0.1579270, (0 missing)
##     Surrogate splits:
##         F.Undergrad < 2046   to the right, agree=0.94, adj=0.625, (0 split)
##         Accept     < 1060.5  to the right, agree=0.90, adj=0.375, (0 split)
##         Top25perc  < 72.5    to the right, agree=0.90, adj=0.375, (0 split)
##         Apps       < 1337.5  to the right, agree=0.88, adj=0.250, (0 split)
##         Top10perc  < 40      to the right, agree=0.88, adj=0.250, (0 split)
##
## Node number 17: 21 observations
##   mean=9060.857, MSE=1983163
##
## Node number 20: 10 observations
##   mean=7980.3, MSE=3952001
##
## Node number 21: 28 observations,    complexity param=0.005124879
##   mean=10284.21, MSE=3240738
##   left son=42 (21 obs) right son=43 (7 obs)
##   Primary splits:
##         Grad.Rate  < 72.5    to the left,  improve=0.3487473, (0 missing)
##         Personal   < 925     to the right, improve=0.3092032, (0 missing)
##         Room.Board < 3598.5  to the left,  improve=0.2268143, (0 missing)
##         perc.alumni < 27.5   to the left,  improve=0.2168475, (0 missing)
##         Books      < 525     to the right, improve=0.1911520, (0 missing)
##     Surrogate splits:
##         perc.alumni < 36.5   to the left,  agree=0.857, adj=0.429, (0 split)
##         Apps       < 1694.5  to the left,  agree=0.786, adj=0.143, (0 split)
##         Books      < 440     to the right, agree=0.786, adj=0.143, (0 split)
##         Personal   < 925     to the right, agree=0.786, adj=0.143, (0 split)
##         Expend     < 10951.5 to the left,  agree=0.786, adj=0.143, (0 split)
##
## Node number 22: 53 observations
##   mean=11241.25, MSE=1834255
##
## Node number 23: 79 observations,    complexity param=0.008768197
##   mean=12849.68, MSE=3587440
##   left son=46 (29 obs) right son=47 (50 obs)
##   Primary splits:
##         perc.alumni < 21      to the left,  improve=0.16160630, (0 missing)
##         Expend     < 10116   to the left,  improve=0.14981690, (0 missing)
##         Personal   < 860     to the right, improve=0.13799190, (0 missing)
##         F.Undergrad < 1607.5 to the right, improve=0.10281960, (0 missing)
##         S.F.Ratio  < 15.95   to the right, improve=0.07857678, (0 missing)
##     Surrogate splits:
##         P.Undergrad < 1167   to the right, agree=0.797, adj=0.448, (0 split)
##         Personal   < 1280    to the right, agree=0.759, adj=0.345, (0 split)
##         Grad.Rate  < 66.5    to the left,  agree=0.759, adj=0.345, (0 split)
##         F.Undergrad < 2680   to the right, agree=0.747, adj=0.310, (0 split)
##         Books      < 615     to the right, agree=0.722, adj=0.241, (0 split)
##
## Node number 26: 24 observations
##   mean=13895.29, MSE=4460892
##
```

```
## Node number 27: 32 observations,    complexity param=0.004448721
##   mean=16007.38, MSE=3051881
##   left son=54 (9 obs) right son=55 (23 obs)
##   Primary splits:
##       perc.alumni < 29       to the left,  improve=0.2812852, (0 missing)
##       P.Undergrad < 31       to the right, improve=0.2567658, (0 missing)
##       S.F.Ratio   < 13.3     to the right, improve=0.1954681, (0 missing)
##       Enroll      < 568.5    to the right, improve=0.1800453, (0 missing)
##       F.Undergrad < 1955.5   to the right, improve=0.1730059, (0 missing)
##   Surrogate splits:
##       Grad.Rate   < 68       to the left,  agree=0.844, adj=0.444, (0 split)
##       P.Undergrad < 115      to the right, agree=0.812, adj=0.333, (0 split)
##       Enroll      < 568.5    to the right, agree=0.781, adj=0.222, (0 split)
##       Personal    < 1150.5   to the right, agree=0.781, adj=0.222, (0 split)
##       F.Undergrad < 2351.5   to the right, agree=0.750, adj=0.111, (0 split)
##
## Node number 28: 11 observations
##   mean=14236.45, MSE=5654830
##
## Node number 29: 9 observations
##   mean=17541.89, MSE=4612830
##
## Node number 32: 8 observations
##   mean=5131.875, MSE=2169977
##
## Node number 33: 42 observations,    complexity param=0.006123922
##   mean=7393.524, MSE=2952998
##   left son=66 (25 obs) right son=67 (17 obs)
##   Primary splits:
##       perc.alumni < 20.5     to the left,  improve=0.3326395, (0 missing)
##       Grad.Rate   < 65.5     to the left,  improve=0.2538424, (0 missing)
##       S.F.Ratio   < 13.15    to the right, improve=0.1818469, (0 missing)
##       Room.Board  < 3075     to the left,  improve=0.1816320, (0 missing)
##       Expend      < 6296     to the left,  improve=0.1300563, (0 missing)
##   Surrogate splits:
##       P.Undergrad < 410.5    to the left,  agree=0.762, adj=0.412, (0 split)
##       S.F.Ratio   < 11.65    to the right, agree=0.714, adj=0.294, (0 split)
##       PhD         < 56.5     to the right, agree=0.690, adj=0.235, (0 split)
##       Grad.Rate   < 65.5     to the left,  agree=0.690, adj=0.235, (0 split)
##       Apps        < 328      to the right, agree=0.643, adj=0.118, (0 split)
##
## Node number 42: 21 observations
##   mean=9670.429, MSE=2050383
##
## Node number 43: 7 observations
##   mean=12125.57, MSE=2291008
##
## Node number 46: 29 observations,    complexity param=0.005272927
##   mean=11849.9, MSE=3564617
##   left son=92 (7 obs) right son=93 (22 obs)
##   Primary splits:
##       P.Undergrad < 225      to the left,  improve=0.3149707, (0 missing)
##       Terminal    < 81       to the left,  improve=0.2348546, (0 missing)
##       PhD         < 82.5     to the left,  improve=0.2255336, (0 missing)
```

```
##         Room.Board  < 5930     to the left,  improve=0.2083413, (0 missing)
##         Accept      < 1709.5  to the left,  improve=0.1950155, (0 missing)
##     Surrogate splits:
##         Accept       < 1301.5  to the left,  agree=0.862, adj=0.429, (0 split)
##         Enroll       < 460.5   to the left,  agree=0.862, adj=0.429, (0 split)
##         Top10perc    < 39.5    to the right, agree=0.862, adj=0.429, (0 split)
##         F.Undergrad < 1741.5   to the left,  agree=0.862, adj=0.429, (0 split)
##         Apps         < 1874.5  to the left,  agree=0.828, adj=0.286, (0 split)
##
## Node number 47: 50 observations,    complexity param=0.008768197
##   mean=13429.56, MSE=2684667
##   left son=94 (30 obs) right son=95 (20 obs)
##   Primary splits:
##         Expend       < 9561.5  to the left,  improve=0.4654939, (0 missing)
##         Personal     < 860     to the right, improve=0.2962261, (0 missing)
##         P.Undergrad < 408      to the right, improve=0.1870012, (0 missing)
##         Top10perc    < 26.5    to the left,  improve=0.1799381, (0 missing)
##         F.Undergrad < 1659     to the right, improve=0.1656782, (0 missing)
##     Surrogate splits:
##         Personal     < 832.5   to the right, agree=0.74, adj=0.35, (0 split)
##         Terminal    < 85.5     to the left,  agree=0.72, adj=0.30, (0 split)
##         Top25perc   < 57.5     to the left,  agree=0.70, adj=0.25, (0 split)
##         PhD         < 78.5     to the left,  agree=0.70, adj=0.25, (0 split)
##         Accept      < 994.5    to the right, agree=0.68, adj=0.20, (0 split)
##
## Node number 54: 9 observations
##   mean=14526.22, MSE=2103569
##
## Node number 55: 23 observations
##   mean=16586.96, MSE=2228595
##
## Node number 66: 25 observations
##   mean=6576.24, MSE=1401347
##
## Node number 67: 17 observations
##   mean=8595.412, MSE=2808019
##
## Node number 92: 7 observations
##   mean=9971.429, MSE=6579498
##
## Node number 93: 22 observations
##   mean=12447.59, MSE=1125348
##
## Node number 94: 30 observations
##   mean=12516.8, MSE=1103459
##
## Node number 95: 20 observations
##   mean=14798.7, MSE=1932239
```

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

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

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%**).

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

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