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
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidymodels)
```

```
## -- Attaching packages ---------------------------------- tidymodels 1.3.0 --
```

```
## v broom        1.0.7     v rsample     1.2.1
## v dials        1.4.0     v tibble      3.2.1
## v dplyr        1.1.4     v tidyr       1.3.1
## v infer        1.0.7     v tune        1.3.0
## v modeldata    1.4.0     v workflows   1.2.0
## v parsnip      1.3.0     v workflowsets 1.1.0
## v purrr        1.0.4     v yardstick   1.3.2
## v recipes      1.1.1


## -- Conflicts ------------------------------------------- tidymodels_conflicts() --
## x purrr::discard()        masks scales::discard()
## x dplyr::filter()         masks stats::filter()
## x dplyr::lag()            masks stats::lag()
## x purrr::lift()           masks caret::lift()
## x yardstick::precision()  masks caret::precision()
## x yardstick::recall()     masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()         masks stats::step()
```

```r
library(ISLR)
```

```
##
## Attaching package: 'ISLR'

## The following object is masked _by_ '.GlobalEnv':
##
##     College
```

```r
library(mlbench)
library(caret)
library(tidymodels)
library(rpart)
```

```
##
## Attaching package: 'rpart'

## The following object is masked from 'package:dials':
##
##     prune
```

```r
library(rpart.plot)
library(party)
```

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4
```

```
##
## Attaching package: 'modeltools'

## The following object is masked from 'package:workflows':
##
##     fit

## The following object is masked from 'package:tune':
##
##     parameters

## The following object is masked from 'package:parsnip':
##
##     fit

## The following object is masked from 'package:infer':
##
##     fit

## The following object is masked from 'package:dials':
##
##     parameters

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

##
## Attaching package: 'party'

## The following object is masked from 'package:dplyr':
##
##     where
```

```r
library(partykit)
```

```
## Loading required package: libcoin

##
## Attaching package: 'partykit'
```

```
## The following objects are masked from 'package:party':
##
##      cforest, ctree, ctree_control, edge_simple, mob, mob_control,
##      node_barplot, node_bivplot, node_boxplot, node_inner, node_surv,
##      node_terminal, varimp
```

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##      cov, smooth, var
```

```r
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       George Fox College    809    726    294        27        52        1271
## 2           Averett College    627    556    172        16        40         777
## 3    Northwestern College    860    811    366        22        56        1040
## 4       Trinity College DC    247    189    100        19        49         309
## 5 Lebanon Valley College   1386   1060    320        28        56         965
## 6     New York University  13594   7244   2505        70        86       12408
##    P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1           43    12500       4130   400     1050  53       53      13.5
## 2          538     9925       4135   750     1350  59       67      22.4
## 3           52     9900       3075   300     1800  68       68      14.9
## 4          639    11412       6430   500      900  89       93       8.3
## 5          502    13850       4755   400     1125  84       84      12.3
## 6         2814    17748       7262   450     1000  87       98       7.8
##    perc.alumni Expend Grad.Rate
## 1           22   7136        52
## 2           11   6523        48
## 3           34   6357        68
## 4           37  11806        96
## 5           30   8196        85
## 6           16  21227        71
```

```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
```

4

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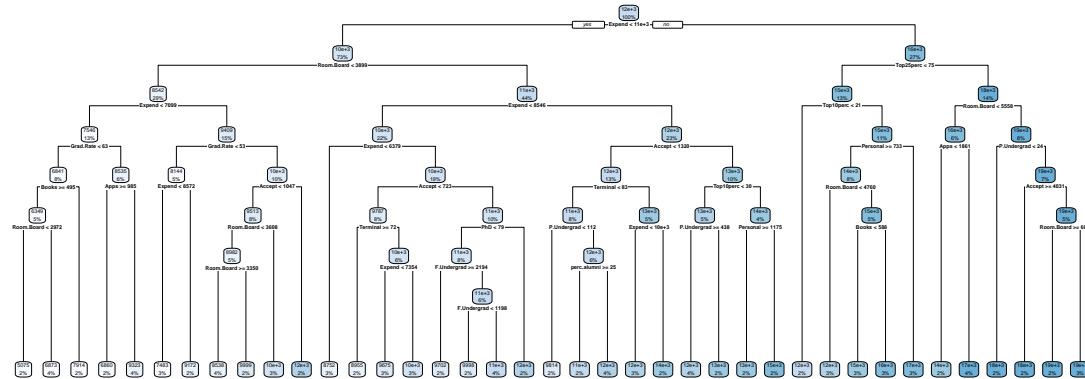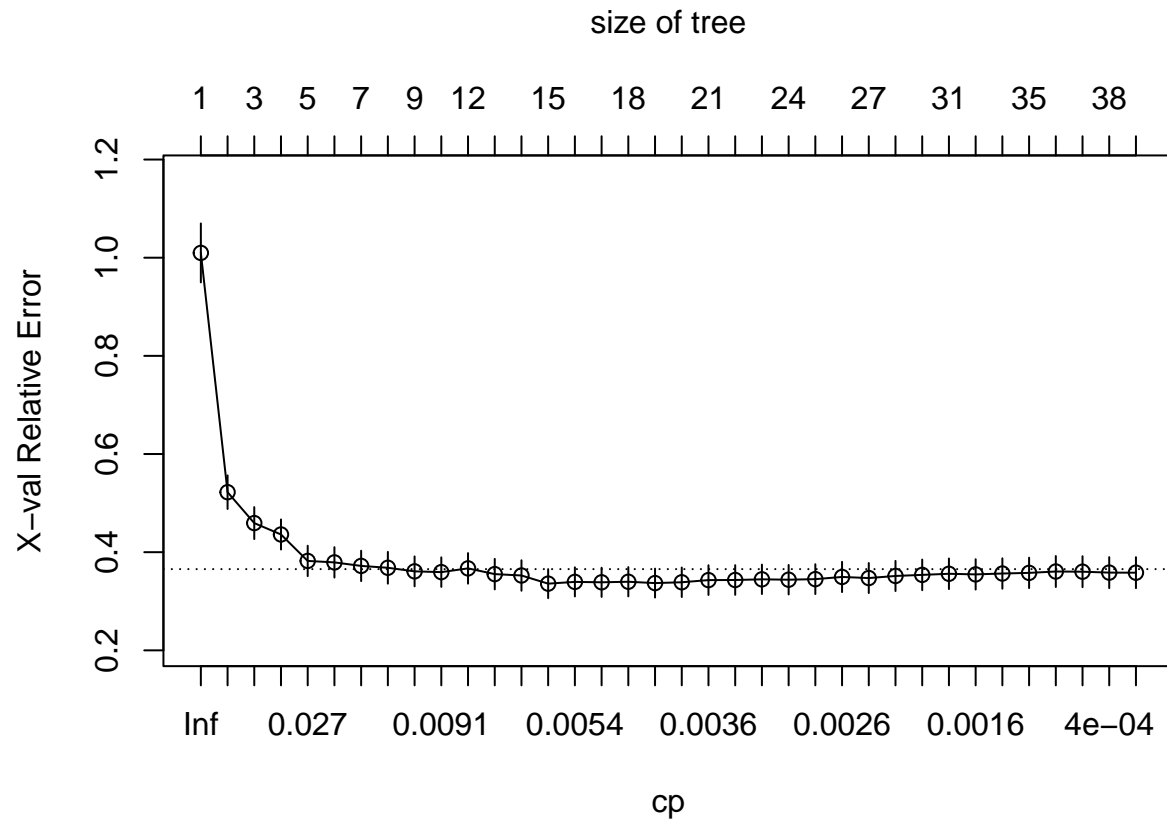

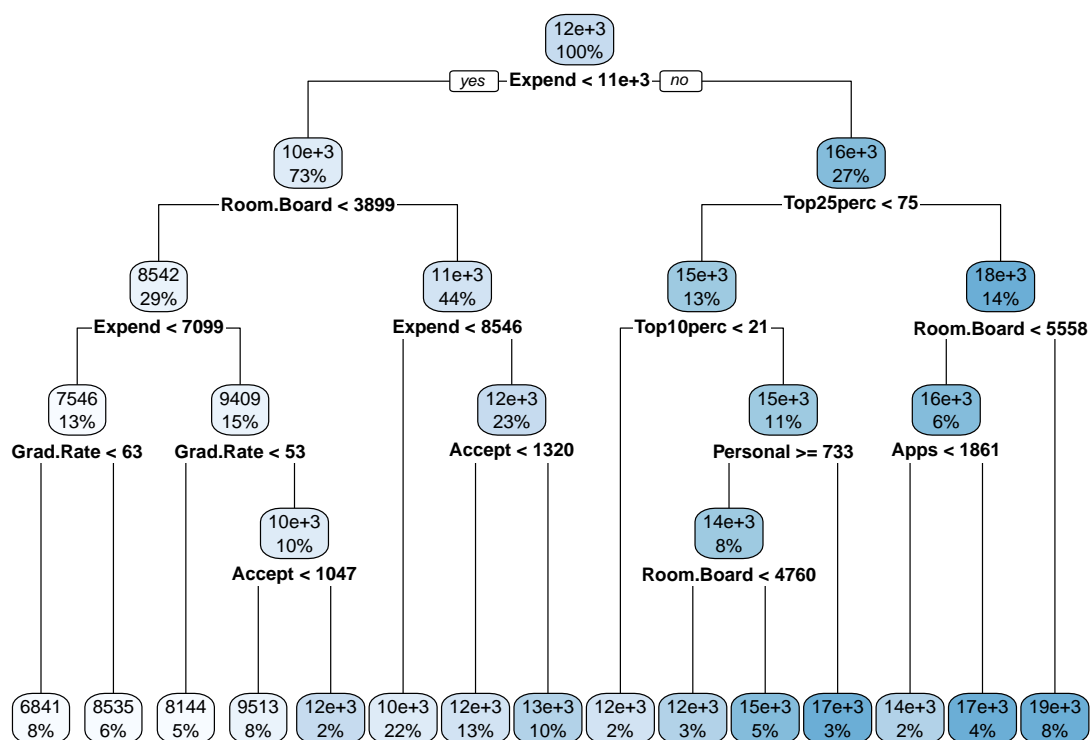**of the tree (10pts).**

```
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       Expend      F.Undergrad Grad.Rate
##  [7] P.Undergrad perc.alumni Personal    PhD         Room.Board  Terminal
## [13] Top10perc   Top25perc
##
## Root node error: 6.163e+09/452 = 13635025
##
## n= 452
##
##             CP nsplit rel error  xerror      xstd
## 1  0.51733456      0   1.00000 1.00977 0.060075
## 2  0.09866047      1   0.48267 0.52223 0.034192
## 3  0.04300806      2   0.38400 0.45927 0.032588
## 4  0.04077284      3   0.34100 0.43608 0.030675
## 5  0.01805822      4   0.30022 0.38218 0.030979
## 6  0.01516963      5   0.28217 0.37922 0.030979
## 7  0.01020228      6   0.26700 0.37202 0.030978
## 8  0.01012268      7   0.25679 0.36824 0.032379
```

5

```
## 9  0.00929824       8    0.24667 0.36091 0.030524
## 10 0.00895071      10    0.22807 0.35937 0.030074
## 11 0.00752059      11    0.21912 0.36700 0.031158
## 12 0.00747651      12    0.21160 0.35532 0.031100
## 13 0.00679012      13    0.20413 0.35249 0.031038
## 14 0.00539238      14    0.19734 0.33590 0.029580
## 15 0.00535343      15    0.19194 0.33948 0.029734
## 16 0.00528035      16    0.18659 0.33865 0.029754
## 17 0.00424742      17    0.18131 0.33975 0.029856
## 18 0.00403991      18    0.17706 0.33702 0.029481
## 19 0.00385669      19    0.17302 0.33871 0.030115
## 20 0.00337285      20    0.16917 0.34311 0.030320
## 21 0.00299730      21    0.16579 0.34331 0.030310
## 22 0.00291331      22    0.16280 0.34470 0.030019
## 23 0.00284619      23    0.15988 0.34392 0.030034
## 24 0.00260256      24    0.15704 0.34505 0.030422
## 25 0.00253490      25    0.15443 0.34951 0.030835
## 26 0.00224677      26    0.15190 0.34731 0.030609
## 27 0.00223341      27    0.14965 0.35155 0.030701
## 28 0.00192136      28    0.14742 0.35375 0.031225
## 29 0.00168598      30    0.14358 0.35599 0.031209
## 30 0.00153869      31    0.14189 0.35476 0.030908
## 31 0.00104705      32    0.14035 0.35653 0.030971
## 32 0.00097410      34    0.13826 0.35785 0.031022
## 33 0.00073253      35    0.13728 0.36062 0.031575
## 34 0.00061604      36    0.13655 0.36022 0.031580
## 35 0.00026169      37    0.13594 0.35837 0.031582
## 36 0.00000000      38    0.13567 0.35824 0.031584
```
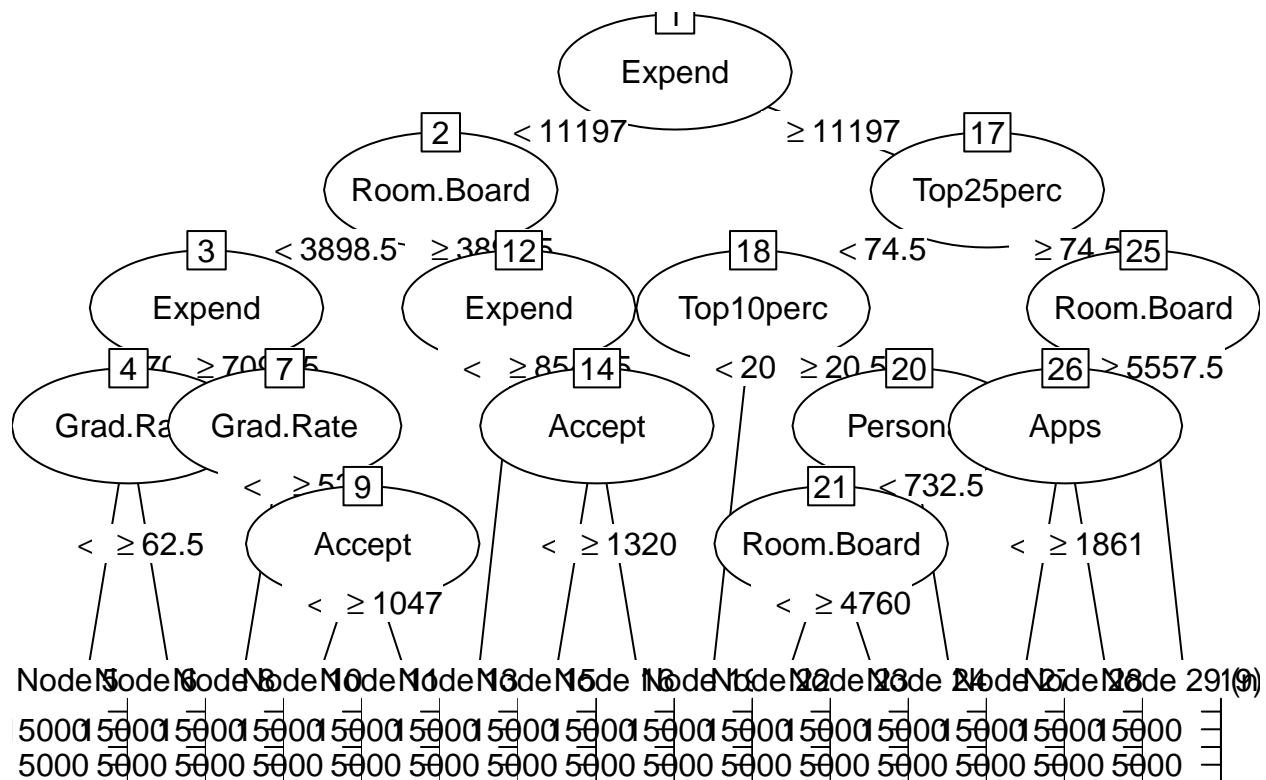
```
cpTable = tree1$cptable
plotcp(tree1)
```

size of tree

1 3 5 7 9 12   15   18   21   24   27   31   35   38

X−val Relative Error

1.2
1.0
0.8
0.6
0.4
0.2

Inf   0.027   0.0091   0.0054   0.0036   0.0026   0.0016   4e−04

cp

```r
# 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  0.517334558      0 1.0000000 1.0097670 0.06007480
## 2  0.098660468      1 0.4826654 0.5222307 0.03419201
## 3  0.043008059      2 0.3840050 0.4592716 0.03258816
## 4  0.040772836      3 0.3409969 0.4360763 0.03067484
## 5  0.018058223      4 0.3002241 0.3821769 0.03097897
## 6  0.015169632      5 0.2821659 0.3792180 0.03097928
## 7  0.010202282      6 0.2669962 0.3720175 0.03097782
## 8  0.010122676      7 0.2567939 0.3682352 0.03237898
## 9  0.009298243      8 0.2466713 0.3609138 0.03052369
## 10 0.008950709     10 0.2280748 0.3593704 0.03007410
## 11 0.007520594     11 0.2191241 0.3669969 0.03115803
## 12 0.007476511     12 0.2116035 0.3553204 0.03110010
## 13 0.006790121     13 0.2041270 0.3524918 0.03103792
## 14 0.005392380     14 0.1973368 0.3359015 0.02958012
##
## Variable importance
##     Expend  Top10perc   Terminal  Top25perc       PhD  S.F.Ratio
##         28         12         12         11        11          8
```

9

```
##  Room.Board    Grad.Rate         Apps        Accept       Enroll F.Undergrad
##           6            3            3            2            1            1
## P.Undergrad
##           1
##
## Node number 1: 452 observations,    complexity param=0.5173346
##   mean=11851.7, MSE=1.363502e+07
##   left son=2 (330 obs) right son=3 (122 obs)
##   Primary splits:
##       Expend    < 11197   to the left,  improve=0.5173346, (0 missing)
##       Terminal  < 84.5    to the left,  improve=0.3703580, (0 missing)
##       PhD       < 77.5    to the left,  improve=0.3504552, (0 missing)
##       Room.Board < 3954.5 to the left,  improve=0.3060686, (0 missing)
##       Top10perc < 35.5    to the left,  improve=0.3012348, (0 missing)
##   Surrogate splits:
##       Terminal  < 93.5    to the left,  agree=0.850, adj=0.443, (0 split)
##       Top10perc < 39.5    to the left,  agree=0.845, adj=0.426, (0 split)
##       PhD       < 89.5    to the left,  agree=0.836, adj=0.393, (0 split)
##       Top25perc < 74.5    to the left,  agree=0.825, adj=0.352, (0 split)
##       S.F.Ratio < 10.35   to the right, agree=0.812, adj=0.303, (0 split)
##
## Node number 2: 330 observations,    complexity param=0.09866047
##   mean=10236.83, MSE=6164406
##   left son=4 (129 obs) right son=5 (201 obs)
##   Primary splits:
##       Room.Board < 3898.5 to the left,  improve=0.2989044, (0 missing)
##       Expend    < 8412    to the left,  improve=0.2961452, (0 missing)
##       Grad.Rate < 64.5    to the left,  improve=0.1811026, (0 missing)
##       Terminal  < 77.5    to the left,  improve=0.1774559, (0 missing)
##       Accept    < 1240    to the left,  improve=0.1593531, (0 missing)
##   Surrogate splits:
##       Expend      < 7091.5 to the left,  agree=0.709, adj=0.256, (0 split)
##       Grad.Rate   < 54.5   to the left,  agree=0.688, adj=0.202, (0 split)
##       Accept      < 587.5  to the left,  agree=0.673, adj=0.163, (0 split)
##       Terminal    < 66.5   to the left,  agree=0.673, adj=0.163, (0 split)
##       P.Undergrad < 80.5   to the left,  agree=0.670, adj=0.155, (0 split)
##
## Node number 3: 122 observations,    complexity param=0.04300806
##   mean=16219.78, MSE=7708428
##   left son=6 (57 obs) right son=7 (65 obs)
##   Primary splits:
##       Top25perc < 74.5    to the left,  improve=0.2818503, (0 missing)
##       Apps      < 1580.5  to the left,  improve=0.2660971, (0 missing)
##       Expend    < 14711.5 to the left,  improve=0.2217110, (0 missing)
##       Top10perc < 46.5    to the left,  improve=0.2196460, (0 missing)
##       Accept    < 921.5   to the left,  improve=0.1956108, (0 missing)
##   Surrogate splits:
##       Top10perc < 39      to the left,  agree=0.918, adj=0.825, (0 split)
##       Grad.Rate < 81.5    to the left,  agree=0.779, adj=0.526, (0 split)
##       Expend    < 13214   to the left,  agree=0.746, adj=0.456, (0 split)
##       PhD       < 88.5    to the left,  agree=0.721, adj=0.404, (0 split)
##       Apps      < 1348.5  to the left,  agree=0.713, adj=0.386, (0 split)
##
## Node number 4: 129 observations,    complexity param=0.01805822
```

```
##    mean=8542.434, MSE=4865784
##    left son=8 (60 obs) right son=9 (69 obs)
##    Primary splits:
##        Expend     < 7098.5  to the left,  improve=0.1773074, (0 missing)
##        Grad.Rate  < 63.5    to the left,  improve=0.1288581, (0 missing)
##        perc.alumni < 18.5   to the left,  improve=0.1265462, (0 missing)
##        S.F.Ratio  < 12.95   to the right, improve=0.1212700, (0 missing)
##        Room.Board < 3529    to the left,  improve=0.1159078, (0 missing)
##    Surrogate splits:
##        S.F.Ratio  < 13.55   to the right, agree=0.752, adj=0.467, (0 split)
##        Room.Board < 3252.5  to the left,  agree=0.729, adj=0.417, (0 split)
##        Top10perc  < 21.5    to the left,  agree=0.643, adj=0.233, (0 split)
##        P.Undergrad < 171    to the right, agree=0.643, adj=0.233, (0 split)
##        Top25perc  < 40.5    to the left,  agree=0.620, adj=0.183, (0 split)
##
## Node number 5: 201 observations,    complexity param=0.04077284
##    mean=11324.28, MSE=3972738
##    left son=10 (98 obs) right son=11 (103 obs)
##    Primary splits:
##        Expend   < 8545.5  to the left,  improve=0.3146874, (0 missing)
##        Accept   < 1320    to the left,  improve=0.1719002, (0 missing)
##        Terminal < 77.5    to the left,  improve=0.1714990, (0 missing)
##        Apps     < 1190.5  to the left,  improve=0.1673124, (0 missing)
##        PhD      < 75.5    to the left,  improve=0.1457601, (0 missing)
##    Surrogate splits:
##        PhD      < 71.5    to the left,  agree=0.726, adj=0.439, (0 split)
##        Terminal < 72.5    to the left,  agree=0.706, adj=0.398, (0 split)
##        Top25perc < 49.5   to the left,  agree=0.637, adj=0.255, (0 split)
##        S.F.Ratio < 14.05  to the right, agree=0.632, adj=0.245, (0 split)
##        Apps     < 1654    to the left,  agree=0.622, adj=0.224, (0 split)
##
## Node number 6: 57 observations,    complexity param=0.01012268
##    mean=14645.75, MSE=6939037
##    left son=12 (7 obs) right son=13 (50 obs)
##    Primary splits:
##        Top10perc   < 20.5   to the left,  improve=0.15773050, (0 missing)
##        Apps        < 953.5  to the left,  improve=0.15494790, (0 missing)
##        Personal    < 732.5  to the right, improve=0.12117480, (0 missing)
##        F.Undergrad < 1138.5 to the left,  improve=0.10723830, (0 missing)
##        Accept      < 836.5  to the left,  improve=0.09989591, (0 missing)
##    Surrogate splits:
##        Top25perc   < 42     to the left,  agree=0.982, adj=0.857, (0 split)
##        Apps        < 452.5  to the left,  agree=0.912, adj=0.286, (0 split)
##        F.Undergrad < 354.5  to the left,  agree=0.912, adj=0.286, (0 split)
##        Accept      < 250.5  to the left,  agree=0.895, adj=0.143, (0 split)
##        Enroll      < 108.5  to the left,  agree=0.895, adj=0.143, (0 split)
##
## Node number 7: 65 observations,    complexity param=0.01516963
##    mean=17600.08, MSE=4305278
##    left son=14 (27 obs) right son=15 (38 obs)
##    Primary splits:
##        Room.Board < 5557.5 to the left,  improve=0.3340833, (0 missing)
##        Apps       < 1762   to the left,  improve=0.3000966, (0 missing)
##        Accept     < 1170.5 to the left,  improve=0.1728499, (0 missing)
```

```
##       Expend     < 14711.5 to the left,  improve=0.1721922, (0 missing)
##       Terminal   < 90.5    to the left,  improve=0.1587814, (0 missing)
##   Surrogate splits:
##       Expend      < 14746.5 to the left,  agree=0.754, adj=0.407, (0 split)
##       Apps        < 2771    to the left,  agree=0.738, adj=0.370, (0 split)
##       Terminal    < 93.5    to the left,  agree=0.738, adj=0.370, (0 split)
##       Enroll      < 507.5   to the left,  agree=0.723, adj=0.333, (0 split)
##       F.Undergrad < 1713.5  to the left,  agree=0.708, adj=0.296, (0 split)
##
## Node number 8: 60 observations,    complexity param=0.006790121
##   mean=7546.367, MSE=3605279
##   left son=16 (35 obs) right son=17 (25 obs)
##   Primary splits:
##       Grad.Rate   < 62.5    to the left,  improve=0.1934558, (0 missing)
##       S.F.Ratio   < 11.55   to the right, improve=0.1778754, (0 missing)
##       Books       < 495     to the right, improve=0.1630936, (0 missing)
##       Apps        < 879     to the right, improve=0.1575230, (0 missing)
##       perc.alumni < 18.5    to the left,  improve=0.1403775, (0 missing)
##   Surrogate splits:
##       F.Undergrad < 1186    to the left,  agree=0.700, adj=0.28, (0 split)
##       Accept      < 568.5   to the left,  agree=0.683, adj=0.24, (0 split)
##       Top25perc   < 57.5    to the left,  agree=0.683, adj=0.24, (0 split)
##       Enroll      < 557.5   to the left,  agree=0.667, adj=0.20, (0 split)
##       S.F.Ratio   < 11.55   to the right, agree=0.667, adj=0.20, (0 split)
##
## Node number 9: 69 observations,    complexity param=0.008950709
##   mean=9408.58, MSE=4348928
##   left son=18 (23 obs) right son=19 (46 obs)
##   Primary splits:
##       Grad.Rate   < 52.5    to the left,  improve=0.1838318, (0 missing)
##       F.Undergrad < 642     to the left,  improve=0.1155065, (0 missing)
##       Top25perc   < 47.5    to the left,  improve=0.1142047, (0 missing)
##       Top10perc   < 33.5    to the left,  improve=0.1035070, (0 missing)
##       Accept      < 1047    to the left,  improve=0.1013709, (0 missing)
##   Surrogate splits:
##       Top10perc   < 15.5    to the left,  agree=0.768, adj=0.304, (0 split)
##       Top25perc   < 41.5    to the left,  agree=0.768, adj=0.304, (0 split)
##       Enroll      < 106.5   to the left,  agree=0.739, adj=0.217, (0 split)
##       F.Undergrad < 394     to the left,  agree=0.725, adj=0.174, (0 split)
##       Terminal    < 50.5    to the left,  agree=0.725, adj=0.174, (0 split)
##
## Node number 10: 98 observations
##   mean=10178, MSE=2371877
##
## Node number 11: 103 observations,    complexity param=0.01020228
##   mean=12414.91, MSE=3056234
##   left son=22 (59 obs) right son=23 (44 obs)
##   Primary splits:
##       Accept   < 1320    to the left,  improve=0.1997413, (0 missing)
##       Apps     < 1158.5  to the left,  improve=0.1791167, (0 missing)
##       Enroll   < 324.5   to the left,  improve=0.1712155, (0 missing)
##       Expend   < 10712   to the left,  improve=0.1317385, (0 missing)
##       Terminal < 82.5    to the left,  improve=0.1251798, (0 missing)
##   Surrogate splits:
```

```
##       Apps        < 1634.5   to the left,   agree=0.981, adj=0.955, (0 split)
##       Enroll      < 446.5    to the left,   agree=0.932, adj=0.841, (0 split)
##       F.Undergrad < 1814.5   to the left,   agree=0.903, adj=0.773, (0 split)
##       PhD         < 80.5     to the left,   agree=0.718, adj=0.341, (0 split)
##       P.Undergrad < 794.5    to the left,   agree=0.699, adj=0.295, (0 split)
##
## Node number 12: 7 observations
##   mean=11849.71, MSE=3622916
##
## Node number 13: 50 observations,    complexity param=0.009298243
##   mean=15037.2, MSE=6155567
##   left son=26 (36 obs) right son=27 (14 obs)
##   Primary splits:
##       Personal    < 732.5    to the right, improve=0.1473277, (0 missing)
##       Apps        < 953.5    to the left,  improve=0.1325652, (0 missing)
##       Room.Board  < 5547.5   to the left,  improve=0.1317621, (0 missing)
##       Top10perc   < 25.5     to the right, improve=0.1246337, (0 missing)
##       Top25perc   < 53.5     to the right, improve=0.1154867, (0 missing)
##   Surrogate splits:
##       P.Undergrad < 14.5     to the right, agree=0.78, adj=0.214, (0 split)
##       perc.alumni < 42       to the left,  agree=0.78, adj=0.214, (0 split)
##       Books       < 362.5    to the right, agree=0.76, adj=0.143, (0 split)
##       PhD         < 96       to the left,  agree=0.76, adj=0.143, (0 split)
##       Apps        < 594.5    to the right, agree=0.74, adj=0.071, (0 split)
##
## Node number 14: 27 observations,    complexity param=0.007520594
##   mean=16177.3, MSE=5888431
##   left son=28 (8 obs) right son=29 (19 obs)
##   Primary splits:
##       Apps        < 1861     to the left,  improve=0.2915299, (0 missing)
##       Accept      < 1412.5   to the left,  improve=0.1825446, (0 missing)
##       P.Undergrad < 51       to the right, improve=0.1739010, (0 missing)
##       Grad.Rate   < 90.5     to the left,  improve=0.1272623, (0 missing)
##       PhD         < 89.5     to the left,  improve=0.1124694, (0 missing)
##   Surrogate splits:
##       Enroll      < 326.5    to the left,  agree=0.926, adj=0.750, (0 split)
##       Accept      < 938      to the left,  agree=0.889, adj=0.625, (0 split)
##       F.Undergrad < 1165     to the left,  agree=0.889, adj=0.625, (0 split)
##       Terminal    < 85.5     to the left,  agree=0.815, adj=0.375, (0 split)
##       Grad.Rate   < 62.5     to the left,  agree=0.815, adj=0.375, (0 split)
##
## Node number 15: 38 observations
##   mean=18611, MSE=720119.6
##
## Node number 16: 35 observations
##   mean=6840.543, MSE=2041590
##
## Node number 17: 25 observations
##   mean=8534.52, MSE=4120534
##
## Node number 18: 23 observations
##   mean=8144.087, MSE=3070591
##
## Node number 19: 46 observations,    complexity param=0.007476511
```

```
##    mean=10040.83, MSE=3788889
##    left son=38 (36 obs) right son=39 (10 obs)
##    Primary splits:
##        Accept      < 1047    to the left,  improve=0.2643769, (0 missing)
##        F.Undergrad < 1311    to the left,  improve=0.2618172, (0 missing)
##        Enroll      < 440.5   to the left,  improve=0.2270294, (0 missing)
##        Apps        < 1263    to the left,  improve=0.2213238, (0 missing)
##        Top10perc   < 33.5    to the left,  improve=0.2138226, (0 missing)
##    Surrogate splits:
##        Apps        < 1216.5  to the left,  agree=0.957, adj=0.8, (0 split)
##        Enroll      < 440.5   to the left,  agree=0.935, adj=0.7, (0 split)
##        F.Undergrad < 1352.5  to the left,  agree=0.891, adj=0.5, (0 split)
##        S.F.Ratio   < 16.3    to the left,  agree=0.848, adj=0.3, (0 split)
##        P.Undergrad < 635.5   to the left,  agree=0.826, adj=0.2, (0 split)
##
## Node number 22: 59 observations
##    mean=11740.19, MSE=2538318
##
## Node number 23: 44 observations
##    mean=13319.66, MSE=2321690
##
## Node number 26: 36 observations,    complexity param=0.009298243
##    mean=14443.33, MSE=5665795
##    left son=52 (12 obs) right son=53 (24 obs)
##    Primary splits:
##        Room.Board  < 4760    to the left,  improve=0.3395937, (0 missing)
##        Apps        < 953.5   to the left,  improve=0.2409709, (0 missing)
##        F.Undergrad < 1138.5  to the left,  improve=0.2037114, (0 missing)
##        Books       < 588     to the left,  improve=0.1489560, (0 missing)
##        Accept      < 758     to the left,  improve=0.1302411, (0 missing)
##    Surrogate splits:
##        Top25perc   < 71.5    to the right, agree=0.778, adj=0.333, (0 split)
##        Apps        < 953.5   to the left,  agree=0.750, adj=0.250, (0 split)
##        Top10perc   < 43      to the right, agree=0.722, adj=0.167, (0 split)
##        F.Undergrad < 1138.5  to the left,  agree=0.722, adj=0.167, (0 split)
##        P.Undergrad < 26.5    to the left,  agree=0.722, adj=0.167, (0 split)
##
## Node number 27: 14 observations
##    mean=16564.29, MSE=4176103
##
## Node number 28: 8 observations
##    mean=14158.12, MSE=2121575
##
## Node number 29: 19 observations
##    mean=17027.47, MSE=5035020
##
## Node number 38: 36 observations
##    mean=9513.333, MSE=2856555
##
## Node number 39: 10 observations
##    mean=11939.8, MSE=2537495
##
## Node number 52: 12 observations
##    mean=12481.67, MSE=4962777
```

```
## 
## Node number 53: 24 observations
##   mean=15424.17, MSE=3131202
```

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).