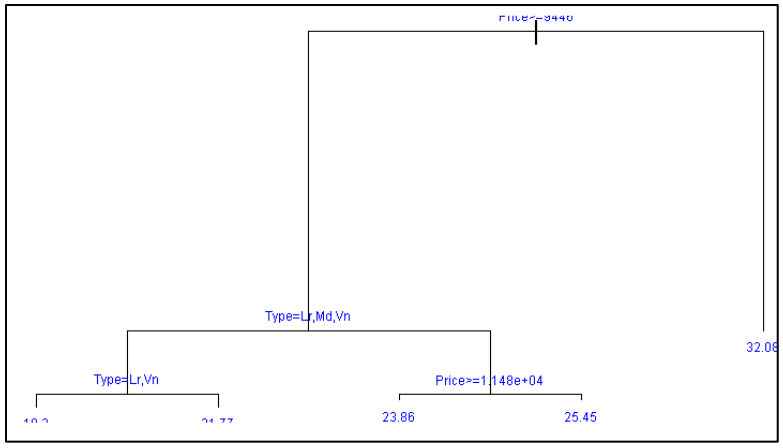
CS5900/STAT 46700 Topics in Data Science Spring 2025

Lab 8

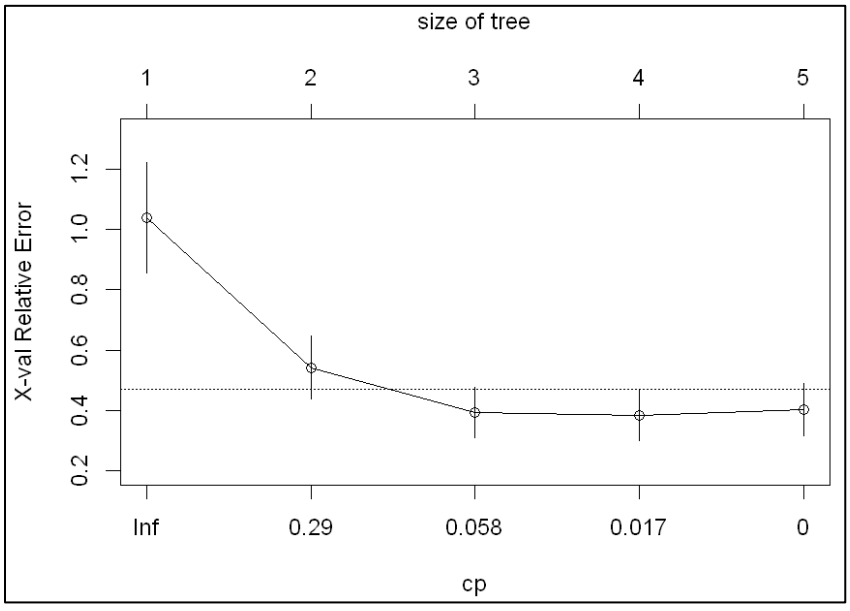
[Vaishak Balachandra – 037831852]

1. Data on makes of cars taken from the April, 1990 issue of Consumer Reports are provided in `cu.summary` in `rpart` library**.**
   a. Access the data and print the names of the variables included in the dataset.
   b. Construct the regression tree to model mileage using Price, Country, Reliability and Type.

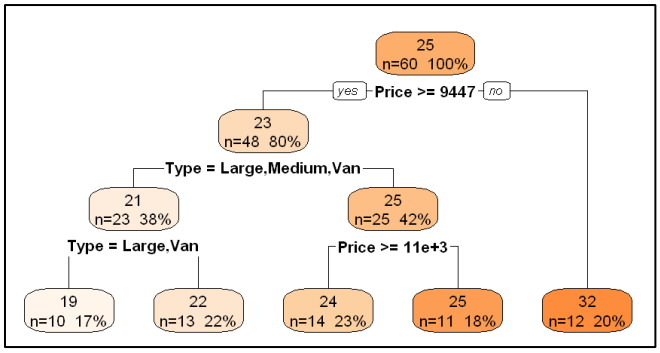
```
> # 1
>
> # a
>
> install.packages("rpart")
Error in install.packages : Updating loaded packages
> install.packages("rpart")
> library(rpart)
> data(cu.summary, package = "rpart")
> data = cu.summary
> head(data)
               Price Country Reliability Mileage  Type
Acura Integra 4 11950   Japan Much better      NA Small
Dodge Colt 4     6851   Japan        <NA>      NA Small
Dodge Omni 4     6995     USA  Much worse      NA Small
Eagle Summit 4   8895     USA      better      33 Small
Ford Escort   4  7402     USA       worse      33 Small
Ford Festiva 4   6319   Korea      better      37 Small
> names(data)
[1] "Price"       "Country"     "Reliability" "Mileage"     "Type"
> attach(data)
>
>
> # b
> install.packages("ISLR")
> library(ISLR)
> install.packages("tree")
> library(tree)
> install.packages("rpart.plot")
> library(rpart.plot)
> tree_model = rpart(Mileage ~ Price + Country + Reliability + Type, data = data, method = "anova")
> tree_model
n=60 (57 observations deleted due to missingness)

node), split, n, deviance, yval
      * denotes terminal node

 1) root 60 1354.58300 24.58333
   2) Price>=9446.5 48  407.91670 22.70833
     4) Type=Large,Medium,Van 23   66.86957 20.69565
       8) Type=Large,Van 10   22.10000 19.30000 *
       9) Type=Medium 13   10.30769 21.76923 *
     5) Type=Compact,Small,Sporty 25  162.16000 24.56000
      10) Price>=11484.5 14  107.71430 23.85714 *
      11) Price< 11484.5 11   38.72727 25.45455 *
   3) Price< 9446.5 12  102.91670 32.08333 *

> cat("Thus, there are 5 terminals in tree.")
Thus, there are 5 terminals in tree.
> plot(tree_model)
```




```
> text(tree_model, col = "blue", minlength = 2L, cex = 0.5)
> printcp(tree_model)

Regression tree:
rpart(formula = Mileage ~ Price + Country + Reliability + Type,
    data = data, method = "anova", cp = 0)

Variables actually used in tree construction:
[1] Price Type

Root node error: 1354.6/60 = 22.576

n=60 (57 observations deleted due to missingness)

        CP nsplit rel error  xerror     xstd
1 0.622885      0   1.00000 1.03978 0.183094
2 0.132061      1   0.37711 0.54223 0.104312
3 0.025441      2   0.24505 0.39395 0.084164
4 0.011604      3   0.21961 0.38466 0.085717
5 0.000000      4   0.20801 0.40367 0.086170
> plotcp(tree_model)
```



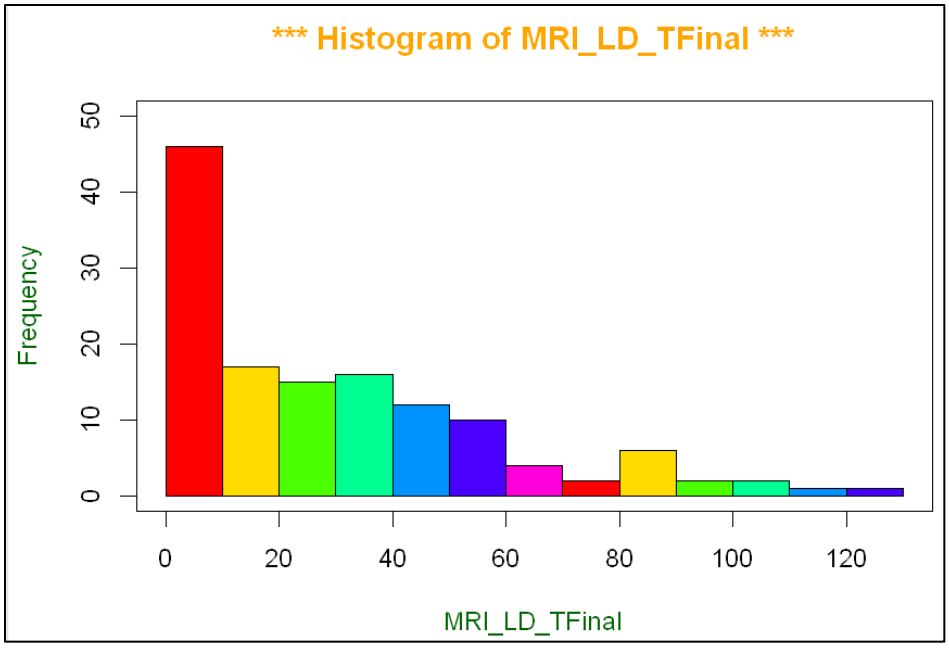

```
> ?rpart.plot
> rpart.plot(tree_model, type = 2, extra = 101, tweak = 1, box.palette = "Orange")
```



2. Use the data provided with this assignment
   (a) Read in the breast cancer imaging data "ispy1doctored.csv" into a data frame called dat
   (b) Generate a histogram of the MRI_LD_Tfinal variable that will be our outcome to predict
   (c) Split the dataset into a training set of size 70 and a test set consisting of the remaining data
   (d) Fit a regression tree to the training data with MRI_LD_Tfinal as outcome and all other variables as candidate predictors. Make sure that you specify the correct method for regression
   (e) Plot the fitted tree and add text labels

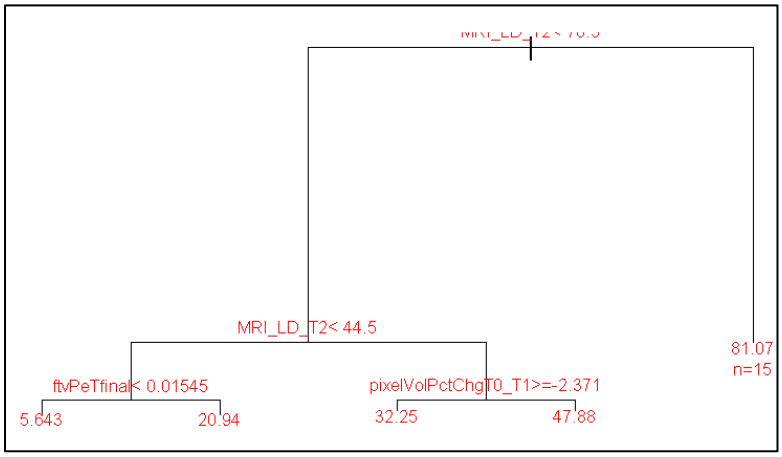
```
> # 2
>
> # a
> dat <- read.csv("C:/Users/PNW_checkout/Downloads/sem 2/0. Coursework/0. Coursework/Data science/L
ab/Lab 8/ispy1doctored.csv")
> head(dat)
   pixelVolT0  pixelVolT1  pixelVolT2 pixelVolTfinal pixelVolPctChgT0_T1   ftvPeT0     ftvPeT1
1 0.001220703 0.001220703 0.001220703    0.001220703                   0 21.156006  0.01098633
2 0.001220703 0.001220703 0.001220703    0.001220703                   0  5.310059  5.13061520
3 0.001220703 0.001220703 0.001220703    0.001220703                   0 13.099365  9.20898440
4 0.000741577 0.000741577 0.000988770    0.000741577                   0 42.757855 53.58117400
5 0.001220703 0.001220703 0.001220703    0.001220703                   0  3.208008  1.89941410
6 0.001220703 0.001220703 0.001220703    0.001220703                   0 25.770264 11.01318400
     ftvPeT2 ftvPeTfinal ftvPePctChgT0_T1   age race HR_HER2status MRI_LD_T0 MRI_LD_T1
1  0.0000000   0.0000000        -99.94807 38.73    1  HRposHER2neg        88        78
2  3.3630371   1.2145996         -3.37931 37.79    1  HRposHER2neg        29        26
3  5.2856445   1.9055176        -29.69900 49.83    1  HRposHER2neg        50        64
4 73.6119140  12.2286070         25.31305 48.28    1     TripleNeg        91        90
5  2.0129395   0.6811523        -40.79148 64.51    1  HRposHER2neg        45        49
6  0.8093262   0.1074219        -57.26399 40.66    4     TripleNeg        75        66
  MRI_LD_T2 MRI_LD_Tfinal
1        30            14
2        66            16
3        54            46
4        99            43
5        47            32
6        57             7
```

```
> dim(dat)
[1] 134  17
> names(dat)
 [1] "pixelVolT0"           "pixelVolT1"           "pixelVolT2"           "pixelVolTfinal"
 [5] "pixelVolPctChgT0_T1" "ftvPeT0"              "ftvPeT1"              "ftvPeT2"
 [9] "ftvPeTfinal"          "ftvPePctChgT0_T1"     "age"                  "race"
[13] "HR_HER2status"        "MRI_LD_T0"            "MRI_LD_T1"            "MRI_LD_T2"
[17] "MRI_LD_Tfinal"
>
>
> # b
> hist(dat$MRI_LD_Tfinal, main = "*** Histogram of MRI_LD_TFinal ***", col = rainbow(7), xlab = "MR
I_LD_TFinal", col.main = "orange", col.lab = "darkgreen", ylim = c(0,50))
> box()
```

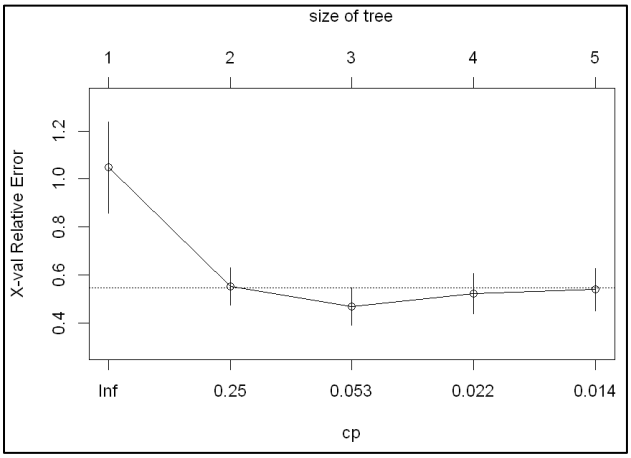



```
> # c
> set.seed(0037831852)
> index <- sample(1:nrow(dat), 70)
> train_dat <- dat[index,]
> test_dat <- dat[-index,]
> dim(dat)
[1] 134  17
> dim(train_dat)
[1] 70 17
> dim(test_dat)
[1] 64 17
>
> # d
> install.packages("rpart")
> library(rpart)
> install.packages("rpart.plot")
> library(rpart.plot)
> model_tree <- rpart(train_dat$MRI_LD_Tfinal~., data = train_dat, method = "anova")
> model_tree
n= 70
node), split, n, deviance, yval
      * denotes terminal node
 1) root 70 68325.1400 36.428570
   2) MRI_LD_T2< 76.5 55 18668.4400 24.254550
     4) MRI_LD_T2< 44.5 31  5416.9680 14.032260
       8) ftvPeTfinal< 0.01544952 14   669.2143  5.642857 *
       9) ftvPeTfinal>=0.01544952 17  2950.9410 20.941180 *
```

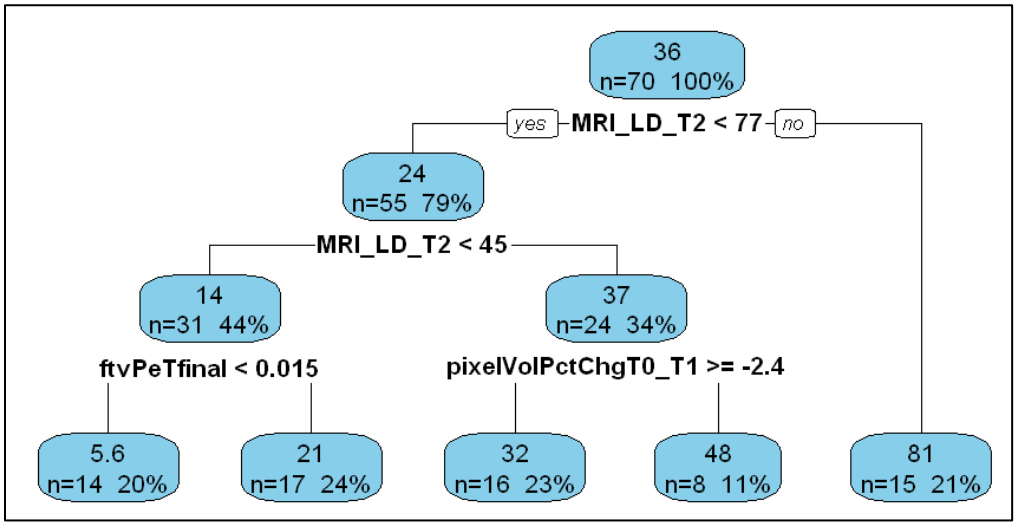
```
     5) MRI_LD_T2>=44.5 24  5827.9580 37.458330
      10) pixelVolPctChgT0_T1>=-2.371395 16  4149.0000 32.250000 *
      11) pixelVolPctChgT0_T1< -2.371395 8   376.8750 47.875000 *
   3) MRI_LD_T2>=76.5 15 11616.9300 81.066670 *
>
> # e
> plot(model_tree)
> text(model_tree, use.n = TRUE, cex = 0.7, col = "red")
```



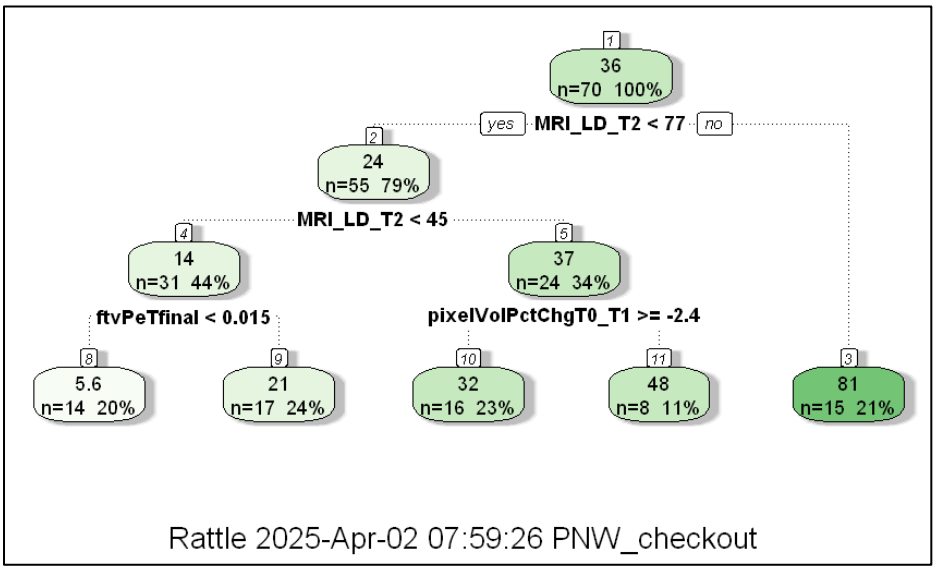

```
> printcp(model_tree)

Regression tree:
rpart(formula = train_dat$MRI_LD_Tfinal ~ ., data = train_dat,
    method = "anova")

Variables actually used in tree construction:
[1] ftvPeTfinal         MRI_LD_T2           pixelVolPctChgT0_T1

Root node error: 68325/70 = 976.07

n= 70

        CP nsplit rel error  xerror     xstd
1 0.556746      0   1.00000 1.04810 0.189792
2 0.108650      1   0.44325 0.55169 0.077558
3 0.026298      2   0.33460 0.46860 0.078385
4 0.019057      3   0.30831 0.52292 0.084729
5 0.010000      4   0.28925 0.54008 0.088212
> plotcp(model_tree)
```



```
> # or
> rpart.plot(model_tree, type = 2, extra = 101, tweak = 1, box.palette = "skyblue")
```




```
> install.packages("rattle")
> library(rattle)
> fancyRpartPlot(model_tree)
> ?fancyRpartPlot
```




```
> # extra
> predicted = predict(model_tree, test_dat, type = "vector")
> # print(predicted)
> actual = test_dat$MRI_LD_Tfinal
>
>
> MAE = sum(abs(predicted - actual))/length(actual)
> MAE
[1] 13.18504
> # VALUE of MAE can be analysed, only by comparison
```