Submitted by:

Aysha Emelda C P

School of Management Studies, University of Hyderabad

TASK 2

Refer to the dataset “Tree Making”.

Analyze the dataset and analyze the nature of the variables.

Find out the course of action using a suitable decision tree.

****

Analysis of Dataset and the Nature of the Variables

It is given in the dataset ‘Tree Making’ that,

‘This dataset contains information collected by the US Census Service concerning housing in the area of Boston Massachusetts. It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston). The dataset has 506 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. ‘

There are 14 attributes in each case of the dataset. They are:

CRIM per capita crime rate by town

ZN proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS proportion of non-retail business acres per town.

CHAS Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX nitric oxides concentration (parts per 10 million)

RM average number of rooms per dwelling

AGE proportion of owner-occupied units built prior to 1940

DIS weighted distances to five Boston employment centres

RAD index of accessibility to radial highways

TAX full-value property-tax rate per $10,000

PTRATIO pupil-teacher ratio by town

B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT % lower status of the population

MEDV Median value of owner-occupied homes in $1000 ‘

Use the `head()` function to view the first 6 observations in the dataset.

```
library(readxl)

Tree_making <- read_excel("C:/Users/Aysha
Emelda/Downloads/Tree Making.xlsx")

View(Tree_making)

head(Tree_making)
```

```
37  head(Tree_making)
38  ```
```

CRIM <dbl>	ZN <dbl>	INDUS <dbl>	CHAS <dbl>	NOX <dbl>	RM <dbl>	AGE <dbl>	DIS <dbl>	RAD <dbl>	TAX <dbl>
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222

6 rows | 1-10 of 15 columns

The column B is not considered like other variables in the tree making because of the racial discrimination the usage of such a variable can cause.

The structure of 'Tree Making' dataset is obtained with the `str()` function.

```
str(Tree_making)
```

```
```{r}
str(Tree_making)
```

tibble [506 x 15] (S3: tbl_df/tbl/data.frame)
 $ CRIM   : num [1:506] 0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN   : num [1:506] 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS   : num [1:506] 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : num [1:506] 0 0 0 0 0 0 0 0 0 0 ...
 $ NOX   : num [1:506] 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524
...
 $ RM   : num [1:506] 6.58 6.42 7.18 7 7.15 ...
 $ AGE   : num [1:506] 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS   : num [1:506] 4.09 4.97 4.97 6.06 6.06 ...
 $ RAD   : num [1:506] 1 2 2 3 3 3 5 5 5 5 ...
 $ TAX   : num [1:506] 296 242 242 222 222 222 311 311 311 311 ...
 $ PTRATIO : num [1:506] 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ B   : num [1:506] 397 397 393 395 397 ...
 $ LSTAT   : num [1:506] 4.98 9.14 4.03 2.94 5.33 ...
 $ MEDV   : num [1:506] 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
 $ CAT. MEDV: num [1:506] 0 0 1 1 1 0 0 0 0 0 ...
```

[506 x 15] indicates that there are 506 rows and 15 columns.

It is checked whether there is any missing value using the `is.na`() function.

```
is.na('Tree_making')
```

```
is.na('Tree_making')
```

[1] FALSE
```

Therefore, missing values are not there. All variables are numeric.
```

The matrices if coefficients of correlation and p-values of each relations can be obtained through the following steps.

```
data("Tree_making")

my_data <- Tree_making[,
c(1,2,3,4,5,6,7,8,9,10,11,13,14)]

install.packages("Hmisc")

library("Hmisc")

res2 <- rcorr(as.matrix(my_data))

res2
```

```
        CRIM    ZN INDUS  CHAS   NOX    RM   AGE   DIS   RAD   TAX PTRATIO LSTAT
CRIM     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29  0.46
ZN      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39 -0.41
INDUS    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38  0.60
CHAS    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12 -0.05
NOX      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19  0.59
RM      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36 -0.61
AGE      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26  0.60
DIS     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23 -0.50
RAD      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46  0.49
TAX      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46  0.54
PTRATIO  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00  0.37
LSTAT    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37  1.00
MEDV    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51 -0.74
         MEDV
CRIM    -0.39
ZN       0.36
INDUS   -0.48
CHAS     0.18
NOX     -0.43
RM       0.70
AGE     -0.38
DIS      0.25
RAD     -0.38
TAX     -0.47
PTRATIO -0.51
LSTAT   -0.74
MEDV     1.00

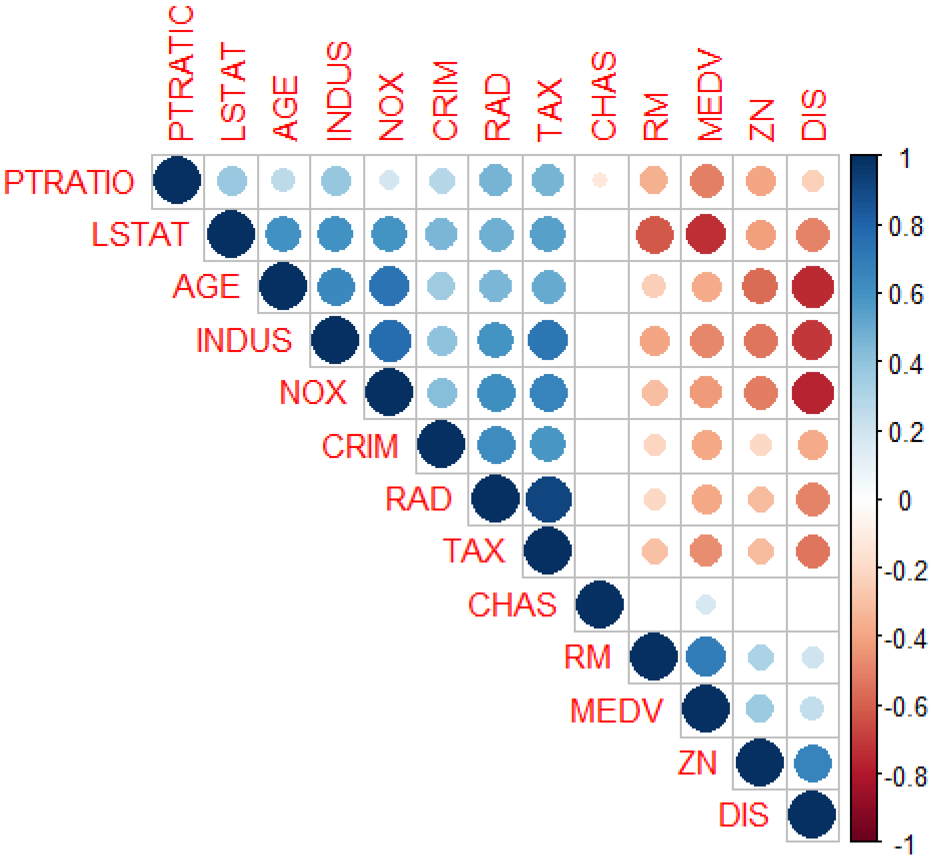
n= 506
```

```
P
        CRIM   ZN     INDUS  CHAS   NOX    RM     AGE    DIS    RAD    TAX   
CRIM           0.0000 0.0000 0.2094 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
ZN      0.0000        0.0000 0.3378 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
INDUS   0.0000 0.0000        0.1575 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
CHAS    0.2094 0.3378 0.1575        0.0403 0.0402 0.0518 0.0257 0.8687 0.4244
NOX     0.0000 0.0000 0.0000 0.0403        0.0000 0.0000 0.0000 0.0000 0.0000
RM      0.0000 0.0000 0.0000 0.0402 0.0000        0.0000 0.0000 0.0000 0.0000
AGE     0.0000 0.0000 0.0000 0.0518 0.0000 0.0000        0.0000 0.0000 0.0000
DIS     0.0000 0.0000 0.0000 0.0257 0.0000 0.0000 0.0000        0.0000 0.0000
RAD     0.0000 0.0000 0.0000 0.8687 0.0000 0.0000 0.0000 0.0000        0.0000
TAX     0.0000 0.0000 0.0000 0.4244 0.0000 0.0000 0.0000 0.0000 0.0000       
PTRATIO 0.0000 0.0000 0.0000 0.0062 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
LSTAT   0.0000 0.0000 0.0000 0.2259 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
MEDV    0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
        PTRATIO LSTAT  MEDV  
CRIM    0.0000  0.0000 0.0000
ZN      0.0000  0.0000 0.0000
INDUS   0.0000  0.0000 0.0000
CHAS    0.0062  0.2259 0.0000
NOX     0.0000  0.0000 0.0000
RM      0.0000  0.0000 0.0000
AGE     0.0000  0.0000 0.0000
DIS     0.0000  0.0000 0.0000
RAD     0.0000  0.0000 0.0000
TAX     0.0000  0.0000 0.0000
PTRATIO         0.0000 0.0000
LSTAT   0.0000         0.0000
MEDV    0.0000  0.0000       
```

Here is a correlogram with data from res2.

```
corrplot(res2$r, type="upper", order="hclust",
```

```
         p.mat = res2$P, sig.level = 0.01, insig =
"blank")
corrplot(res2$r, type="upper", order="hclust",
         p.mat = res2$P, sig.level = 0.01, insig =
"blank")
```

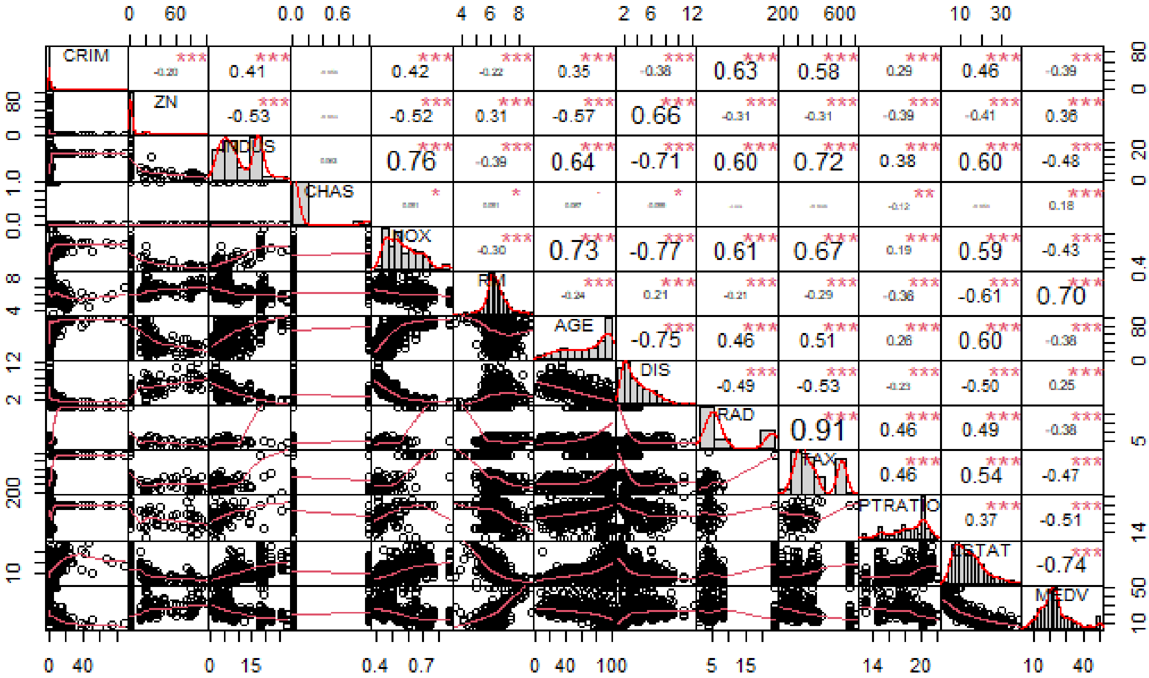


Here is a correlation plot with histograms, density functions, smoothed regression lines and correlation coefficients with the corresponding significance levels.

```
install.packages("PerformanceAnalytics")

library(PerformanceAnalytics)

chart.Correlation(my_data, histogram = TRUE, method =
"pearson")
```



We get good impression about the data from correlation matrix and plot which have provided us with p values, coefficients of correlation, significances, relevant scatter plots and histograms. We can identify correlations which are strong eough to lead to collinearity. Identifying these help in making linear models and variable selection. CHAS, DIS, ZN, and MEDV are positively correlated with RM. CRIM, INDUS, NOX,AGE, RAD,PTRATIO and LSTAT are positively correlated with TAX. RAD is strongly correlated with TAX indicated by a coefficient of correlation of value 0.91. This property implies that as the accessibility to radial highways increased, the full value property tax rate per 1000 $ also increased. CRIM, INDUS, NOX, AGE, RAD, TAX, PTRATIO, LSTAT are in negative correlation with MEDV. LSTAT is negatively correlated with MEDV with a considerable strength indicated by a coefficient of correlation, -0.74.

Using multiple linear regression modelling, the features of relationship all variables other than B is found here.

```
regressor<-lm(formula = MEDV ~ ., data = my_data)

summary(regressor)
```

```
Call:
lm(formula = MEDV ~ ., data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1304  -2.7673  -0.5814   1.9414  26.2526 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
CRIM         -0.121389   0.033000  -3.678 0.000261 ***
ZN            0.046963   0.013879   3.384 0.000772 ***
INDUS         0.013468   0.062145   0.217 0.828520    
CHAS          2.839993   0.870007   3.264 0.001173 ** 
NOX         -18.758022   3.851355  -4.870 1.50e-06 ***
RM            3.658119   0.420246   8.705  < 2e-16 ***
AGE           0.003611   0.013329   0.271 0.786595    
DIS          -1.490754   0.201623  -7.394 6.17e-13 ***
RAD           0.289405   0.066908   4.325 1.84e-05 ***
TAX          -0.012682   0.003801  -3.337 0.000912 ***
PTRATIO      -0.937533   0.132206  -7.091 4.63e-12 ***
LSTAT        -0.552019   0.050659 -10.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.798 on 493 degrees of freedom
Multiple R-squared:  0.7343,	Adjusted R-squared:  0.7278 
F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

In this situation, the variables AGE and INDUS are insignificant to the model as their p-value is greater than 0.05. This linear model is significant to explain the variation in dependent variable as the p-value of the entire model is less than 2.2e-16 which is also less than 0.05 which is considered the threshold value.

****

Course of Action Using a Decision Tree

A decision tree is made in order to predict the dependent variable here, MEDV.

```
library(readxl)

Tree_making <- read_excel("C:/Users/Aysha
Emelda/Downloads/Tree Making.xlsx")

View(Tree_making)

install.packages("tree")

library(tree)

tree1 <- tree(MEDV ~
CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+LST
AT, data = Tree_making)

summary(tree1)
```

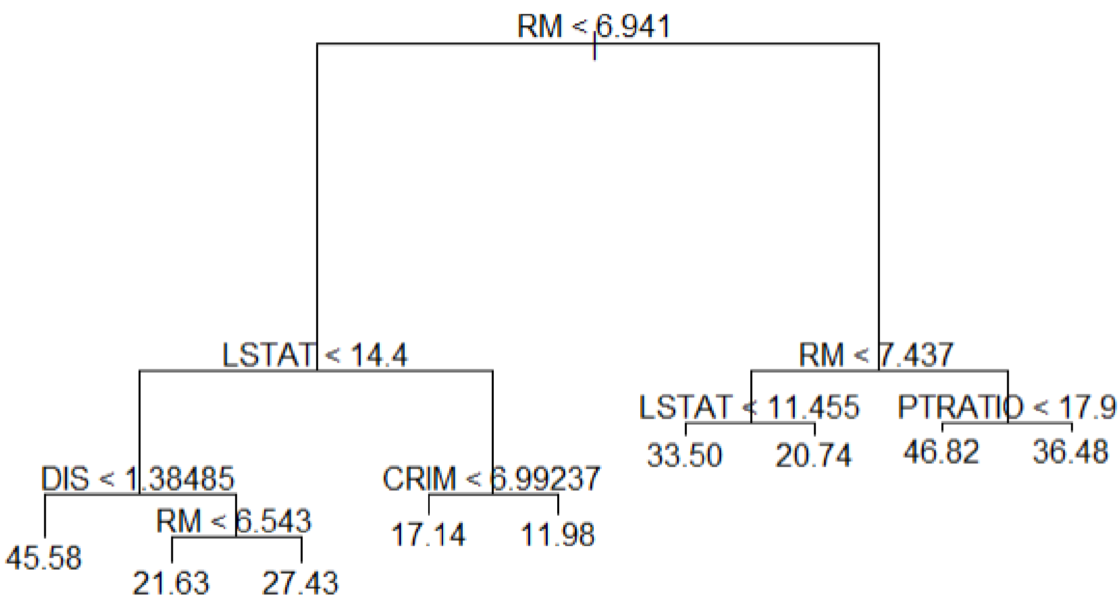
```
Regression tree:
tree(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
    DIS + RAD + TAX + PTRATIO + LSTAT, data = Tree_making)
Variables actually used in tree construction:
[1] "RM"      "LSTAT"   "DIS"     "CRIM"    "PTRATIO"
Number of terminal nodes:  9
Residual mean deviance:  13.55 = 6734 / 497
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-17.68000  -2.23000   0.07026   0.00000   2.22100  16.50000
```

A regression tree is the suitable decision tree for the continuous data we have. The dataset is repeatedly broken into smaller subsets. As a result the decision tree has grown. R algorithm selected the suitable variables RM, LSTAT, DIS, CRIM and PTRATIO for the construction of the regression tree.

Here is the decision tree.

```
plot(tree1)

text(tree1)
```



Let us look into how MEDV is predicted using this regression tree model. Observations with the value of RM less than 6.941 will go to the left of the tree, where the data is again split on the basis of the value of LSTAT of those particular observations. The remaining values from the first split, those with RM greater than 6.941 will go to the node on the right of the tree. There the data is split again based on the value of RM of those particular observations.

Data from first split with LSTAT less than 14.4 is sent to left where the data is split based on DIS. If DIS is less than 1.38485, the predicted MEDV is 45.58. Here if DIS is greater than 1.38485, it should be again split on the basis if RM. If the data which reached that node has a value of RM less than 6.543, predicted MEDV is 21.63. If it is greater than 6.543, the predicted MEDV is 27.43. At the node where LSTAT<14.4 is tested, the observations tested false will be again tested with CRIM<6.99237. Observations which are tested true will have a predicted value of MEDV as 17.14. Observations which are tested false will have a predicted value if MEDV as 11.98.

Now look at the branch of false grown from the first split. In the next node of that branch, observations which have RM<7.437 will form a new branch. At the next node if LSTAT is less than 11.455, the predicted value of MEDV is 33.5 and if LSTAT is greater than 11.455, the predicted value of MEDV is 20.74. The observations with condition RM <7.437 as false, the next condition PTRATIO<17.9 is checked. Those observations with condition is TRUE, the predicted value of MEDV is 46.82. If false, the predicted value of MEDV is 36.48.

Please note that MEDV is median value of owner-occupied homes in $1000.

****