# Module 3 CT Option 1

Didem Aykurt

2022-12-02

## R Markdown

### LOAD AND PREPARE DATA

```
bostonHousing.df <-read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master DA
/MIS510-1 Data Mining_4 term/Module 3/MIS510BostonHousing.csv", header=TRUE)
```

Show all the data in a new tab

```
View(bostonHousing.df)
```

Check to null object(result False means there isn't null)

```
is.null(bostonHousing.df)

## [1] FALSE
```

There are number of rows and column

```
nrow(bostonHousing.df)

## [1] 506

ncol(bostonHousing.df)

## [1] 14
```

Display the first ten rows of each column

```
bostonHousing.df[1:10, ]

##       CRIM   ZN INDUS CHAS   NOX    RM   AGE    DIS RAD TAX PTRATIO LSTAT
MEDV
## 1  0.00632 18.0  2.31    0 0.538 6.575  65.2 4.0900   1 296    15.3  4.98
24.0
## 2  0.02731  0.0  7.07    0 0.469 6.421  78.9 4.9671   2 242    17.8  9.14
21.6
## 3  0.02729  0.0  7.07    0 0.469 7.185  61.1 4.9671   2 242    17.8  4.03
34.7
## 4  0.03237  0.0  2.18    0 0.458 6.998  45.8 6.0622   3 222    18.7  2.94
33.4
## 5  0.06905  0.0  2.18    0 0.458 7.147  54.2 6.0622   3 222    18.7  5.33
36.2
## 6  0.02985  0.0  2.18    0 0.458 6.430  58.7 6.0622   3 222    18.7  5.21
```

```
28.7
## 7  0.08829 12.5  7.87    0 0.524 6.012  66.6 5.5605   5 311    15.2 12.43
22.9
## 8  0.14455 12.5  7.87    0 0.524 6.172  96.1 5.9505   5 311    15.2 19.15
27.1
## 9  0.21124 12.5  7.87    0 0.524 5.631 100.0 6.0821   5 311    15.2 29.93
16.5
## 10 0.17004 12.5  7.87    0 0.524 6.004  85.9 6.5921   5 311    15.2 17.10
18.9
##    CAT..MEDV
## 1          0
## 2          0
## 3          1
## 4          1
## 5          1
## 6          0
## 7          0
## 8          0
## 9          0
## 10         0
```

Print the list in a useful column format

```
t(t(names(bostonHousing.df)))
```

```
##       [,1]
##  [1,] "CRIM"
##  [2,] "ZN"
##  [3,] "INDUS"
##  [4,] "CHAS"
##  [5,] "NOX"
##  [6,] "RM"
##  [7,] "AGE"
##  [8,] "DIS"
##  [9,] "RAD"
## [10,] "TAX"
## [11,] "PTRATIO"
## [12,] "LSTAT"
## [13,] "MEDV"
## [14,] "CAT..MEDV"
```

Display summary statistics for each column

```
summary(bostonHousing.df)
```

```
##       CRIM                ZN              INDUS            CHAS
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
```

```
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       NOX               RM              AGE              DIS
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       RAD              TAX           PTRATIO          LSTAT
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97
##       MEDV         CAT..MEDV
##  Min.   : 5.00   Min.   :0.000
##  1st Qu.:17.02   1st Qu.:0.000
##  Median :21.20   Median :0.000
##  Mean   :22.53   Mean   :0.166
##  3rd Qu.:25.00   3rd Qu.:0.000
##  Max.   :50.00   Max.   :1.000
```

Display the correlation among the attributes of the housing data set

```
round(cor(bostonHousing.df),2)
```

```
##            CRIM    ZN INDUS  CHAS   NOX    RM   AGE   DIS   RAD   TAX PTRA
TIO
## CRIM       1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0
.29
## ZN        -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0
.39
## INDUS      0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0
.38
## CHAS      -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0
.12
## NOX        0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0
.19
## RM        -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0
.36
## AGE        0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0
.26
## DIS       -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0
.23
## RAD        0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0
.46
## TAX        0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0
.46
## PTRATIO    0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1
```

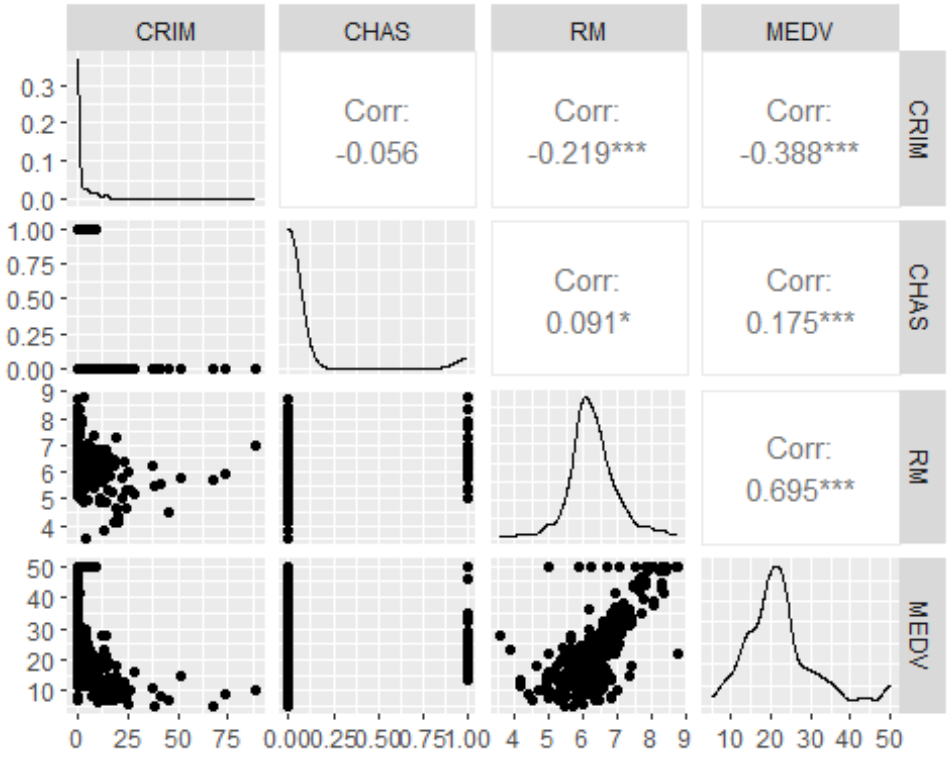
```
.00
## LSTAT      0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0
.37
## MEDV      -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0
.51
## CAT..MEDV -0.15  0.37 -0.37  0.11 -0.23  0.64 -0.19  0.12 -0.20 -0.27   -0
.44
##            LSTAT  MEDV CAT..MEDV
## CRIM        0.46 -0.39     -0.15
## ZN         -0.41  0.36      0.37
## INDUS       0.60 -0.48     -0.37
## CHAS       -0.05  0.18      0.11
## NOX         0.59 -0.43     -0.23
## RM         -0.61  0.70      0.64
## AGE         0.60 -0.38     -0.19
## DIS        -0.50  0.25      0.12
## RAD         0.49 -0.38     -0.20
## TAX         0.54 -0.47     -0.27
## PTRATIO     0.37 -0.51     -0.44
## LSTAT       1.00 -0.74     -0.47
## MEDV       -0.74  1.00      0.79
## CAT..MEDV -0.47  0.79      1.00
```

Display the variable that want the table on the console and View a scatter matrix of attributes of the housing data set at columns of 1,4,6 and 13

```
library(ggplot2)
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

bhousing<- data.frame(bostonHousing.df[,c(1,4,6,13)])
ggpairs(bhousing)
```




Apply a multiple linear regression model to the median house price(MEDV) as a function of CRIM, CHAS, and RM

```
library(mlr3)
lm(MEDV ~ CRIM+CHAS+RM, data=bostonHousing.df)

##
## Call:
## lm(formula = MEDV ~ CRIM + CHAS + RM, data = bostonHousing.df)
##
## Coefficients:
## (Intercept)         CRIM         CHAS           RM
##    -28.8107      -0.2607       3.7630       8.2782
```

Fit the model

```
model<- lm(MEDV ~ CRIM+CHAS+RM, data=bostonHousing.df)
```

Display of the model result with summary() function

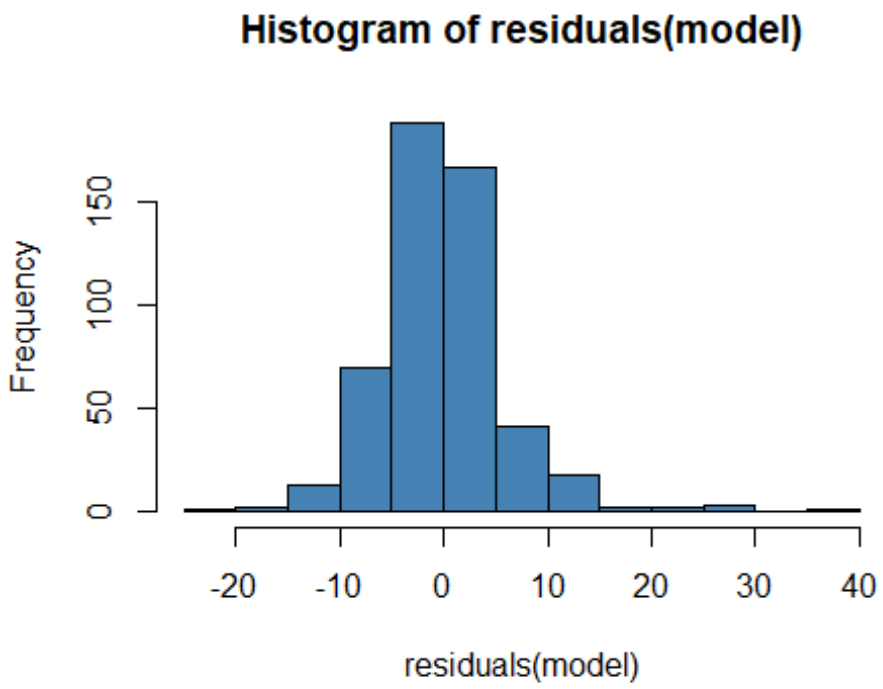
```
summary(model)

##
## Call:
## lm(formula = MEDV ~ CRIM + CHAS + RM, data = bostonHousing.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -24.829  -2.968  -0.415   2.433  38.945
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.81068    2.56331 -11.240  < 2e-16 ***
## CRIM         -0.26072    0.03274  -7.964 1.12e-14 ***
## CHAS          3.76304    1.08620   3.464 0.000577 ***
## RM            8.27818    0.40182  20.602  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.17 on 502 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:   0.55
## F-statistic: 206.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

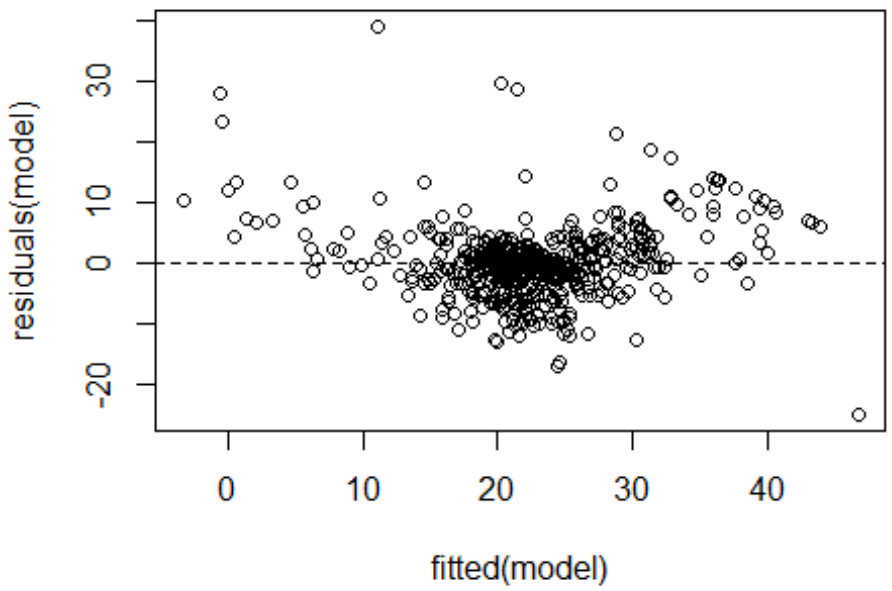
1-The distribution of model residuals should be approximately normal

```
hist(residuals(model), col="steelblue")
```




2-The variance of the residuals should be consistent for all observations. Create fitted value vs. residual plot

```
plot(fitted(model), residuals(model))
#Add horizontal line at 0
abline(h=0, lty=2)
```




Linear regression of MEDV on all the predictors in the dataset except CAR..MEDV

```
model<- lm(formula = MEDV~ CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+
LSTAT, data =bostonHousing.df )

summary(model)

##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + TAX + PTRATIO + LSTAT, data = bostonHousing.df)
##
## 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
```

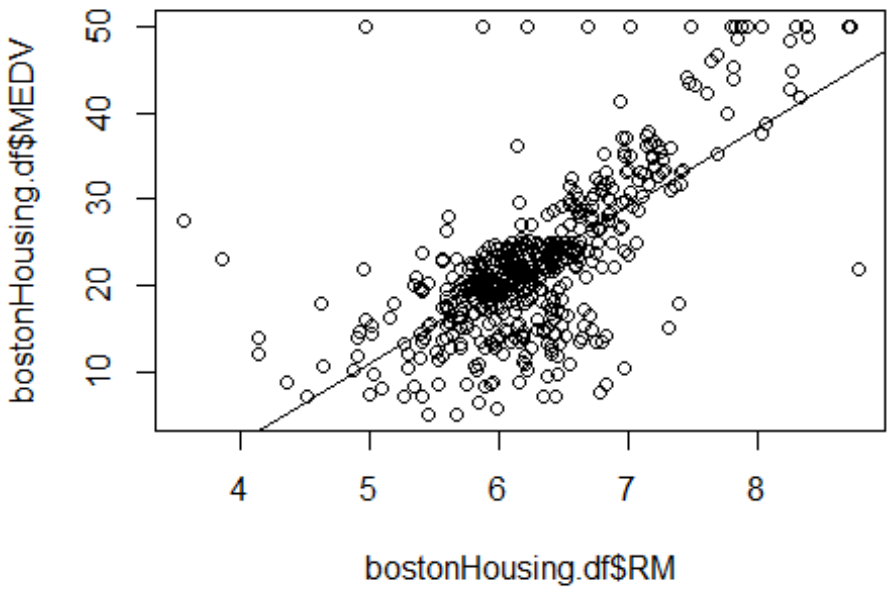
Extract standardized Coefficient

```
lm(data.frame(scale(model$model)))

##
## Call:
## lm(formula = data.frame(scale(model$model)))
##
## Coefficients:
## (Intercept)         CRIM           ZN        INDUS         CHAS           N
OX
##  -1.309e-16   -1.135e-01    1.191e-01    1.005e-02    7.843e-02   -2.363e-
01
##          RM          AGE          DIS          RAD          TAX      PTRAT
IO
##   2.795e-01    1.105e-02   -3.413e-01    2.740e-01   -2.324e-01   -2.207e-
01
##       LSTAT
##  -4.286e-01
```

Most significant attribute RM because p value<0.05 and the highest correlation .70, linear relation and result of standardized regression coefficient at 2.795e-01

```
plot(bostonHousing.df$RM,bostonHousing.df$MEDV)
#ablien() function add linear regression line.
abline(reg=lm(bostonHousing.df$MEDV ~ bostonHousing.df$RM))
```




### The Boston Housing's Exploratory Data Analysis

The Boston Housing dataset describes house selling and price with houses properties. The dataset is available at CSUGlobal. Five hundred six intakes represent aggregated data and 12 attributes(features) with the dependent variable cost of the house as follows: CRIM (crime rate), ZN (percentage land zone for lost over 25,000squarfeet), INDUS ( percentage land occupied by nonretail), CHAS (track bound Charles river, one is yes,0 is other), NOX(nitric oxide per 10 million), RM (average number of rooms per house), AGE (percentage of units built before 1940), DIS (Weighted distances to Boston employment center), RAD( accessibility to radial highways), TAX(total value property tax rate per $10,000), PTRATIO( Pupil to teacher ratio y town), LSTAT(Percentage of the lower status of the population), MEDV(Median value of owner-occupied homes in $1000s), CAT..MEDV(median value of owner-occupied dwellings above $30,000).

The retail business is top of the world, and companies must study the elements contributing to higher prices and understand what the community of Boston has to offer home. To examine the relationship between the cost of the house and the average number of rooms. I used R studio to analyze house prices and find the most significant attribute. I used descriptive statistics to see all the data detail and see how the relationship between independent variables and the dependent variable is MEDV.

For the analysis, the MLR result of the group of independent variables is CRIM, CHAS, and RM, which look relatively normal distribution and fit the horizontal line. The CHAS variable can affect the result because CHAS and MEDV variables do not have a linear relation.

Next, multi-linear regression with all variables except CAR..MEDV model summary result is that most variables' p-value was lower than 0.05, excluding INDUS and AGE, which means the model's statistical significance. Linear regression has validated assumptions that:

- 1. Linear relation between the target variable and input variables.
- 2. Independence data points are any data point that can't tell for the next point.
- 3.Normal distribution.
- 4. Equal error variance doesn't show any patterns in dataset variability.

Predictive analytics helps to find which independent variable most affects the dependent variable. In this case, RM and MEDV have a linear relation; the scatter plot shows a particular data point and the result of a p-value lower than 0.05, correlation of 0.70, and error variance of 2.795e-01. Thus, RM is the most significant variable for the price of the house.

## *Reference*

Data Mining for Business Analytics… concepts, Techniques, and Applications in R by Galit Shmueli: Peter C. Bruce: Inbal Yahav: Nitin R. Patel: Kenneth C. Lichtendahl, Jr. Page 56, chapter 3, Figures 3.2 and 3.3. Page 94, chapter 4, Figures 4.3 and 4.4, chapter 6, Figure 6.3, page 157.