Boston Housing dataset is fetched from UCI machine learning repository. It contains 506 observations, 13 continuous attributes and 1 binary-valued attribute. Here is the detailed attribute information:

1. CRIM per capita crime rate by town

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

3. INDUS proportion of non-retail business acres per town

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

5. NOX nitric oxides concentration (parts per 10 million)

6. RM average number of rooms per dwelling

7. AGE proportion of owner-occupied units built prior to 1940

8. DIS weighted distances to five Boston employment centers

9. RAD index of accessibility to radial highways

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

11. PTRATIO pupil-teacher ratio by town

12. B 1000(Bk - 0.63)^2

13. LSTAT % lower status of the population

14. MEDV Median value of owner-occupied homes in $1000's

The objective is to explain the Median home value in Boston as a function of its characteristics.

1. Import Boston Housing dataset from UCI machine learning repository: assignment4.csv

**data <- read.csv(file.choose(), stringsAsFactors = F)**

1. Use str() to inspect the type and data frame of variables. How many observations and how many variables in this dataset? Learn the meaning of each variable.

**str(data)**

**506 obs.**

**14 variables**

1. Reform the CHAS to nominal (factor) variable. How many levels does CHAS have? How many instances belong to level 0?

**data$CHAS <- factor(data$CHAS)**

**str(data$CHAS)**

**table(data$CHAS)**

**CHAS has 2 levels**

**471 instances belong to Charles river dummy variable**

1. What is the mean value of age variable? Since age is the proportion of homes built before 1940, how do you interpret this mean value? What is the maximum value of MEDV? Build a histogram for MEDV, which value range has the highest frequency?

**summary(data$AGE)**

**mean: 68.57, this mean explains the average age of homes built before 1940**

**hist(data$MEDV)**

**summary(data$MEDV)**

**Max MEDV: 50**

**20-30 has the highest frequency**

1. Check the correlation between all the variables except for CHAS. Which pairs have correlations larger than 0.6? What variables have correlations with MEDV larger than 0.6?

**cor(data[c("CRIM", "ZN", "INDUS", "NOX", "RM", "AGE", "DIS", "RAD", "TAX", "PTRATIO", "B", "LSTAT", "MEDV")])**

**Correlations larger than 0.6: AGE and INDUS, INDUS and NOX, AGE and NOX, DIS and ZN, INDUS and DIS, DIS and NOX, DIS and AGE, RAD and CRIM, NOX and RAD, TAX and INDUS, TAX and NOX, LSTAT and INDUS, LSTAT and RM, LSTAT and AGE, MEDV and RM, RAD and TAX, MEDV and LSTAT**

**RM and LSTAT have a 0.6+ correlation with MEDV**

1. What are your assumptions of the relations between crime and home value, between DIS and home value? Does higher pupil-teacher ratio lead to lower home value? Evaluate your assumptions with the results from correlation.

**CRIM AND MEDV: negative correlation as crime goes up median value of houses goes down.**

**MEDV and DIS: positive correlation between distance from Boston center and median value of homes.**

**MEDV and PTRATIO: negative correlation between higher pupil-teacher ratio and home value. Therefore, as the pupil-teacher ratio increases home value decreases.**

1. Use pairs.panels() to visualize the relations between “RM”, “LSTAT” and “MEDV”. What are the histograms depicting on the diagonal? Does more room lead to higher home values? Does higher percentage of lower status of the population cause higher home values?

**install.packages('psych')**

**library(psych)**

**pairs.panels(data[c("RM", "LSTAT", "MEDV")])**

**Diagonal histograms are showing the curve of each relation.**

**More room doesn’t always lead to higher home values. It peaks in the middle of the histogram according to the relation**

**No, lower percentage of LSTAT causes higher MEDV**

1. **Randomize** the dataset and use the first 80% of observations for training. Set seed as 1234.

**Is CHAS an ID and should it be removed from the data? I wasn’t sure because question 14 and on require CHAS.**

**set.seed(1234)**

**data.r <- sample(data)**

**str(data.r)**

**nrows <- nrow(data.r)**

**train.size <- floor(nrows \* 0.8)**

**data.train <- data.r[1:train.size,]**

1. Build a linear regression model with MEDV as dependent variable and the rest variables as independent variables. Which variables have insignificant impact on MEDV?

**model <- lm(MEDV ~ B + TAX + RM + NOX + RAD + PTRATIO + ZN + DIS + CHAS + LSTAT + INDUS + AGE + CRIM, data = data.r)**

**summary(model)**

**Insignificant impact: INDUS and CRIM have an insignificant impact according to the rule that if P-value is less than 5% it’s significant**

1. How to interpret the estimate for variable RM? What is the p-value? Is this variable statistically significant? Why?

**P-Value is < 2e-16**

**RM is statistically significant at a 5% significance level**

**Because it falls below the significance level we have set.**

1. What is the R-squared value and adjusted R-squared value?

**Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338**

1. Try Crim^2 as one of the dependent variables and build a liner regression model include this variable (all other variables included). Does this Crim^2 have significant impact on MEDV?

**crim\_model <- lm(CRIM^2 ~., data.r)**

**summary(crim\_model)**

**Crim^2 does have a significant impact on MEDV**

1. Also create a new tax2 variable with tax2 = 1 if TAX >=350 and tax2 = 0 if TAX < 350. Build a linear regression model include this tax2 (all other variables included). What is the adjusted R-squared value?

**data.r$tax2 <- ifelse(data.r$TAX >= 350, 1, 0)**

**tax2\_model <- lm(tax2 ~., data.r)**

**summary(tax2\_model)**

**Adjusted R-squared: 0.7871**

1. Create a new simple regression model and estimate MEDV using only CHAS and CRIM.

**new\_crim\_model <- lm(CRIM ~ MEDV + CHAS, data.r)**

**summary(new\_crim\_model)**

1. How to interpret the estimate for variable CHAS1? What is the p-value? Is this variable statistically significant? Why?

**CHAS1 P-Value is 0.764**

**It’s not statistically significant because it’s above 5% and 10% general rules of significance.**

1. Calculate MAE using the function as discussed in class using training dataset. Also, calculate the correlation between predicted and the actual using training dataset

**MAE <- function(actual, predicted) {**

**mean(abs(actual - predicted))**

**}**

**library(rpart)**

**m <- rpart(MEDV ~ CRIM + CHAS, data = data.train)**

**p <- predict (m, data.train)**

**MAE(p, data.train$MEDV)**

**cor(p, data.train$MEDV)**

1. Calculate MAE and correlation using testing dataset.

**m <- rpart(MEDV ~ CRIM + CHAS, data = data.test)**

**p <- predict(m, data.test)**

**MAE(p, data.test$MEDV)**

**cor(p, data.test$MEDV)**

1. Estimate MEDV via regression tree using CHAS and CRIM.
   1. Plot the tree using rpart.plot

**m <- rpart(MEDV ~ CHAS + CRIM, data=data.r)**

**rpart.plot(m)**

* 1. What is the most important feature? What is the rule for the first leaf node?   
       
     **CRIM is the most important feature**

**First leaf’s rule is CRIM >= 6.7**

* 1. How many examples/records/data points are there in the first leaf node? What is the predicted value for these data points?

**23 examples are in the first leaf node**

**100% is the predicted value of these data points**

* 1. Calculate the correlation & MAE between predicted and the actual using training dataset

**p <- predict(m, data.train)**

**MAE(p, data.train$MEDV)**

**cor(p, data.train$MEDV)**

* 1. Calculate the correlation & MAE between predicted and the actual using testing dataset

**p <- predict(m, data.test)**

**MAE(p, data.test$MEDV)**

**cor(p, data.test$MEDV)**

1. Estimate MEDV via model tree using CHAS and CRIM.
   1. What is the most important feature? What is the rule for the first leaf node?

**CRIM is the most important feature**

**CRIM <= 2.08 is the first rule**

* 1. What is the linear model for the first leaf node?

**LM1 = 3.9137\*CRIM + 24.7293**

* 1. How many examples/records/data points are there in the first leaf node?

**357 examples are in the first node**

* 1. Calculate the correlation & MAE between predicted and the actual using training dataset

**m <- M5P(MEDV ~ CHAS + CRIM, data = data.train)**

**p <- predict(m, data.train)**

**MAE(p, data.train$MEDV)**

**cor(p, data.train$MEDV)**

* 1. Calculate the correlation & MAE between predicted and the actual using testing dataset

**m <- M5P(MEDV ~ CHAS + CRIM, data = data.test)**

**p <- predict(m, data.test)**

**MAE(p, data.test$MEDV)**

**cor(p, data.test$MEDV)**

* 1. Compare MAE and correlation of model tree vs regression tree vs simple regression model

**simple\_reg\_model <- lm(MEDV ~ CRIM + CHAS, data.r)**

**predicted\_simple <- predict(simple\_reg\_model, data.r)**

**MAE(predicted\_simple, data.r$MEDV)**

**cor(predicted\_simple, data.r$MEDV)**

**Simple Reg. MAE: 6.096066**

**Simple Cor: 0.4176532**

**reg\_tree\_model <- rpart(MEDV ~ CRIM + CHAS, data.r)**

**reg\_tree\_predict <- predict(reg\_tree\_model, data.r)**

**MAE(reg\_tree\_predict, data.r$MEDV)**

**cor(reg\_tree\_predict, data.r$MEDV)**

**Reg. Tree MAE: 5.241381**

**Reg. Tree Cor: 0.5992906**

**model\_tree <- M5P(MEDV ~ CRIM + CHAS, data = data.r)**

**model\_tree\_pred <- predict(model\_tree, data.r)**

**MAE(model\_tree\_pred, data.r$MEDV)**

**cor(model\_tree\_pred, data.r$MEDV)**

**Model Tree MAE: 5.838835**

**Model Tree Cor: 0.4710983**

**The MAE comparison shows us that simple regression models have the highest value and Regression Trees have the lowest value.**

**The Cor comparison show us that Regression Trees have the highest value and the Simple Regressions have the lowest value.**