Exercise 6 – Random Forest (Boston Housing Data)

```
##############   Random Forests   ##########################

# Growing a random forest proceeds in exactly the same way,
#     except that we use a smaller value of the mtry argument.
# By default, randomForest() uses $p/3$ variables when building
#     a random forest of regression trees, and $\sqrt{p}$ variables
#     when building a random forest of classification trees.

# Building a random forest on the same data set using mtry = 6.
# Comment on the difference from the test MSE from using the random
# forest compared to bagging.

set.seed(1)
rf.boston=randomForest(medv~.,data=Boston,subset=train, mtry=6,importance =TRUE)
yhat.rf = predict(rf.boston ,newdata=Boston[-train ,])
mean((yhat.rf-boston.test)^2)
    # We see that the test MSE for a random forest is 11.48;
    # this indicates that random forests yielded an improvement over bagging
    # in this case (versus 13.34)

# Investigating variable importance
importance(rf.boston)
    # Two measures of variable importance are reported:
    # 1) The first is based upon the mean *decrease of accuracy*
    # in predictions on the out of bag samples when a given variable
    # is excluded from the model.
    # 2) THe second is a measure of the total *decrease in node impurity*
    # that results from splits over that variable, averaged over all trees.
varImpPlot (rf.boston)
    # The results indicate that across all of the trees considered in the
    # random forest, the wealth level of the community (lstat) and the house size (rm)
    # are by far the two most important variables for median house prices (which makes sense).
```