(Textbook) 8.4 Exercises, Problem 8. (a), (b), (d), and (e). Use set.seed(1126). When you split the data set in (a), randomly choose half of observations as a training set and remaining half of observations as test set. (You must copy your R code together with your answer)

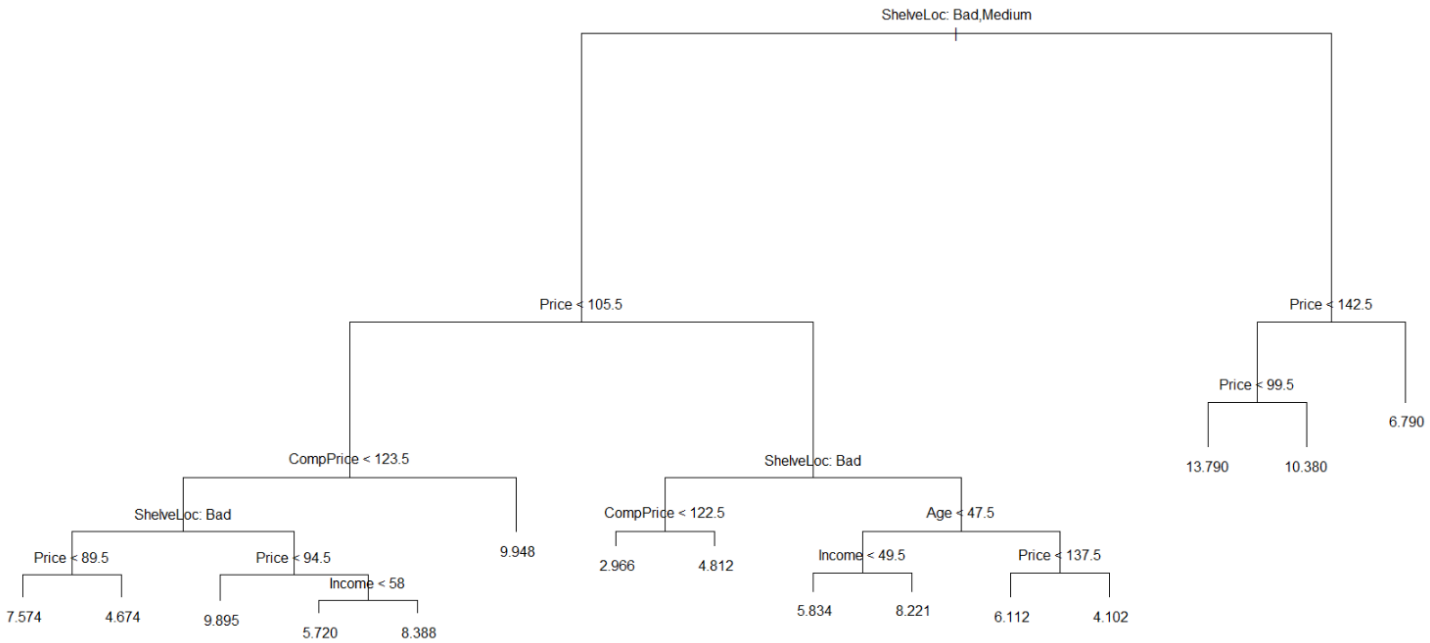
In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set.

```
train <- sample(1:nrow(Carseats), nrow(Carseats)/2)
test.Sales <- Carseats[-train, 1]
```

(b) Fit a regression tree to the training set. Plot the tree and interpret the results. What test MSE do you obtain?

```
tree.Carseats <- tree(Sales ~., subset = train, data = Carseats)
summary(tree.Carseats) #15 terminal nodes
#variables actually used in tree construction:
#ShelveLoc, Price, CompPrice, Income, Age
plot(tree.Carseats)
text(tree.Carseats, pretty = 0)
yhat <- predict(tree.Carseats, newdata = Carseats[-train, ])
test.Sales <- Carseats[-train, 1]
cbind(yhat, test.Sales)[1:10,]
mean((yhat - test.Sales)^2) #5.055247
```



I also tried a pruned model, but the results were the same (in terms of test MSE and number of nodes.

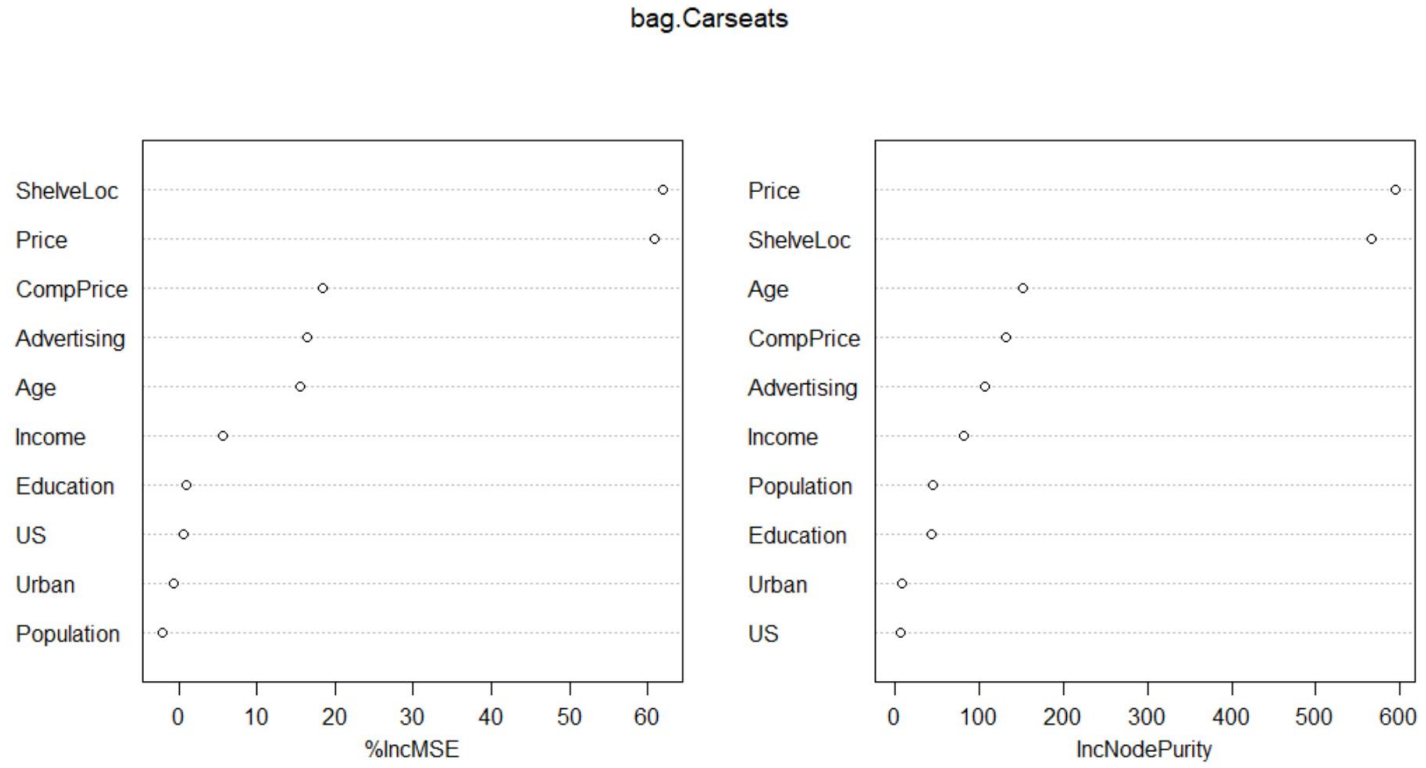
```
cv.Carseats <- cv.tree(tree.Carseats)
```

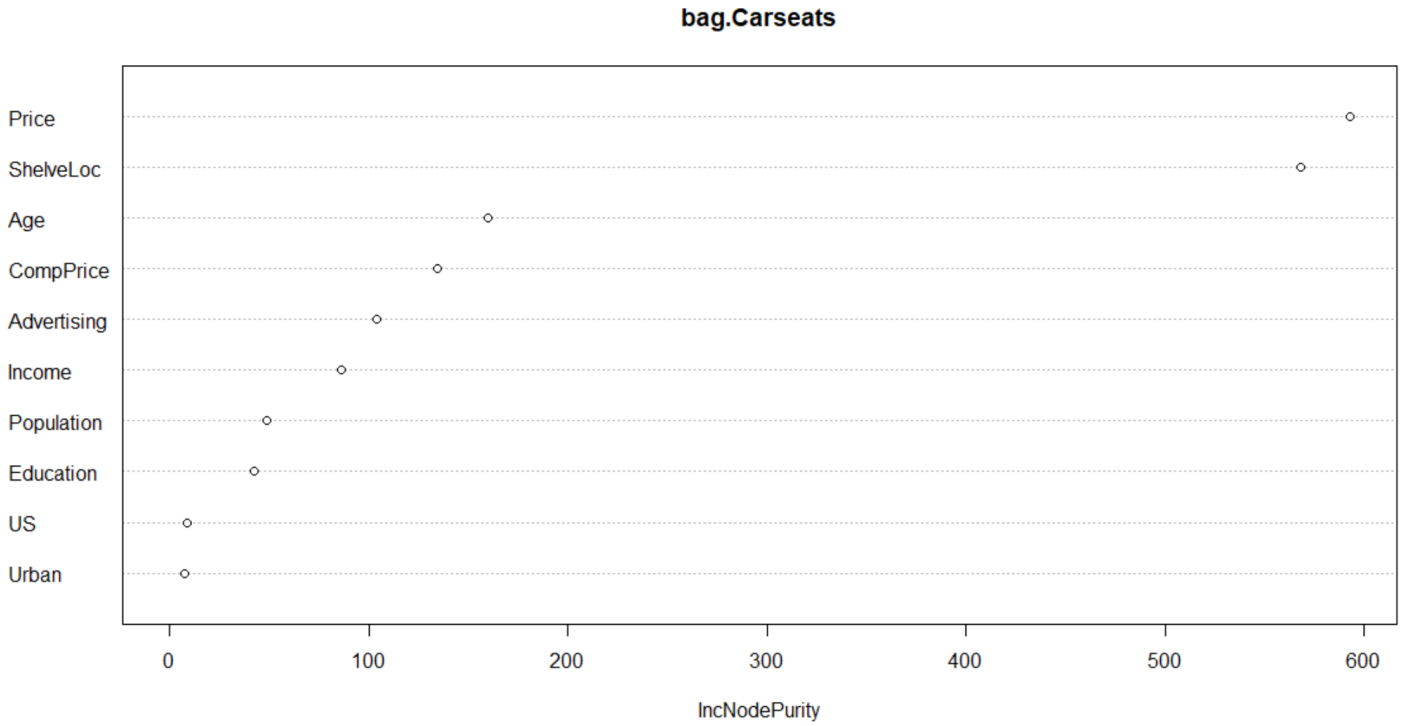
```
cv.Carseats
which.min(cv.Carseats$dev)

cv.Carseats$size[1]
prune.Carseats <- prune.tree(tree.Carseats, best = 15)
plot(prune.Carseats)
text(prune.Carseats, pretty = 0)

yhat <- predict(prune.Carseats, newdata = Carseats[-train, ])
test.Sales <- Carseats[-train, 1]
cbind(yhat, test.Sales)[1:10,]
mean((yhat - test.Sales)^2)
```

The test MSE for this model is 5.055247. The number of nodes is 15. The variables used for this model are: ShelveLoc, Price, CompPrice, Income, Age out of the 10 predictors possible.

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

```
set.seed(1126)
train <- sample(1:nrow(Carseats), nrow(Carseats)/2)
bag.Carseats <- randomForest(Sales ~., subset = train, data = Carseats,
                             mtry = 10, importance = TRUE)
bag.Carseats
test.Sales <- Carseats[-train, 1]
yhat.bag <- predict(bag.Carseats, newdata = Carseats[-train,])
cbind(yhat.bag, test.Sales)[1:10,]
mean((yhat.bag - test.Sales)^2)
varImpPlot(bag.Carseats)
```

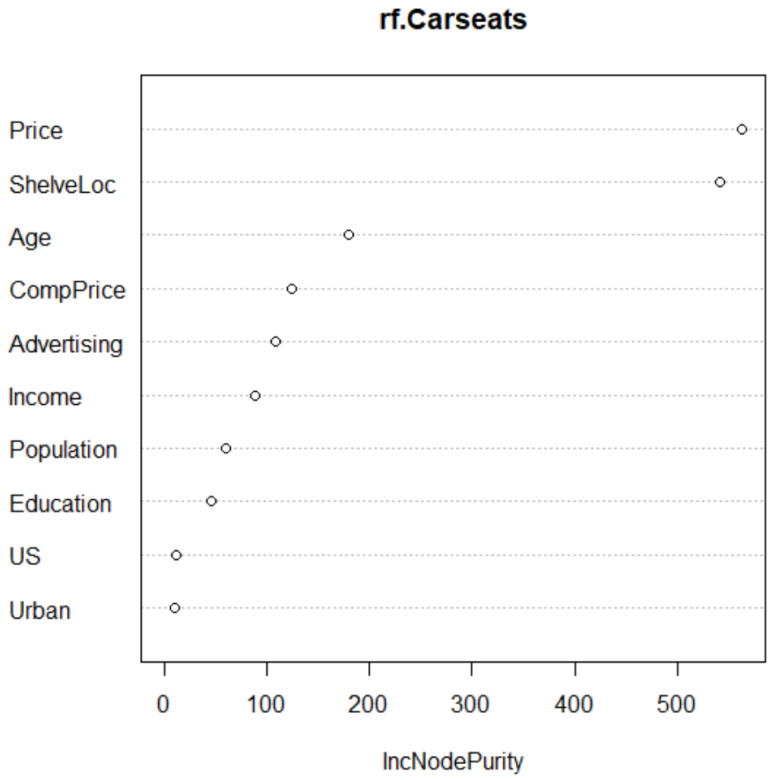




The test MSE for bagging is much lower than for regression decision tree, at 2.588667. The important predictors are Price and ShelveLoc.

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

```
rf.Carseats <- randomForest(Sales ~., subset = train, data = Carseats, mtry = 7)
yhat.rf <- predict(rf.Carseats, newdata = Carseats[-train,])
cbind(yhat.rf, test.Sales)[1:10,]
mean((yhat.rf - test.Sales)^2) #2.396501
varImpPlot(rf.Carseats)
```



The test MSE for Random Forest at m = 2 is 3.022149.
The test MSE for Random Forest at m = 3 is 2.682835.
The test MSE for Random Forest at m = 4 is 2. 485148.
The test MSE for Random Forest at m = 5 is 2. 38904.
The test MSE for Random Forest at m = 6 is 2.41555.
The test MSE for Random Forest at m = 7 is 2. 396501.
The test MSE for Random Forest at m = 8 is 2. 481931.
The test MSE for Random Forest at m = 9 is 2. 5727.
The test MSE for Random Forest at m = 10 is 2. 600736.

The best test MSE is when m = 7, which is makes sense because the last few variables on the importance chart are pretty close to zero. Before 7, the closer m gets to 7, the better the test MSE. After m = 7, the closer m gets to 10 (the total number of predictors, the worse the test MSE.

The two most important predictors are still Price and ShelveLoc