CS4100/CS5100: Data Analysis - Lab worksheet 2: Answers for Exercises

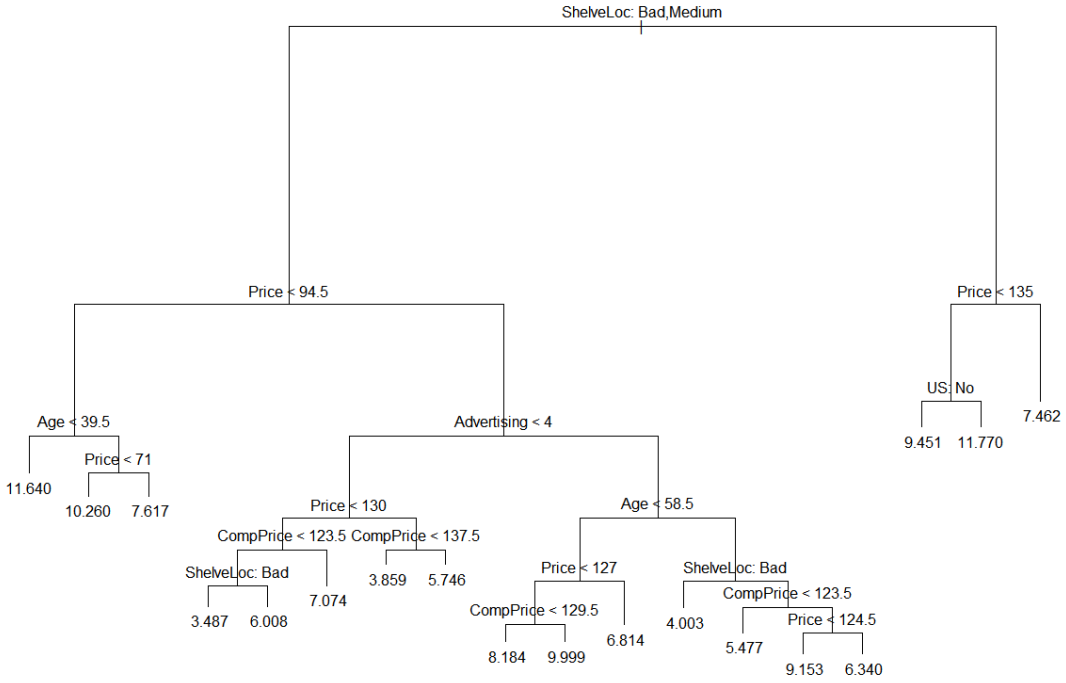
1.

```
> library(ISLR)
> library(tree)
> data(Carseats)
> fix(Carseats)
> names(Carseats)
 [1] "Sales"       "CompPrice"   "Income"
 [4] "Advertising" "Population"  "Price"
 [7] "ShelveLoc"   "Age"         "Education"
[10] "Urban"       "US"
> set.seed(2)
> train <- sample(1:nrow(Carseats), 200)
> Carseats.test <- Carseats[-train,]
> High.test <- High[-train]
```

2.

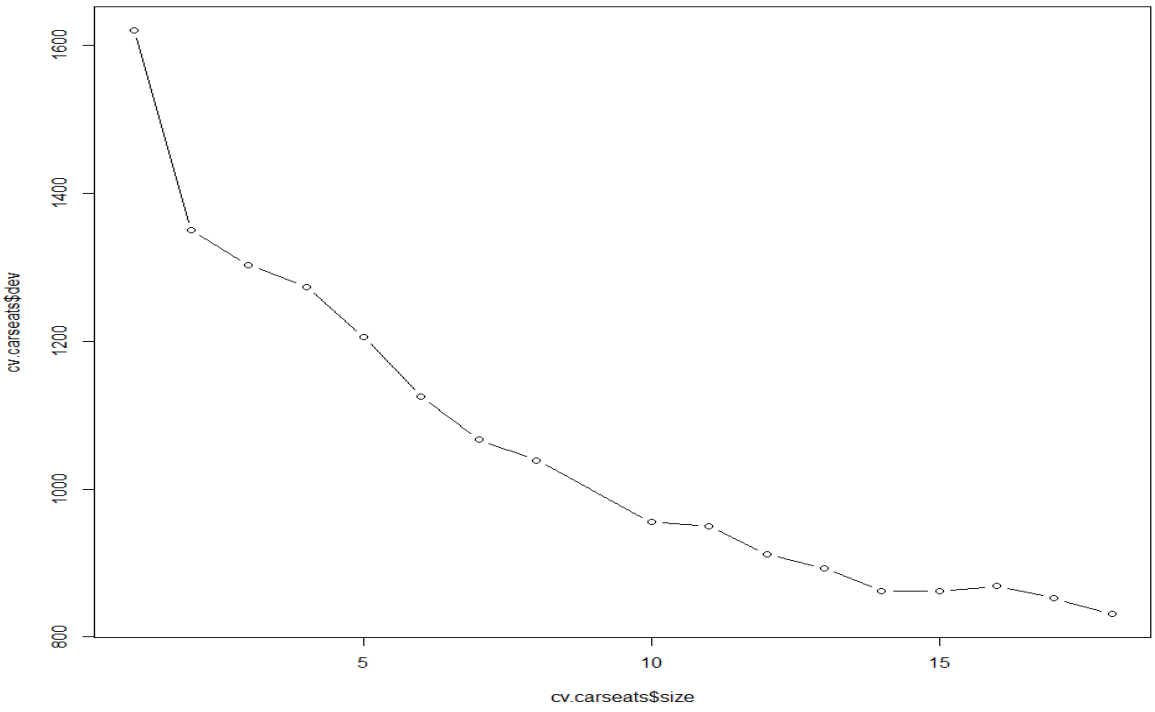
```
> set.seed(1)
> train <- sample(1:nrow(Carseats), nrow(Carseats)/2)
> tree.carseats <- tree(Sales ~ ., Carseats, subset=train)
> summary(tree.carseats)

Regression tree:
tree(formula = Sales ~ ., data = Carseats, subset = train)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Age"
[4] "Advertising" "CompPrice"   "US"
Number of terminal nodes:  18
Residual mean deviance:  2.167 = 394.3 / 182
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.88200 -0.88200 -0.08712  0.00000  0.89590  4.09900
> plot(tree.carseats)
> text(tree.carseats,pretty = 0)
```

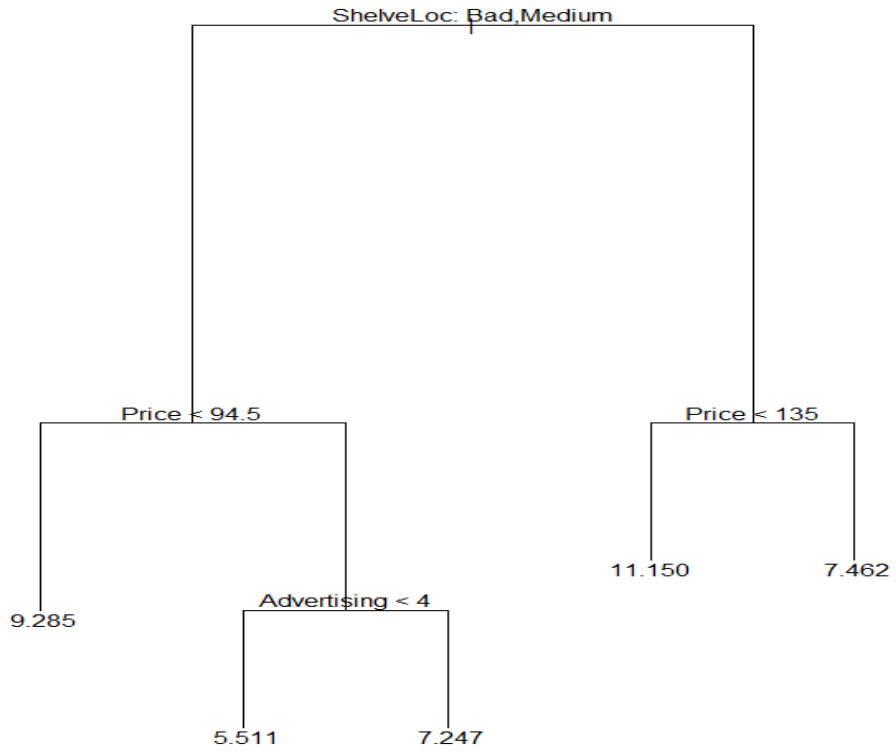


3.

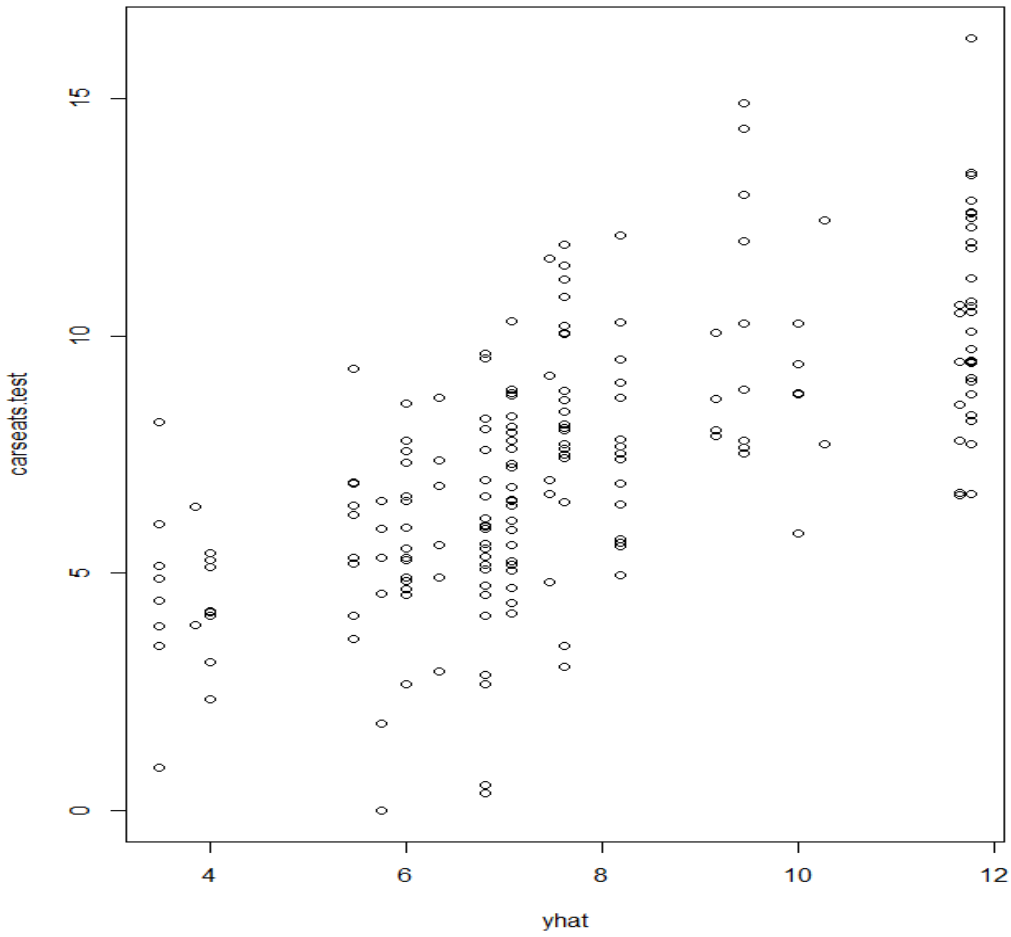
```
> cv.carseats <- cv.tree(tree.carseats)
> plot(cv.carseats$size, cv.carseats$dev, type = 'b')
```



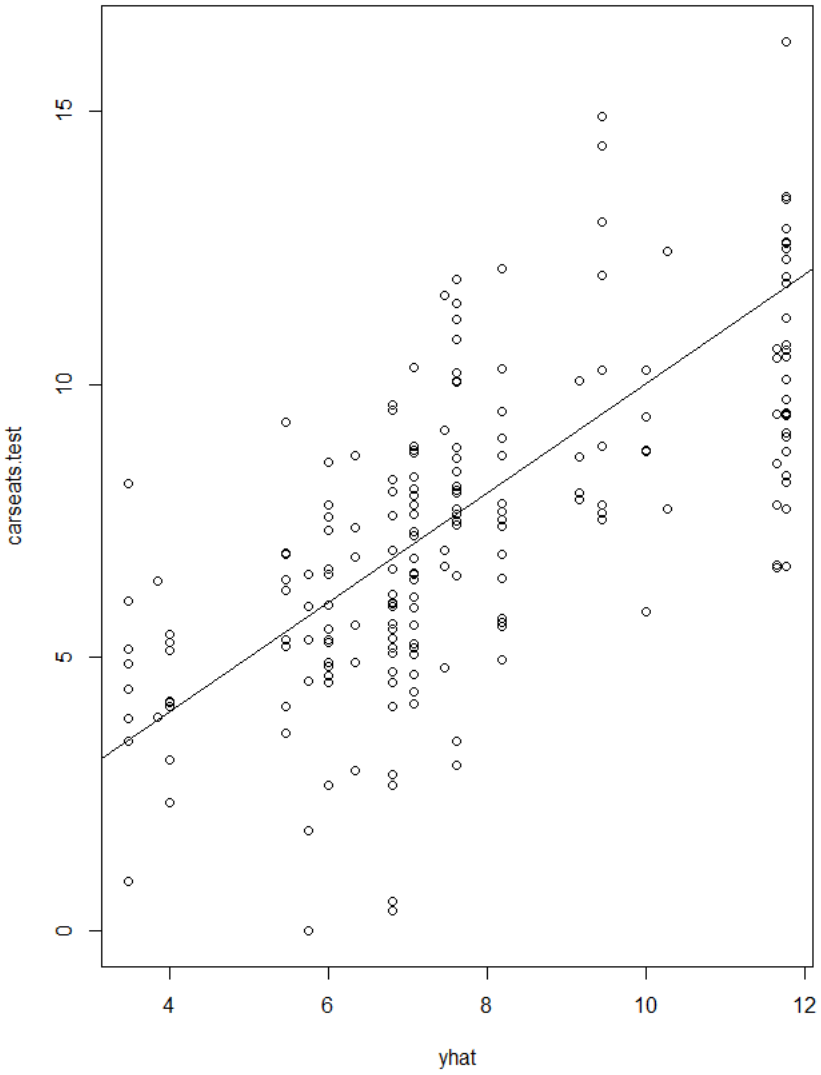
```
> prune.carseats <- prune.tree(tree.carseats, best=5)
> plot(prune.carseats)
> text(prune.carseats, pretty=0)
```




```
> yhat <- predict(tree.carseats, newdata = Carseats[-train,])
> carseats.test = Carseats[-train,"Sales"]
> plot(yhat,carseats.test)
```




```
> abline(0,1)
> mean((yhat - carseats.test)^2)
[1] 4.922039
```



Pruning the tree improved the test MSE.