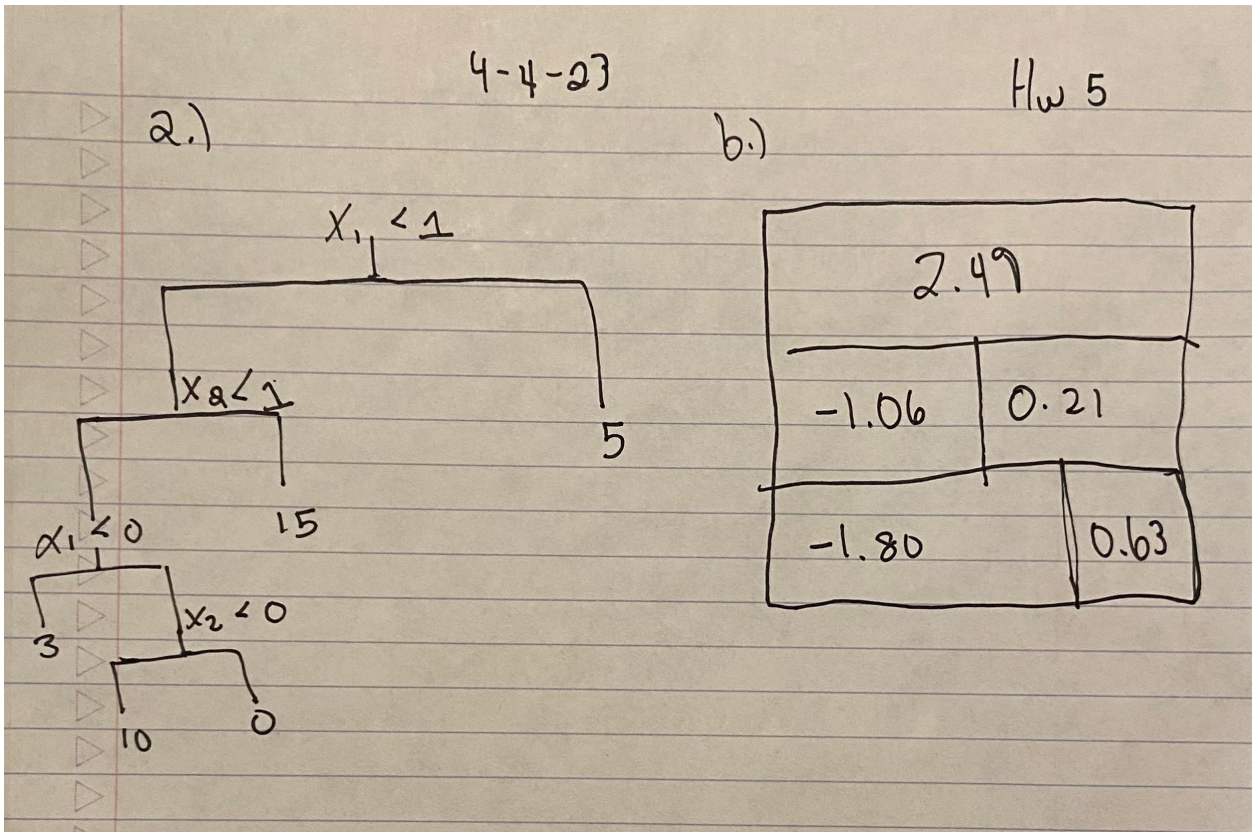
1. **Problem 1:**
    a. Sketch the tree corresponding to the partition of the predictor space illustrated on the left-hand plot. The numbers inside the boxes indicate the mean of Y within each region.
    b. Create a diagram similar to the left-hand plot using the tree illustrated in the right-hand plot. You should divide up the predictor space inot the correct regions, and indicate the mean for each region.



2. **Problem 2:**
    a. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of P(Class is Red|X):

        0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

        There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

        Answer: 6 and 0.45
3. **Problem 3:**
    a. Provide a detailed explanation of the algorithm that is used to ft a regression tree.
        i. **Answer:**

ii. Starting from the root node, the algorithm searches for the predictor variable that provides the best split of the data into two groups. The quality of a split is determined by a splitting criterion such as the reduction in variance or mean squared error (MSE) of the response variable. The split that yields the greatest reduction in variance or MSE is chosen.
iii. The data is partitioned into two groups based on the chosen split: one group for which the value of the predictor variable is less than or equal to a chosen threshold, and the other group for which it is greater than the threshold.
iv. The same process of splitting is then recursively applied to each of the two subgroups, until some stopping criterion is met. This stopping criterion could be based on a minimum number of observations in a terminal node, a maximum depth of the tree, or a lack of further improvement in the splitting criterion.
v. Once the tree has been fully grown, it may be pruned to prevent overfitting. This involves removing nodes that do not contribute much to the accuracy of the predictions on the test data. One way to prune the tree is to use cost complexity pruning, as described in the previous answer.
vi. To make a prediction for a new observation, the algorithm traverses the tree from the root node to a terminal node that corresponds to the observation's values of the predictor variables. The predicted value for the response variable is the mean value of the training observations in that terminal node.
vii. The accuracy of the predictions is typically evaluated using a measure such as mean squared error, which measures the average squared difference between the predicted and actual values of the response variable.
viii. In summary, the algorithm for fitting a regression tree involves recursively partitioning the predictor space into smaller and more homogeneous subspaces based on the values of the predictor variables, and then using the resulting subspaces to make predictions for new observations. By using a splitting criterion and a stopping criterion, the algorithm can balance between the complexity of the tree and its ability to fit the data well. Pruning can then be used to simplify the tree and prevent overfitting.

4. **Problem 4:** This problem involves the OJ data set which is part of the ISLR2 package.
    a. Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
    b. Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics

about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

i. **Answer:** Training error rate: 16% Number of terminal nodes: 8

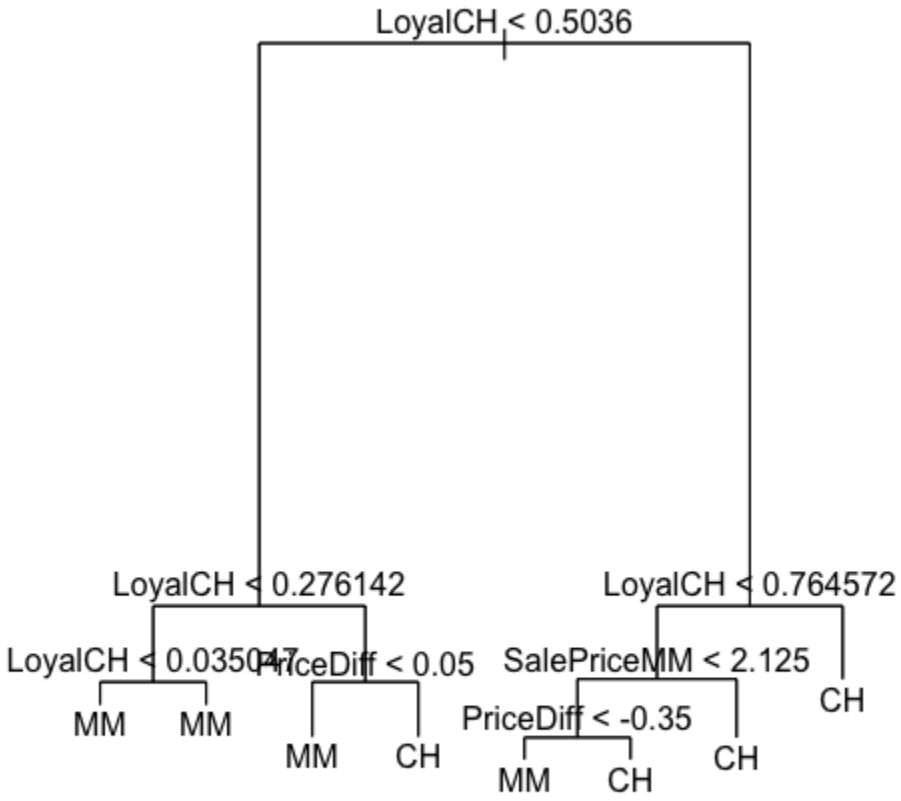
c. Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

i. **Answer:** From my node 2): If LoyalCH < 0.5036 there are 353 customers with this criteria the deviance is 422.6, the

ii. chance that the customer will by Minute Made is 71.388%.

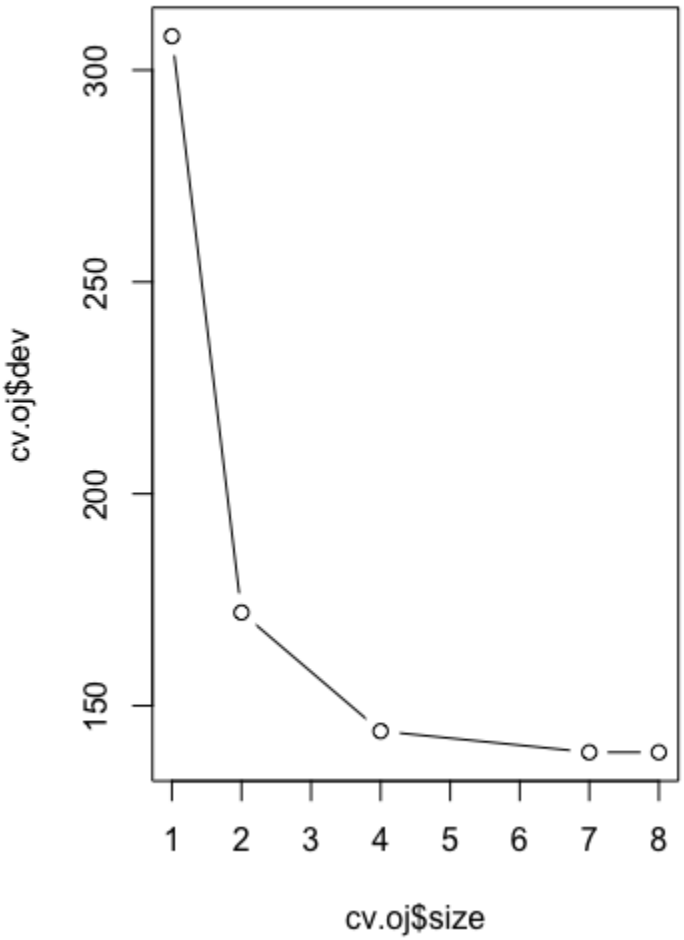
d. Create a plot of the tree, and interpret the results.

i. **Answer:**

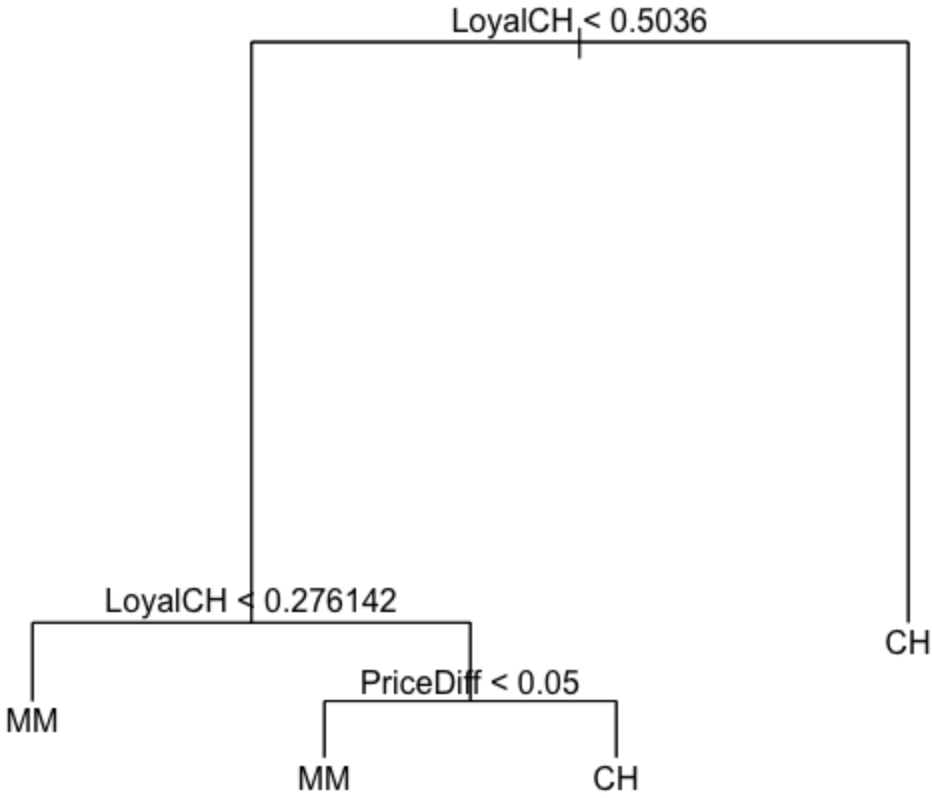


ii. The variables that appear to be used to predict if they will but MM or CH "LoyalCH", "SalePriceMM," and "PriceDiff"

e. Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
   i. **Answer:** 0.1814815
f. Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.
   i. **Answer:**



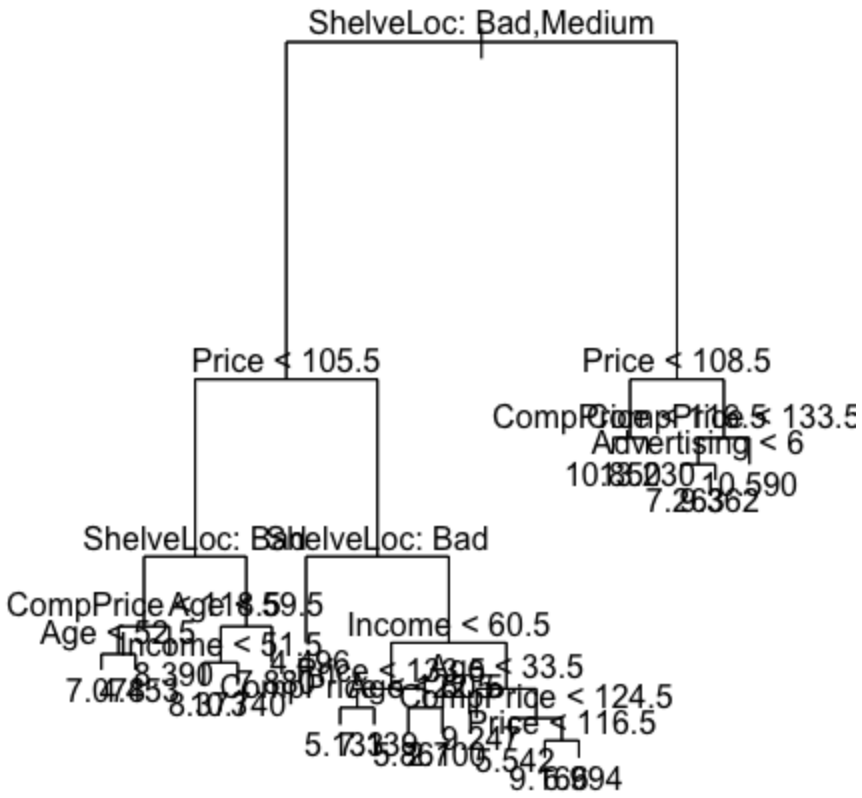
   ii. **Answer:** The Optimal Tree size is 4



g.

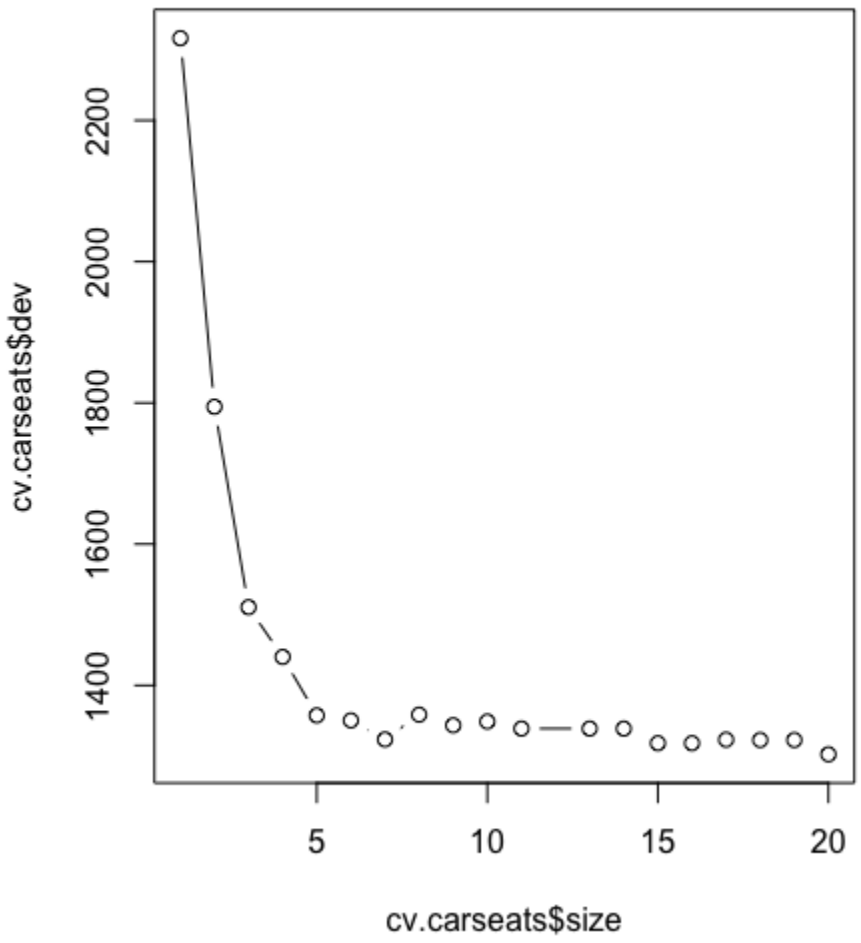
i. Answers: 0.16, 0.17, 0.1814815, 0.2037037

5. **Problem 5:** We will use the Carseats data set that is in the ISLR package to predict Sales using regression trees and related approaches.
   a. Split the data set into a training set and a test set.
   b. Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?



i. **Answer:**5.157823

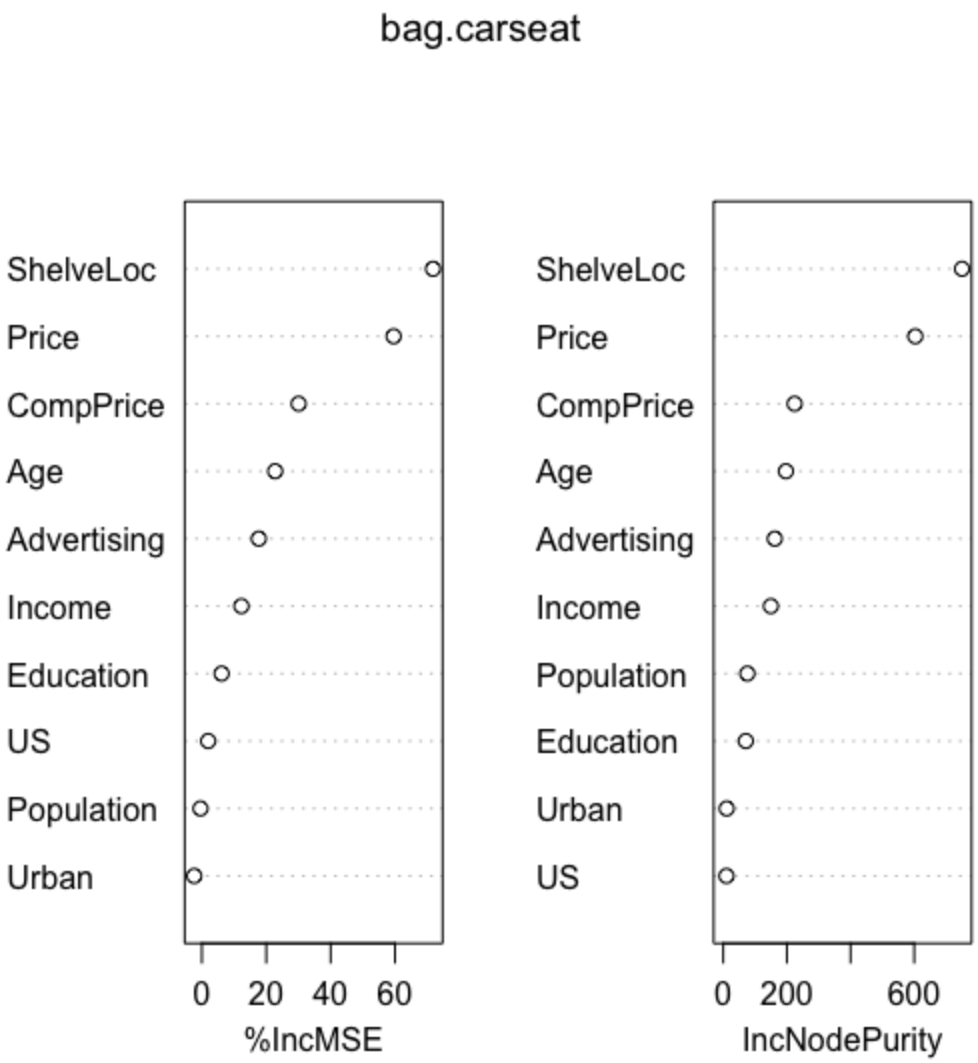
c. Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the MSE test?



i. **Answer:**It appears that pruning to 7 would be best.
ii. **Answer:** 4.679617

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.

i. **Answer:** 2.587705

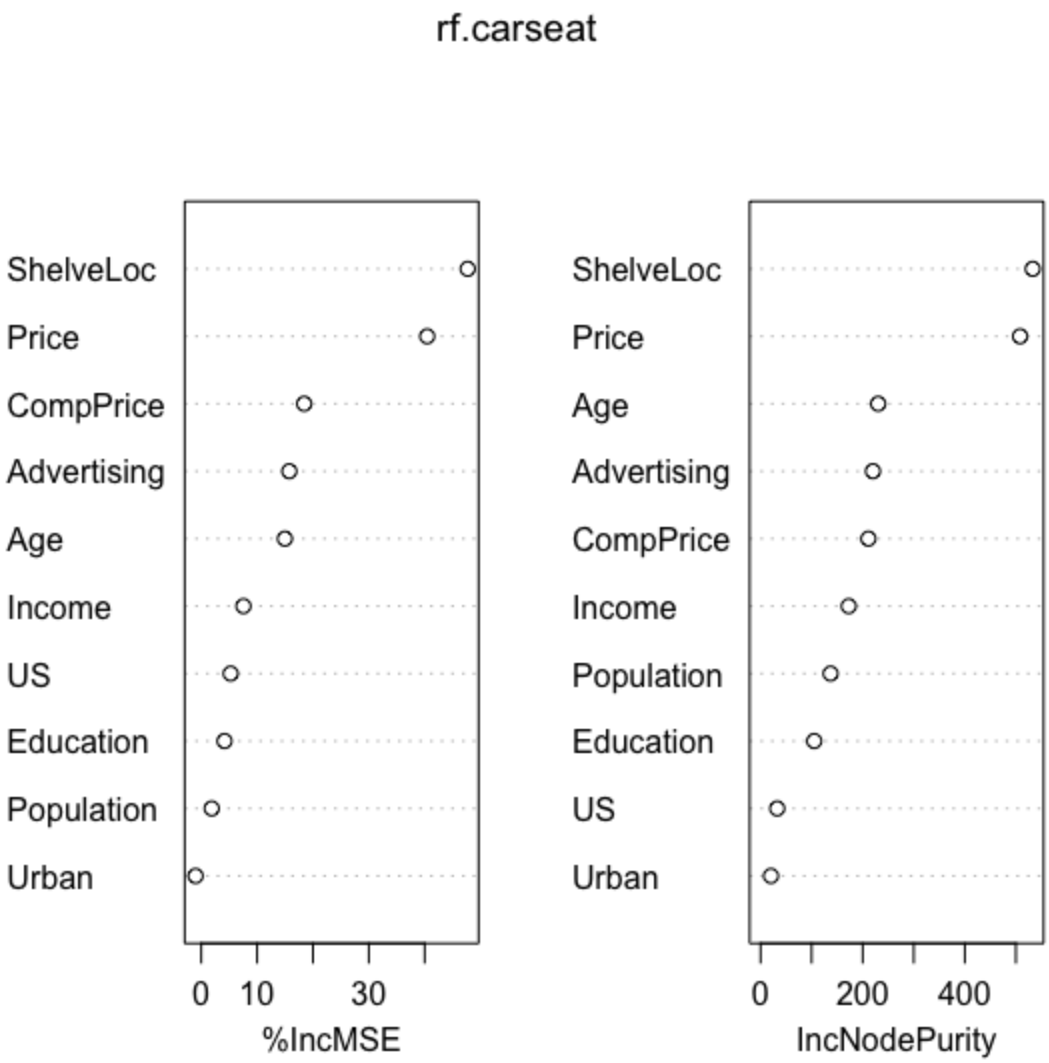


ii.

iii. Answer: The two most important variables are ShelveLoc and Price.

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.

i. Answer:2.89114



ii.

iii. **Answer:** For my random samples, the random forests did not yield much of an improvement over the bagging.

6. **Problem 6:**This question uses the Caravan data set in the ISLR2 package.
   a. Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.
   b. Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?
   c. (c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one?
      i. **Answer:** 0.2298137