1. You are asked to build a **decision Tree** model using the data Table 1 below for this question 1. Show your work. Stopping grow the tree if a node contains 2 or less points.

Heights	**Gender**	**Weight**
1.6	M	88
1.7	F	82
1.5	F	60
1.8	M	73
1.5	M	77
1.4	F	55
1.7	M	80

Table 1: Response variable is Weight. Two features with seven points for the following questions

.

   (a) Determine the root, indicate variable and the corresponding cutpoint, and corresponding leafs and their association SSR.

   (b) Determine the second split, indicate the variable and the cutpoint, and corresponding leafs and their associate SSR.

   (c) Finish up the decision tree. what are the leafs and their corresponding RSS for each region?

   (d) depict the whole tree.

   (e) What is the weight for male who has height 1.45m?

   (f) What is the sequence of the turning parameters α_T based on the (d)? Show your work.

2. Please refer to dataset Boston which is in package MASS for the question. The median home value, medv, is the response and the rest are predictors.

   The packages used for the question: tree; randomForest; gbm.

   (a) Randomly split Boston dataset into equal two parts, one for training and the other for the testing datasets. You will use the training dataset to fit the model and the test dataset for the test error for the following questions for the 2. Please set random seed to set.seed(1).

   (b) Build a decision tree.

      i. Fit a regression tree to the training set. Plot the tree, and interpret the results, what important variables are? how are they reflect in your tree?

      ii. What test MSE (mean square error) do you obtain?

      iii. Use cross-validation to determine the optimal size of the subtree. Plot Deviance against size.

      iv. Prune the tree and plot the pruned tree.

v. Calculate the test error for the pruned tree.

(c) Build a Bagging (ntree=500) model (to the training dataset that you create in (a)) to predict medv on the test dataset and report the test MSE. What are first three important variables?

(d) Build a random forest (with ntree=500, use the default value for mtry by omitting mtry) model to predict medv on the test dataset and report the test MSE.

(e) Build a Boosting (with n.trees=1000).

i. Build a boosting model with interaction.depth = 4 and default shrinkage. Please report the test MSE.

ii. Determine the optimal interaction.depth d by a for loop limiting the range of d from 2 to 10. Please use the default shrinkage parameter. Report the the test MSE.

iii. Determine the optimal shrinkage parameter by a for loop limiting the range of λ (shrinkage parameter) from 0.001 to 0.2. Use the interaction.depth value you found in (d).ii. above. Report the test MSE.

iv. Compare the test MSEs in (e).i. - (e).iii. above.

3. You are asked to build a **gradient boost** model using the data Table 1 in Question 1 above. Show your work.

Assume that: the interaction.depth $d = 2$ (or 3 leaves);
Shrinkage $\lambda = 0.1$; and
number of trees built $B = 3$

(a) Create the first tree. What is the residuals?

(b) Create the second tree (based on the the residuals from (a)).

i. Determine the root for the second tree.

ii. Determine the second internal node for the second tree.

iii. Determine the new leaves, new prediction and the new residuals for the first second tree.

(c) Create the third tree (based on the the residuals you found in (b).iii) above. Repeats (b).i - iii. above for the third tree.

(d) Assume that the whole gradient boost tree is consist of three trees. Draw the tree along with the average weight on each leaf.

(e) What is the weight for male who has height 1.45m? Compare this result with the result you obtained in (d).