Demonstration of use of regression trees

Use the Boston housing data. Get the data from MASS library of R.

1. Split the data set into a training and a test set, with the test set size being 20% of the number of observations in the dataset. Fit a regression tree to the training data with 'medv' as the response and all other variables as predictors.
2. Calculate and report the training MSE and test MSE.
3. Apply cross-validation to the training set to determine the optimal tree size with the weakest link pruning.
4. Produce a plot of the average cross-validated MSE of the sequence of pruned trees vs effective αs (cost complexity parameters). Which effective α corresponds to the lowest cross-validated MSE (in the training set)?
5. Plot the number of nodes and tree-depths of the sequence of pruned trees vs effective αs.
6. Plot the training and test MSE of the sequence of pruned trees vs effective αs.
7. Fit the optimally chosen tree to the train data.
8. Report the depth, number of nodes, training, and test MSE of the optimal tree.
9. Compare the performance of the pruned and the original tree with a multiple linear regression model based on training and test MSE.
10. Repeat the entire task using the R^2-score as a performance metric of the fitted regression model.
11. Perform the entire task on the California Housing data set of sklearn.