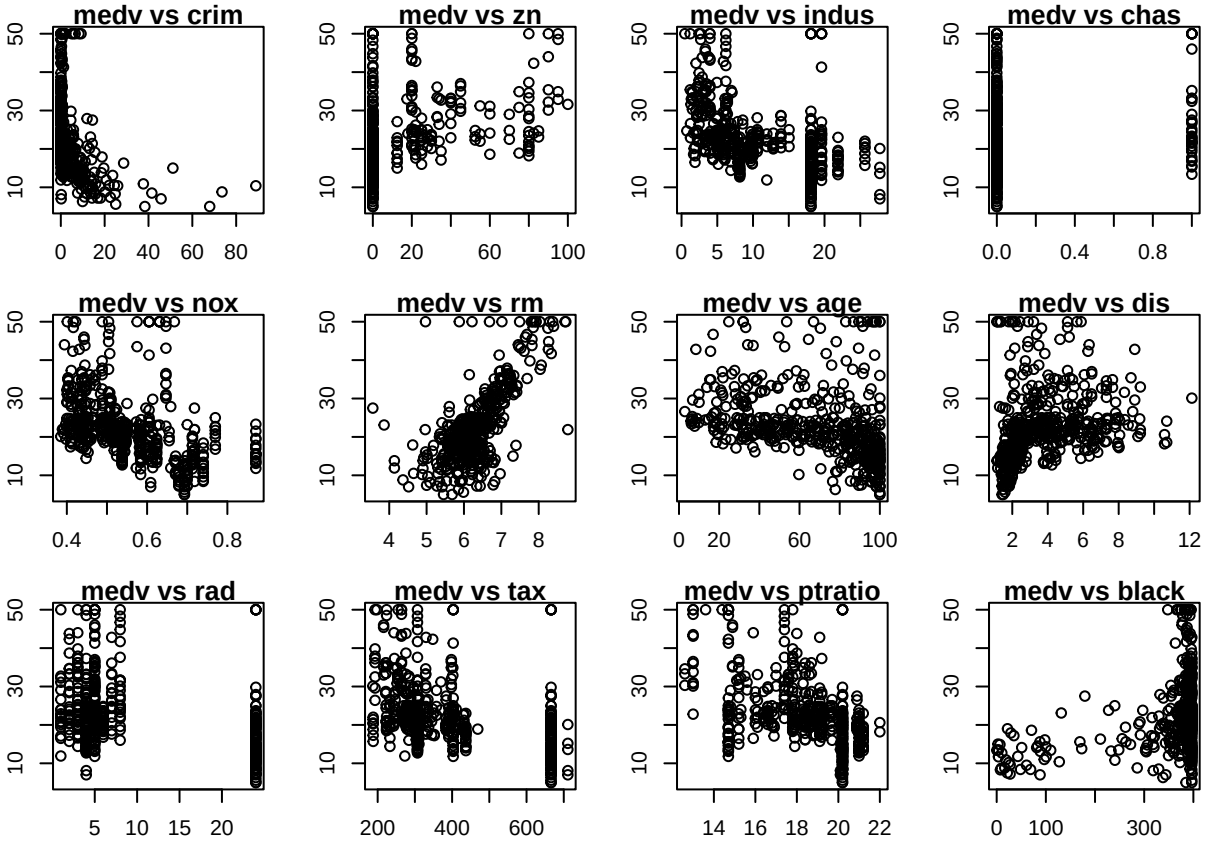
Homework 7

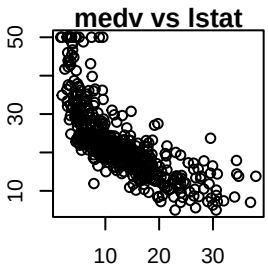
Alex Ojemann

2023-11-15

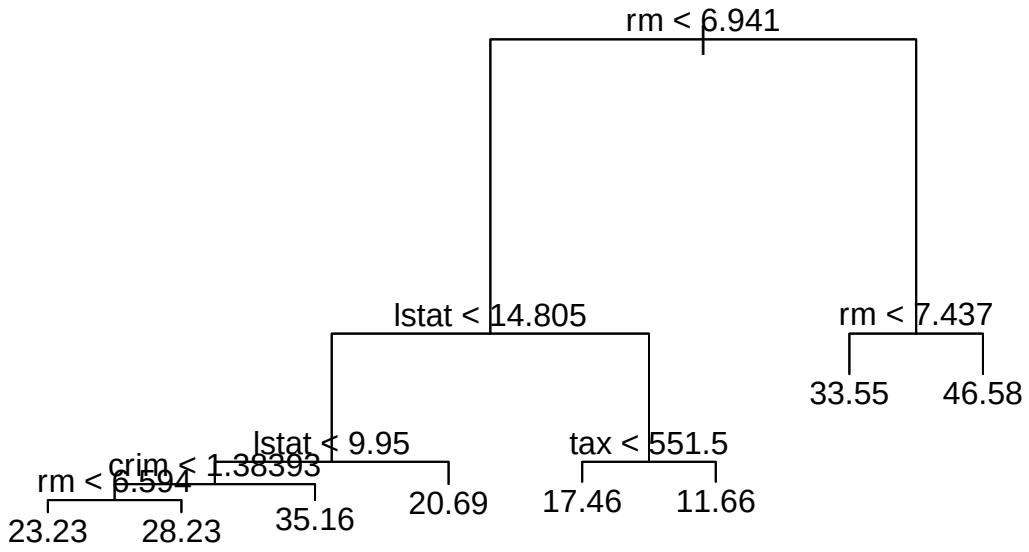
Problem 1

a/b. The median value decreases exponentially as the crime rate increases and as the percentage of the population which is lower status increases. It also appears to increase linearly as the number of rooms increases and decrease linearly as the age and pupil-teacher ratio increase at a lower rate. For all of the other predictor variables there is no apparent visual relationship.





c. In the tree shown below, the most important features are the number of rooms (3 splits) and the percentage of the population that is lower status which is defined as the proportion of adults without some high school education and proportion of male workers classified as laborers (2 splits). These make heuristic sense because houses with more rooms cost more and houses is lower income neighborhoods cost less, both of which are the case in the tree splits. The RMSE of the tree on the test set is 4.848972.



d. The RMSE on the test set using 500 bagged regression trees is 4.22869. There are no measures of feature importance in the tree package. The package doesn't even grant access to which feature the tree is splitting on at each node, so no measure of feature importance can be calculated.

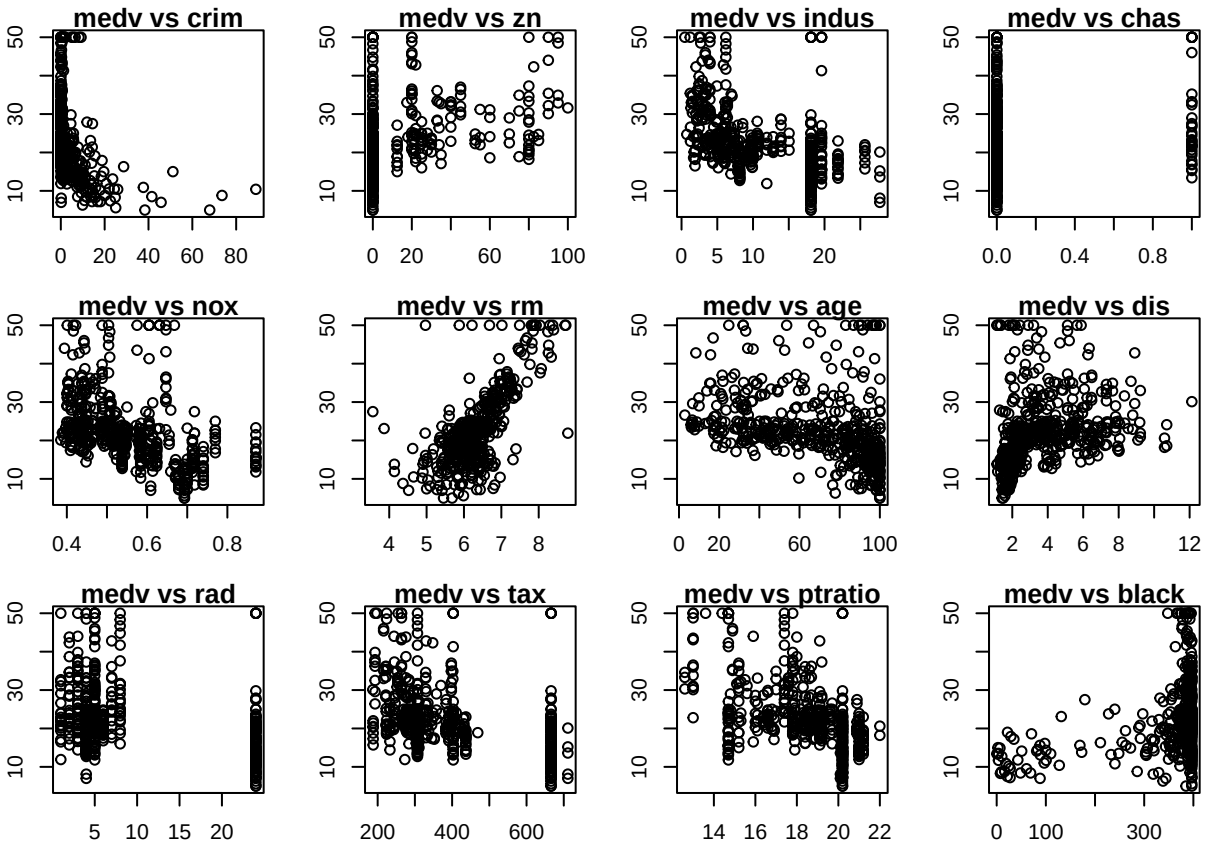
e. The RMSE on the test set using a random forest with 500 trees is 3.612896. Lstat and rm are by far the most important features, as they were in the single tree from part c. The full table of importances is shown below.

crim zn indus chas nox 9.5923642 0.6740533 6.9426407 1.3561459 7.0554987 rm age dis rad tax 33.4928644 3.8954976 6.1444968 2.5487704 4.5915755 ptratio black lstat 7.1166190 1.4370530 47.6766840

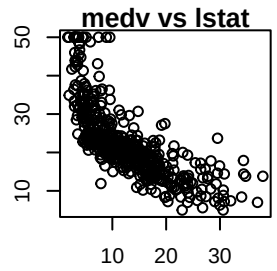
Appendix

```
#Part A
library(MASS)
data(Boston)
par(mfrow = c(3, 4), mar = c(3, 3, 1, 1))


for (col in names(Boston)) {
  if (col != 'medv') {
    plot(Boston[, col], Boston$medv, xlab = col, ylab = "medv", main = paste("medv vs", col))
  }
}
```

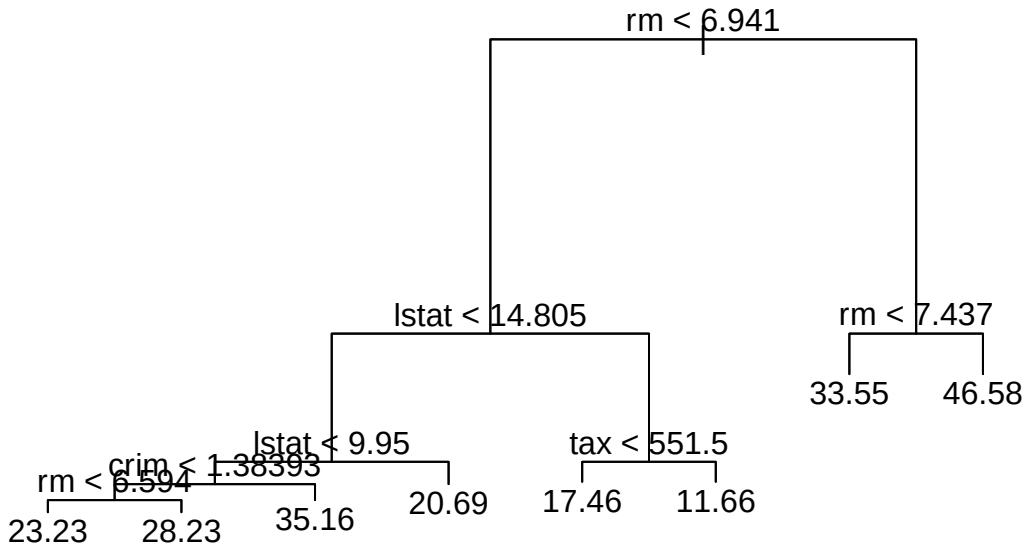



```
par(mfrow = c(1, 1))
```




```
#Part B
library(rsample)
set.seed(123)
split_data <- initial_split(Boston, prop = 0.5)
training_data <- training(split_data)
testing_data <- testing(split_data)
```

```
#Part c
library(tree)
fit <- tree(medv~.,data=training_data)
pred <- predict(fit,newdata=testing_data)
rmse <- sqrt(mean((testing_data$medv-pred)^2))
plot(fit)
text(fit)
```




```
cat("RMSE on the test set: ",rmse)
```

```
#Part d
N <- 500
PRED.boot <- matrix(nr=length(testing_data$medv),nc=N)

set.seed(180)
for(i in 1:N){
  bag.indices <- sample(1:dim(training_data)[1],size=dim(training_data)[1],replace=TRUE)
  out <- tree(medv~.,data=training_data[bag.indices,])
  PRED.boot[,i] <- predict(out,newdata=testing_data)
}
# average the predictions from the bootstrap-resampled data tree fits
PRED.bagged <- apply(PRED.boot,1,mean)

cat("RMSE on the test set: ",sqrt(mean( (testing_data$medv - PRED.bagged)^2 )))
```

```
#Part e
library(ranger)
fit <- ranger(medv~.,data=training_data,num.trees=500,importance="permutation")
pred <- predict(fit,data=testing_data)
cat("RMSE on the test set: ",sqrt(mean((testing_data$medv - pred$predictions)^2)))
print('')
print('Feature Importances:')
print(importance(fit))
```