Problem Set 10

1.

a.

Basically a decision tree is a type of statistical model that presents a number of attributes and gives them respective weights based on their historical data. You start with the "Best" attribute, or the one most statistically significant, and create branches that lead to other attributes. These lnodes have a weight and lead to other attributes based on their significance.

b.

Basically bagging is an algorithm aimed at minimizing the level of variance. By doing this it improves overall accuracy of a model. Source used: https://en.wikipedia.org/wiki/Bootstrap_aggregating

c.

Bootstrapping is when you have multiple subsets of a greater data set to make inferences. You do this by sampling the dataset with replacements. https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

D.

You take a sample of your predictors and you make multiple trees to map on your data set in order to make predictions.

2.

a.

```
library('ISLR')
Auto <- Auto

set.seed(1861)
```

b.

```
trainSize <- 0.75
trainInd <- sample(1:nrow(Auto), size = floor(nrow(Auto) * trainSize))
autoTrain <- Auto[trainInd,]
autoValidate <- Auto[-trainInd,]
str(autoTrain)
```

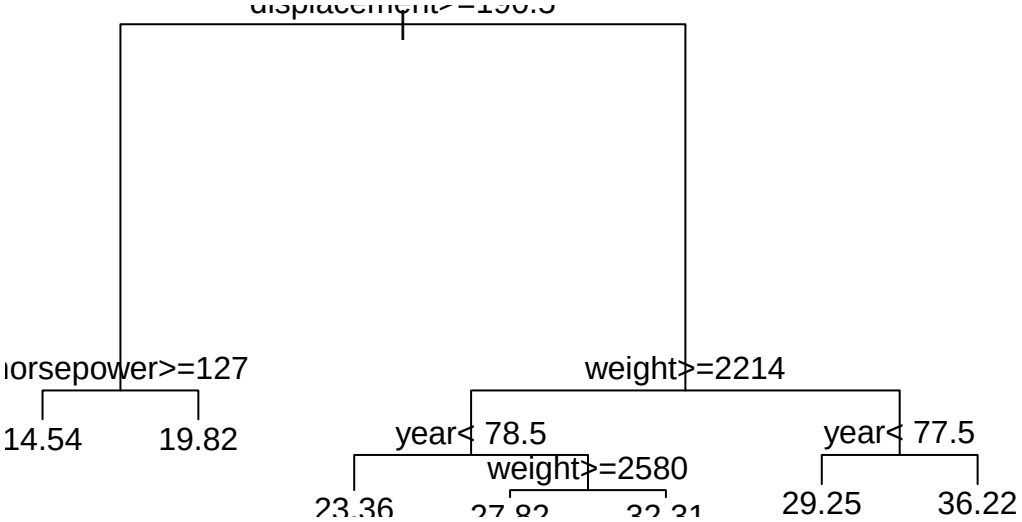
```
## 'data.frame':    294 obs. of  9 variables:
##  $ mpg         : num  44 26.8 25.5 12 18 32.3 34.1 33.5 27 33.7 ...
##  $ cylinders   : num  4 6 4 8 6 4 4 4 4 4 ...
##  $ displacement: num  97 173 140 400 250 97 86 85 101 107 ...
##  $ horsepower  : num  52 115 89 167 88 67 65 70 83 75 ...
##  $ weight      : num  2130 2700 2755 4906 3139 ...
##  $ acceleration: num  24.6 12.9 15.8 12.5 14.5 17.8 15.2 16.8 15.3 14.4 ...
##  $ year        : num  82 79 77 73 71 81 79 77 76 81 ...
##  $ origin      : num  2 1 1 1 1 3 3 3 2 3 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 301 205 155 128 151 261 173 94 25
```

c.

```
library(rpart)
rpart.mpg<- rpart(mpg~.-name, data=autoTrain)
```

d.

```
plot(rpart.mpg); text(rpart.mpg,pretty=0)
```



e.

The right side of the tree splits between year and proceeds. If displacement is greater than 190.5 and horsepower is above 127 then our predicted horsepower is 14.54. If the displacement is less than 190.5 and greater than 2214, and below 78.5 then our expected mpg is 23.36.

f.

```
MSE <- function(truth, predict) {mean((truth - predict)^2)}
MSE(predict(rpart.mpg, newdata = autoTrain), autoTrain$mpg)
```

```
## [1] 9.67135
```

```
MSE(predict(rpart.mpg, newdata = autoValidate), autoValidate$mpg)
```

```
## [1] 9.549045
```

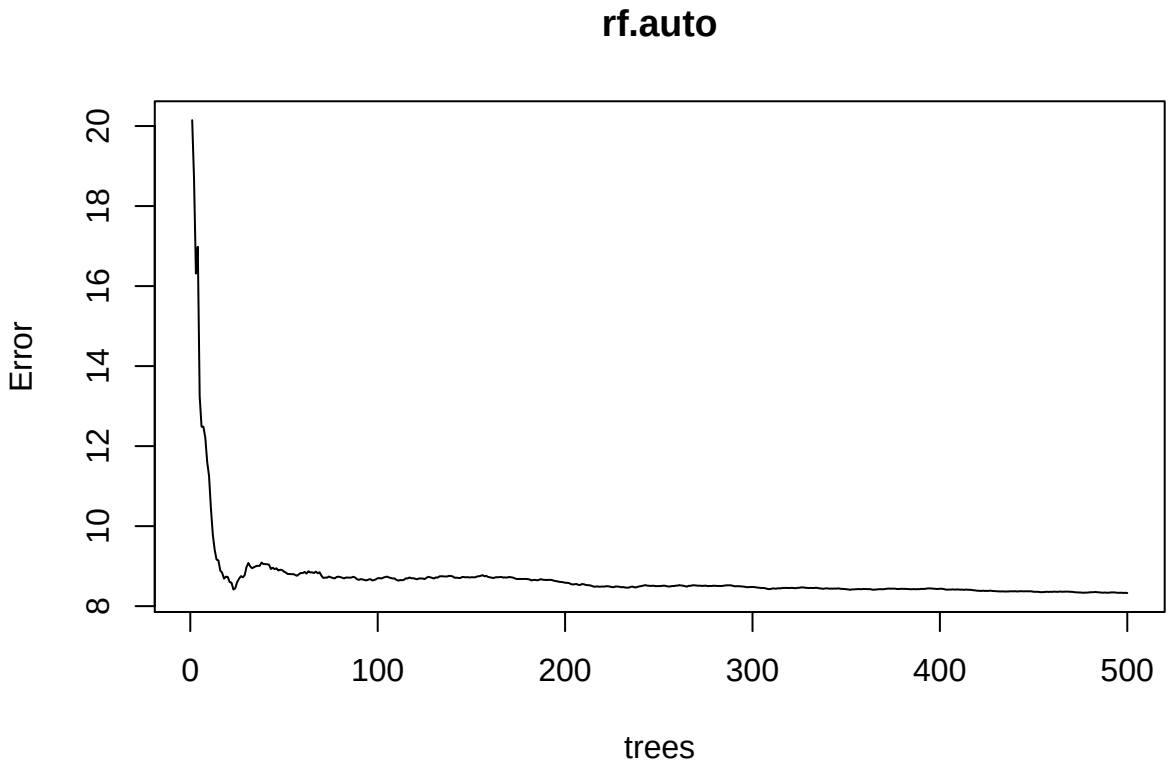
g.

```
#install.packages("randomForest")
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
rf.auto = randomForest(mpg~.-name,data=autoTrain,mtry = 3)
plot(rf.auto)
```



h.

The mtry parameter is an attempt to find the lowest mse. The parameters in place are sort of the boundary lines that the the mtry value can be. Sort of like limitations.

i.

```
MSE(predict(rf.auto, newdata = autoTrain), autoTrain$mpg)
```

```
## [1] 1.772985
```

```
MSE(predict(rf.auto, newdata = autoValidate), autoValidate$mpg)
```

```
## [1] 5.787653
```

j.

Having multiple trees is better than just having one. By having multiple trees we can average out the results of our predicitons and get a better prediction than just one really good tree. So I would say the second model is better.