Random Forest Exercise on Boston Dataset

```
suppressMessages(library(ipred))
```

```
## Warning: package 'ipred' was built under R version 3.5.2
```

```
suppressMessages(library(randomForest))
```

```
## Warning: package 'randomForest' was built under R version 3.5.2
```

```
suppressMessages(library(tidyverse))
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
suppressMessages(library(ggplot2))
suppressMessages(library(tree))
```

```
## Warning: package 'tree' was built under R version 3.5.2
```

```
suppressMessages(library(ISLR))
```

```
## Warning: package 'ISLR' was built under R version 3.5.2
```

```
suppressMessages(library(adabag))
```

```
## Warning: package 'adabag' was built under R version 3.5.2
```

```
## Warning: package 'rpart' was built under R version 3.5.2
```

```
## Warning: package 'caret' was built under R version 3.5.2
```

```
## Warning: package 'foreach' was built under R version 3.5.2
```

```
## Warning: package 'doParallel' was built under R version 3.5.2
```

```
## Warning: package 'iterators' was built under R version 3.5.2
```

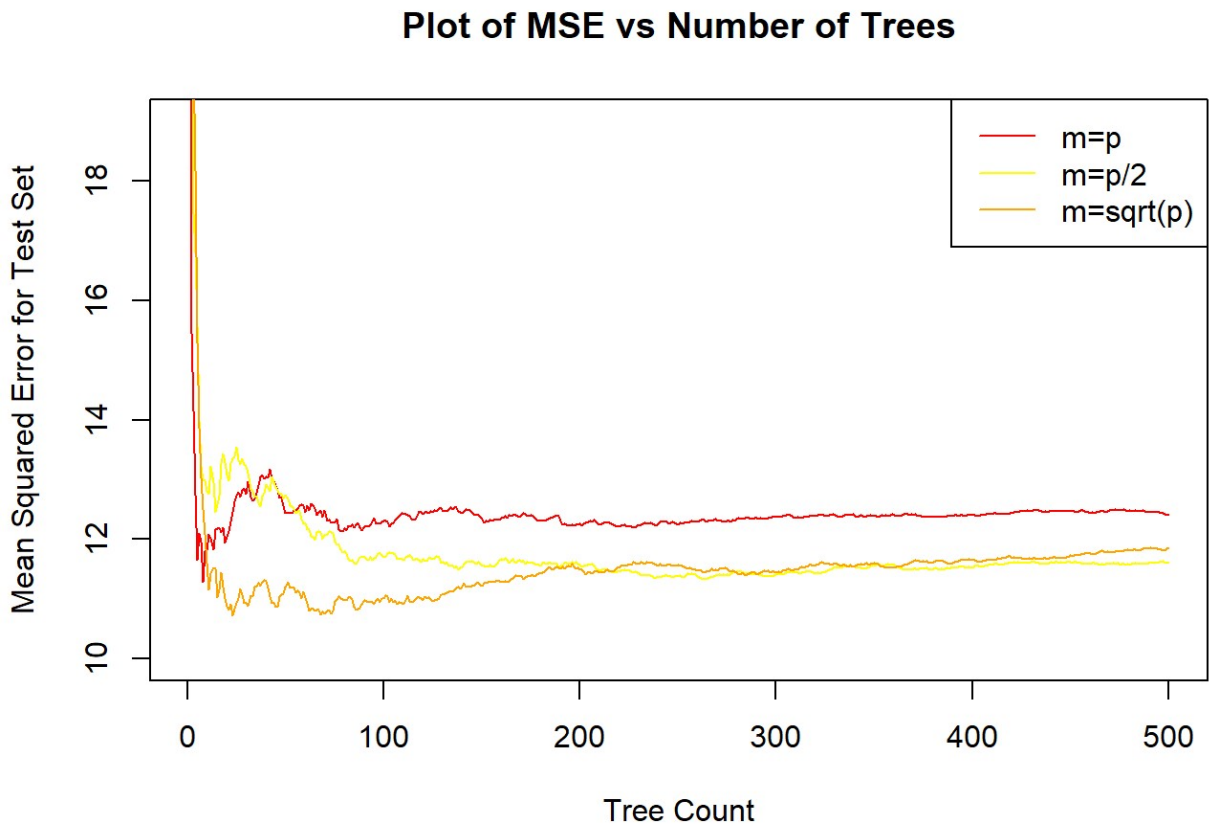
```
suppressMessages(library(rpart))
suppressMessages(attach(Carseats))
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
plot(1:500, rf.boston.a$test$mse, col = "red", type = "l", xlab = "Tree Count",
    ylab = "Mean Squared Error for Test Set", main = "Plot of MSE vs Number of Trees",
ylim = c(10, 19))
lines(1:500, rf.boston.a.half$test$mse, col = "yellow", type = "l")

lines(1:500, rf.boston.a.squared$test$mse, col = "orange", type = "l")
legend("topright", c("m=p", "m=p/2", "m=sqrt(p)"), col = c("red", "yellow", "orange"),
    cex = 1, lty = 1)
```



In the case where all the variables are included we can see from the plot that the test MSE is distinctly higher. This is in comparison to the case where we take half the number variables or square root of the number of variables. From the plot we see that the MSE (Mean Squared Error) for each singular tree starts of at a high value of approximately 16. As the added number of trees increases to about 110, we see that the MSE tapers off and becomes more constant.