Question2

Yufei Yin

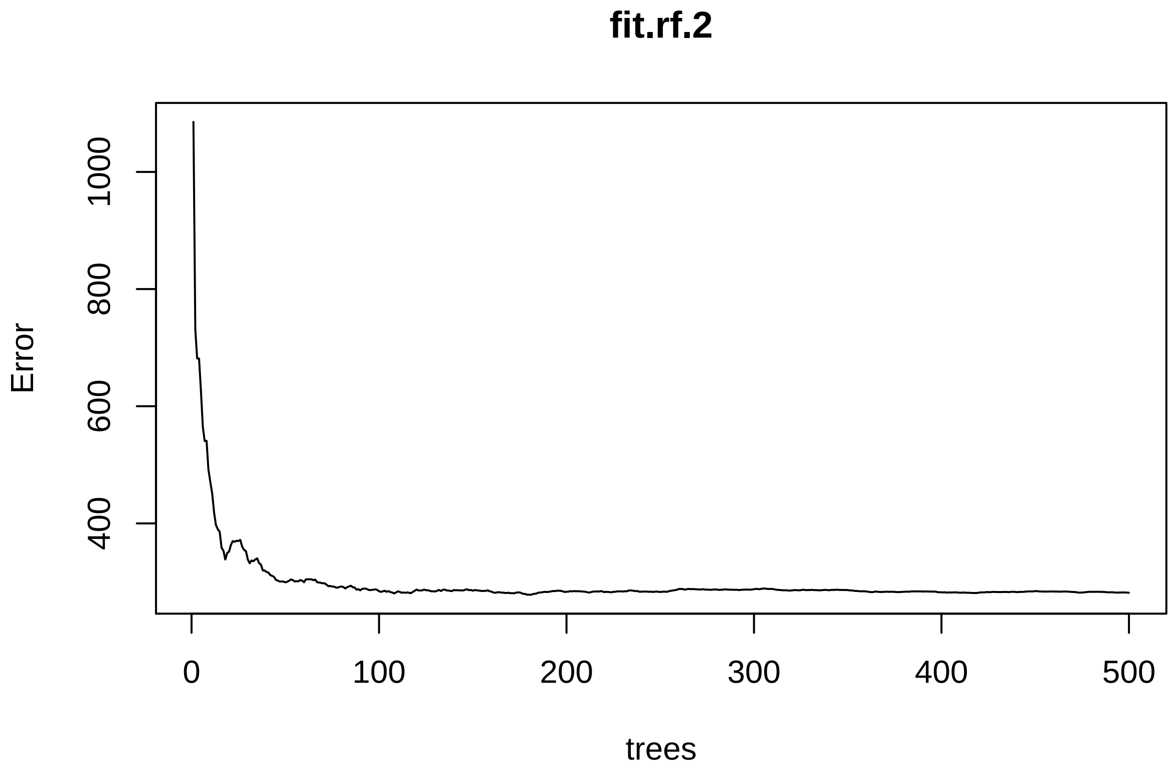
2.

(a)

```
library(randomForest)
set.seed(452)
```

```
###############################
### Import and process data ###
###############################

### Import and clean the air quality data
data("airquality")
AQ.raw = na.omit(airquality[,1:4])

### Construct new variables
AQ = AQ.raw
AQ$TWcp = with(AQ.raw, Temp * Wind)
AQ$TWrat = with(AQ.raw, Temp / Wind)

########################
### Helper Functions ###
########################

### Create function to compute MSPEs
get.MSPE = function(Y, Y.hat){
  return(mean((Y - Y.hat)^2))
}
```

```
fit.rf.2 = randomForest(Ozone ~ ., data = AQ, importance = T)

# OOB error
plot(fit.rf.2)
```



```
get.MSPE(AQ$Ozone, predict(fit.rf.2))
```

```
## [1] 281.5672
```

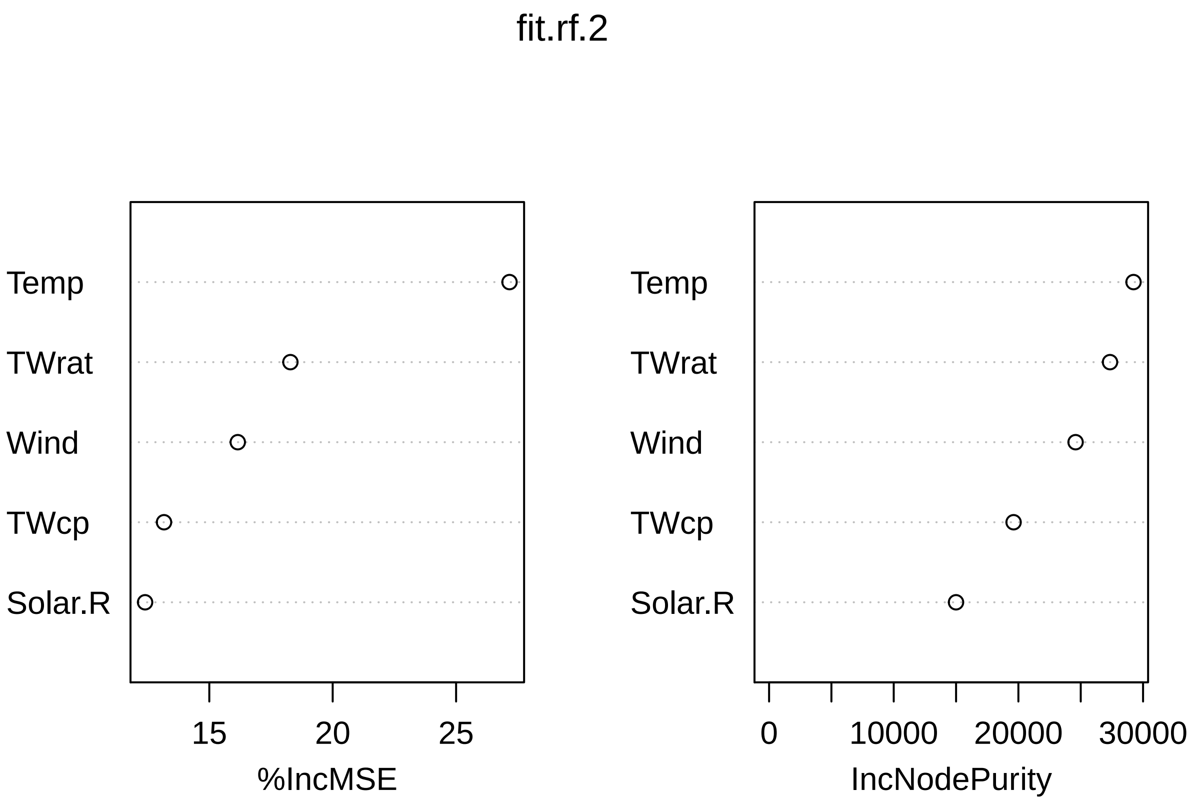
The previous OBB error is 282.3182, and the OBB error from RF analysis adding the two engineered features is 281.5672. It just improved a little bit.

(b)

```
importance(fit.rf.2)
```

```
##          %IncMSE IncNodePurity
## Solar.R 12.39780      14995.75
## Wind    16.15382      24588.00
## Temp    27.16516      29235.17
## TWcp    13.16517      19614.69
## TWrat   18.28606      27348.87
```

```
varImpPlot(fit.rf.2)
```



Both methods suggest that the ratio of temperature and wind speed is particular important, but the product of temperature and wind speed is not that important.