Question1

Yufei Yin

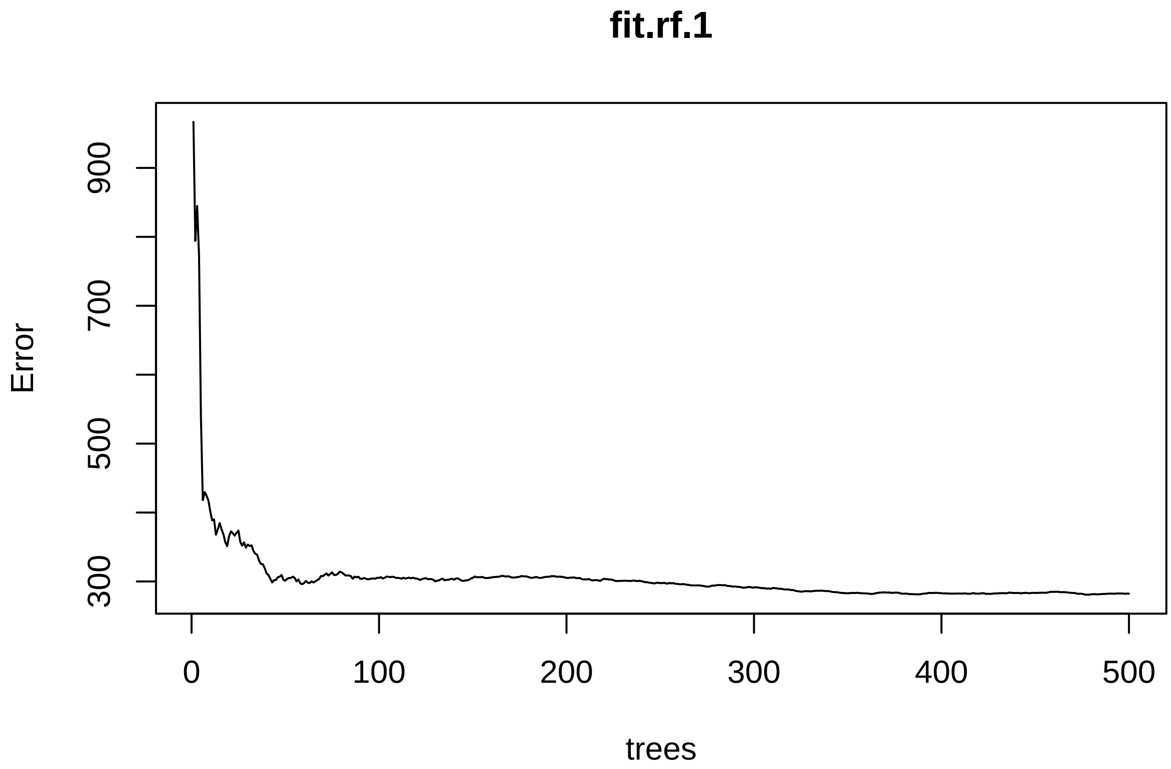
1.

(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.1 = randomForest(Ozone ~ Temp + Wind + Solar.R, data = AQ, importance = T)

# OOB error
plot(fit.rf.1)
```



```
get.MSPE(AQ$Ozone, predict(fit.rf.1))
```

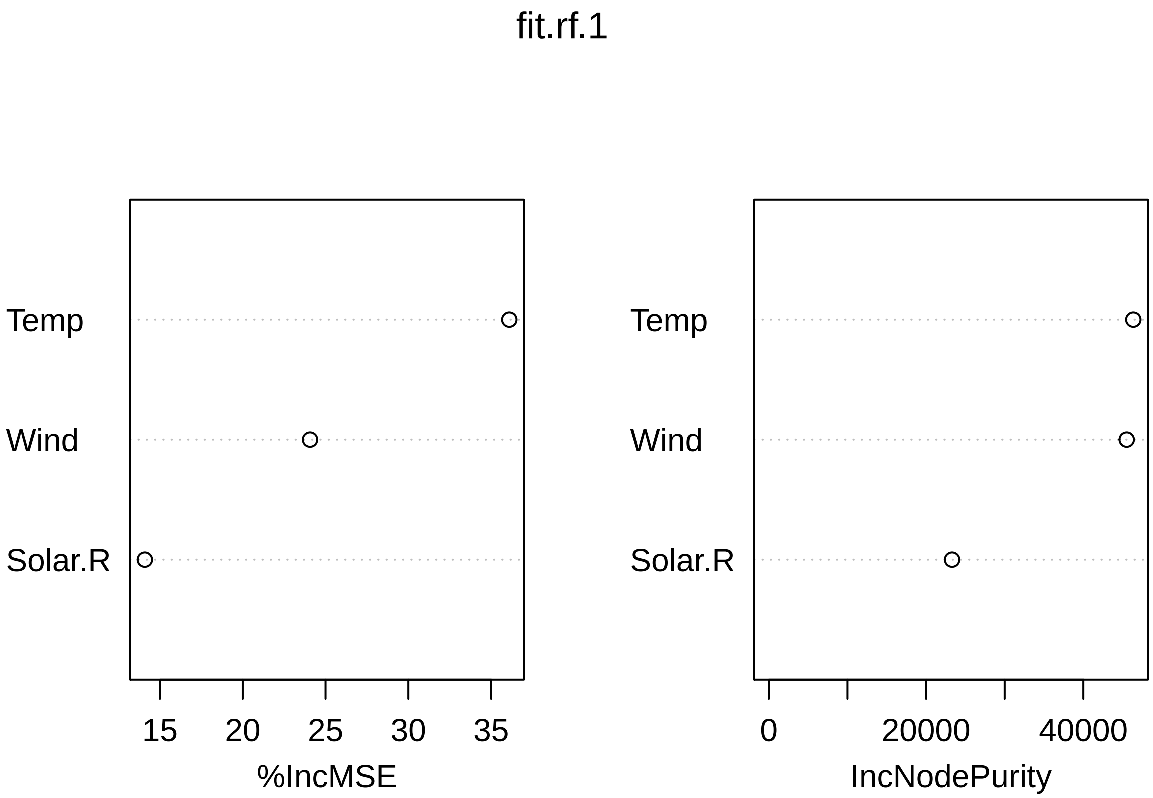
```
## [1] 282.3182
```

(b)

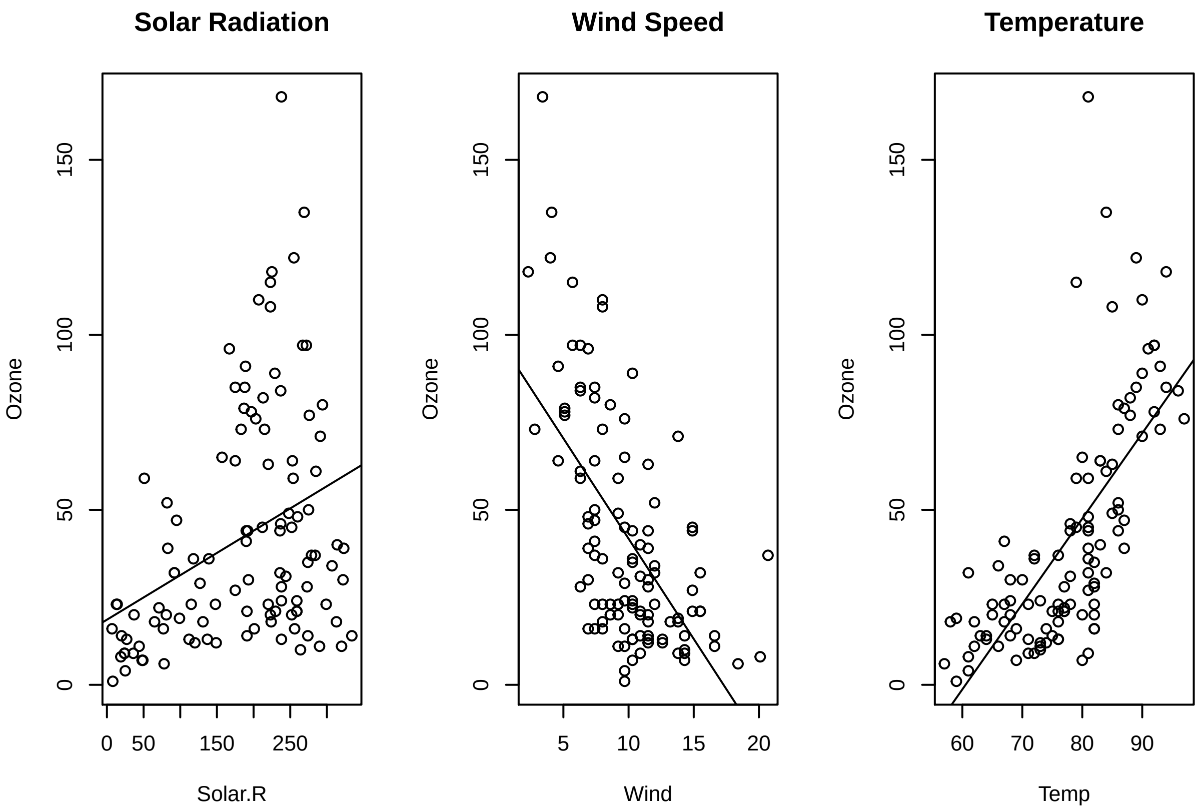
```
importance(fit.rf.1)
```

```
##           %IncMSE IncNodePurity
## Temp    36.09553      46355.77
## Wind    24.06412      45516.97
## Solar.R 14.08747      23297.06
```

```
varImpPlot(fit.rf.1)
```



Based on the result of mean decrease in RSS and mean decrease in accuracy, both methods suggest that variables Temp and Wind are more significant compared to Solar radiation. This is similar to the analysis we did in Lecture 2. We fit simple linear regression for these 3 variables respectively. The line for temperature provides a strong fit to the data. The line for wind speed provides a moderate-strong fit, and the line for solar radiation provides a weak-moderate fit.


Solar Radiation
Wind Speed
Temperature
Ozone
0
50
100
150
0
50
150
250
Solar.R
5
10
15
20
Wind
60
70
80
90
Temp