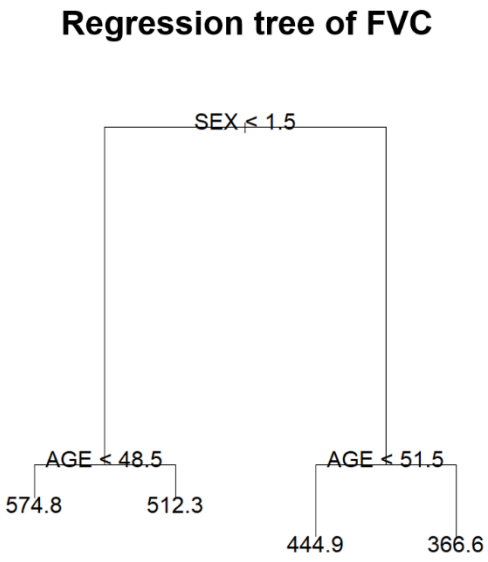
The data set "**FHS_data.csv**" includes variables SEX (1=male, 2 = female), Age, Smoke (=1 for current smoker), FVC (score of pulmonary function), SPF (systolic blood pressure), T2D (1=history of diabetes) and other variables. The data were used to generate the regression tree in the figure below.




```
Regression tree:
snip.tree(tree = tree.0, nodes = c(5L, 7L, 4L, 6L))
Variables actually used in tree construction:
[1] "SEX" "AGE"
Number of terminal nodes:  4
Residual mean deviance:  6194 = 15690000 / 2533
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-309.800  -48.300    1.213    0.000   48.700  271.100
node), split, n, deviance, yval
      * denotes terminal node

1) root 2537 29370000 466.5
  2) SEX < 1.5 1038  8712000 543.7
    4) AGE < 48.5 521  4095000 574.8 *
    5) AGE > 48.5 517  3604000 512.3 *
  3) SEX > 1.5 1499 10210000 413.1
    6) AGE < 51.5 890  4424000 444.9 *
    7) AGE > 51.5 609  3565000 366.6 *
```

Write down the 4 rules described by the 4 branches of the tree (i.e. the rules defining each group and their prediction).

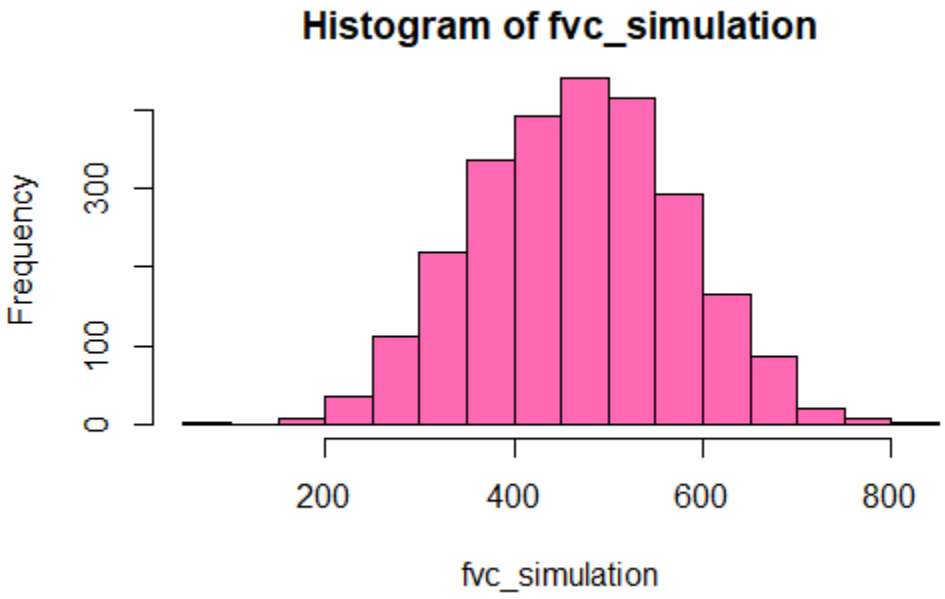
Branch 1: The average FVC of male patients younger than 48.5 years is 574.8
Branch 2: The average FVC of male patients older than 48.5 years is 512.3
Branch 3: The average FVC of female patients younger than 51.5 years is 444.9
Branch 4: The average FVC of female patients older than 51.5 366.6

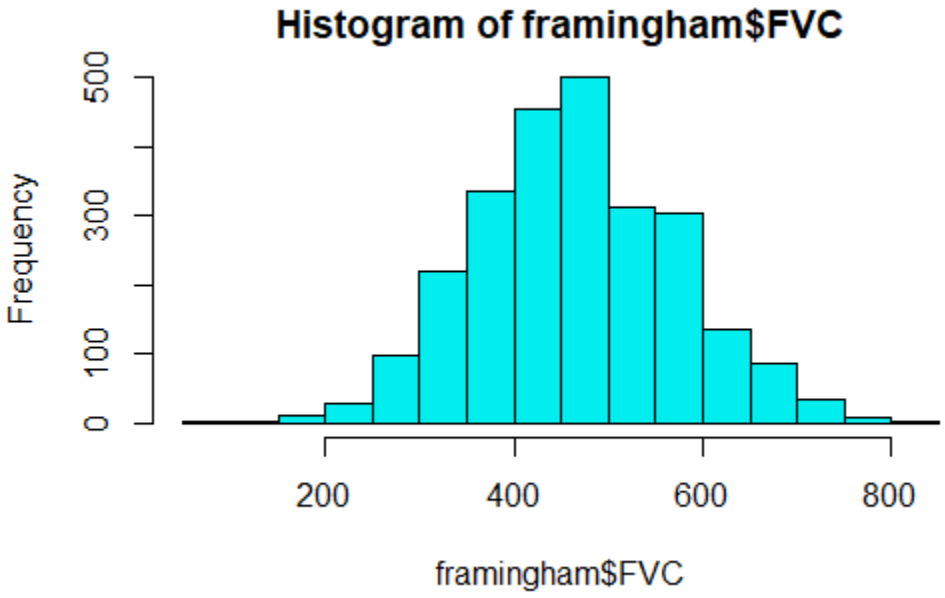
What is the estimate of the variance of FVC from this tree?

estimated variance of FVC = 6194

Simulate a sample of **FVC values** from the displayed tree above (i.e. use the estimated mean, sd and the structure information. Let's assume the sd is the same for all four leaves so you may use the information from previous question) using the existing variable values SEX and AGE in the data set "FHS_data.csv" as predictors. Use the same sample size (i.e. we will generate 2,537 FVC values from this simulation) reported in the original tree/branch, and assume that FVC follows a normal distribution (i.e. use ***rnorm()*** function).

Compare the distribution of the simulated data for FVC and the real FVC values in the data set.

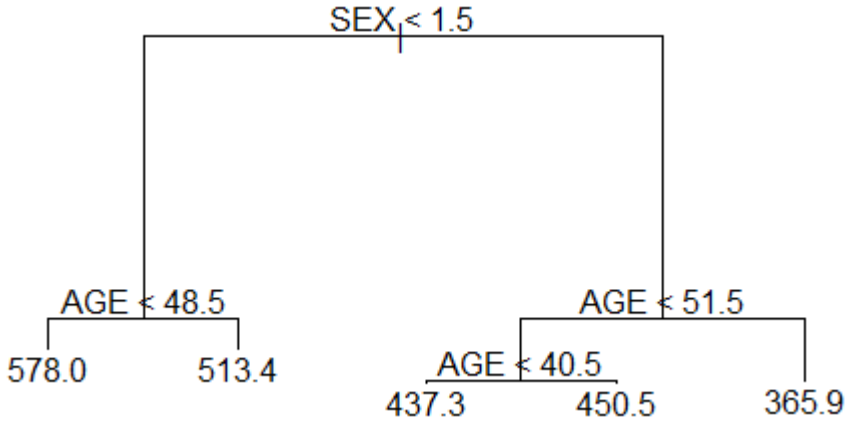




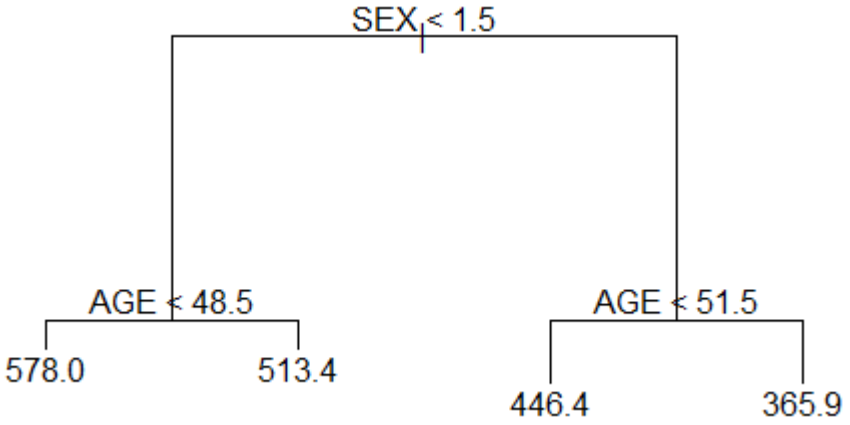
The distributions of the FVC scores of the simulated and real data are both approximately normally distributed.

Use the simulated data for FVC and the real data for the covariates SEX, AGE, Smoke, SPF, T2D to construct a new regression tree, and control the depth of the tree using the option control=tree.control(nobs=nrow(data.frame), mindev = 0.001) where "data.frame" is the data frame created from the data set "FHS_data.csv". Use the **cv.tree()** function to select the best regression tree. Build the best regression tree and plot it. Include the plot in your write-up.

Using the simulated FVC scores and the real data for all other variables, only the variables sex and age were used to create 5 terminal nodes. The residual mean deviance was 5999.



The tree was pruned by cross-validation using k=5. The lowest deviance belonged to trees with 4 and 5 terminal nodes, so the final pruned tree contains 4 terminal nodes.



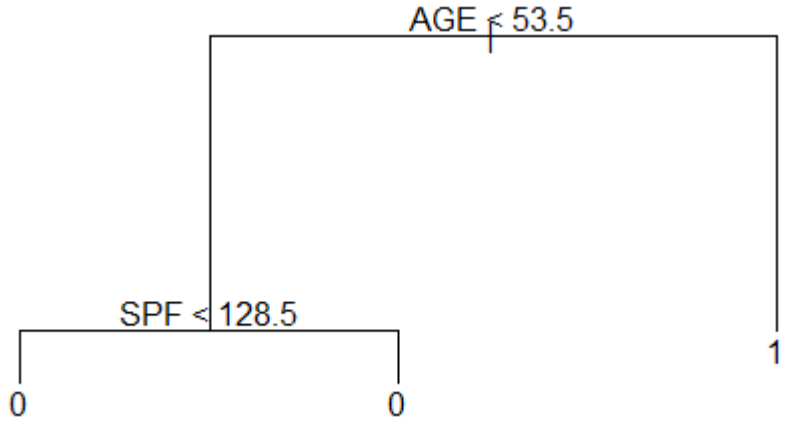
Branch 1: The average FVC of male patients younger than 48.5 years is 578.8
Branch 2: The average FVC of male patients older than 48.5 years is 513.4
Branch 3: The average FVC of female patients younger than 51.5 years is 446.4
Branch 4: The average FVC of female patients older than 51.5 365.9

Compare the tree used to simulate the data (i.e. the displayed tree above) with the tree generated using your simulated data.

The pruned tree from the simulated tree was very similar to the tree given in the problem set. Both trees used the same exact rules to create 4 leaves, and the prediction for FVC scores at each node nearly identical.

Build a classification tree to predict death within 20 years, using the function tree() with default parameters and plot it. How many branches are represented by the tree?

The variables age and systolic blood pressure were used to create 3 branches. The residual mean deviance was 1.113.



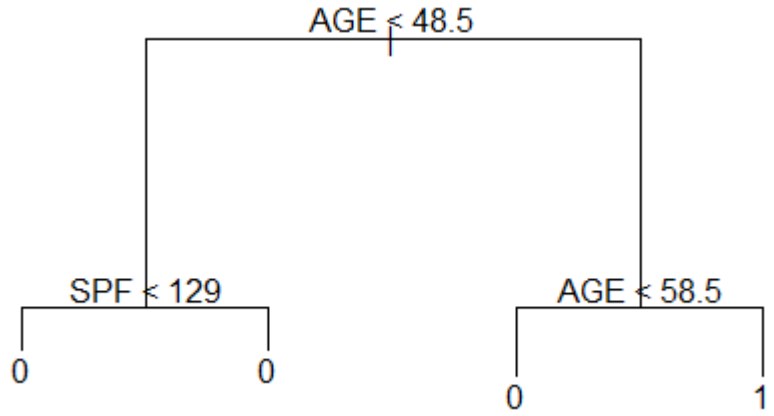
What is the probability of death within 20 years for a person aged > 60 years based on the tree you constructed?

The probably of death within 20 years for someone over the age of 60 is 0.5567.

What is the predicted value for a person aged > 60 years based on the tree you constructed?

It's predicted that someone over the age of 60 will be dead within 20 years of measurements.

Generate a classification tree to predict death within 20 years using a training set of 2/3 of the observations and then predict the outcome in the test set comprising the remaining 1/3 observations. What is the misclassification error of the prediction? (Hints: using predict() function with type ="class")



The misclassification error of the prediction is 25.53%.