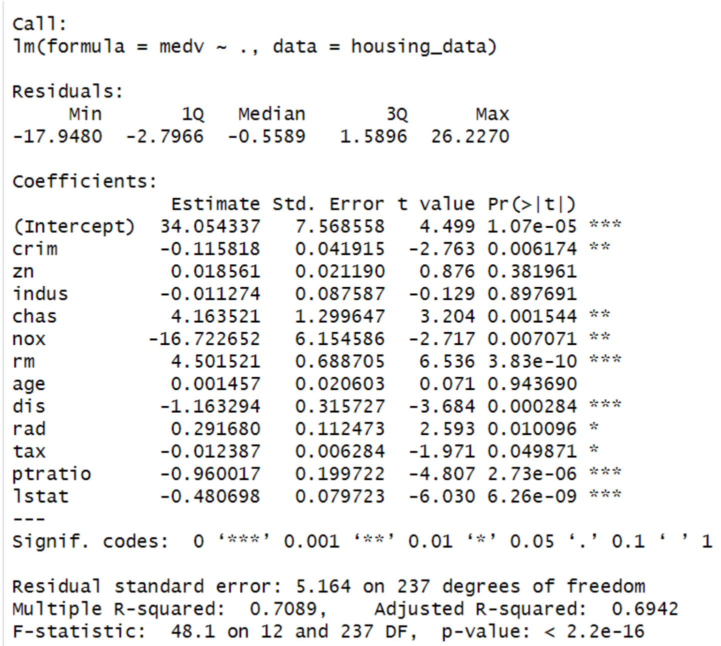
Question 1 (1.1 - 1.7) Analyse a regression dataset

The file housing.2023.csv contains the data that we will use for this question. This dataset is a modified version of the Boston housing data which was collected to study house prices in the metropolitan region of Boston. In this data set, each observation represents a particular suburb from the Boston region. The outcome, medv, is the median value of owner-occupied homes in $1,000 in the suburb. The variables are summarised in Table 1. The data consists of p = 12 variables measured on n = 250 suburbs. We are interested in discovering which predictors are good determinants of housing price, and how these variables effect the median house price.

1.1

Fit a multiple linear model to the housing data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with median house value, and why? Which three variables appear to be the strongest predictors of housing price, and why?

Output:



```
Call:
lm(formula = medv ~ ., data = housing_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.9480  -2.7966  -0.5589   1.5896  26.2270 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  34.054337   7.568558   4.499 1.07e-05 ***
crim         -0.115818   0.041915  -2.763 0.006174 ** 
zn            0.018561   0.021190   0.876 0.381961    
indus        -0.011274   0.087587  -0.129 0.897691    
chas          4.163521   1.299647   3.204 0.001544 ** 
nox         -16.722652   6.154586  -2.717 0.007071 ** 
rm            4.501521   0.688705   6.536 3.83e-10 ***
age           0.001457   0.020603   0.071 0.943690    
dis          -1.163294   0.315727  -3.684 0.000284 ***
rad           0.291680   0.112473   2.593 0.010096 *  
tax          -0.012387   0.006284  -1.971 0.049871 *  
ptratio      -0.960017   0.199722  -4.807 2.73e-06 ***
lstat        -0.480698   0.079723  -6.030 6.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.164 on 237 degrees of freedom
Multiple R-squared:  0.7089,    Adjusted R-squared:  0.6942 
F-statistic:  48.1 on 12 and 237 DF,  p-value: < 2.2e-16
```

Explanation:

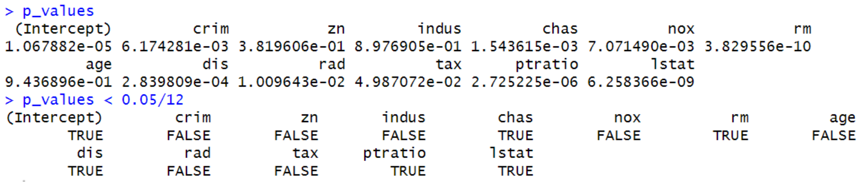
We can see that the predictors crim, chas, nox, rm, dis,rad, tax, ptratio, and lstat seem to be borderline (p<0.01) associated and are probably our best guesses, based on the p-values, of which variables might be associated with median house value. P-values of 0.01 means that the chance of seeing an association just by chance, if there was no association at the population level is around 1%, which is not super likely. Hence, it is strong evidence that against null that the coefficient for these variables is 0.

The stronger predictors in this case will be rm, dis, ptratio and lstat. These variables have the p-value < 0. However, the variable, dis, has a higher value to the other 3 variables, which is 0.000284. Thus, the 3 variables that appear to be the strongest predictors of housing price are rm,ptratio and lstat as these 3 variables have the smallest p-values (smaller than 0.)

1.2

How would your assessment of which predictors are associated change if you used the Bonferroni procedure with α = 0.05?

Output:



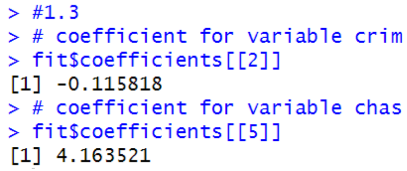
```
> p_values
 (Intercept)         crim           zn        indus         chas          nox           rm 
1.067882e-05 6.174281e-03 3.819606e-01 8.976905e-01 1.543615e-03 7.071490e-03 3.829556e-10 
         age          dis          rad          tax      ptratio        lstat 
9.436896e-01 2.839809e-04 1.009643e-02 4.987072e-02 2.725225e-06 6.258366e-09 
> p_values < 0.05/12
(Intercept)        crim          zn       indus        chas         nox          rm         age 
       TRUE       FALSE       FALSE       FALSE        TRUE       FALSE        TRUE       FALSE 
        dis         rad         tax     ptratio       lstat 
       TRUE       FALSE       FALSE        TRUE        TRUE 
```

There are 12 predictors in total. Alpha = 0.05.
Bonferroni threshold = 0.05/12 = 0.004166667. The predictors that are possibly associated with median house value : **chas, rm, dis, ptratio** and **lstat** have a p-value small enough to pass the Bonferroni threshold.

1.3

Describe what effect the per-capita crime rate (crim) appears to have on the median house price. Describe what effect a suburb having frontage on the Charles River has on the median house price for that suburb.

Output:



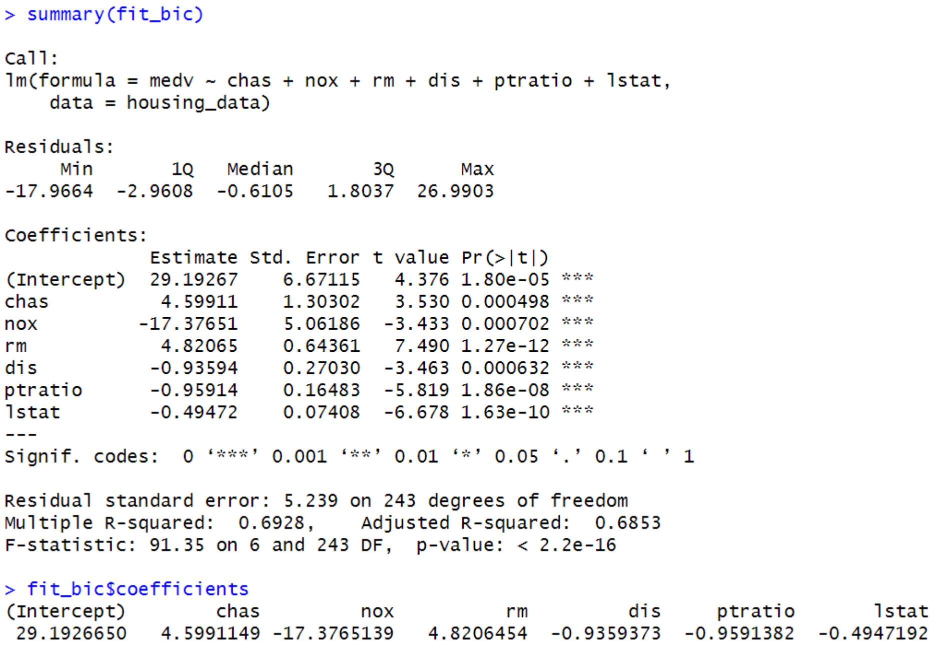
```
> #1.3
> # coefficient for variable crim
> fit$coefficients[[2]]
[1] -0.115818
> # coefficient for variable chas
> fit$coefficients[[5]]
[1] 4.163521
```

Explanation:
Intercept =34.054337
The coefficient of the variable crim = -0.115818.
The coefficient of the variable chas = 4.163521.
Therefore with the information above, we know that for every growth of the crime rate per capita,the median house price decreases by 0.115818 units.
For every suburb that fronts on the Charles River, the median house price increases by 4.163521. In other words, being in a suburb that fronts on the Charles River is associated with an increase in the median house price by approximately 4.163521 units.

1.4

Use the stepwise selection procedure, with the BIC criterion (use direction="both"), to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning.

Output:



```
> summary(fit_bic)

Call:
lm(formula = medv ~ chas + nox + rm + dis + ptratio + lstat, 
    data = housing_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.9664  -2.9608  -0.6105   1.8037  26.9903 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.19267    6.67115   4.376 1.80e-05 ***
chas          4.59911    1.30302   3.530 0.000498 ***
nox         -17.37651    5.06186  -3.433 0.000702 ***
rm            4.82065    0.64361   7.490 1.27e-12 ***
dis          -0.93594    0.27030  -3.463 0.000632 ***
ptratio      -0.95914    0.16483  -5.819 1.86e-08 ***
lstat        -0.49472    0.07408  -6.678 1.63e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.239 on 243 degrees of freedom
Multiple R-squared:  0.6928,	Adjusted R-squared:  0.6853 
F-statistic: 91.35 on 6 and 243 DF,  p-value: < 2.2e-16

> fit_bic$coefficients
(Intercept)        chas         nox          rm         dis     ptratio       lstat 
 29.1926650   4.5991149 -17.3765139   4.8206454  -0.9359373  -0.9591382  -0.4947192 
```

Explanation:

Regression Equation

E[medv] = 29.1926650 + 4.5991149*chas -17.3765139*nox +4.8206454*rm -0.9359373*dis -0.9591382*ptratio -0.494719*lstat

1.5

If a council wanted to try and improve the median house value in their suburb, what does the model that we found in Question 1.4 suggest they could try and do?

chas;
Having frontage on the Charles River (chas = 1) is associated with an increase in the median house price of approximately 4.5991149 units.Therefore,the council is suggested to promote the development and preservation of riverfront areas, creating attractive parks, recreational spaces, and amenities along the riverbanks, thus enhance the overall appeal of the suburb.

Nox;
For every unit increase in the nitrogen oxide concentration (nox, parts per million), the median house price is estimated to decrease by approximately 17.3765139 units.This is a negative relationship, indicating that higher nitrogen oxide levels are associated with lower house prices. Therefore, the council is suggested to implement stricter environmental regulations, promoting clean energy, and reducing industrial emissions in order to reduce air pollution and improve air quality.

Rm;
For every additional room (rm) per dwelling, the median house price is estimated to increase by approximately 4.8206454 units.This is a positive relationship, showing that more rooms are associated with higher house prices.Therefore,the council are suggested to build housing properties that have more rooms.

Dis;
For every unit increase in the weighted distance to five Boston employment centres (dis), the median house price is estimated to decrease by approximately 0.9359373 units.This is a negative relationship,indicating that greater distances to Boston employment centres are associated with lower house prices.Therefore, the council can work to promote the creation of job opportunities and also improve public transportation to make it easier for residents to access nearby employment centres.

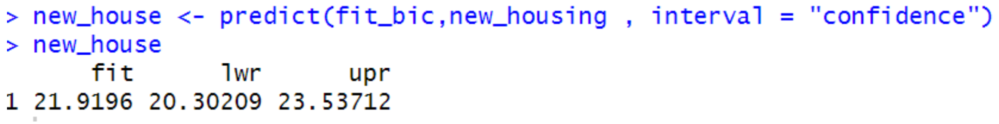
Ptratio;
For every unit increase in the pupil-teacher ratio (ptratio), the median house price is estimated to decrease by approximately 0.9591382 units.This is a negative relationship,indicating that the higher pupil-teacher ratio is associated with lower house prices.Therefore,the council is suggested to invest more in the education system, reducing class sizes and providing a better learning environment. This can attract parents looking for good schools for their children and thus improve the suburb's overall desirability.

Lstat;
For each unit increase in the percentage of lower status population (lstat), the median house price is estimated to decrease by approximately 0.4947192 units. This is a negative relationship, indicating that higher percentages of lower-status population are associated with lower house prices. Therefore, the council is suggested to focus on economic and social development initiatives in order to improve the economic well-being and social status of residents in the particular area.

1.6

Use the model found in Question 1.4 to predict the median house price for this suburb. Provide a 95% confidence interval for this prediction.

Output:



```
> new_house <- predict(fit_bic,new_housing , interval = "confidence")
> new_house
      fit      lwr      upr
1 21.9196 20.30209 23.53712
```

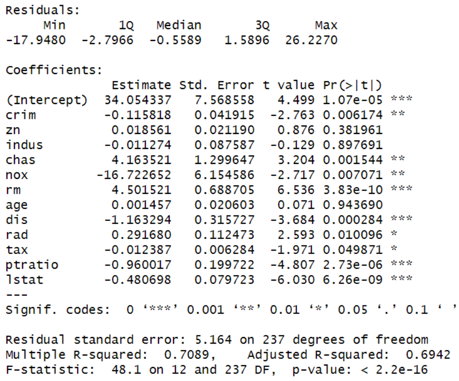
Explanation:
The median of the house price for this new suburb is 21.9196 units($21,919.60). We are 95% confident that the median house price for this suburb falls within the range of approximately 20.30209 units ($20,302.09) to 23.53712 units($23,537.12).

Count using regression equation found in 1.4
E[medv] = 29.1926650 + 4.5991149*0 -17.3765139*0.573 +4.8206454*6.03 -0.9359373*2.505 -0.9591382*21 -0.494719*7.88 = 21.9196

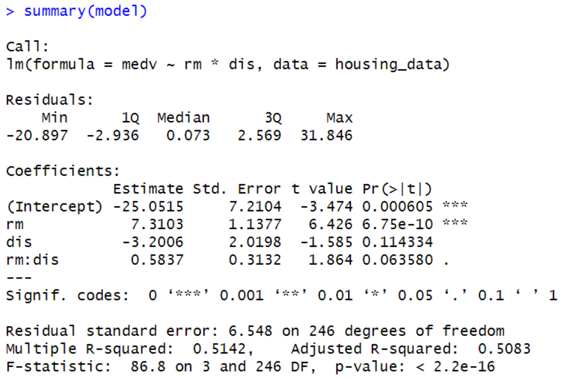
1.7

A friend who works at a local council suggests that they believe there is possibly an interaction effect between the number of rooms a dwelling has and its distance to one of the employment centres. Assess whether you think this is the case, and what effect it has on the model?

Output:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-17.9480  -2.7966  -0.5589   1.5896  26.2270 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  34.054337   7.568558   4.499 1.07e-05 ***
crim         -0.115818   0.041915  -2.763 0.006174 ** 
zn            0.018561   0.021190   0.876 0.381961    
indus        -0.011274   0.087587  -0.129 0.897691    
chas          4.163521   1.299647   3.204 0.001544 ** 
nox         -16.722652   6.154586  -2.717 0.007071 ** 
rm            4.501521   0.688705   6.536 3.83e-10 ***
age           0.001457   0.020603   0.071 0.943690    
dis          -1.163294   0.315727  -3.684 0.000284 ***
rad           0.291680   0.112473   2.593 0.010096 *  
tax          -0.012387   0.006284  -1.971 0.049871 *  
ptratio      -0.960017   0.199722  -4.807 2.73e-06 ***
lstat        -0.480698   0.079723  -6.030 6.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.164 on 237 degrees of freedom
Multiple R-squared:  0.7089,	Adjusted R-squared:  0.6942 
F-statistic:  48.1 on 12 and 237 DF,  p-value: < 2.2e-16
```



```
> summary(model)

Call:
lm(formula = medv ~ rm * dis, data = housing_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.897  -2.936   0.073   2.569  31.846 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -25.0515     7.2104  -3.474 0.000605 ***
rm            7.3103     1.1377   6.426 6.75e-10 ***
dis          -3.2006     2.0198  -1.585 0.114334    
rm:dis        0.5837     0.3132   1.864 0.063580 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.548 on 246 degrees of freedom
Multiple R-squared:  0.5142,	Adjusted R-squared:  0.5083 
F-statistic:  86.8 on 3 and 246 DF,  p-value: < 2.2e-16
```

Explanation:

The coefficient for rm is 7.3103, and it's highly significant (p-value <0). This suggests that the number of rooms has a strong positive effect on median house prices. As the number of rooms increases, the median house price increases by 7.3103 units.

The coefficient for dis is -3.2006, but it's not statistically significant (p-value is large). This implies that the distance to employment centres does not affect median house prices.

The coefficient for the interaction (rm:dis) is 0.5837 with the p-value 0.063580, which is only slightly above the typical significance level of 0.05. This indicates that there is very weak evidence for a potential interaction effect on median house prices.

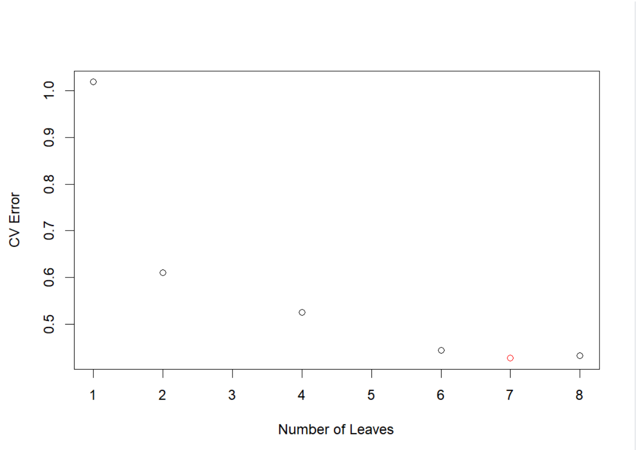
Question 2 (2.1 - 2.9) Analyse the data in heart.train.2023.csv.

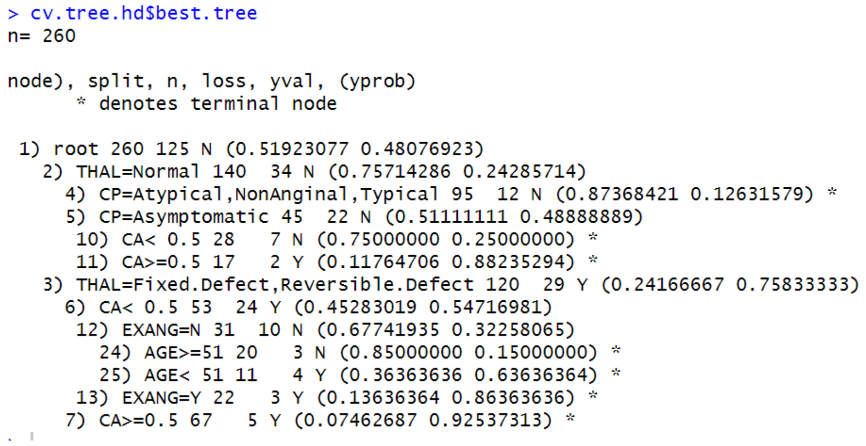
In this dataset, each observation represents a patient at a hospital that reported showing signs of possible heart disease. The outcome is presence of heart disease (HD), or not, so this is a classification problem.

2.1

Using the techniques you learned in Studio 9, fit a decision tree to the data using the tree package. Use cross-validation with 10 folds and 5,000 repetitions to select an appropriate size tree. What variables have been used in the best tree? How many leaves (terminal nodes) does the best tree have?

Output:





```
> cv.tree.hd$best.tree
n= 260

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 260 125 N (0.51923077 0.48076923)
   2) THAL=Normal 140  34 N (0.75714286 0.24285714)
     4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
     5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)
      10) CA< 0.5 28   7 N (0.75000000 0.25000000) *
      11) CA>=0.5 17   2 Y (0.11764706 0.88235294) *
   3) THAL=Fixed.Defect,Reversible.Defect 120  29 Y (0.24166667 0.75833333)
     6) CA< 0.5 53  24 Y (0.45283019 0.54716981)
      12) EXANG=N 31  10 N (0.67741935 0.32258065)
        24) AGE>=51 20   3 N (0.85000000 0.15000000) *
        25) AGE< 51 11   4 Y (0.36363636 0.63636364) *
      13) EXANG=Y 22   3 Y (0.13636364 0.86363636) *
     7) CA>=0.5 67   5 Y (0.07462687 0.92537313) *
```

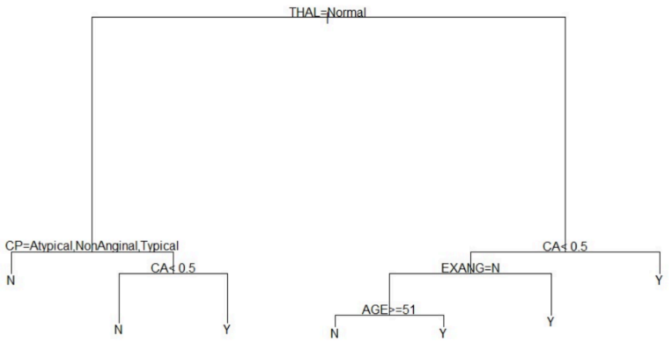
Explanation:
The fitted decision tree shows that there are 7 leaf nodes and 5 variables, the variables use are: THAL, CP, CA, EXANG,AGE.

The CV plot shows that the best tree size is 7. So,the tree should have 7 leaf nodes.

2.2

Plot the tree found by CV, and discuss clearly and thoroughly in plain English what it tells you about the relationship between the predictors and heart disease.

Output:



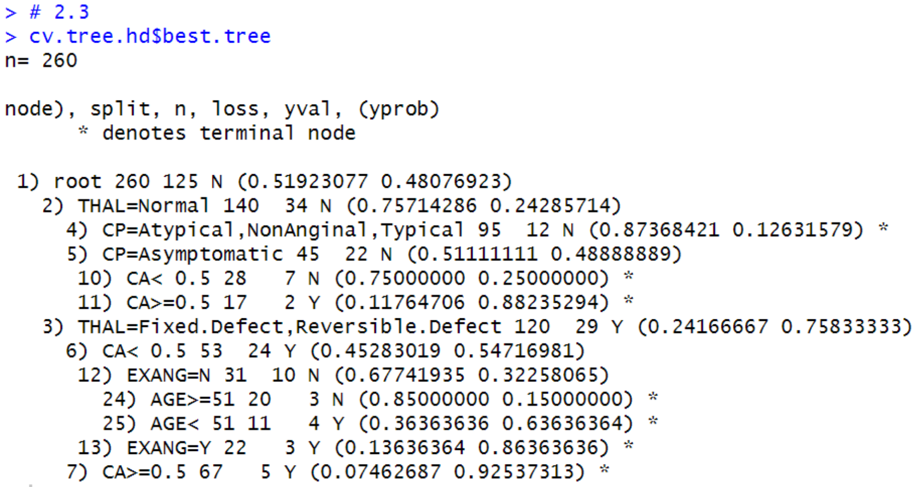
Explanation:

The conditions required for the tree to predict that the presence of heart disease is:

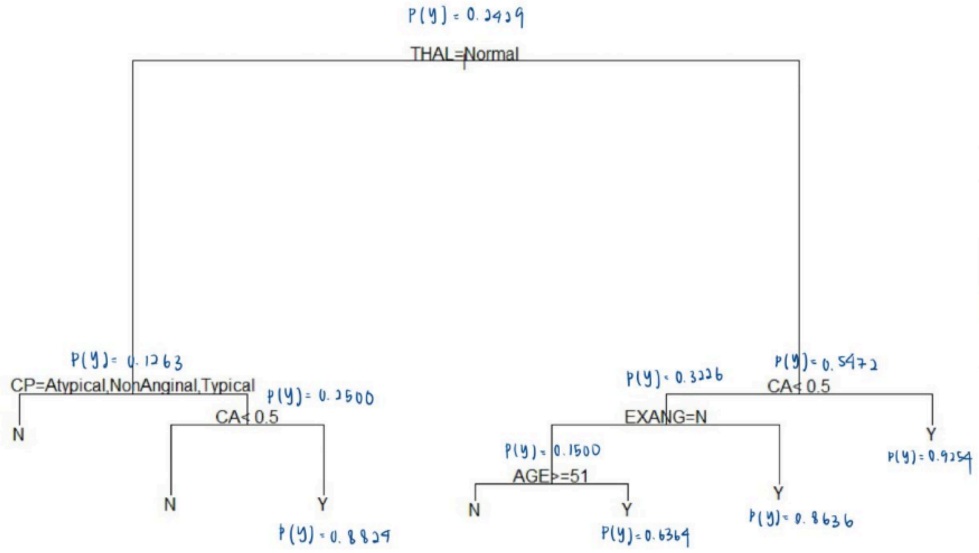
- THAL(Thallium Scanning Result) is Normal, CP (Chest pain type) is Asymptomatic, CA (Number of major vessels coloured by fluoroscopy) is more than or equal to 0.5.
- THAL(Thallium Scanning Result) is Fixed.Defect or Reversible.Defect,CA (Number of major vessels coloured by fluoroscopy) is less than 0.5, without EXANG(Exercise induced angina?) ,AGE (Age of patient in years) less than 51.
- THAL(Thallium Scanning Result) is Fixed.Defect or Reversible.Defect,CA (Number of major vessels coloured by fluoroscopy) is less than 0.5, with EXANG(Exercise induced angina?).
- THAL(Thallium Scanning Result) is Fixed.Defect or Reversible.Defect,CA (Number of major vessels coloured by fluoroscopy) is more than or equal 0.5.

2.3 Classification

Output:



```
> # 2.3
> cv.tree.hd$best.tree
n= 260

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 260 125 N (0.51923077 0.48076923)
   2) THAL=Normal 140  34 N (0.75714286 0.24285714)
     4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
     5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)
      10) CA< 0.5 28   7 N (0.75000000 0.25000000) *
      11) CA>=0.5 17   2 Y (0.11764706 0.88235294) *
   3) THAL=Fixed.Defect,Reversible.Defect 120  29 Y (0.24166667 0.75833333)
     6) CA< 0.5 53  24 Y (0.45283019 0.54716981)
      12) EXANG=N 31  10 N (0.67741935 0.32258065)
        24) AGE>=51 20   3 N (0.85000000 0.15000000) *
        25) AGE< 51 11   4 Y (0.36363636 0.63636364) *
      13) EXANG=Y 22   3 Y (0.13636364 0.86363636) *
     7) CA>=0.5 67   5 Y (0.07462687 0.92537313) *
```



2.4

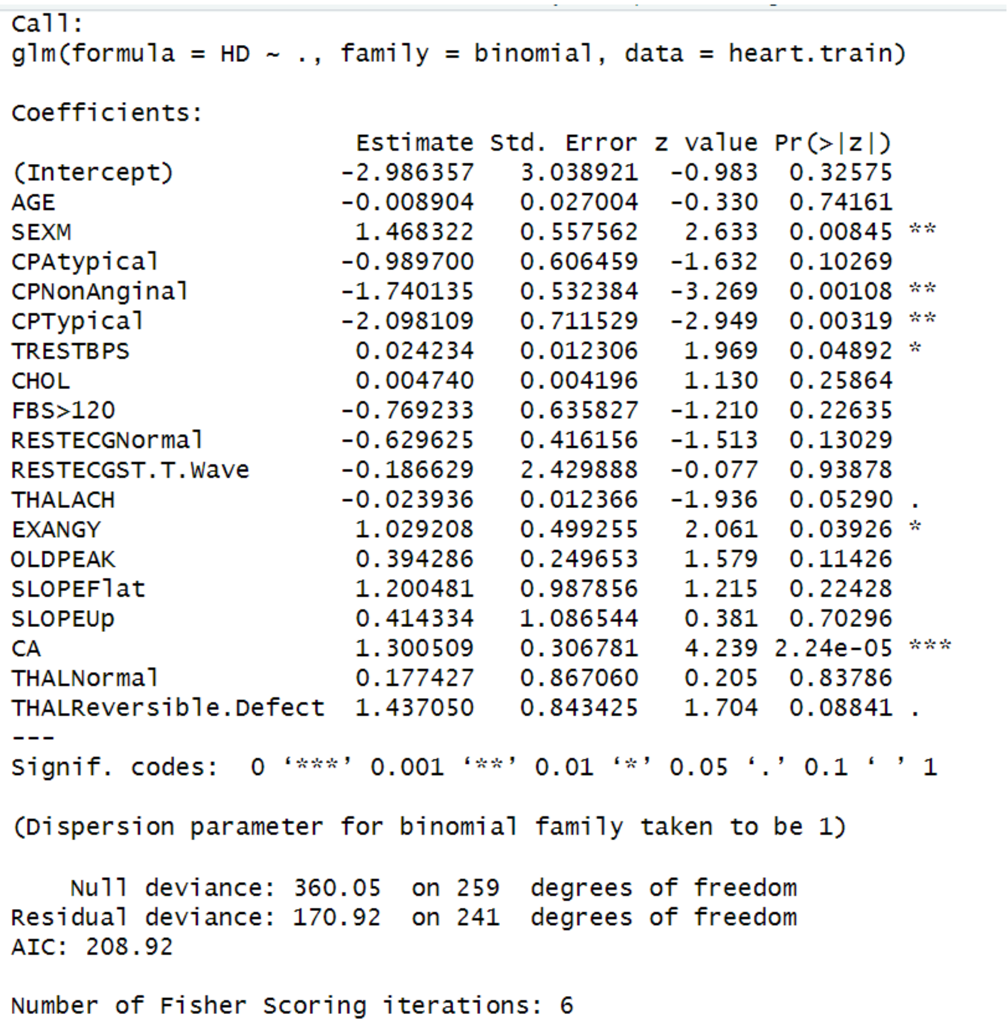
Which predictor combination results in the highest probability of having heart-disease?

The predictor combination results in the highest probability of having heart-disease is: THAL(Thallium Scanning Result) is Fixed.Defect or Reversible.Defect,CA (Number of major vessels coloured by fluoroscopy) is more than or equal 0.5.The probability of having heart-disease for this predictor combination is 0.9254 (highest among all).

2.5 Fit a logistic regression model to the data.

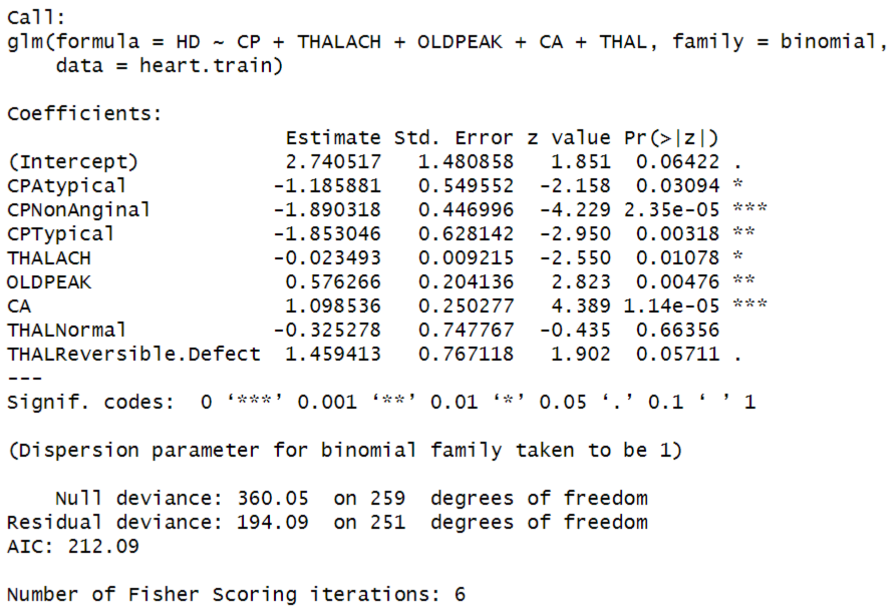
Output:

Logistic regression model before pruning:



```
Call:
glm(formula = HD ~ ., family = binomial, data = heart.train)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.986357   3.038921  -0.983  0.32575    
AGE                   -0.008904   0.027004  -0.330  0.74161    
SEXM                   1.468322   0.557562   2.633  0.00845 ** 
CPAtypical            -0.989700   0.606459  -1.632  0.10269    
CPNonAnginal          -1.740135   0.532384  -3.269  0.00108 ** 
CPTypical             -2.098109   0.711529  -2.949  0.00319 ** 
TRESTBPS               0.024234   0.012306   1.969  0.04892 *  
CHOL                   0.004740   0.004196   1.130  0.25864    
FBS>120               -0.769233   0.635827  -1.210  0.22635    
RESTECGNormal         -0.629625   0.416156  -1.513  0.13029    
RESTECGST.T.Wave      -0.186629   2.429888  -0.077  0.93878    
THALACH               -0.023936   0.012366  -1.936  0.05290 .  
EXANGY                 1.029208   0.499255   2.061  0.03926 *  
OLDPEAK                0.394286   0.249653   1.579  0.11426    
SLOPEFlat              1.200481   0.987856   1.215  0.22428    
SLOPEUp                0.414334   1.086544   0.381  0.70296    
CA                     1.300509   0.306781   4.239 2.24e-05 ***
THALNormal             0.177427   0.867060   0.205  0.83786    
THALReversible.Defect  1.437050   0.843425   1.704  0.08841 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.05  on 259  degrees of freedom
Residual deviance: 170.92  on 241  degrees of freedom
AIC: 208.92

Number of Fisher Scoring iterations: 6
```

Logistic regression model after pruning:



```
Call:
glm(formula = HD ~ CP + THALACH + OLDPEAK + CA + THAL, family = binomial,
    data = heart.train)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)            2.740517   1.480858   1.851  0.06422 .  
CPAtypical            -1.185881   0.549552  -2.158  0.03094 *  
CPNonAnginal          -1.890318   0.446996  -4.229 2.35e-05 ***
CPTypical             -1.853046   0.628142  -2.950  0.00318 ** 
THALACH               -0.023493   0.009215  -2.550  0.01078 *  
OLDPEAK                0.576266   0.204136   2.823  0.00476 ** 
CA                     1.098536   0.250277   4.389 1.14e-05 ***
THALNormal            -0.325278   0.747767  -0.435  0.66356    
THALReversible.Defect  1.459413   0.767118   1.902  0.05711 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.05  on 259  degrees of freedom
Residual deviance: 194.09  on 251  degrees of freedom
AIC: 212.09

Number of Fisher Scoring iterations: 6
```

Explanation:
The total variables used in the final logistic regression model =8 which are:
CPATYPICAL, CPNonAnginal, CPTypical, THALACH, OLDPEAK, CA, THALNormal, THALReversible.Defect
The total variables used in the CV pruned decision tree =18, the variables are:
AGE, SEXM,CPAtypical,CPNonAnginal,CPTypical, TRESTBPS, CHOL, PBS>120, RESTECGNormal, RESTECGST.T.Wave, THALACH, EXANGY,OLDPEAK,SLOPEFlat,SLOPEUp,CA,THALNormal,THALReversible.Defect

The logistic regression model has lesser variables after pruning, by removing the 10 variables which are:
AGE, SEXM,TRESTBPS, CHOL, PBS>120,RESTECGNormal, RESTECGST.T.Wave, EXANGY,OLDPEAK,SLOPEFlat,SLOPEUp compared to the tree model.

The important predictor for logistic regression model is CPNonAnginal and CA as both of these variables have the p-value less than 0.
By looking at the p-value, the most important predictor for the logistic regression model is CPNonAnginal with the lowest p-value of 2.35 × 10^(-5) among all.

2.6

Write down the regression equation for the logistic regression model you found using step-wise selection.

Regression equation for the logistic regression model:

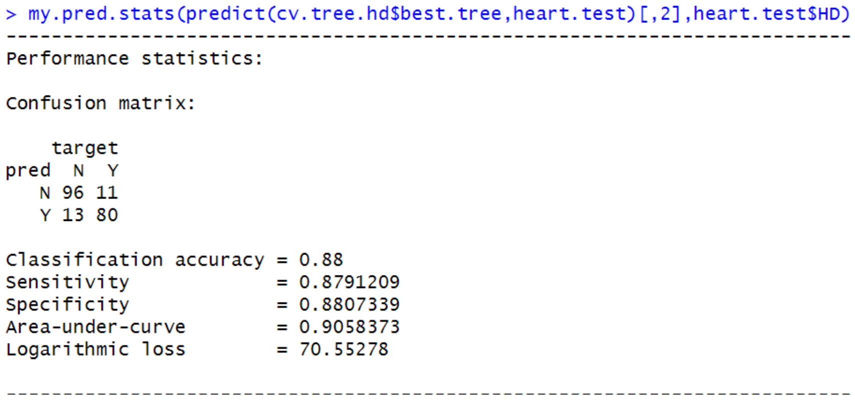
$P(HD = Y)$ = 2.740517 -1.185881 * CPAtypical -1.890318 * CPNonAnginal - 1.853046 * CPTypical -0.023493 * THALACH + 0.576266 * OLDPEAK + 1.098536 * CA - 0.325278 * THALNormal + 1.459413 * THALReversible.Defect

2.7

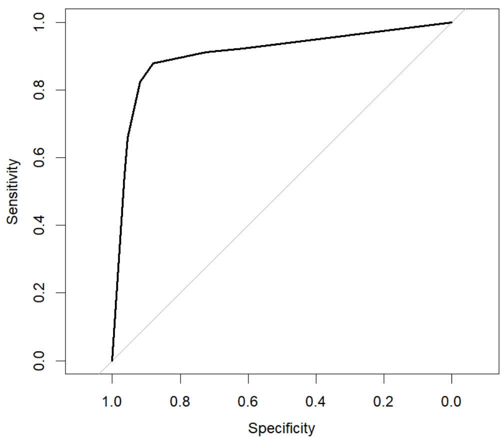
Compute the prediction statistics for both the tree and the step-wise logistic regression model on this test data

Output:

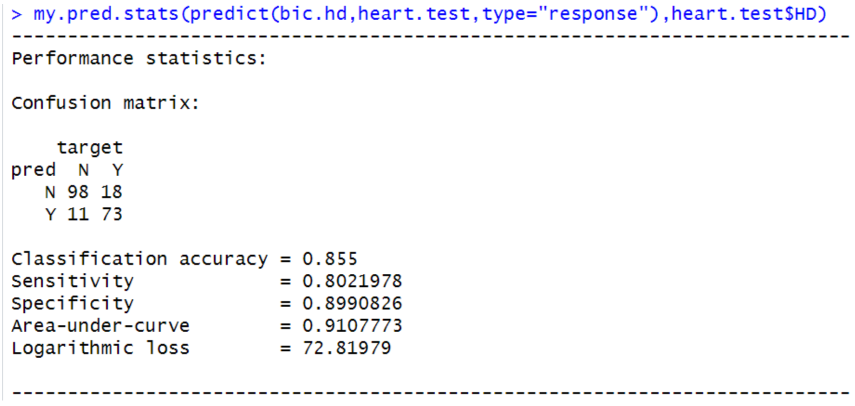
Prediction statistic for cv pruned tree



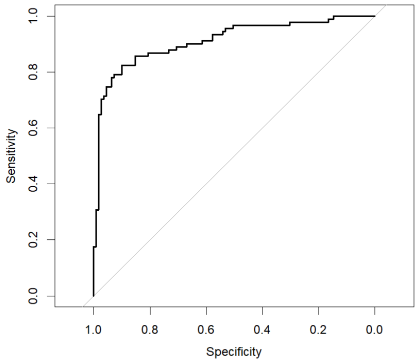
```
> my.pred.stats(predict(cv.tree.hd$best.tree,heart.test)[,2],heart.test$HD)
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 96 11
   Y 13 80

Classification accuracy = 0.88
Sensitivity             = 0.8791209
Specificity             = 0.8807339
Area-under-curve        = 0.9058373
Logarithmic loss        = 70.55278

---------------------------------------------------------------------------
```



Prediction statistic for stepwise logistic regression model



```
> my.pred.stats(predict(bic.hd,heart.test,type="response"),heart.test$HD)
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 98 18
   Y 11 73

Classification accuracy = 0.855
Sensitivity             = 0.8021978
Specificity             = 0.8990826
Area-under-curve        = 0.9107773
Logarithmic loss        = 72.81979

---------------------------------------------------------------------------
```



Explanation:
Classification Accuracy:
CV pruned tree = 0.88
Logistic regression = 0.855
CV pruned tree has a slightly higher classification accuracy compared to the logistic regression model.

Sensitivity:
CV pruned tree = 0.8791
Logistic regression = 0.8022
CV pruned tree also have a higher sensitivity, which is a higher ability to detect people with heart disease.

Specificity:
CV pruned tree = 0.8807
Logistic regression = 0.8991
Logistic regression model has a slightly higher specificity, when the goal is to correctly identify non-diseased individuals.

Area Under Curve (AUC):
CV pruned tree = 0.9058
Logistic regression = 0.9108
The logistic regression model has a higher AUC, we can see that the curve bends further away from the diagonal line from the plot above.

Logarithm loss:
CV pruned tree = 70.5528
Logistic regression =72.8198
CV pruned tree has a lower logarithm loss, indicating better calibration of predicted probabilities.

In conclusion, both models offer distinct advantages and are best suited for particular diagnostic objectives, contingent on the trade-offs between sensitivity and specificity.
For example,
The CV pruned tree is the preferred choice when the primary objective is the accurate identification of individuals with the condition and the precise calibration of predicted

probabilities. This model excels in terms of higher classification accuracy, sensitivity, and lower logarithm loss.

The logistic regression model is the optimal selection when the main aim is to identify individuals without the condition and achieve superior discrimination between classes. This model demonstrates superiority in terms of higher specificity and AUC.

2.8

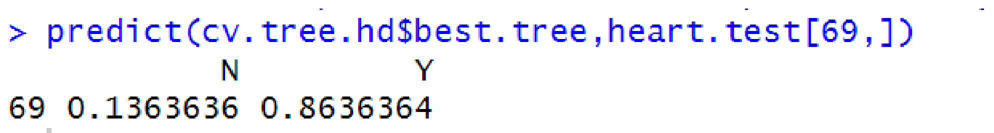
Calculate the odds of having heart disease for the 69th patient in the test dataset.

(a) the tree model found using cross-validation

(b) the step-wise logistic regression model. How do the predicted odds for the two models compare?

Output:

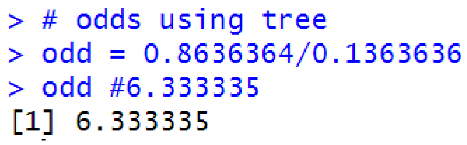
(a)



```
> predict(cv.tree.hd$best.tree,heart.test[69,])
           N         Y
69 0.1363636 0.8636364
```

The conditional probability of having heart disease for the 69th patient predicted by the tree is 0.8636364.

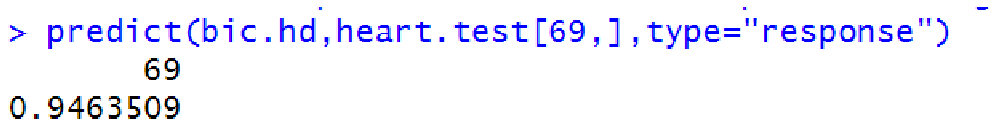
The conditional probability of not having heart disease for the 69th patient predicted by the tree is 0.1363636.



```
> # odds using tree
> odd = 0.8636364/0.1363636
> odd #6.333335
[1] 6.333335
```

The odds of having heart disease for the 69th patient using the tree model is 6.333335.

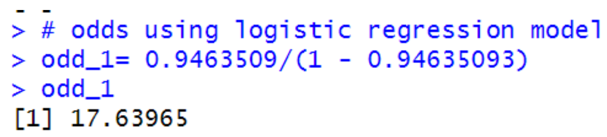
(b)



```
> predict(bic.hd,heart.test[69,],type="response")
       69
0.9463509
```

The conditional probability of having heart disease for the 69th patient predicted by the logistic regression model is 0.9463509.

The conditional probability of not having heart disease for the 69th patient predicted by the logistic regression model is 1 - 0.9463509 = 0.0536491



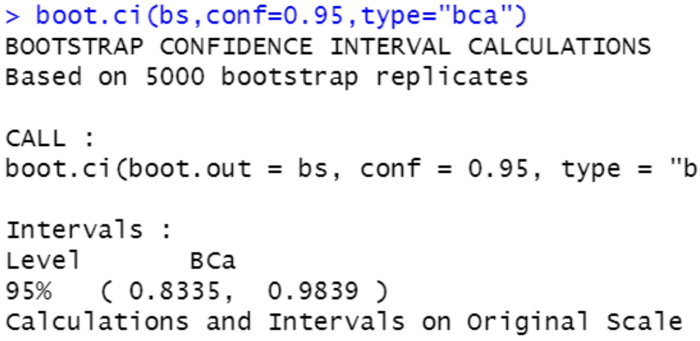
```
> # odds using logistic regression model
> odd_1= 0.9463509/(1 - 0.94635093)
> odd_1
[1] 17.63965
```

The odds of having heart disease for the 69th patient using the logistic regression model is 17.63965.

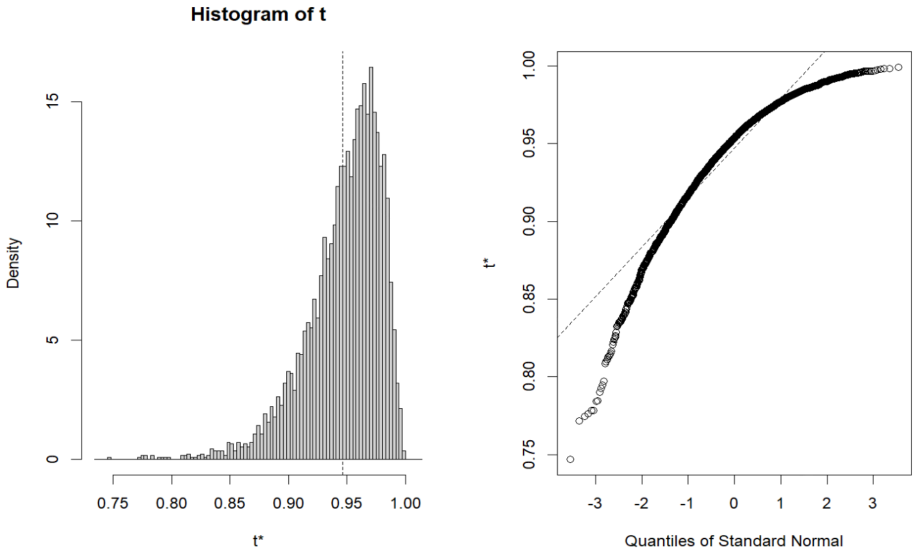
2.9

For the logistic regression model using the predictors selected by BIC in Question 2.6, use the bootstrap procedure (use at least 5,000 bootstrap replications) to find a confidence interval for the probability of having heart disease for the 69th patient in the test data.

Output:



```
> boot.ci(bs,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL : 
boot.ci(boot.out = bs, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   ( 0.8335,  0.9839 )  
Calculations and Intervals on Original Scale
```



Explanation:
We are 95% confident that the confidence interval obtained using the BCa (Bias-Corrected and Accelerated) method is (0.8335, 0.9839).The conditional probability predicted by the tree and logistic regression model are 0.8335 and 0.9839.The conditional probability predicted by both models (tree = 0.8636364,logistic regression = 0.9463509) falls within this interval.

Question 3 (3.1- 3.8)

Data Smoothing

Data “smoothing” is a very common problem in data science and statistics. We are often interested in examining the unknown relationship between a dependent variable (y) and an independent variable (x), under the assumption that the dependent variable has been imperfectly measured and has been contaminated by measurement noise. The model of reality that we use is y =f(x)+ε where f(x) is some unknown, “true”, potentially non-linear function of x, and ε ~ N(0,σ2) is a random disturbance or error. This is called the problem of function estimation, and the process of estimating f(x) from the noisy measurements y is sometimes called “smoothing the data” (even if the resulting curve is not “smooth” in a traditional sense, it is less rough than the original data). In this question you will use the k-nearest neighbours machine learning technique to smooth data. This technique is used frequently in practice (think for example the 14-day rolling averages used to estimate coronavirus infection numbers). This question will explore its effectiveness as a smoothing tool.

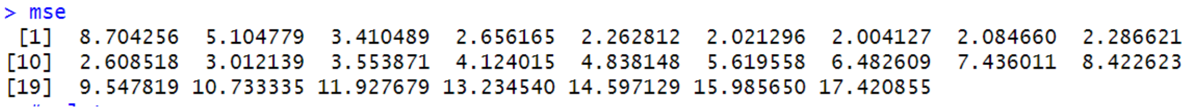
Mass Spectrometry Data Smoothing

The file ms.measured.2023.csv contains n = 443 measurements from a mass spectrometer. Mass spectrometry is a chemical analysis tool that provides a measure of the physical composition of a material. The outputs of a mass spectrometry reading are the intensities of various ions, indexed by their mass-to-charge ratio. The resulting spectrum usually consists of a number of relatively sharp peaks that indicate a concentration of particular ions, along with an overall background level. A standard problem is that the measurement process is generally affected by noise– that is, the sensor readings are imprecise and corrupted by measurement noise. Therefore, smoothing, or removing the noise is crucial as it allows us to get a more accurate idea of the true spectrum, as well as determine the relative quantity of the ions more accurately. However, we would ideally like for our smoothing procedure to not damage the important information contained in the spectrum (i.e., the heights of the peaks). The file ms.measured.2023.csv contains measurements of our mass spectrometry reading; the vari able ms.measured$MZ contains the mass-to-charge ratios of various ions, and ms.measured$intensity are the measured (noisy) intensities of these ions in our material. The file ms.truth.2023.csv contains n =886 different values of MZ along with the “true” intensity values, stored in ms.truth$intensity. These true values have been found by using several advanced statistical techniques to smooth the data, and are being used here to see how close your estimated spectrum is to the truth. For reference, the samples ms.measured$intensity and the value of the true spectrum ms.truth$intensity are plotted in Figure 1 against their respective MZ values.

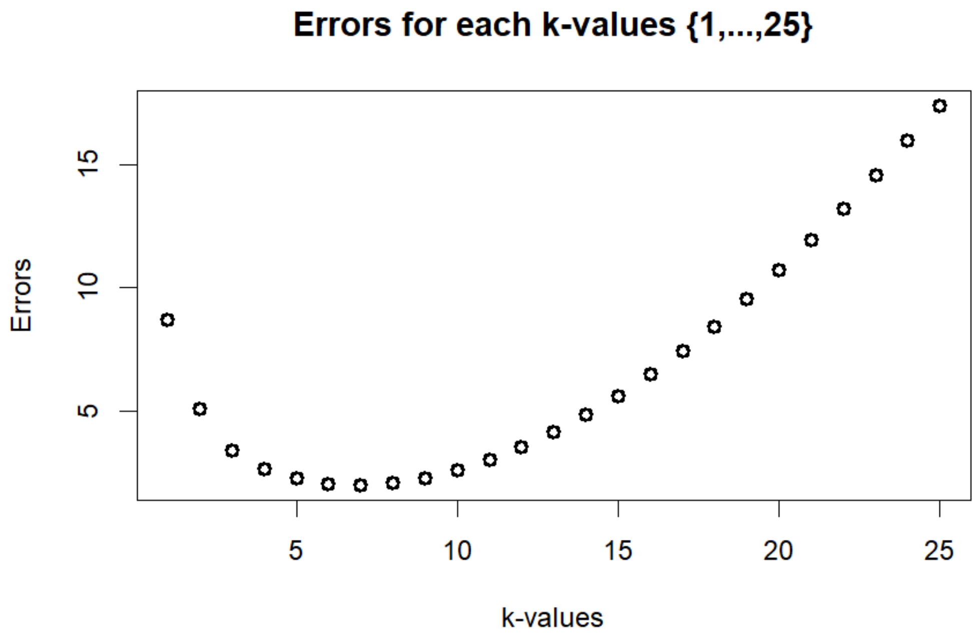
3.1

For each value of k = 1,...,25, use k-NN to estimate the values of the spectrum at each of the MZ values in ms.truth$MZ. Then, compute the mean-squared error between your estimates of the spectrum, and the true values in ms.truth$intensity. Produce a plot of these errors against the various values of k.

Output:



```
> mse
 [1]  8.704256  5.104779  3.410489  2.656165  2.262812  2.021296  2.004127  2.084660  2.286621
[10]  2.608518  3.012139  3.553871  4.124015  4.838148  5.619558  6.482609  7.436011  8.422623
[19]  9.547819 10.733335 11.927679 13.234540 14.597129 15.985650 17.420855
```



Explanation:

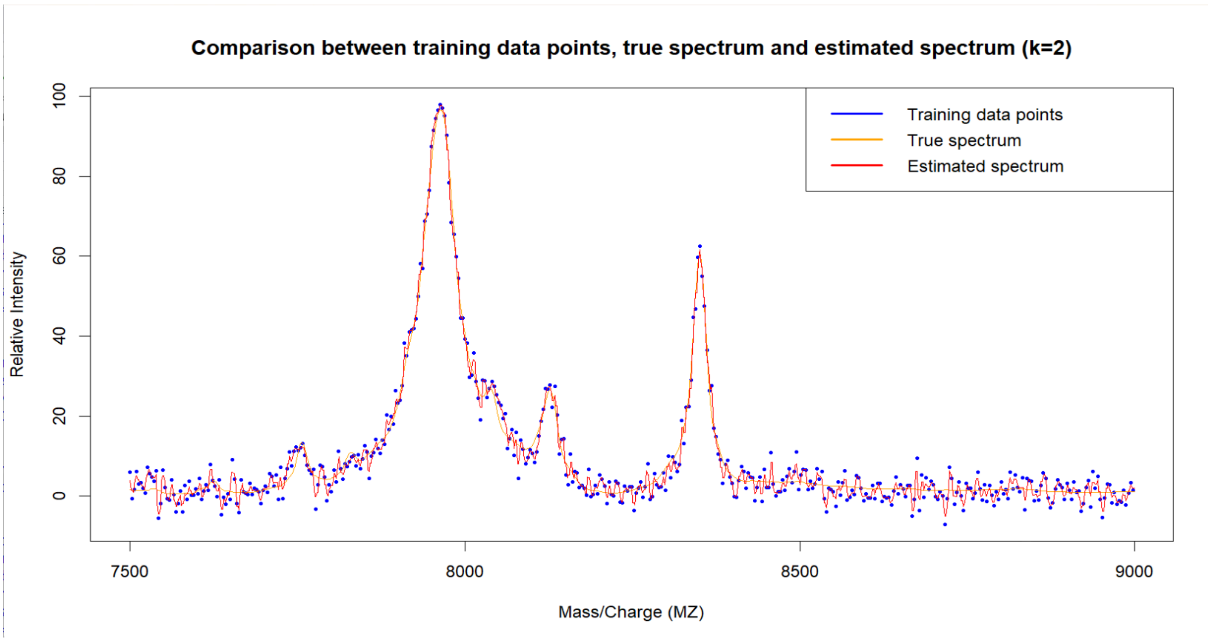
From the plot above, we know that k-value 7 has the smallest error.

3.2

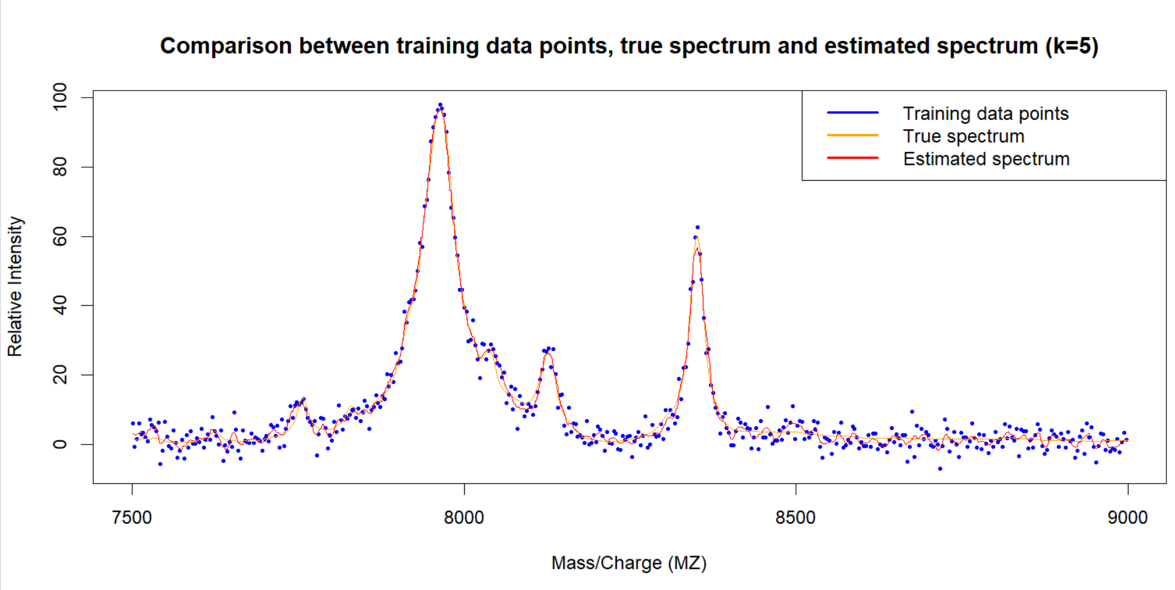
Produce four graphs, each one showing: (i) the training data points (ms.measured$intensity), (ii) the true spectrum (ms.truth$intensity) and (iii) the estimated spectrum (predicted intensity values for the MZ values in ms.truth.csv) produced by the k-NN method for four different values of k; do this for k = 2, k = 5, k = 10 and k = 25.

Output:

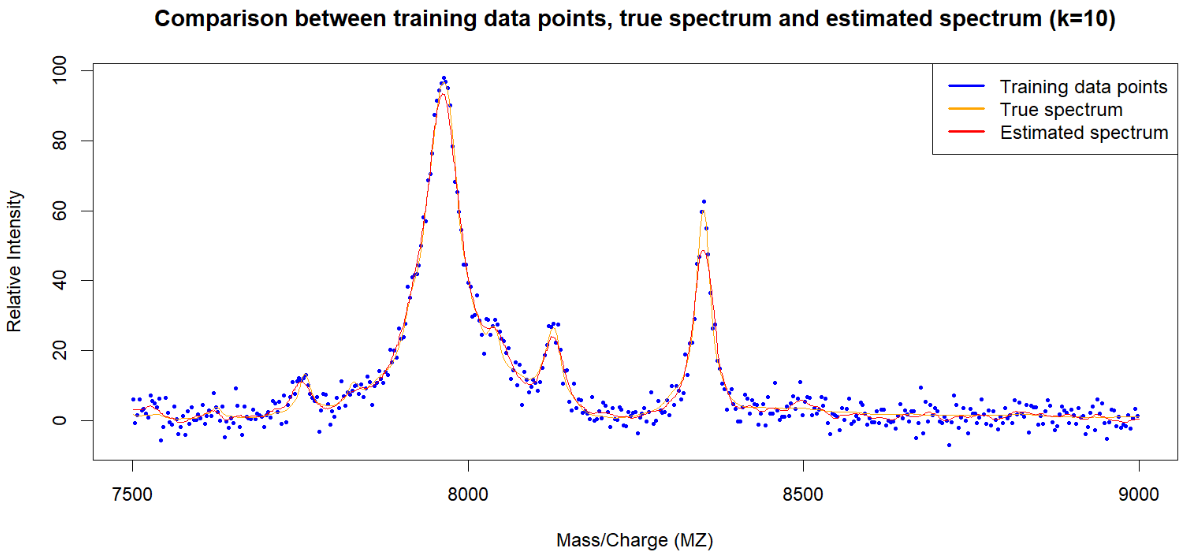
k=2



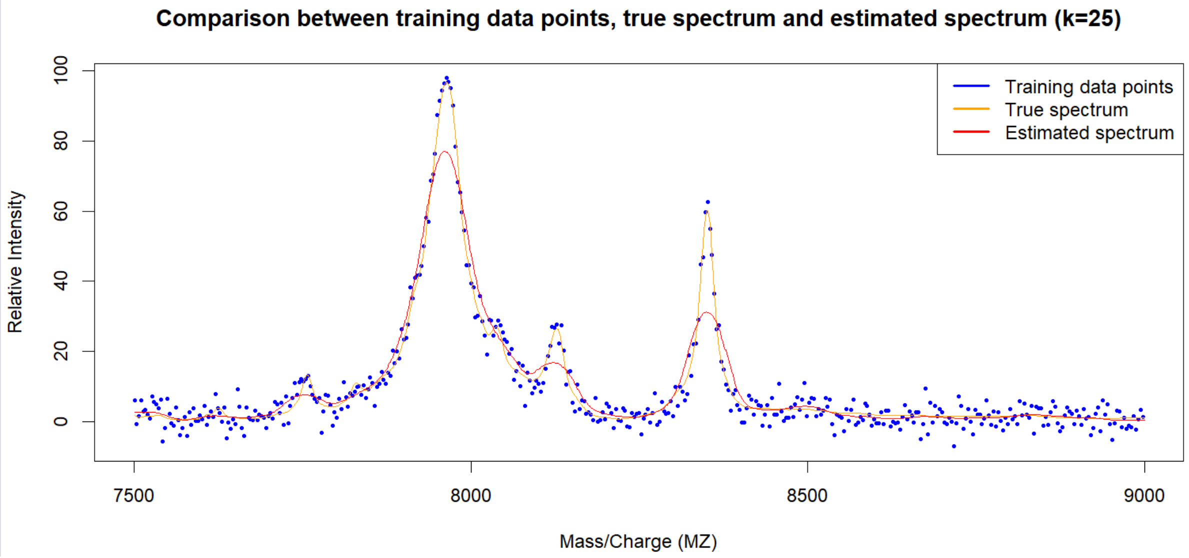
k=5



k=10


Comparison between training data points, true spectrum and estimated spectrum (k=10)
Relative Intensity
0
20
40
60
80
100
7500
8000
8500
9000
Mass/Charge (MZ)
Training data points
True spectrum
Estimated spectrum

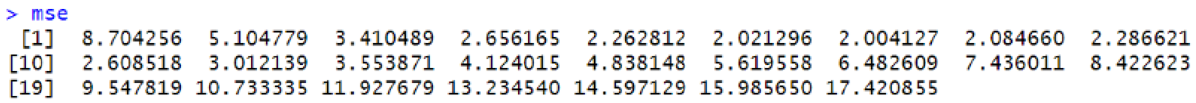
k=25


Comparison between training data points, true spectrum and estimated spectrum (k=25)
Relative Intensity
0
20
40
60
80
100
7500
8000
8500
9000
Mass/Charge (MZ)
Training data points
True spectrum
Estimated spectrum

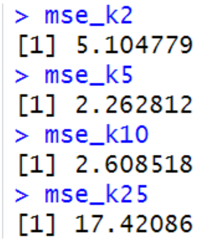
3.3

Discuss, qualitatively (i.e., visually), and quantitatively (i.e., in terms of mean-squared error on the true spectrum) the four different estimates of the spectrum.

Output:



```
> mse
 [1]  8.704256  5.104779  3.410489  2.656165  2.262812  2.021296  2.004127  2.084660  2.286621
[10]  2.608518  3.012139  3.553871  4.124015  4.838148  5.619558  6.482609  7.436011  8.422623
[19]  9.547819 10.733335 11.927679 13.234540 14.597129 15.985650 17.420855
```



```
> mse_k2
[1] 5.104779
> mse_k5
[1] 2.262812
> mse_k10
[1] 2.608518
> mse_k25
[1] 17.42086
```

Explanation:

Qualitatively (graph):
Look at the 4 graphs we plotted in 3.2
When k=5, the estimated spectrum (red line) follows the underlying trend of the training data, it does not fit every single point of the training data (blue dots). The estimated spectrum is almost aligned with the true spectrum , as the red line is having many peaks with similar heights as the orange line. Thus, we can say that the model with the k-values =5 has a good estimate of the spectrum.

When k=25, the estimated spectrum does not follow the shape of the true spectrum. It is less jagged than the true spectrum and has fewer peaks than the true spectrum. Thus, the model is said to be underfitting , leading it to have many systemic errors when predicting new data. The model with k-value =25 is the worst among all.

Quantitatively (MSE):
From the output shown above , we can see that the mean square error is the lowest (2.0213) when k-value = 6, indicating that the estimated spectrum is closest to the true spectrum when k=6. As the k-value increases, the mean square error increases, indicating that higher k values, the estimated spectrum deviates more from the true spectrum, leading to larger errors in the estimate. The mean square error increases when k-value decreases after reaching the minimum MSE (k=6), indicating overfitting or excessive smoothing of the spectrum when using smaller k values.

Therefore, we can tell that as the k value increases, the model's estimated spectrum is less aligned with the true spectrum, mean square error increases and thus the model gets simpler and tends to underfit. When k value decreases, the model is more complex and tends to overfit. K-value =6 is picked to be the best among all whereas k-value =25 is said to be the worst.

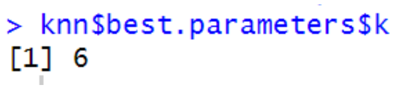
3.4

Do any of the estimated spectra achieve our aim of providing a smooth, low-noise estimate of background level as well as accurate estimation of the peaks? Explain why you think the k-NN method is able to achieve, or not achieve, this aim.

The estimated spectrum plotted with the k value =6 achieves the aims of providing a smooth low-noise estimate of background level and has an accurate estimation of the heights of the peaks.Therefore, the kNN method is able to achieve this aim.

3.5

Use the cross-validation functionality in the kknn package to select an estimate of the best value of k (make sure you still use the optimal kernel). What value of k does the method select? How does it compare to the (in practice, unknown) value of k that would minimise the actual mean-squared error (as computed in Question 3.1a)?



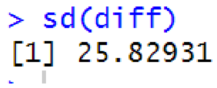
```
> knn$best.parameters$k
[1] 6
```

We estimate the best value of k is 6 using the cross validation functionality in the kknn package.
From 3.1, the plot shows that 7 is the lowest k value with the best performance. However, it's important to note that the values of k, between 6 to 8, all appear to have good performance. Thus,it's valid to estimate the best k value as 6 using the cross-validation functionality in the kknn package.

3.6

Using the estimate of the spectrum produced in Q3.5 using the value of k selected by cross validation, and the values in ms.measured$intensity, see if you can think of a way to find an estimate of the standard deviation of the sensor/measurement noise that has corrupted our intensity measurements.

Output:



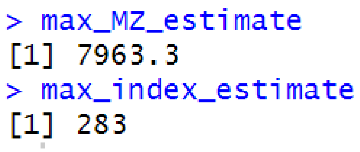
```
> sd(diff)
[1] 25.82931
```

Explanation:
The standard deviation of sensor/measurement noise is 25.82931.

3.7

An important task when processing mass spectrometry signals is to locate the peaks, as this gives information on which elements are present. From the smoothed signal produced using the value of k found in Question 3.5, which value of MZ corresponds to the maximum estimated abundance?

Output:



```
> max_MZ_estimate
[1] 7963.3
> max_index_estimate
[1] 283
```

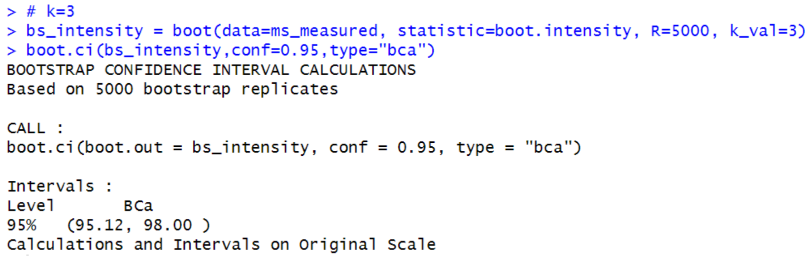
Explanation:
From the smoothed signal produced using the value of k=6 found in 3.5, the value of MZ corresponds to the maximum estimated abundance is 7963.3

3.8

Using the bootstrap procedure (use at least 5,000 bootstrap replications), write code to find a confidence interval for the k-nearest neighbours estimate of relative abundance at a specific MZ value. Use this code to obtain a 95% confidence interval for the estimate of relative abundance at the MZ value you determined previously in Question 3.7 (i.e., the value corresponding to the highest relative intensity). Compute confidence intervals using the k determined in Question 3.5, as well as k = 3 neighbour and k = 20 neighbours. Report these confidence intervals.

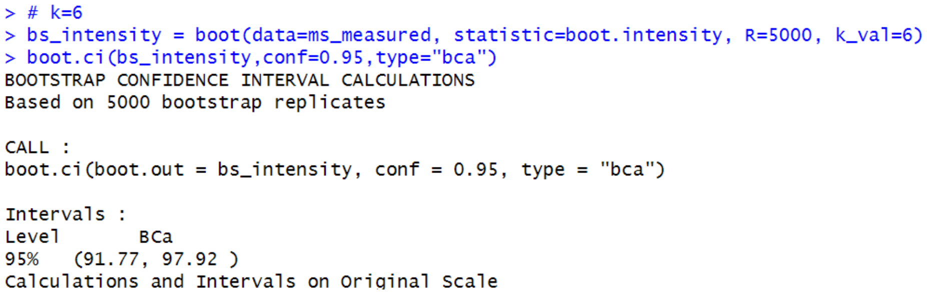
Output:

k=3



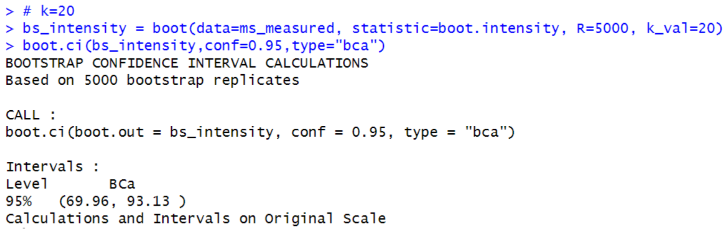
```
> # k=3
> bs_intensity = boot(data=ms_measured, statistic=boot.intensity, R=5000, k_val=3)
> boot.ci(bs_intensity,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL : 
boot.ci(boot.out = bs_intensity, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   (95.12, 98.00 )  
Calculations and Intervals on Original Scale
```

k=6



```
> # k=6
> bs_intensity = boot(data=ms_measured, statistic=boot.intensity, R=5000, k_val=6)
> boot.ci(bs_intensity,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL : 
boot.ci(boot.out = bs_intensity, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   (91.77, 97.92 )  
Calculations and Intervals on Original Scale
```

k=20



```
> # k=20
> bs_intensity = boot(data=ms_measured, statistic=boot.intensity, R=5000, k_val=20)
> boot.ci(bs_intensity,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL : 
boot.ci(boot.out = bs_intensity, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   (69.96, 93.13 )  
Calculations and Intervals on Original Scale
```

Explanation:

From 3.7, we know that the MZ value corresponds to the highest intensity = 7963.3 ,Indice = 283.

From 3.5, we know that the best value of k using the kknn package = 6.

When k = 3, the 95% confidence interval for the k-nearest neighbours estimate of intensity at a MZ value =7963.3 is (95.12, 98.00) . Range of coincidence interval = 98.00-95.01 = 2.99. The smaller confidence interval width (2.99) suggests that the estimate is sensitive to local variations and can result in a narrower but less stable confidence interval.

When k = 6, the 95% confidence interval for the k-nearest neighbours estimate of intensity at a MZ value = 7963.3 is (91.77, 97.92). Range of coincidence interval = 97.96-91.90 = 6.06.

The moderate confidence interval width (6.06) indicates a balance between local sensitivity and stability in the estimate.
When k = 20, the 95% confidence interval for the k-nearest neighbours estimate of intensity at a MZ value = 7963.3 is (69.96, 93.13).Range of coincidence interval = 93.13-69.96 = 23.17.The larger confidence interval width (23.17) implies that the estimate is more stable but less sensitive to local variations.

In conclusion, we should choose the k value wisely based on the trade-off between bias and variance. This is said as smaller k-values can capture fine details but might be sensitive to noise, while larger k-values provide smoother estimates but may over smooth and lose important information.