Assignment 3

Question 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? [3 marks]

```
#import data
rm(list=ls())
housing = read.csv("housing.2023.csv")

#fit simple linear regression
fit = lm(medv ~., data = housing)summary(fit)
summary(fit)
```

```
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
```

crim, chas, nox, rm, dis, rad, tax, ptratio, lstat potentially have an association with median house prices because they have a small p-value (<0.05 significance test). The p value is a good indicator of a predictors’ association because a low p value suggest strong evidence against the null hypothesis that there is no correlation between the predictor and housing price. This means that it is very unlikely for these predictors to have a correlation as large or larger than the one we observed if there was no correlation. This means that it is very unlikely for these predictors to have a correlation as large or larger than the one we observed if there was no correlation. The three strongest predictors are rm (Average number of rooms per dwelling), lstat (Percentage of lower status population) and ptratio (Pupil-teacher ratio). These are the top 3 predictors with the lowest p-value.

2. How would your assessment of which predictors are associated change if you used the Bonferroni procedure with α = 0.05? [1 mark]

```
pvalues = coefficients(summary(fit))[,4]
#12 predictorsp = 12
pvalues < 0.05/psum(pvalues < 0.05/p)
#6 predictors passed the Bonferroni procedure
pvalues < 0.05sum(pvalues < 0.05)
```

```
(Intercept)        crim          zn       indus        chas         nox          rm         age         dis         rad         tax     ptratio
       TRUE       FALSE       FALSE       FALSE        TRUE       FALSE        TRUE       FALSE        TRUE       FALSE       FALSE        TRUE
      lstat
       TRUE
> sum(pvalues < 0.05/p)
[1] 6
```

Using the Bonferroni procedure, only chas, dis, ptratio and lstat passed the bonferroni procedure. This means crim, nox, rad and tax that passed the 0.05 significance test did not pass the Bonferroni procedure.

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. [2 marks]

looking at the coefficients of crim in our linear model, for every unit increase in crime rate, the median house price will decrease by 0.115817989.

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. [1 mark]

```
n = length(housing$medv)
fit_bic = step(fit, direction="both", k = log(n))
summary(fit_bic)
fit_bic$coefficients
```

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

E[medv] = 29.1926650 + 4.5991149*chas -17.3765139*nox + 4.8206454*rm -0.9359373*dis -0.9591382*ptratio -0.4947192* lstat

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? [2 marks]

based on our model nox, dis, ptratio and lstat have a negative association with median house value thus lowering these predictors would increase median house value.Likewise increasing predictor values for chas and rm would increase median house value. It would be impossible to have a frontage to the river charles if the houses does not have one already and neither is it possible to increase the number of rooms for houses built in the suburb or make the suburb have a larger weighed distances to five Boston employment centres. The council can however look into lowering nitric oxides concentration in the air. The council can also look into having smaller class sizes in schools to reduce pupil student ratio. Lastly the council can try to find ways to decrease the percentage of “lower status” of the population

1. Table 2 gives the values of predictors for a new suburb. 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. [1 mark]

```
# Create a data frame with the predictor values for the new suburb
new_suburb_data <- data.frame(  crim = 0.04741,  zn = 0,  indus = 11.93,  chas = 0,  nox = 0.573,  rm = 6.03,  age = 80.8,  dis = 2.505,  r

pred = predict(fit_bic, new_suburb_data)
pred_ci = predict(fit_bic, new_suburb_data, interval="confidence")
```

the predicted value is 21.9196 with a 95% confidence interval of 20.30209 23.53712.

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? [1 mark]

```
fit_interact <- lm(medv ~ rm * dis, data = housing)
summary(fit_interact)
```

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

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

>
```

dis by itself does not suggest to be associated to median house prices with a p-value >0.05 but p-value of interactions between number of rooms a dwelling has and its distance to one of the employment centres is 0.065 which potentially indicates positive effect (p value > 0.05) of the interaction between rm and dis on median house prices.

Question 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? [2 marks]
   fit a decision tree

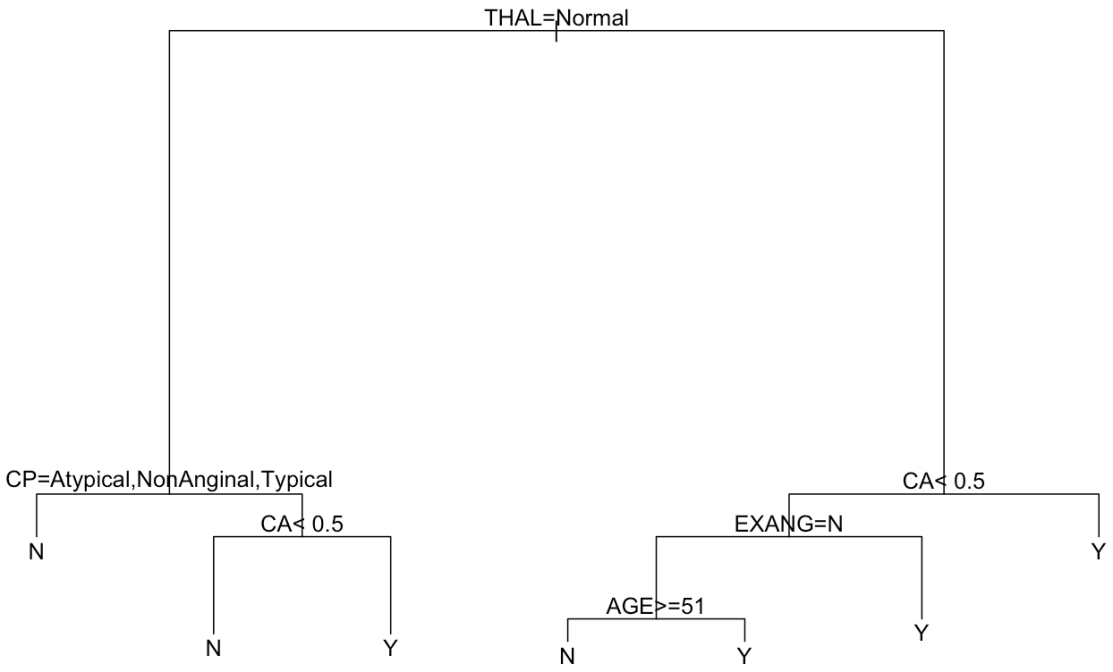
```
heart.train = read.csv("heart.train.2023.csv",stringsAsFactors = T)
heart.test = read.csv("heart.test.2023.csv",stringsAsFactors = T)
#fit a decision tree
tree.heart = rpart(HD~., heart.train)
#decision tree
tree.heart
```

using cross validation with 10 folds and 5000 repetitions

```
cv = learn.tree.cv(HD~.,data=heart.train,nfolds=10,m=5000)
cv$best.tree
plot.tree.cv(cv)
```

```
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) *
```

The tree after cross validation used predictors THAL, CP, CA, EXANG and AGE with 7 terminal nodes.



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. (hint: you can use the text(cv$best.tree,pretty=12) function to add appropriate labels to the tree). [3 marks]

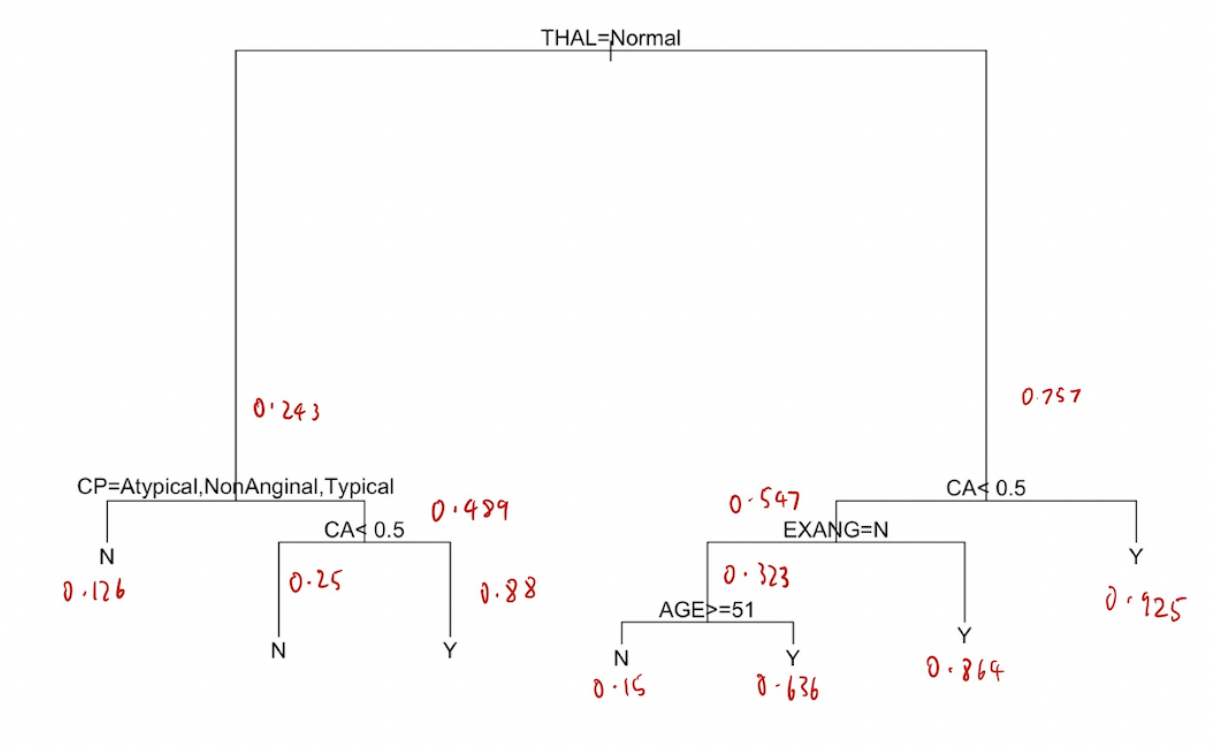
   What it tells about relationship between predictors and heart disease

   - if a patient does not have a normal Thallium result and CA≥0.5 they are predicted to have heart disease
   - if a patient has a normal Thallium, number of blood CA<0.5 and has Exercise induced angina, they are predicted to have heart disease.

- if a patient has a normal Thallium, number of blood CA<0.5, does not have exercise induced angina and are below the age of 51, they are predicted to have heart disease.
- if a patient has a normal Thallium, number of blood CA<0.5, does not have exercise induced angina and are or above the age of 51, they are predicted to not have heart disease.
- if a patient has a normal THAL, CP that is Atypical, NonAnginal or Typical, they are predicted not to have heart disease.
- if a patient has a normal THAL, CP that is Asymptomatic and CA<0.5, they are predicted to not have heart disease.
- if a patient has a normal THAL, CP that is Asymptomatic and CA≥0.5, they are predicted to have heart disease.

3. For classification problems, the rpart package only labels the leaves with the most likely class. However, if you examine the tree structure in its textural representation on the console, you can determine the probabilities of having heart disease (see Question 2.3 from Studio 9 as a guide) in each leaf (terminal node). Take a screen-capture of the plot of the tree (don't forget to use the "zoom" button to get a larger image) or save it as an image using the "Export" button in R Studio. Then, use the information from the textual representation of the tree available at the console and annotate the tree in your favourite image editing software; next to all the leaves in the tree, add text giving the probability of contracting heart disease. Include this annotated image in your report file. [1 mark]

```
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) *
```



4. According to your tree, which predictor combination results in the highest probability of having heart-disease? [1 mark]

   According to the tree, not having a normal Thallium scanning results with greater than or equal to 5 major vessels colored by flourosopy will give a 0.925 probability of heart disease.

5. We will also fit a logistic regression model to the data. Use the glm() function to fit a logistic regression model to the heart data, and use stepwise selection with the BIC score to prune the model (use direction="both"). What variables does the final model include, and how do they compare with the variables used by the tree estimated by CV? Which predictor is the most important in the logistic regression? [3 marks]

   ```
   #Q2.5
   #fit a logistic regression model to the data
   fullmod=glm(HD ~ ., data=heart.train, family=binomial)
   summary(fullmod)
   ```

   The final logistic model includes CP, THALACH, OLDPEAK, CA and THAL. Compared to the decision tree, the logistic model uses THALACH and OLDPEAK which were omitted by the decision tree. The logistic model also does not use EXANG and AGE like the decision tree. The most important predictor for the logistic regression is CA because it has the smallest p-value which suggests it has the strongest association with heart disease among all predictors.

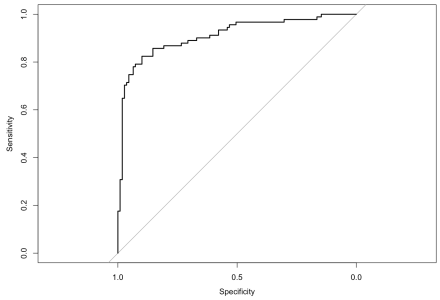
6. Write down the regression equation for the logistic regression model you found using step-wise selection. [1 mark]

   logOdds(HD = Y) = 2.740517 -1.18588127**CPAtypical* *-1.89031823**CPNonAnginal -1.85304564**CPTypical* *-0.02349346**THALACH + 0.57626556**OLDPEAK* + *1.09853619**CA*

```
> fit_bic$coefficients
         (Intercept)           CPAtypical         CPNonAnginal             CPTypical               THALACH
          2.74051704          -1.18588127          -1.89031823          -1.85304564          -0.02349346
             OLDPEAK                   CA          THALNormal THALReversible.Defect
          0.57626556           1.09853619          -0.32527806           1.45941349
> |
```

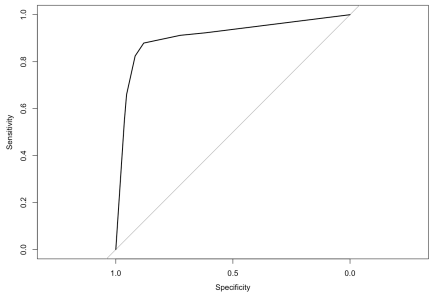
7. The file heart.test.2023.csv contains the data on a further n = 200 individuals. Using the my.pred.stats() function contained in the file my.prediction.stats.R, compute the prediction statistics for both the tree and the step-wise logistic regression model on this test data.Contrast and compare the two models in terms of the various prediction statistics? Would one potentially be preferable to the other as a diagnostic test? Justify your answer. [2 marks]

prediction statistic for step-wise logistic regression




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

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

prediction statistic for tree



```
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

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

Based on the tests, the decision tree is a better model. The decision tree gave a classification accuracy of 0.88 which is higher than the step-wise logistic regression at 0.855. The tree showed better sensitivity at identifying heart disease than the logistic regression. Although the tree performs marginally worse, 0.88 than the logistic regression, 0.899, at correctly identifying people with no heart disease (specificity), classification accuracy and sensitivity are more important compared to specificity for identifying heart disease because it is better to incorrectly identify a false positive than a false negative for heart disease. Additionally the decision tree has a smaller log loss which means it performs better at estimating the probability of an individual being in either classes. The Area under the curve for decision tree is only marginally smaller than the logistic regression too.

8. Calculate the odds of having heart disease for the 69th patient in the test dataset. The odds should be calculated for both: (a) the tree model found using cross-validation; and (b) the step-wise logistic regression model. How do the predicted odds for the two models compare? [2 marks]

```
> heart.test[69,]
   AGE SEX           CP TRESTBPS CHOL  FBS     RESTECG THALACH EXANG OLDPEAK SLOPE CA              THAL HD
69  59   M Asymptomatic      170  326 <120 Hypertrophy     140     Y     3.4  Down  0 Reversible.Defect  Y
>
```

Logistic regression model

odds of having HD using prediction from logistic regression

$$O = P(HD = Y)/(1 - P(HD = Y))$$
$$O = 0.9463509/(1 - 0.9463509)$$
$$O = 17.63966$$

Odds of having heart disease using prediction from decision tree

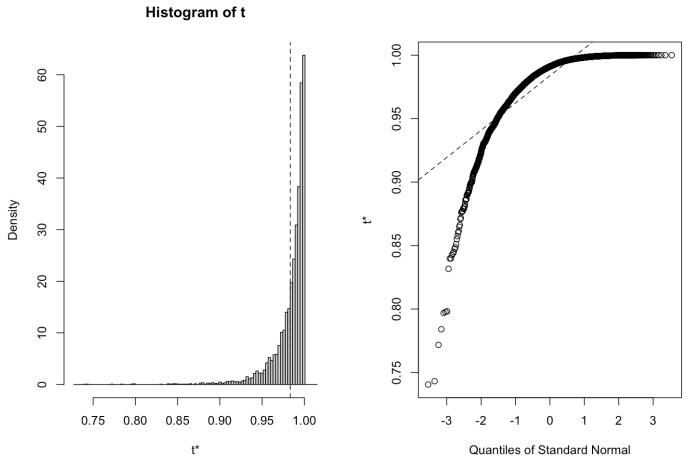
$$O = P(HD = Y)/(1 - P(HD = Y))$$
$$O = 0.8636364/(1 - 0.8636364)$$
$$O = 6.333335$$

Based on these findings, the log Odds estimated by the step wise logistic regression is more than 2 times larger than the log odds estimated by the decision tree. Given that the person has heart disease, both models correctly predicted that the person is highly likely to have heart disease. The log Odds estimated by the step wise logistic regression is more than 2 times larger than the log odds estimated by the decision tree. This suggest that the logistic regression is a better model because it was more confident that the person had heart disease, which the person did.

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. Use the bca option when computing this confidence interval. Discuss this confidence interval in comparison to the predicted probabilities of having heart disease for both the logistic regression model and the tree model. [2 marks]

Using logistic regression model with BIC

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level       BCa
95%   ( 0.7787,  0.9979 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```



The 95% confidence interval for patient 69 having a conditional probability of having heart disease is between 0.7787, 0.9979. The probability predicted by the decision tree was 0.8636364 which is within our confidence interval. This is closer to the expected value (at the centre of the CI) compared to the regression model which estimated a probability of 0.9463509, closer to the right tail end of the distribution.

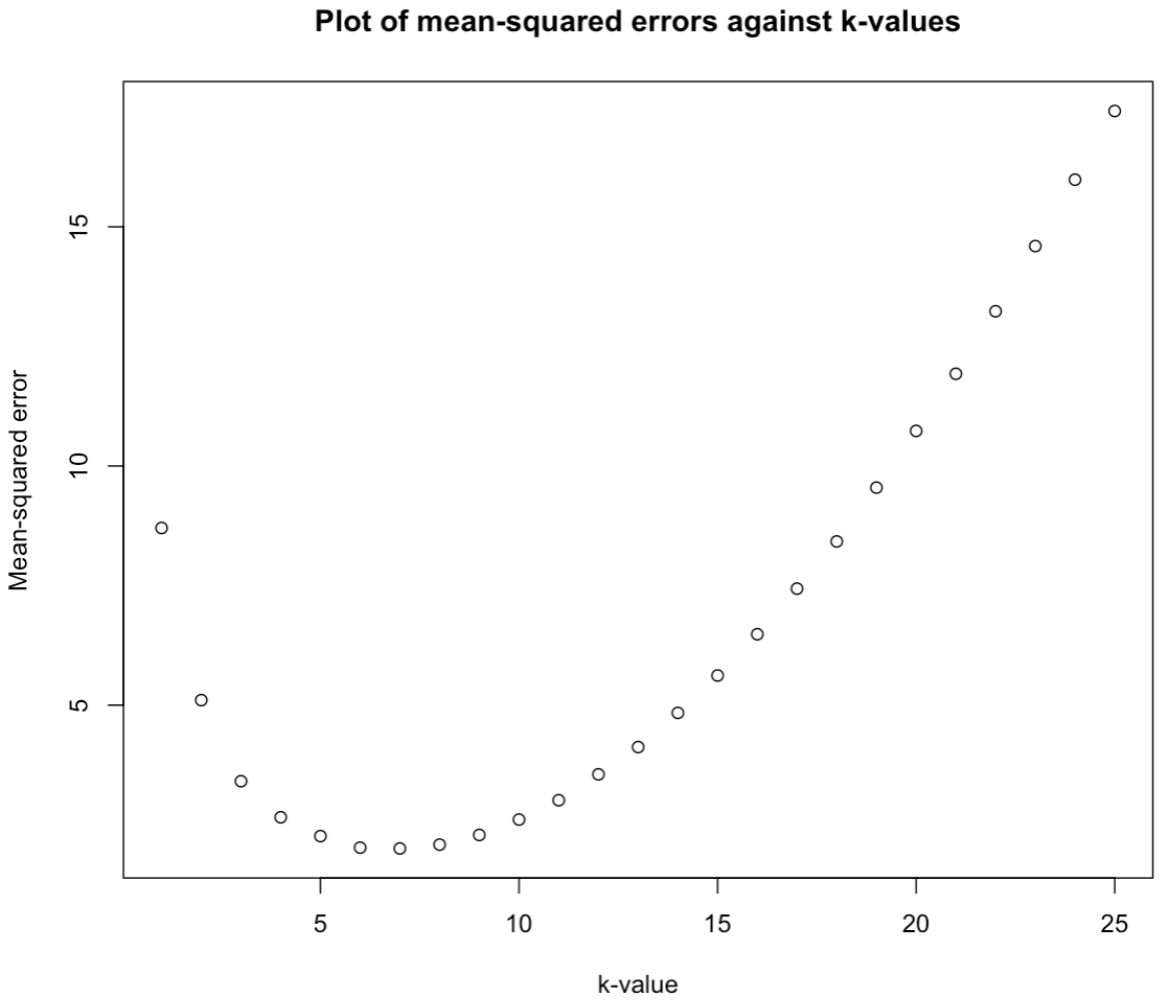
Question 3

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. [1 mark]

   mean squared error for k=1 to k=25

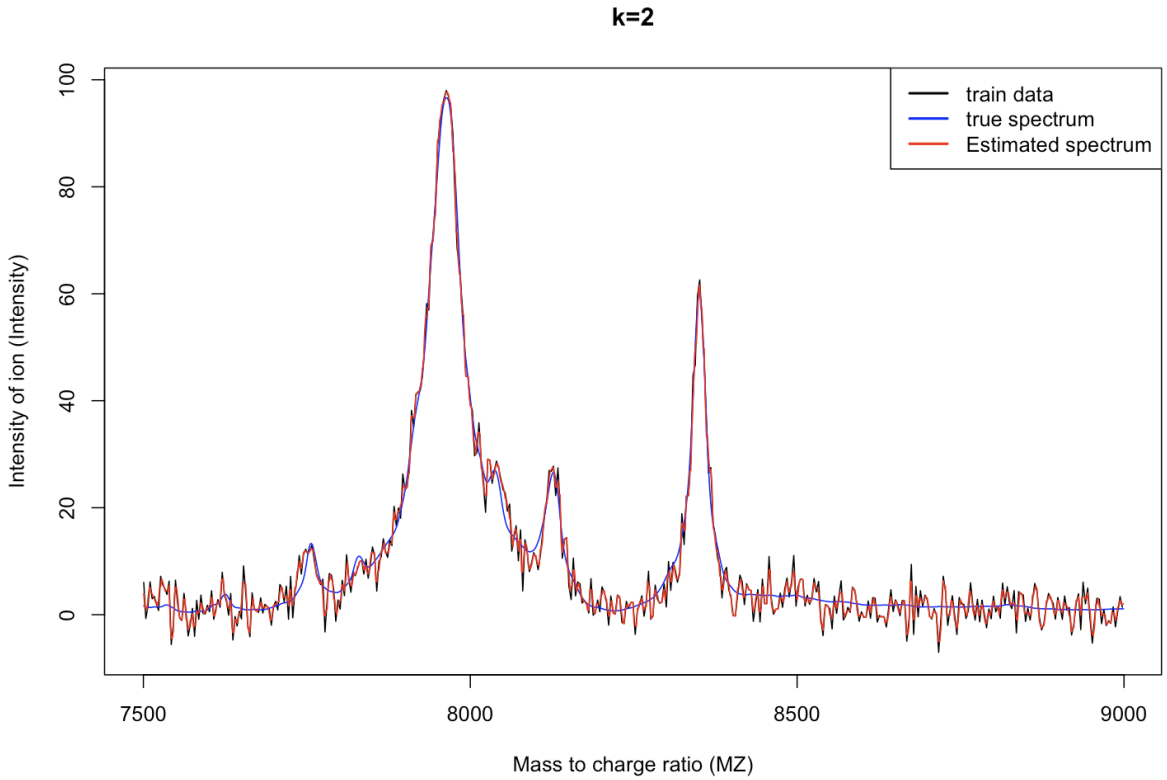
```
k value: 1 , mean-squared error: 8.704256
k value: 2 , mean-squared error: 5.104779
k value: 3 , mean-squared error: 3.410489
k value: 4 , mean-squared error: 2.656165
k value: 5 , mean-squared error: 2.262812
k value: 6 , mean-squared error: 2.021296
k value: 7 , mean-squared error: 2.004127
k value: 8 , mean-squared error: 2.08466
k value: 9 , mean-squared error: 2.286621
k value: 10 , mean-squared error: 2.608518
k value: 11 , mean-squared error: 3.012139
k value: 12 , mean-squared error: 3.553871
k value: 13 , mean-squared error: 4.124015
k value: 14 , mean-squared error: 4.838148
k value: 15 , mean-squared error: 5.619558
k value: 16 , mean-squared error: 6.482609
k value: 17 , mean-squared error: 7.436011
k value: 18 , mean-squared error: 8.422623
k value: 19 , mean-squared error: 9.547819
k value: 20 , mean-squared error: 10.73333
k value: 21 , mean-squared error: 11.92768
k value: 22 , mean-squared error: 13.23454
k value: 23 , mean-squared error: 14.59713
k value: 24 , mean-squared error: 15.98565
k value: 25 , mean-squared error: 17.42086
```

plot of mean squared error

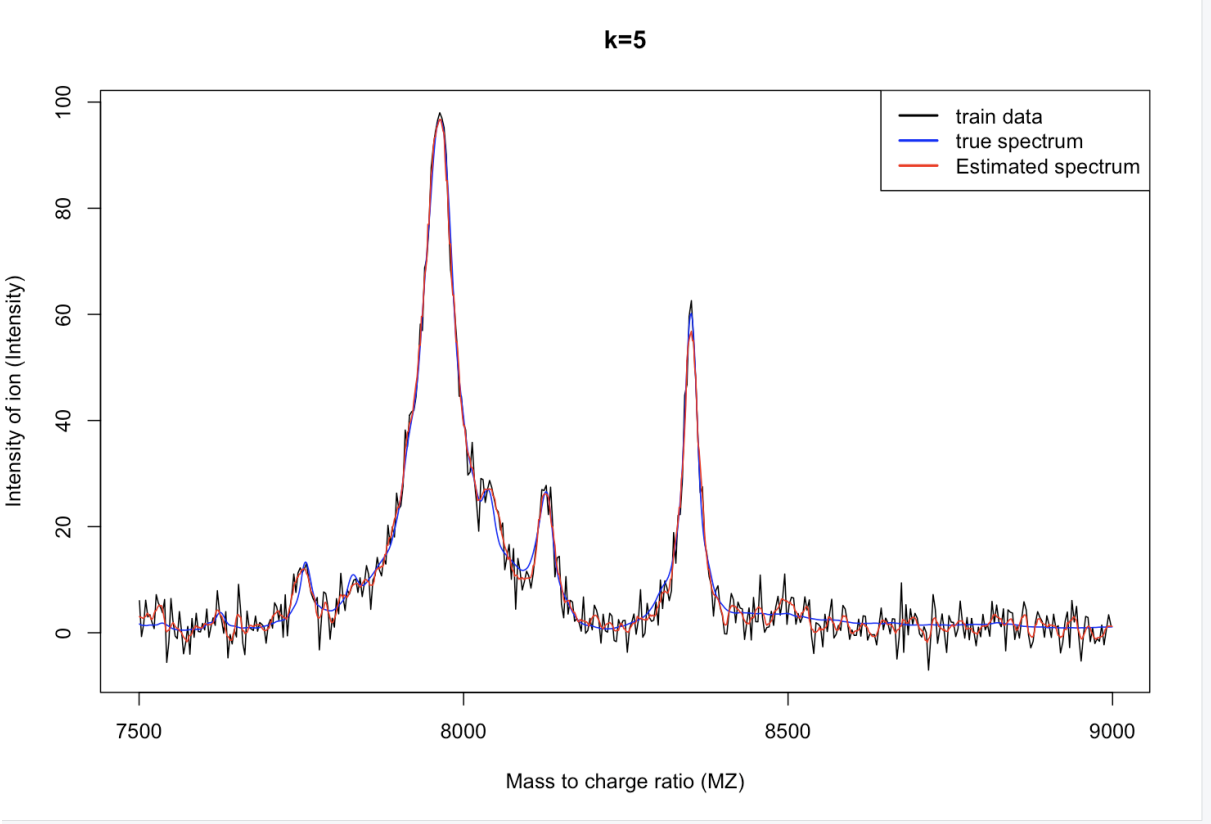


2. Produce four graphs, each one showing: (i) the training data points (ms.measured$intensity) (ii) the true spectrum (ms.truth$intensity) and (iii) the estimated spectrum (predicted intensity values for the MZ values in ms.truth.csv) produced by the k-NN method for four different values of k; do this for k = 2, k = 5, k = 10 and k = 25. Make sure the graphs have clearly labelled axis' and a clear legend. Use a different colour for your estimated curve. [3 marks]

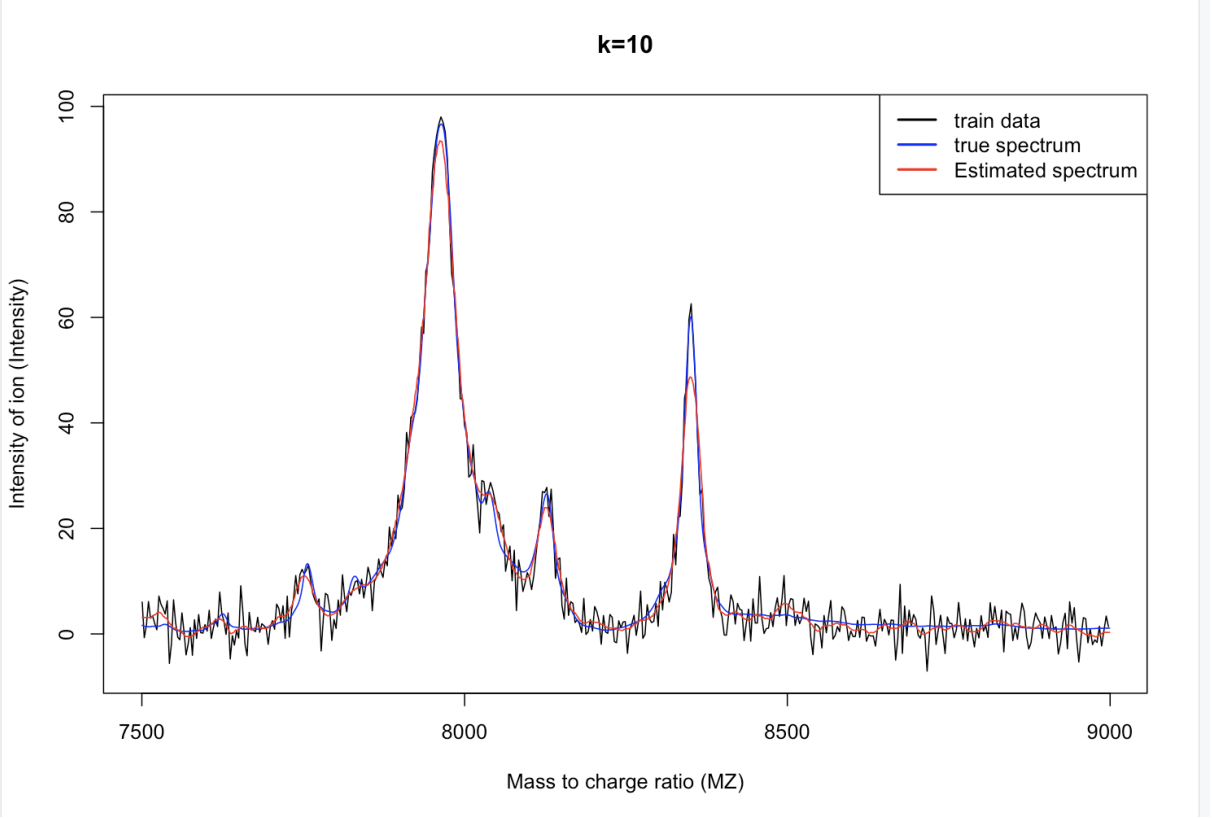
   k=2


k=2
train data
true spectrum
Estimated spectrum
Intensity of ion (Intensity)
0
20
40
60
80
100
7500
8000
8500
9000
Mass to charge ratio (MZ)

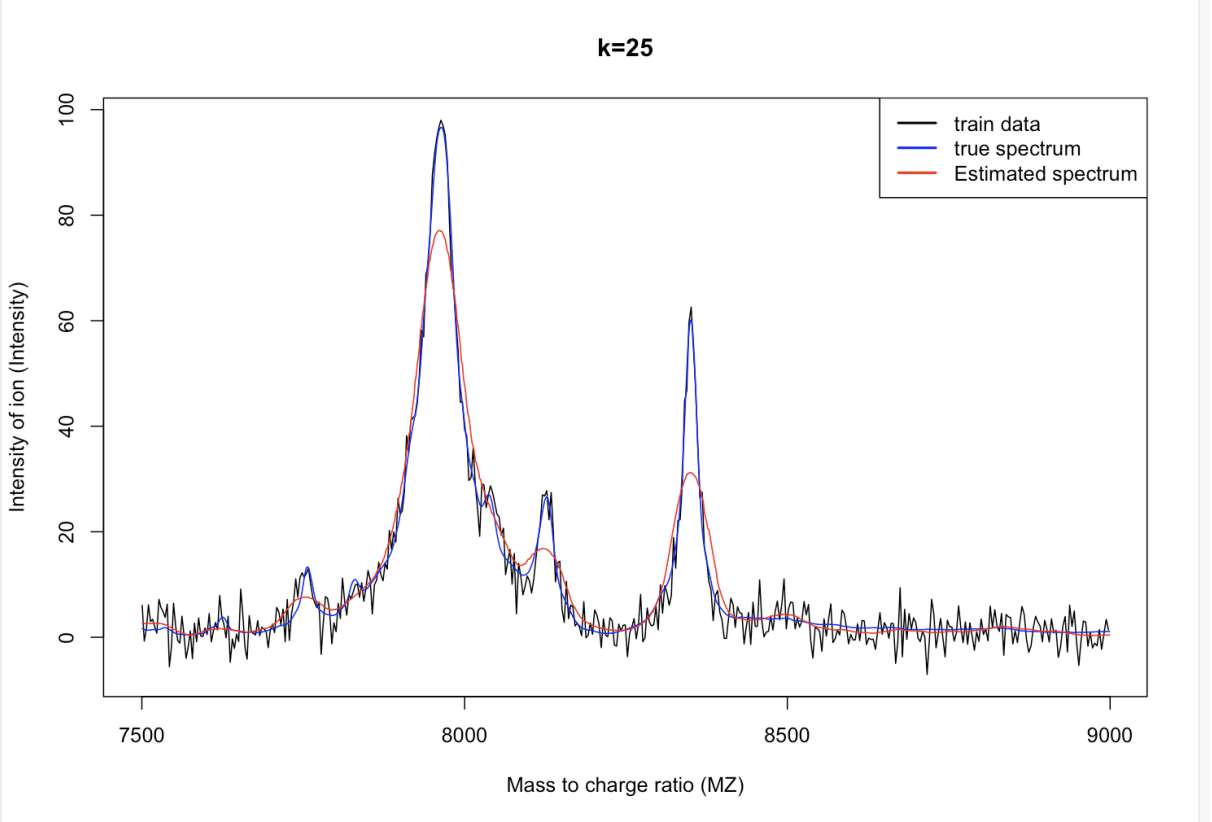
k =5


k=5
train data
true spectrum
Estimated spectrum
Intensity of ion (Intensity)
0
20
40
60
80
100
7500
8000
8500
9000
Mass to charge ratio (MZ)

k=10


k=10
train data
true spectrum
Estimated spectrum
Intensity of ion (Intensity)
0
20
40
60
80
100
7500
8000
8500
9000
Mass to charge ratio (MZ)

`k=25



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. [2 marks]

   With k=5, the predicted spectrum has the lowest MSE of 2.26. The estimated spectrum does not predicts true values of high intensity as well as k=2 with lower peaks compared to k=2 but does a better job at predicting true values of low intensity of ions compared to k=2 as shown by the flatter curve on the lower parts of the plot. By using only 2 training data point, k=2 with MSE of 5.1 is likely learning noise as it has a predicted spectrum similar to the training data. This can be shown on the plot where the estimates closely follow the training data. At k=10, the MSE yields 2.608. The estimated spectrum is more smooth compared k=5 and predicts true low intensity levels closer than the others but badly predicts high peaks of intensity levels as shown by its lower peaks. This suggests that by taking more neighbouring data points it is more robust to outliers from our training data that causes worse predictions for low intensity levels but is also overgeneralising the data. This makes it unable to capture the high peaks of the spectrum as well as k=5. Similarly at k=25, the MSE becomes very large (17.42) and the spectrum becomes even more smoother but is not predicting data well. Overall it seems that k=5 gives the best predictions as it is more robust to noise compared to k=2, allowing it to have relatively good performance in low intensity levels but also able to predict high intensity levels with peaks.

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. [2 marks].

   The estimated spectra achieved the aim of providing a smooth low noise estimate of background level as well as accurate estimation of peaks. This is shown in 8500 to 9000 MZ where although the training data is sporadic, the predicted estimate appears more flat, smoother and closes to the true spectrum. It also captures high peaks in the data. k-NN is able to achieve this because it calculates its prediction based on multiple similar data points. This allows it to capture non-linear characteristics

of the data such as identifying the general pattern in the data. This also makes it less affected by noise since it will try to fit data within the general proximity of other similar predicted data.

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)?

   Using cross validation, we found the best k to be 6 while the actual value of k that would minimise the actual mean-squared error is 7.

```
#5#train knn
knn = train.kknn(intensity ~ ., data = ms_train, kmax=25, kernel="optimal")
#best value for k = 6
knn$best.parameters

bestk_ms = fitted( kknn(intensity ~ ., ms_train, ms_test,
kernel = knn$best.parameters$kernel,
k = knn$best.parameters$k) )
```

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. [1 mark]

   To find the estimated standard deviation of the noise, we find the standard deviation of our predicted data from our model. This provides us an estimate for the noise because we make the assumption that differences in our reading and the actual value is caused by noise in our training data. With this we use the equation below:

$$xi = y - yi$$
$$s = sqrt(\sum(xi - x)^2/(N-1))$$

where xi is the residual from our model, x is the mean of the residual and N is the sample size of our test data (433). From this we calculated the standard deviation of the noise to be 1.421134.

```
> #Q3.6
> #cal residuals
> residuals = ms_test$intensity - bestk_ms
> #mean of residuals
> mean_residuals = mean(residuals)
> # Sum of squared error SSE
> sum_squared_residuals = sum((residuals - mean_residuals)^2)
> # Divide by (N-1) where N is the number of data points in the test set
> n = length(ms_test$intensity)
> var = sum_squared_residuals / (n - 1) # using unbiased
> # Take the square root to get the standard deviation
> sd_noise = sqrt(var)
> sd_noise
[1] 25.74217
>
```

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? [1 mark]

maximum estimated abundance of MZ = 7963.3

```
# find the index in which max peaks occur
idx = which.max(bestk_ms)
# find the MZ for the index found
mz = ms_test$MZ[idx]
```

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. Explain why you think these confidence intervals vary in size for different values of k. [3 marks]

From Q3.5 the data is shown below for the highest relative intensity

Row no	MZ	Intensity
283	7963.3	96.638

k=3

```
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%   (95.04, 98.00 )
Calculations and Intervals on Original Scale
>
```

The true intensity fits within the 95% confidence interval of (95.11, 98)

k=6 - from Q3.5

```
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%   (91.83, 97.96 )
Calculations and Intervals on Original Scale
>
```

The true intensity fits within the 95% confidence interval of (91.83, 97.96)

k=20

```
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%   (69.48, 92.77 )
Calculations and Intervals on Original Scale
>
```

The true intensity does not fits within the estimated 95% confidence interval of (69/48, 92.77)

The confidence interval vary in size when the k changes. k=3 gave the smallest interval because there is only 3 training data to predict from which means less variance in the resulting prediction. On the other hand k=6 and k=20 gave the 2nd largest and largest confidence interval respectively because it used the more data points for calculating a prediction, this introduces more variance as the prediction now relies on more data points than when k=3. Secondly, we observe that the confidence interval is decreasing as k increases. As shown in Q3.2, the estimates have lower peaks when k increases which aligns with these results. This is because the estimates are becoming more smoother and thus suggest that a high k decreases the noise in our data. This however leads to over generalisation as shown in k=20 where the true intensity is outside the 95% confidence interval.