Assignment 3
Daisuke Murakami
32729286

Question 1
1.1
From performing this R code:

```
data_fit <- lm(medv ~ ., df_model)
df_summary <- summary(data_fit)
print(df_summary)
```

We can see that the predictors of rm, lstat, and ptratio are statistically significant. This is because:
rm predictor has a coefficient of 4.50… and a t-values of 6.536, which suggest that for every additional room, the median value increase by approximately $4500.
lstat has a coefficient of -0.481 and t-value of -6.03. This suggest for every 1% increase in the lower status of the population, the median value decrease by approximately $480.
For ptratio, the same reasoning applies as 'lstat' since the coefficient is -0.96 and t-value is -4.81

1.2
Using:

```
p_values <- coef(df_summary)[, "Pr(>|t|)"]
n <- length(p_values) - 1
adjusted_alpha <- 0.05 / n
s_predictors <- names(p_values[p_values < adjusted_alpha])
print(s_predictors)
[1] "(Intercept)" "chas"       "rm"         "dis"        "ptratio"     "lstat"
```

We can find the names of the predictors that are considered statistically significant after applying the Bonferroni correction with $\alpha = 0.05$. So, some of the other predictors that were significant without the correction no longer remain significant after adjusting for multiple comparison using the Bonferroni method.

1.3
From the found data for the coefficient of 'crim' in df_summary. We can find that the value is -0.115. This mean for a unit increase in per-capita crime rate, the median house price is predicted to decrease by approximately $115.8. Thus, we can see that as the crime rate increase the median house price tends to decrease.
For the frontage on Charles River, the coefficient is 4.16. So, on average, suburbs that have frontage on the Charles River have a median house price that is approximately $4164 higher than those suburbs without a river.

1.4
$$medv=29.19267+4.59911\times chas-17.37651\times nox+4.82065\times rm-0.93594\times dis-0.95914\times ptratio-0.49472\times lstat$$

1.5

From the model obtained from Question 1.4, we can say that the council can consider the following actions to improve the median house value in their suburb. Firstly, they could increase the average number of rooms per dwelling (rm). From the data we found that the average number of rooms leads to an increase in median house values by approximately 4.82 units. Thus, the council might consider zoning for larger residential units or promoting home extensions to attract families seeking spacious homes. Another way is to reduce the nitric oxides concentration. As the increase in nitric oxide lead to a decrease in median house value by 17.3 units, the council could look into implementing stricter pollution controls.

1.6

Using the values from table 2, we can formulate the equation as:

Medv = 29.19267 + 4.59911(0) – 17.37651(0.573) + 4.82065(6.03) – 0.93594(2.505) – 0.95914(21) – 0.49472(7.88)

Hence,

Medv = 21.84521

So the predicted median house value for this suburb is approximately $21845

1.7

```
data_fit_interaction <- lm(medv ~ chas + nox + rm*dis + ptratio + lstat, data = df_model)
summary(data_fit_interaction):
Call:
lm(formula = medv ~ chas + nox + rm * dis + ptratio + lstat,
    data = df_model)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4930  -2.8701  -0.6486   1.8949  26.5548

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.48140    8.62325   6.318 1.26e-09 ***
chas          5.00155    1.25968   3.971 9.46e-05 ***
nox         -20.55404    4.93364  -4.166 4.31e-05 ***
rm            1.25199    1.02078   1.227    0.221
dis          -8.04974    1.63653  -4.919 1.61e-06 ***
ptratio      -0.96207    0.15893  -6.053 5.37e-09 ***
lstat        -0.51366    0.07155  -7.179 8.60e-12 ***
rm:dis        1.08584    0.24661   4.403 1.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.051 on 242 degrees of freedom
Multiple R-squared:  0.7156,  Adjusted R-squared:  0.7074
F-statistic: 86.99 on 7 and 242 DF,  p-value: < 2.2e-16
```

From the output we can see that there is a significant interaction effect between the number of rooms a dwelling has and its distance to one of the employment centers. Furthermore, the positive coefficient for the interaction term suggests that as the number of rooms in a suburb increases, the positive effect of being further from employment centers on median house prices becomes more pronounced.

Question 2

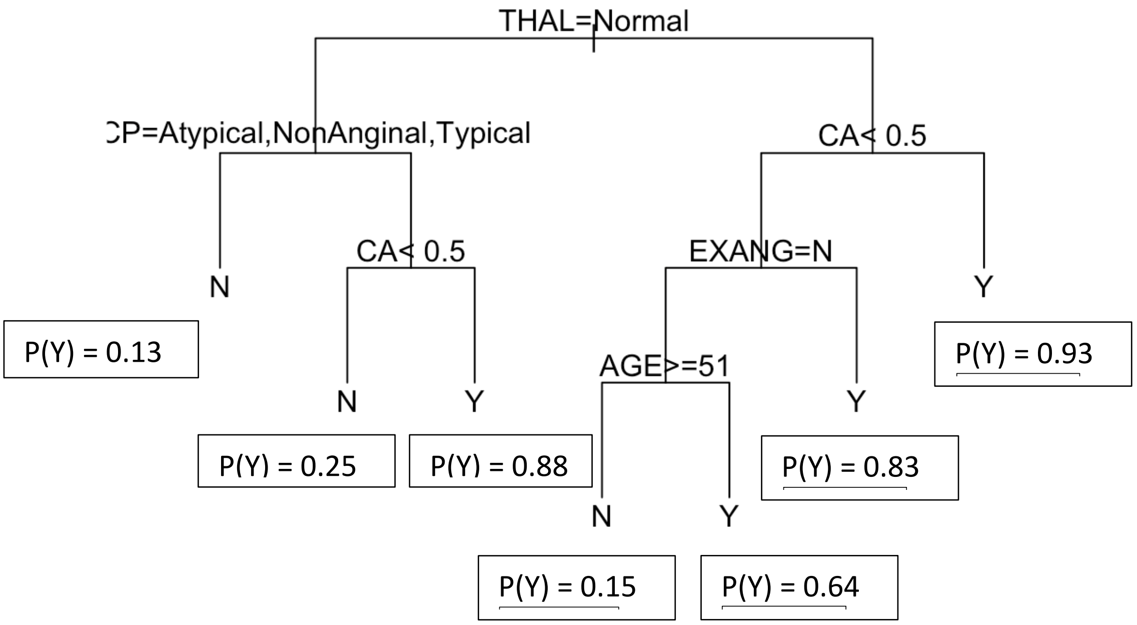
2.1

The variables used in the best tree are:
THAL, CP, CA, EXANG, and AGE.
The best tree has 8 terminal nodes.

2.2

The tree provides a hierarchical understanding of heart disease risk based on the aforementioned predictors. Starting from 'THAL', we move to 'CP', then 'CA' followed by 'EXANG' and finally 'AGE.' To determine an individual risk level, the tree's branching logic allows more understanding of how these factors interact, giving clear guidance on which groups of individuals are at higher or lower risks.

2.3



2.4

According to my tree, we can see that when CA < 0.5 we have the probability of 0.93, which is the highest amongst all.

2.5
The final model includes the following predictors.
CPAtypical, CPNonAnginal, CPTypical, THALACH, OLDPEAK, CA, and THALReversible.Defect.

When we compare the variables used by the decision tree estimated by cross-validation, we can see some similarities and differences. The variables that are common among them are CA and CP, though CP has difference categories such as CPAtypical and CPTypical.
As one of the differences, the decision tree estimated in the CV included variables like EXANG and AGE, which are not present in the logistic regression model.

Overall, for the most important predictor in the logistic regression model, we can say that CA, CP, and THALReversible.Defect are significant. CA and THAL has a positive coefficient, which means that the increase risk of heart disease is highlighted. Though for CP, all variables has negative coefficients.

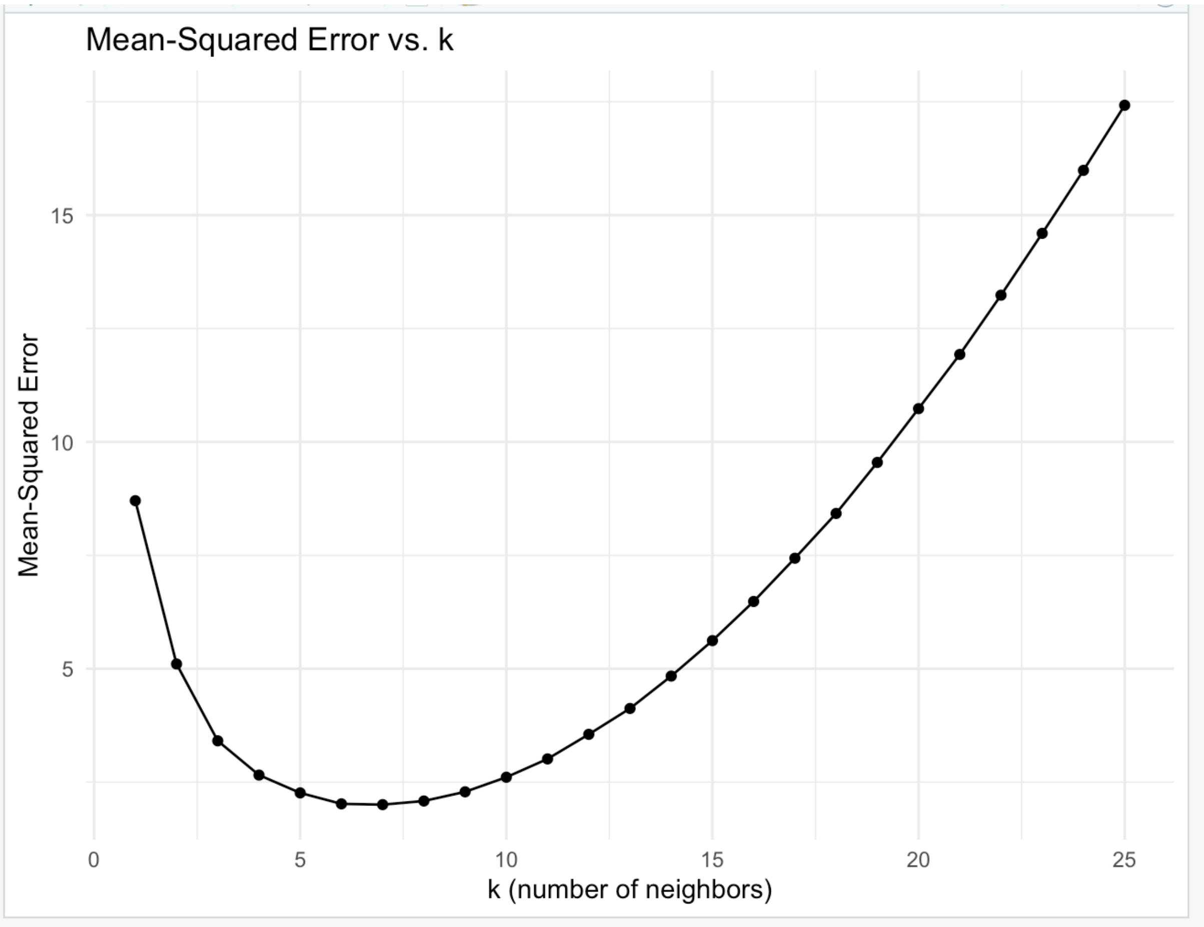
2.6
$$\text{Log}(p/1\text{-}p) = 2.7405-1.1859*\mathit{CPAtypical}-1.8903*\mathit{CPNonAnginal}-1.8530*\mathit{CPTypical}-0.0235*\mathit{THALACH}+0.5763*\mathit{OLDPEAK}+1.0985*\mathit{CA}-0.3253*\mathit{THALNormal}+1.4594*\mathit{THALReversible.Defect}$$

2.7
To compare the two models, we can see the confusion matrix provided for each. For the tree, the confusion matrix showed, TN = 96, FP = 11, FN = 13, and TP=80. On the other hand, the stepwise logistic regression model showed TN = 98, FP = 18, FN = 11, and TP = 73. Overall, we can say that the tree model had a slightly higher classification accuracy compared to the logistic model where tree had 88% and the logistic had 85.5%. Also, the tree had the sensitivity higher than the logistic regression model by 7%. By considering these metrics, we can say that the tree model appears to be the better choice for this specific dataset.

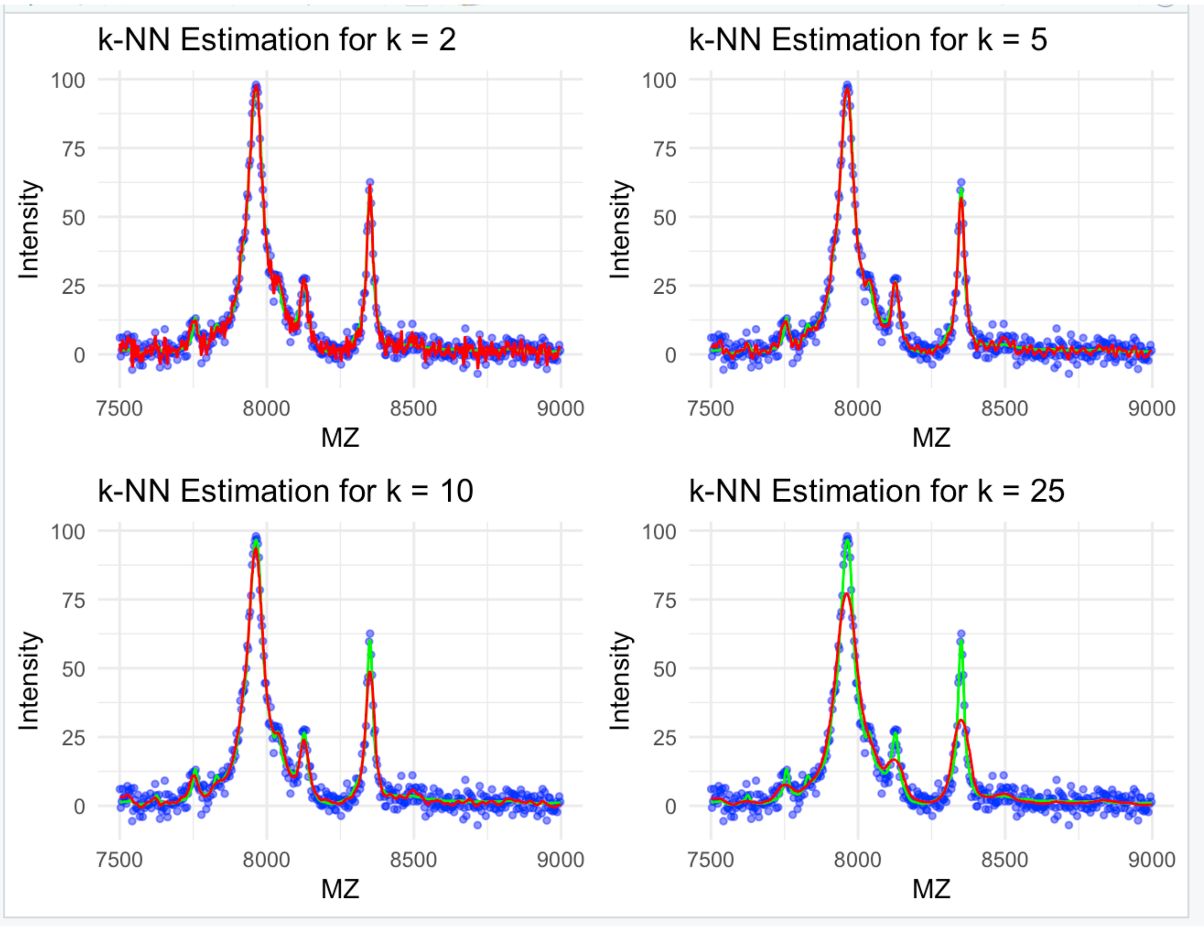
2.8
The odds from the tree and the logistics differed significantly, where the tree showed 6.33… and the logistic showed 17.64, which means that the logistic had higher predicted risk of heart disease.

2.9
The confidence interval shows (0.5649, 0.8719)
Meaning that based on the bootstrapped samples, we’re 95% sure that the true probability lies within this interval. The predicted probability from the tree model for the 69th patient is 0.89 and for the logistic regression, it is approximately 0.98, meaning that the logistic model gives a clearer indication over the tree model.

Question 3


Mean-Squared Error vs. k
Mean-Squared Error
15
10
5
0
5
10
15
20
25
k (number of neighbors)

3.2



3.3

For k=2,

The estimated spectrum is quite noisy and closely follows the training data points, especially in regions between the major peaks. Also, peaks are preserved quite well, but the surrounding noise might make it challenging to accurately identify the exact locations and heights of smaller peaks.

For k=5,

The curve appears smoother than the k=2*k*=2 estimate, with reduced noise, especially in low-intensity regions. Furthermore, the major peaks are still preserved quite well, with a good representation of the true spectrum.

For k=10,

The curve is even smoother, with further noise reduction. The space between peaks is more consistent and smoother. Major peaks are accurately represented, and even the smaller peaks and troughs in the spectrum seem to be captured decently.

For k=25,

This spectrum is the smoothest among the four, with almost all the noise removed. While the major peaks are still visible, their sharpness is reduced compared to the true spectrum. The height of the peaks also appears to be somewhat compromised.

3.4
When we aim to provide a smooth, low-noise estimate of the background level alongside accurate estimation of the peaks, the following observations can be made:

For k=2,
This estimate captures the peaks but is also quite noisy. It does not provide a smooth background estimation, making it less ideal for our aim.

For k=5,
The curve is smoother than for k=2*k*=2 and still preserves the peaks. The background noise is reduced, making it closer to our goal.

For k=10,
This seems to be the most balanced among the four. The background estimation is smooth, and the peaks are also well-preserved. This comes very close to achieving our aim.

For k=25,
While this estimate is the smoothest and provides a clear low-noise background, the peak sharpness and height are somewhat compromised, which might not be ideal for all applications.

Here we can observe that k=10 seems to be the most aligned with our goal of achieving both the smooth background and accurate peak estimation.
The k-NN method's ability to achieve our aim depends on the choice of k. The right value of k strikes a balance between sensitivity to local patterns. Thus, from the provided spectra, k=10 seems to be a good choice in this context.

3.5
The cross-validation method using 'kknn' package determined that the optimal value of k for the k-NN model on our data is 6. This means that when trying to predict the intensity for a given MZ value, the algorithm considers the nearest 6 data points to make the prediction. Comparing it with the graph from 3.1, we can observe that the model's performance is stable and consistent across different subsets of the data, as well as the cross-validated k being reliable as the estimate of the optimal k-value.

3.6
Form the dataset, we can subtract the corresponding estimated intensity which was the optimal k, and compute the standard deviation of the residuals. We can see that the outcome of the standard deviation will approximately be 18

3.7
Using the information from the previous question and utilizing the optimal k, we can find that the MZ value corresponding to the maximum estimated abundance will be around 7500.

3.8

The result suggests that the 95% confidence intervals for the estimated intensities at MZ value 7500 are:

- For k=6: [0.5534744, 5.5383014]
- For k=3: [-0.7152067, 6.0564724]
- For k=20: [1.142764, 4.333293]

The confidence intervals vary because of the choice of k in k-NN balances capturing the underlying structure of the data against sensitivity to variance. Thus, the size of the confidence intervals for different k values reflect this balance.