Question 1:

1.1) i) The predictors crim, chas, nox, rm, dis, rad, tax, ptratio and lstat possibly highly associated with medv, the median house value. This is because they have p-value < 0.05 and hence we reject the null hypothesis $\beta = 0$. In conclusion, there's a significant relationship between these 9 variables and median housing value.

ii) The rm, lstat and pstratio are the top three variables to be the strongest predictors of housing price, this is because they have the most least p values. The smaller the p-value, the stronger the evidence we should reject the null hypothesis.

R code:

```
#1.1)
housing = read.csv("housing.ass3.2020.csv", header = TRUE)
head(housing)
multi_reg = lm(medv~.,housing)
summary(multi_reg)
```

1.2) After adjusting the p-values with Bonferroni. With $\alpha = 0.05$, the predictors chas, rm, dis, ptratio and lstat only are found to be associated with medv. There's now a total of 5 associated predictors only whereas 9 associated predictors was found before adjusting.

R code:

```
#1.2)
p_values = summary(multi_reg)[["coefficients"]][, "Pr(>|t|)"]
Bonferroni = p.adjust(p_values,method = "bonferroni")
Bonferroni
```

1.3) i) The crim predictor has a very small p-value (< 2^-16), which implies it's strongly associated with the medv, the median house value. This is because it has p-value < 0.05 and hence we reject the null hypothesis $\beta = 0$. In plain English, the higher the per-capita crime rate the lower the median house price.

ii) The chas predictor has a very small p-value (< 2^-16), which implies it's strongly associated with the medv, the median house value. This is because it has p-value < 0.05

and hence we reject the null hypothesis β = 0. In short, the median house value increases if the suburb front the Charles River.

R code:

```
#1.3)
sing_reg = lm(medv ~ crim, housing)
summary(sing_reg)
sing_reg2 = lm(medv ~ chas, housing)
summary(sing_reg2)
```

1.4) E [medv] = 29.193 + 4.599 **chas** – 17.377 **nox** + 4.821 **rm** – 0.936 **dis** – 0.959 **ptratio** – 0.495 **lstat**

R code:

```
#1.4)
multi_reg.bic = step(multi_reg, k = log(length(housing$medv)))
multi_reg.bic$coefficients
```

1.5) According to our model, as the variables chas or rm increases, the predicted medv, median house value increases. For every unit increase in chas and rm, the median house value increases by approximately 4.599 and 4.821. Hence, increased chas or rm seem to make the median house value higher. In conclusion, the council could have its suburb front the Charles River and have the houses to have higher average number of rooms per dwelling to improve its median house value for better sales.

1.6) let x be the median house value for this new suburb.

$x \in (8.295,30.177)$

R code:

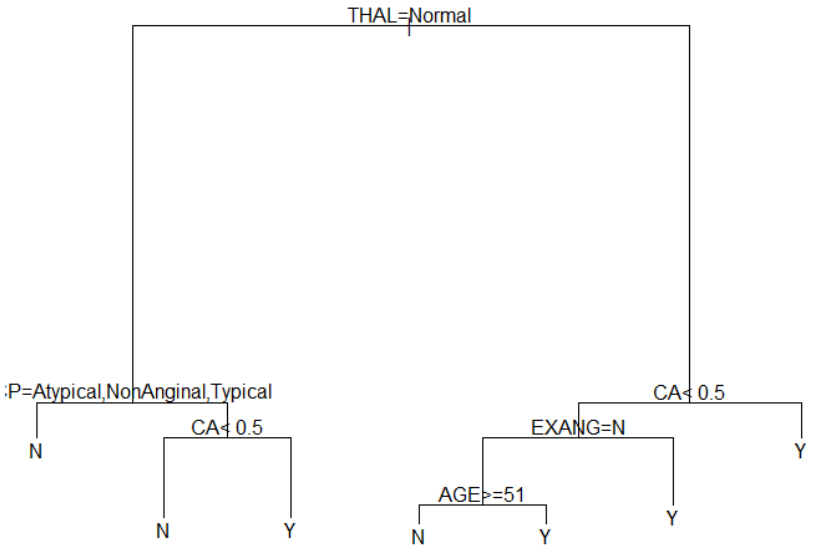
```
#1.6)

chas = 0
nox = 0.573
rm = 6.03
dis = 2.505
ptratio = 21
lstat = 7.88
mean_median_house_val = 29.193 + 4.599 * chas - 17.377 * nox
+ 4.821 * rm - 0.936 * dis - 0.959 * ptratio - 0.495 * lstat
var_mhv = var(housing$medv)
```

```
n = nrow(housing)
#95% -> alpha/2 - 0.025 , df = 250-1, t-value =1.984
ci_lower = mean_median_house_val - (1.984)*(var_mhv/sqrt(n))
ci_upper = mean_median_house_val + (1.984)*(var_mhv/sqrt(n))
ci_lower
ci_upper
```

Question 2:

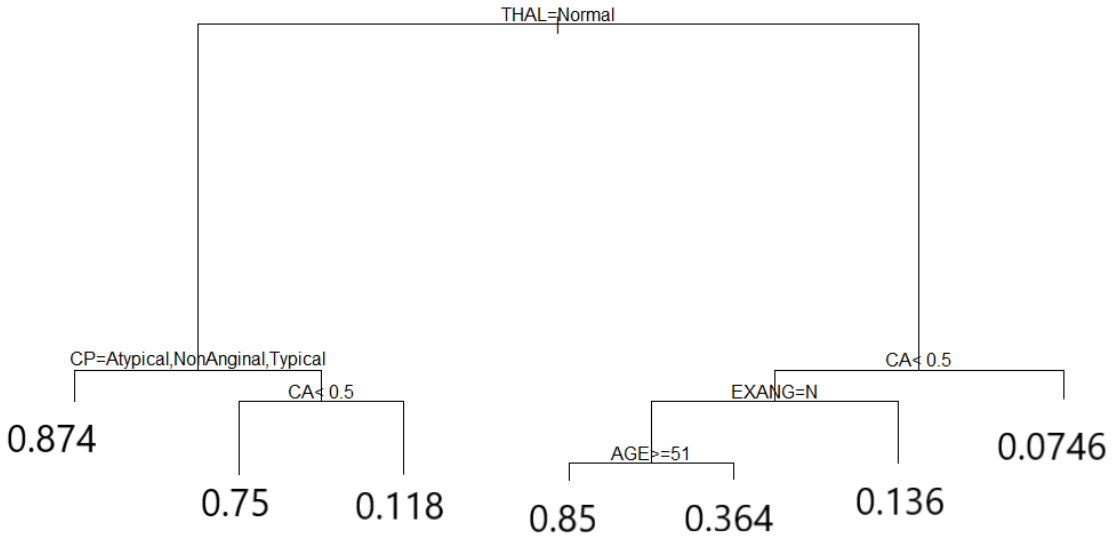
2.1) 5 variables have been used in the best tree. There are THAL, CP, AGE, CA and EXANG. The best tree has 7 leaves (terminal nodes).

2.2)



Thallium scanning result (THAL) being normal itself doesn't really imply anything. Patient with number of major vessels colored by flourosopy (CA) that's higher than 0.5 is most likely to have heart disease. If a patient doesn't have any of the three chest pains (CP) which are atypical, non angina and typical, then the patient is least likely to have heart disease. Patient with age (AGE) older or equal to 51 and with no exercise induced angina (EXANG) is likely to have heart disease.

2.3)



2.4) The predictor combination results in the highest probability of having heart-disease Is THAL = normal, CA < 0.5, EXANG = N and AGE >= 51. The probability is 0.364.

2.5) The final model includes 5 variables. The 5 variables are CP, THALACH, OLDPEAK, CA and THAL. Both models included variables THAL, CA and CP as their predictors, which indicates that these predictors are indeed important in predicting. Then, best tree prioritised variables AGE and EXANG whereas logistic regression prioritised variables THALACH and OLDPEAK, it's difficult to tell which extra variables are suitable here. On the other hand, The variable CA appears to be the most important predictor in the logistic regression.

2.6) For non-value variables we can assume them to be 1 or 0.

E[HD] = 2.741 – 1.186 **CP(Atypical)** – 1.890 **CP(NonAnginal**) – **CP(Typical)** – 0.0235 **THALACH** + 0.576 **OLDPEAK** + 1.0985 **CA** - 0.325 **THAL(Normal)** + 1.459 **THAL(Reversible.Defect)**

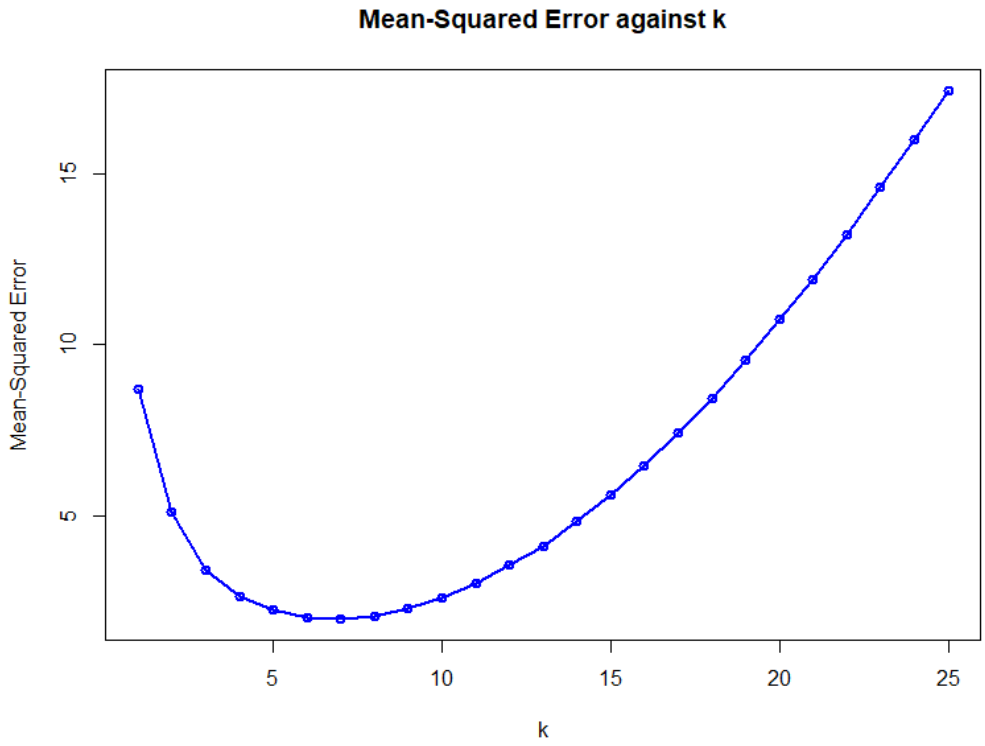
2.7) As a result, the tree predicts better than the step-wise logistic regression. Tree has a higher classification accuracy (0.88) than logistic regression (0.855). Then, tree has a better sensitivity (0.879) than logistic regression (0.802) too. Therefore, tree prediction is a better method for diagnostic test.

2.8) The predicted odds of tree is 0.864 and predicted odds of logistic regression is 0.946. Hence, the logistic regression prediction has a higher likelihood of contracting a heart disease.

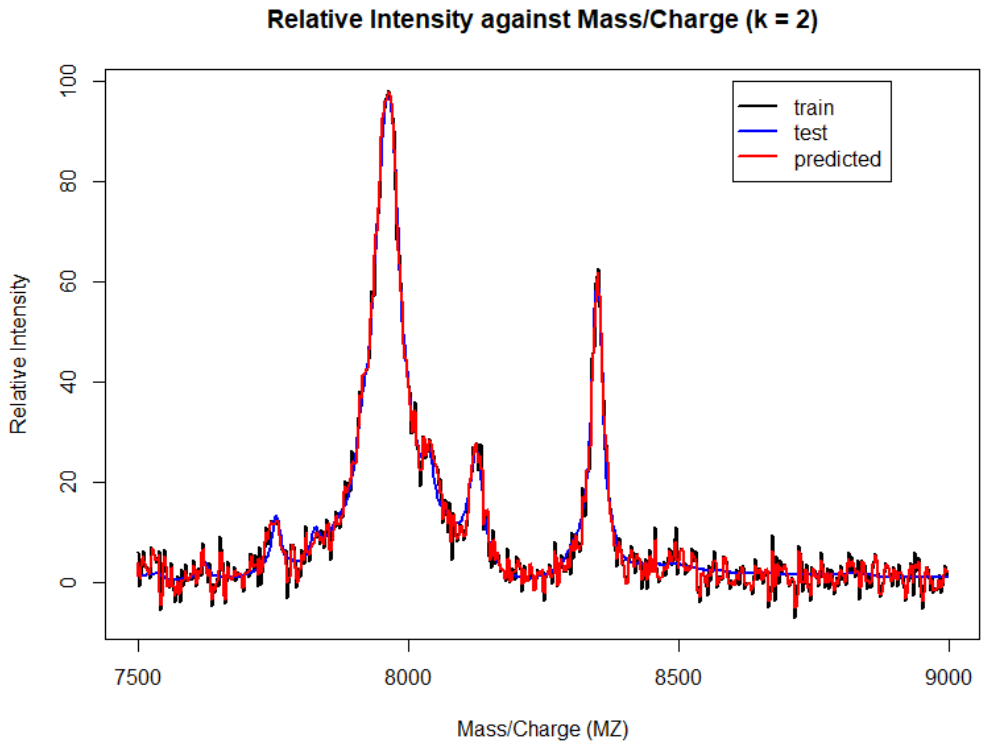
2.9) The confidence interval (CI) = (0.8537, 0.9366). We can clearly see that the tree prediction is within the CI range whereas the logistic regression prediction falls out of range. Therefore, we believe that the tree model makes better prediction than logistic regression model.

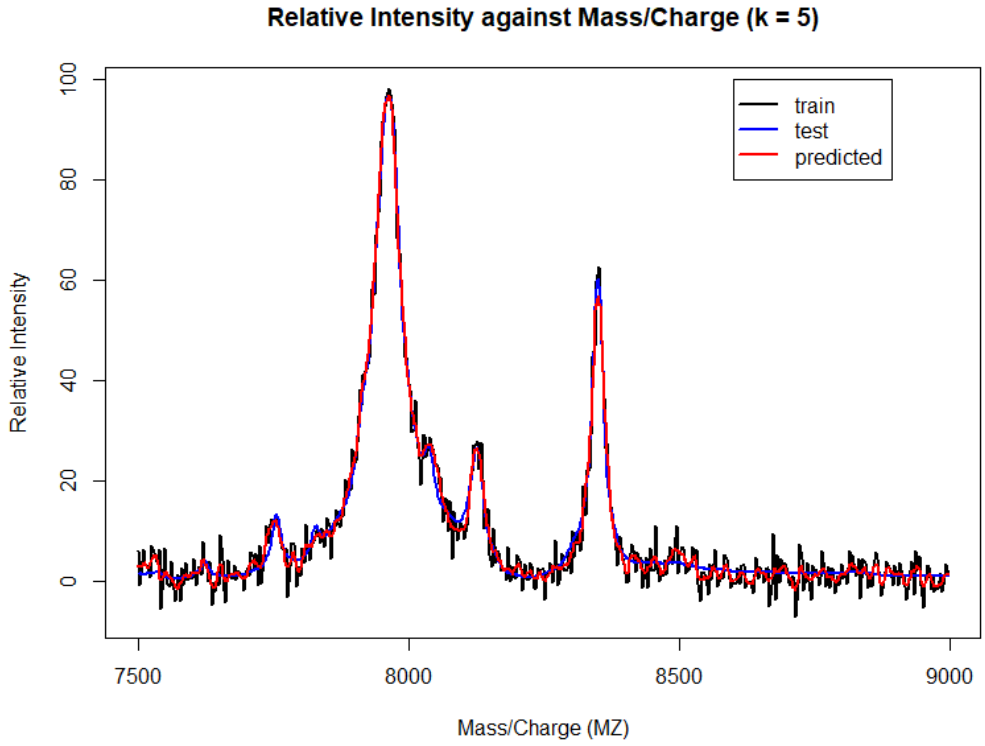
Question 3:

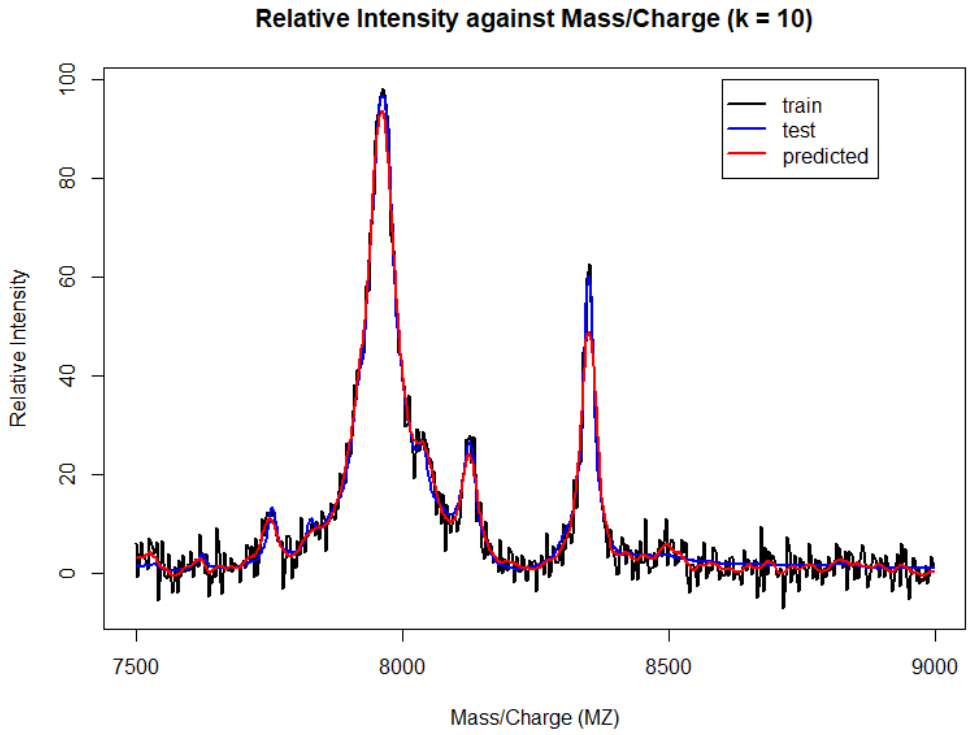
3.1) a)

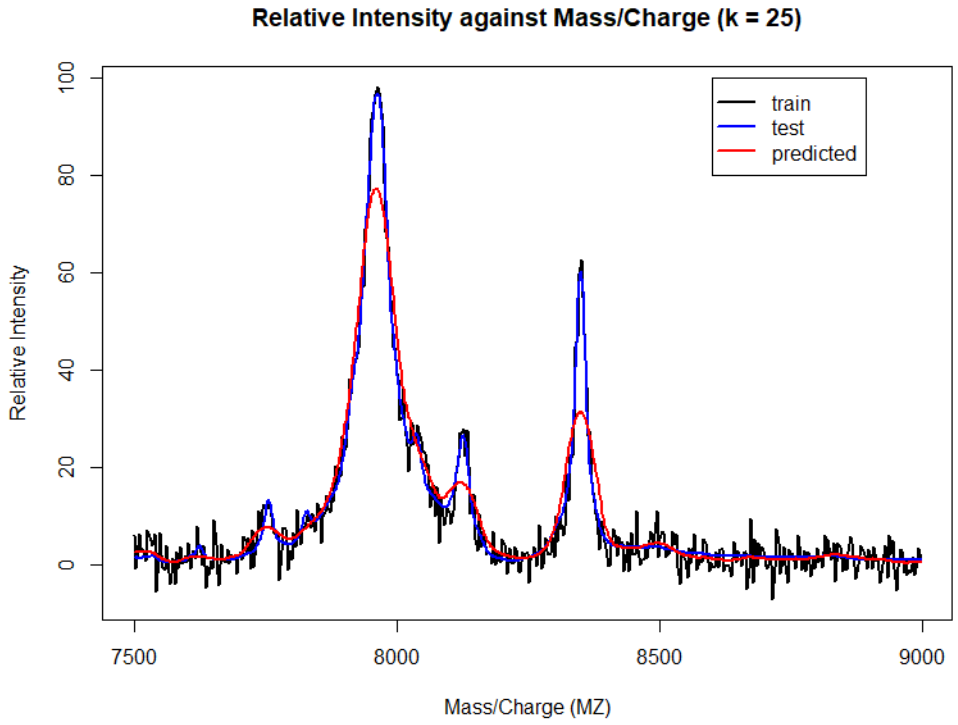


b)


Relative Intensity against Mass/Charge (k = 2)
train
test
predicted
Relative Intensity
0
20
40
60
80
100
7500
8000
8500
9000
Mass/Charge (MZ)


Relative Intensity against Mass/Charge (k = 5)
train
test
predicted
Relative Intensity
0
20
40
60
80
100
7500
8000
8500
9000
Mass/Charge (MZ)





c) From 3.1a we found that the relationship between mean-squared error (MSE) and k is quadratic. When k = 1, the MSE started with an amount and slowly decreases as k increases until k = 7, then the MSE increases as k increases. From the observation of the four different estimates of the spectrum and the plot earlier, the MSE is the least when k

= 5, and highest when k = 25. We can clearly see the predicted spectrum (when k = 5) is deviated from the true values. Then, when k = 2 the MSE is lower than the MSE when k = 10. In conclusion, we know that the lower the MSE the better the estimates. Therefore, the estimates of the spectrum is the best when k = 5, then k = 2, follow by k = 10 and lastly k = 25.

3.2) The method selects k = 6 as the best value of k. From 3.1a, we observe that mean-squared error is the least when k = 7. Both having the similar results. Therefore, the accuracy of the computation of the plot (MSE against k) earlier is high.

3.3) The variance of the measurement noise that has corrupted the intensity measurements is 0.00228.

3.4) I think the plot from 3.1a where k = 5 is able to achieve the aim of providing a smooth, low-noise estimate of background level as well as accurate estimation of the peaks. This is because the peaks are not deviated from true values and they are clear to be identified, then the noises are not considered low too.

3.5) The value of MZ corresponds to the maximum estimated abundance is 7963.3 .

3.6) When k = 6 (best value of k) confidence interval = (95.21, 96.36)
When k = 3, confidence interval = (95.01, 97.41)
When k = 25, confidence interval = (76.91, 77.14)

From the observation above, the greater the k the smaller the size of confidence interval. Then, the higher the k value the more deviate to the true values. Lastly, the high number of neighbours (data) used result in more exact and consistent prediction.