Case-1: Boston Housing Prices

QUESTIONS (10 marks)
[Answer these in your Answer document]
1. After performing EDA on the Boston Housing dataset, which variable shows the strongest negative correlation with the target variable 'medv' (median home value)? (1 marks)

'lstat' shows the strongest negative correlation with the target variable 'medv'.
The correlation matrix shows that there is a correlation coefficient of -0.7376627 between these two variables which is the most negative.

2. Share your final fine-tuned trained model summary below and provide interpretation of the summary in bullet points including variable coefficients.
(1 + 2 marks)

```
> summary(full_mdl)

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

Residuals:
    Min      1Q  Median      3Q     Max
-15.491  -2.753  -0.499   1.942  24.549

Coefficients:
              Estimate Std. Error t value
(Intercept)  33.593363   5.735441   5.857
crim         -0.091542   0.038713  -2.365
zn            0.029760   0.016233   1.833
indus        -0.043242   0.070898  -0.610
chas          2.911364   0.997890   2.918
nox         -16.335169   4.467780  -3.656
rm            3.964627   0.458191   8.653
age           0.001477   0.015580   0.095
dis          -1.271695   0.230636  -5.514
rad           0.276750   0.074258   3.727
tax          -0.010817   0.004152  -2.605
ptratio      -0.928753   0.152052  -6.108
b             0.008713   0.002988   2.916
lstat        -0.484062   0.062091  -7.796
             Pr(>|t|)
(Intercept) 1.11e-08 ***
crim         0.018605 *
zn           0.067624 .
indus        0.542320
chas         0.003762 **
nox          0.000296 ***
rm           < 2e-16 ***
age          0.924507
dis          6.93e-08 ***
rad          0.000227 ***
tax          0.009589 **
ptratio      2.74e-09 ***
b            0.003781 **
lstat        7.77e-14 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 4.611 on 342 degrees of freedom
Multiple R-squared: 0.7482, Adjusted R-squared: 0.7386
F-statistic: 78.15 on 13 and 342 DF, p-value: < 2.2e-16

- intercept estimate of 33.593363 means that when all other predictors are 0, the baseline of medv is about $33 593.36
- crim estimate of -0.091542 shows that when the crime rate rises by 1%, the median home value decreases by $92.54, assuming all other predictors are constant and crim variable is signficant
- ptratio estimate of -0.928753 means that larger class sizes would have a negative effect on home values
- very low Pr(>|t|) valuptratioes (p<0.01) for rm, lstat, ptratio,nox and chas means they are statistically significant to medv
- most predictors have very low p values (p < 0.05) except for age, indus and zn meaning most are statistically significant
- F-statistic of 78.15 with a p value less than 2.2e-16 (p < 0.001) means that the overall model is statistically significant (predictors as a whole explain a significant portion of the variation in medv)
- model's Adjusted R-squared value of 0.7386 indicates that the predictors together can explain about 73.86% of the variation in medv, which shows a strong fit for the real world data
- residual standard error of 4.611 means that the predicted medv is expected to deviate from the actual ones by $4611

3. What is the difference between your train and test accuracy. What does it tell you? (1 +1 marks)

```
> train_pred <- predict(full_mdl, newdata = train)
> test_pred <- predict(full_mdl, newdata = test)
> train_r2 <- cor(train_pred, train$medv)^2
> test_r2 <- cor(test_pred, test$medv)^2
> train_r2
[1] 0.7481554
> test_r2
[1] 0.7207317
> train_r2 - test_r2
[1] 0.02742374
```

- The R squared for the training set is 0.7481554 which means that the model explains 74.8% of the variation in medv for the data it was trained on
- The R squared for the testing set is 0.7207317 which means that the model explains 72.1% of the variation in the unseen data
- a difference of 0.02742374 between these two values which is quite small means that the model performs well on the train and unseen data so it generalizes well and not overfitting
- this model can be used to predict and estimate values for new properties well

4. If you retain ‘rm’ in your final model, what is the coefficient for 'rm' (average rooms per dwelling). What is the correct business interpretation? (1 marks)

- based on the summary given above, the coefficient estimate for rm is 3.964627
- this means that for every extra room per dwelling, the medv or median home value increases by about $3964.63
- this assumes that all other factors / predictors / independent variables remain constant
- so this means that houses with more rooms will become more valuable
- this means that it makes more business sense to build more houses with more rooms
- buyers are more likely to pay more for it and they can expect to sell it at a way higher price
- this can be marketed to potential buyers as a lucrative investment

5. Identify the top 2 predictors of the medv price based on variable importance analysis of your final model. (1 marks)

- the top 2 predictors are rm (8.652778) and lstat (7.796054)

6. What is the R-squared of your final model? Provide a business interpretation.
(1 + 1 marks)

- Multiple R-squared: 0.7482, Adjusted R-squared: 0.7386 from the summary
- roughly about 74% (73.86%) of the variation in the median house value (medv) can be explained by the 13 predictors in the model
- from the business perspective, it means that the model is quite accurate, reliable and trustworthy
- but it must be noted that about 26% of the variation is unexplained probably due to other hidden factors or randomness

Case-2: Heart Disease Risk Prediction

QUESTIONS (10 marks)
[Answer these in your Answer document]
1. After examining the target variable 'AHD' (heart disease), what is the exact class distribution (percentages for Yes/No)? Is this a balanced or imbalanced dataset? (1 mark)

```
> table(df2$AHD)

 No Yes
164 139
> prop.table(table(df2$AHD))

       No       Yes
0.5412541 0.4587459
> prop.table(table(df2$AHD)) * 100

      No      Yes
54.12541 45.87459
```

- This implies that the dataset is approximately balanced because the class distribution is almost 50:50
- No AHD is 54% while Yes AHD is 46%, and they are quite balanced and not so skewed
- the two classes are nearly equal in size so they are quite balanced

2. Share your final fine-tuned logistic regression model summary below and provide interpretation of the summary in bullet points including significant predictors and their odds ratios. (1 + 2 marks)

```
> summary(mdl2)

Call:
glm(formula = AHD ~ ., family = binomial, data = train2[, -1])

Coefficients:
                  Estimate Std. Error z value
(Intercept)       -3.821249   3.123882  -1.223
Age               -0.020449   0.027698  -0.738
Sex1               1.148934   0.598034   1.921
ChestPainnonanginal -2.227295   0.597094  -3.730
ChestPainntypical -0.792243   0.649839  -1.219
ChestPaintypical    -2.489071   0.803574  -3.098
RestBP             0.019707   0.013030   1.513
```

```
Chol                0.001748   0.004975   0.351
Fbs1               -0.654345   0.634776  -1.031
RestECG1           -0.066187   4.614172  -0.014
RestECG2            0.878659   0.447962   1.961
MaxHR              -0.011852   0.012825  -0.924
ExAng1              0.657982   0.504000   1.306
Oldpeak             0.545917   0.280377   1.947
Slope2              1.384122   0.556719   2.486
Slope3              0.259785   1.116505   0.233
Ca1                 2.481900   0.580161   4.278
Ca2                 3.240210   0.904466   3.582
Ca3                 2.427108   0.953875   2.544
Thalnormal          0.275711   1.000963   0.275
Thalreversable      1.736081   0.975786   1.779
                   Pr(>|z|)
(Intercept)        0.221240
Age                0.460351
Sex1               0.054708 .
ChestPainnonanginal 0.000191 ***
ChestPainnontypical 0.222792
ChestPaintypical    0.001952 **
RestBP             0.130400
Chol               0.725303
Fbs1               0.302622
RestECG1           0.988555
RestECG2           0.049826 *
MaxHR              0.355409
ExAng1             0.191716
Oldpeak            0.051525 .
Slope2             0.012911 *
Slope3             0.816012
Ca1                1.89e-05 ***
Ca2                0.000340 ***
Ca3                0.010944 *
Thalnormal         0.782974
Thalreversable     0.075213 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.58  on 237  degrees of freedom
Residual deviance: 146.10  on 217  degrees of freedom
AIC: 188.1

Number of Fisher Scoring iterations: 6

> exp(coef(mdl2))
       (Intercept)              Age             Sex1 ChestPainnonanginal
        0.02190042        0.97975900        3.15482668        0.10781968
ChestPainnontypical    ChestPaintypical            RestBP              Chol
        0.45282780        0.08298706        1.01990296        1.00174964
             Fbs1          RestECG1          RestECG2             MaxHR
        0.51978259        0.93595552        2.40766853        0.98821766
           ExAng1           Oldpeak            Slope2            Slope3
        1.93089224        1.72619118        3.99132135        1.29665108
              Ca1               Ca2               Ca3        Thalnormal
       11.96397321       25.53907461       11.32608006        1.31746679
    Thalreversable
```

5.67505655

- Chest Pain non-anginal has p value of 0.000191 and an estimate of -2.227295
- this means that there is a strong negative association with AHD
- Chest Pain typical: has a negative association
- Rest ECG2 has a positive association
- Slope2: has a positive association
- Ca1,2,3 have strong positive association
- list of significant predictors (p value < 0.05) and their odds ratio:
  - ChestPainnonanginal: 0.10781968
  - ChestPaintypical: 0.08298706
  - RestECG2: 2.40766853
  - Slope2: 3.99132135
  - Ca1,2,3: 11.96397321, 25.53907461, 11.32608006

3. Is your final model generalizable? (2 marks)

```
> library(caret)
> confusionMatrix(
+   factor(pred_classes, levels = c("No", "Yes")),
+   test2$AHD,
+   positive = "Yes"
+ )
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  31   7
       Yes  1  20

               Accuracy : 0.8644
                 95% CI : (0.7502, 0.9396)
    No Information Rate : 0.5424
    P-Value [Acc > NIR] : 1.458e-07

                  Kappa : 0.722

 Mcnemar's Test P-Value : 0.0771

            Sensitivity : 0.7407
            Specificity : 0.9688
         Pos Pred Value : 0.9524
         Neg Pred Value : 0.8158
             Prevalence : 0.4576
         Detection Rate : 0.3390
   Detection Prevalence : 0.3559
      Balanced Accuracy : 0.8547

       'Positive' Class : Yes
```

- The model is generalizable because it has
  - an accuracy of 86.44% on the test data
  - a sensitivity of 74.07% which is quite good
  - a specificity of 96.9% which is super good
- these imply that the model captures the underlying patterns in the data

- instead of overfitting to the training set
- this means it makes more trustworthy predictions on unseen data

4. If you retain ‘MaxHR’ in your final model, what is the odds ratio for ‘MaxHR’ (maximum heart rate)? What is the correct clinical interpretation? (1 mark)

```
> summary(mdl2)

Call:
glm(formula = AHD ~ ., family = binomial, data = train2[, -1])

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.821249   3.123882  -1.223 0.221240
Age               -0.020449   0.027698  -0.738 0.460351
Sex1               1.148934   0.598034   1.921 0.054708 .
ChestPainnonanginal -2.227295   0.597094  -3.730 0.000191 ***
ChestPainnontypical -0.792243   0.649839  -1.219 0.222792
ChestPaintypical    -2.489071   0.803574  -3.098 0.001952 **
RestBP             0.019707   0.013030   1.513 0.130400
Chol               0.001748   0.004975   0.351 0.725303
Fbs1               -0.654345   0.634776  -1.031 0.302622
RestECG1           -0.066187   4.614172  -0.014 0.988555
RestECG2            0.878659   0.447962   1.961 0.049826 *
MaxHR              -0.011852   0.012825  -0.924 0.355409
ExAng1             0.657982   0.504000   1.306 0.191716
Oldpeak            0.545917   0.280377   1.947 0.051525 .
Slope2             1.384122   0.556719   2.486 0.012911 *
Slope3             0.259785   1.116505   0.233 0.816012
Ca1                2.481900   0.580161   4.278 1.89e-05 ***
Ca2                3.240210   0.904466   3.582 0.000340 ***
Ca3                2.427108   0.953875   2.544 0.010944 *
Thalnormal         0.275711   1.000963   0.275 0.782974
Thalreversable      1.736081   0.975786   1.779 0.075213 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.58  on 237  degrees of freedom
Residual deviance: 146.10  on 217  degrees of freedom
AIC: 188.1

Number of Fisher Scoring iterations: 6

> exp(coef(mdl2)["MaxHR"])
    MaxHR
0.9882177

> (1 - exp(coef(mdl2)["MaxHR"])) * 100
   MaxHR
1.178234
```

- odds ratio for Max HR is .9882177 which means that each 1 bpm increase in max heart rate
- odds of having heart diseases falls by 1.18% assuming all variables are constant
- so in simple words if you exervise more and have a higher max heart rate you are less likely to die from heart disease

5. Identify the top 2 predictors of heart disease risk based on variable importance analysis of your final model. (1 mark)

```
> summary(mdl2)

Call:
glm(formula = AHD ~ ., family = binomial, data = train2[, -1])

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.821249   3.123882  -1.223 0.221240
Age               -0.020449   0.027698  -0.738 0.460351
Sex1               1.148934   0.598034   1.921 0.054708
ChestPainnonanginal -2.227295   0.597094  -3.730 0.000191
ChestPainnontypical -0.792243   0.649839  -1.219 0.222792
ChestPaintypical   -2.489071   0.803574  -3.098 0.001952
RestBP             0.019707   0.013030   1.513 0.130400
Chol               0.001748   0.004975   0.351 0.725303
Fbs1               -0.654345   0.634776  -1.031 0.302622
RestECG1           -0.066187   4.614172  -0.014 0.988555
RestECG2           0.878659   0.447962   1.961 0.049826
MaxHR              -0.011852   0.012825  -0.924 0.355409
ExAng1             0.657982   0.504000   1.306 0.191716
Oldpeak            0.545917   0.280377   1.947 0.051525
Slope2             1.384122   0.556719   2.486 0.012911
Slope3             0.259785   1.116505   0.233 0.816012
Ca1                2.481900   0.580161   4.278 1.89e-05
Ca2                3.240210   0.904466   3.582 0.000340
Ca3                2.427108   0.953875   2.544 0.010944
Thalnormal         0.275711   1.000963   0.275 0.782974
Thalreversable     1.736081   0.975786   1.779 0.075213

(Intercept)
Age
Sex1               .
ChestPainnonanginal ***
ChestPainnontypical
ChestPaintypical   **
RestBP
Chol
Fbs1
RestECG1
RestECG2           *
MaxHR
ExAng1
Oldpeak            .
Slope2             *
Slope3
Ca1                ***
Ca2                ***
Ca3                *
Thalnormal
Thalreversable     .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.58  on 237  degrees of freedom
```

Residual deviance: 146.10 on 217 degrees of freedom
AIC: 188.1

Number of Fisher Scoring iterations: 6

>
> exp(coef(mdl2))
```
       (Intercept)               Age              Sex1 ChestPainnonanginal
        0.02190042         0.97975900         3.15482668         0.10781968
ChestPainnontypical    ChestPaintypical             RestBP               Chol
        0.45282780         0.08298706         1.01990296         1.00174964
              Fbs1           RestECG1           RestECG2              MaxHR
        0.51978259         0.93595552         2.40766853         0.98821766
            ExAng1            Oldpeak             Slope2             Slope3
        1.93089224         1.72619118         3.99132135         1.29665108
               Ca1                Ca2                Ca3         Thalnormal
       11.96397321        25.53907461        11.32608006         1.31746679
    Thalreversable
        5.67505655
```

- based on the above output, 3 predictors are statistically significant ($p < 0.05$, with ***)
- these are:
  - ChestPainnonanginal
  - Ca1
  - Ca2
- but if we were to take a look at their odds ratio:
  - Ca1 is 11.96397321
  - Ca2 is 25.53907461
  - ChestPainnonanginal is 0.10781968
- this implies that Ca1 and Ca2 both have the most impact due to them having the top two odds ratio
- so in simple terms: the number of major blood vessels affected are top predictors of heart disease

6. What is the AUC (Area Under Curve) of your final model? Provide a clinical interpretation of this performance metric. (1 + 1 marks)

> auc(roc_obj)
Area under the curve: 0.912

- The AUC of the final model is 0.912
- this means there is very good discrimination between patients with and without heart disease
- so the model can correctly tell apart patients with heart disease from those who do not about 91.2% of the time
- this indicates that the model is reliable and can be used for health screening especially those at a great risk of getting it
- clinically, it allows doctors to quickly identify patients early and give early intervention and do more testing
- in simple terms, it allows for people more likely to get the disease to be identified earlier so that they can be treated faster so they do not die