STA 380: INTRO TO MACHINE LEARNING – Problem Set 1

Exploratory Data Analysis

Question 10

Question (a)

This exercise involves the Boston housing data set.
(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library.
> library(ISLR2)
Now the data set is contained in the object Boston.
> Boston
Read about the data set:
> ?Boston
How many rows are in this data set? How many columns? What do the rows and columns represent?

```
> cat("The number of rows in the Boston dataset:", nrow(Boston), "\n")
The number of rows in the Boston dataset: 506
> cat("The number of columns in the Boston dataset:", ncol(Boston), "\n")
The number of columns in the Boston dataset: 14
```

The Boston dataset contains 506 rows and 14 columns.

```
> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"
[10] "tax"     "ptratio" "black"   "lstat"   "medv"
```

The dataset contains the housing values in 506 suburbs of Boston. Each row represents a suburb in Boston.

Description of each of the columns in the Boston dataset is as below.

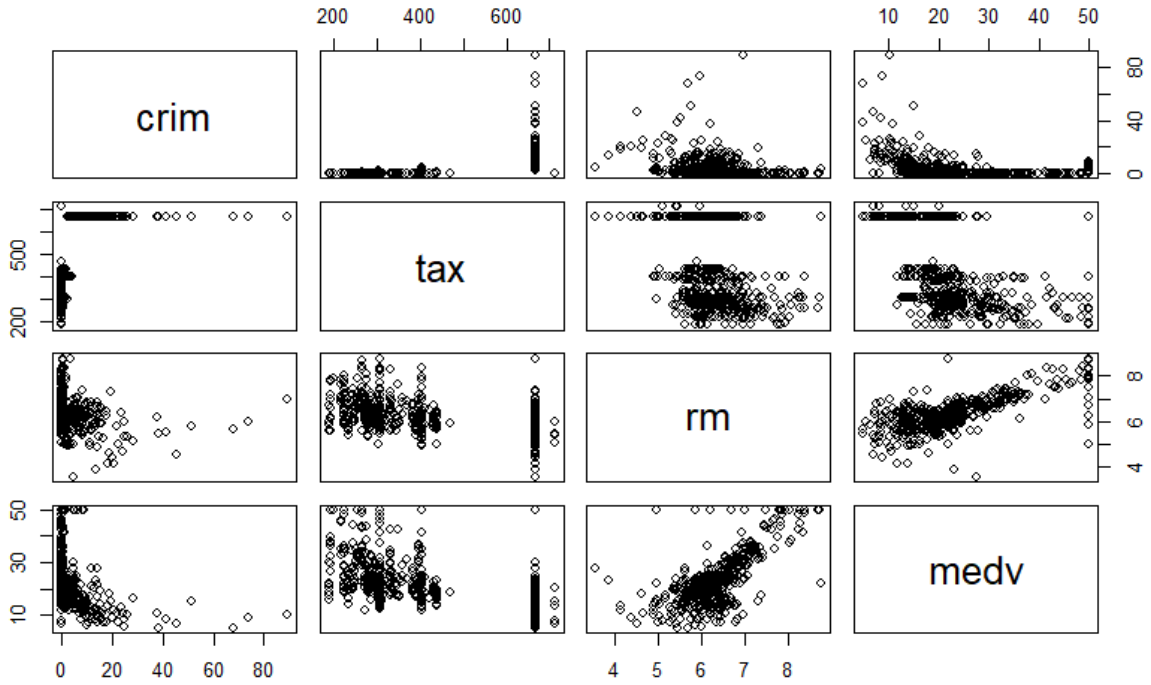
- crim: per capita crime rate by town.
- zn: proportion of residential land zoned for lots over 25,000 sq.ft.
- indus: proportion of non-retail business acres per town.
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- nox: nitrogen oxides concentration (parts per 10 million).
- rm: average number of rooms per dwelling.
- age: proportion of owner-occupied units built prior to 1940.
- dis: weighted mean of distances to five Boston employment centres.

- rad: index of accessibility to radial highways.
- tax: full-value property-tax rate per $10,000.
- ptratio: pupil-teacher ratio by town.
- lstat: lower status of the population (percent).
- medv: median value of owner-occupied homes in $1000s.

Source - https://islp.readthedocs.io/en/latest/datasets/Boston.html

Question(b)

Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



From the above graph, we can interpret the following findings:

- crim vs tax - As the property tax rate increases, the crime rate also increases, indicating a positive correlation.
- crim vs rm - The crime rate decreases as the average rooms per dwelling increase, implying wealthier neighborhoods might have lower crime rates.
- rm vs medv - As the number of rooms increases, the median value of the owner-occupied houses also increases, indicating a positive correlation.

- crim vs medv - The crime rate decreases as the median value of the owner-occupied houses increases, indicating a negative relationship.

Question (c)

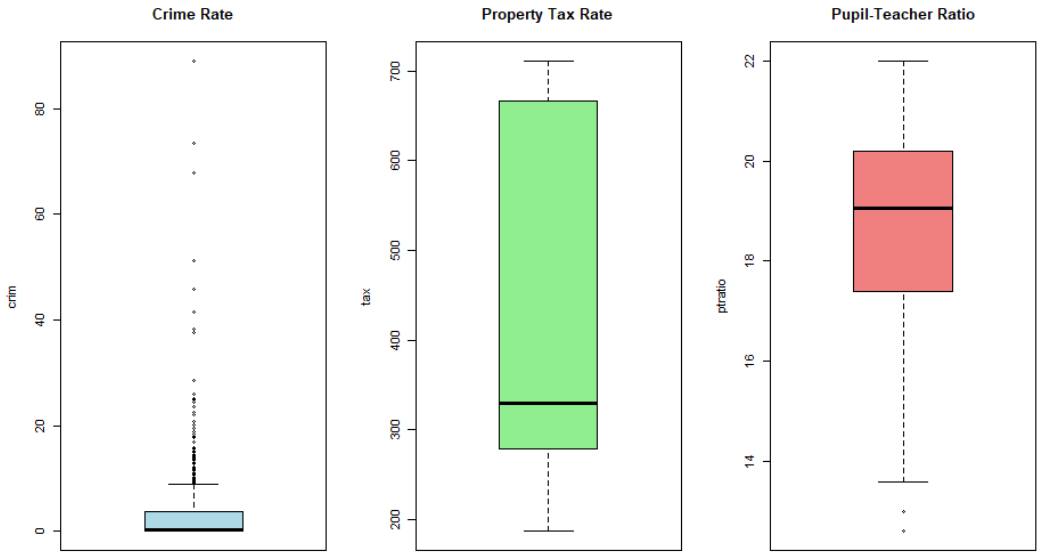
Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

- By looking at the pairwise plots, it was evident that as crime rates increased, tax rates and the median value of owner-occupied houses also increased, indicating a positive correlation.
- Crime rates decreased as the average number of rooms per dwelling increased, indicating a negative correlation.

Question (d)

Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil teacher ratios? Comment on the range of each predictor.

```
> range(Boston$crim)
[1]  0.00632 88.97620
> range(Boston$tax)
[1] 187 711
> range(Boston$ptratio)
[1] 12.6 22.0
```



- The highest crime rate in the Boston data set is 88.97620. The boxplot reveals that there are several outliers in the upper quartile.

- The highest tax rate goes up to 711, and the highest pupil-teacher ratio is 22.
- The crime rate and tax rate have the widest range, from 0.006 to 88.97 for the former, and 187 to 711 for the latter.

Question (e)

How many of the census tracts in this data set bound the Charles river?

```
> table(Boston$chas)

  0   1
471  35
```

The number of census tracts that bounds Charles river is : 35

Question (f)

What is the median pupil-teacher ratio among the towns in this data set?

```
> cat("The median of the pupil teacher ratio among towns is: ", median(Boston$ptratio))
The median of the pupil teacher ratio among towns is:  19.05
```

Question (g)

Which census tract of Boston has the lowest median value of owner occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
> min(Boston$medv)
[1] 5
>
> Boston [Boston$medv == 5, ]
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
>
```

- The above two observations have the lowest medv values.
- The crime rate differs between the two observations, even though a few parameters have similar values. The lstat and black values differ, which may also explain the variation in crime rates.
- They are not located near the Charles River.
- This census tract shows a higher crime rate when the median value of owner-occupied houses is at its minimum. The rad value is 24, the maximum in the range, indicating high accessibility to highways.

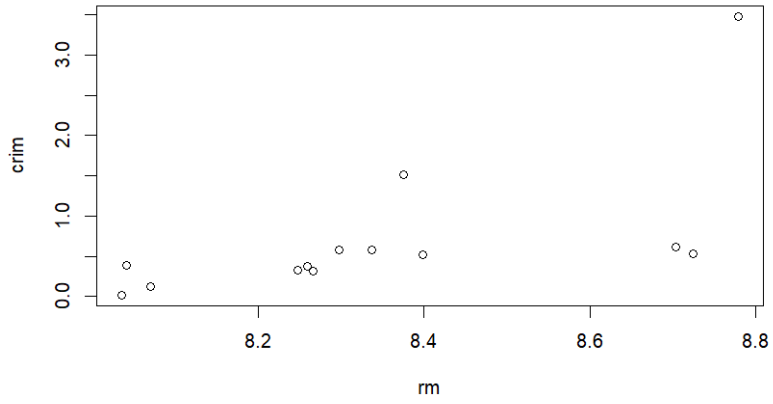
- The tax rates are also close to the maximum value.
- The ptratio is high and close to its maximum value, indicating a crowded school system.

Question (h)

In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

The number of census tracts which averages more than seven rooms per dwelling: 64

The number of census tracts which average more than eight rooms per dwelling: 13



- As the average number of rooms per dwelling increases, the crime rate stabilizes at a lower level, suggesting that these are wealthier neighborhoods.
- However, there is one outlier with a crime rate of 3.4, indicating that even in wealthier areas, higher crime can still occur.

Linear Models

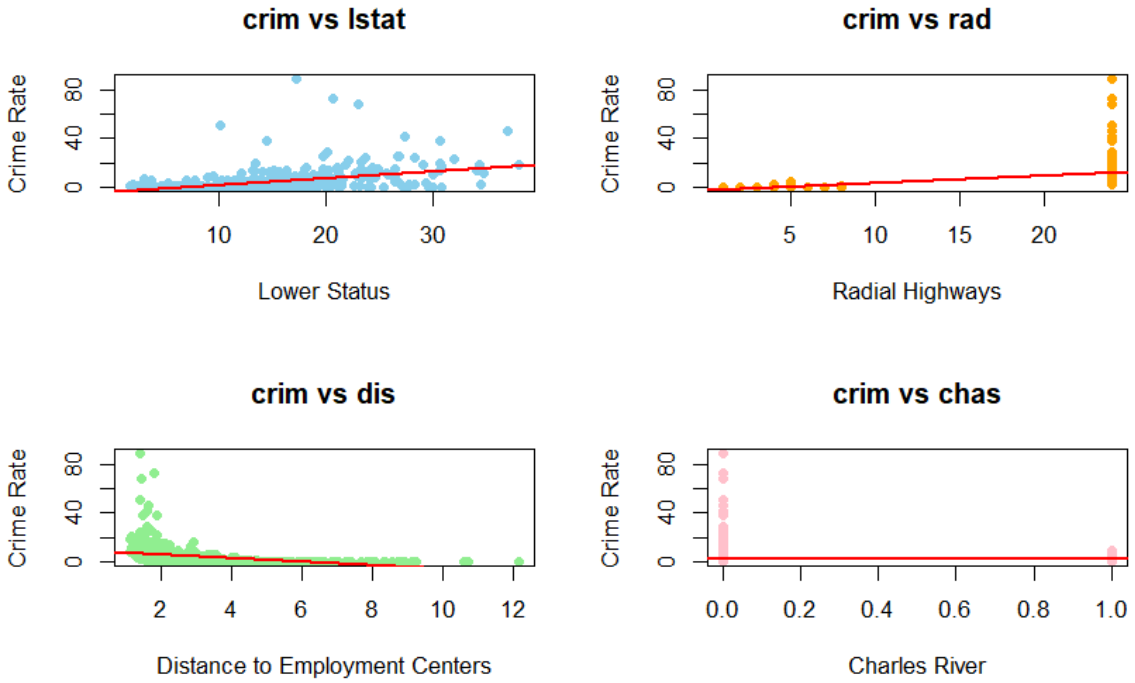
Question 15

This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

Question (a)

For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

- When running all the predictors separately in linear regression, all the predictors are statistically significant in predicting the average crime rate except for the chas predictor. There is a strong association between the 12 predictors and the crime rate since the p-value is less than 0.05.
- Predictors such as tax, lstat, age, ptratio, rad, indus, nox had a positive correlation with the crime rate (i.e) as the predictor values increases the crime rate also increased.
- On the other hand, predictors such as rm, zn, medv, dis, black had a negative correlation with the crime rate parameter (i.e) as the predictor values increases the crime rate decreased.



The above plots show a positive correlation between lstat and crime, similarly crim and rad. Where else, it shows a negative correlation between dis and crim. Lastly, there is no relationship between crim and chas.

Question (b)

Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

Below is the regression table, where using all the predictors the multiple regression model was fit.

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

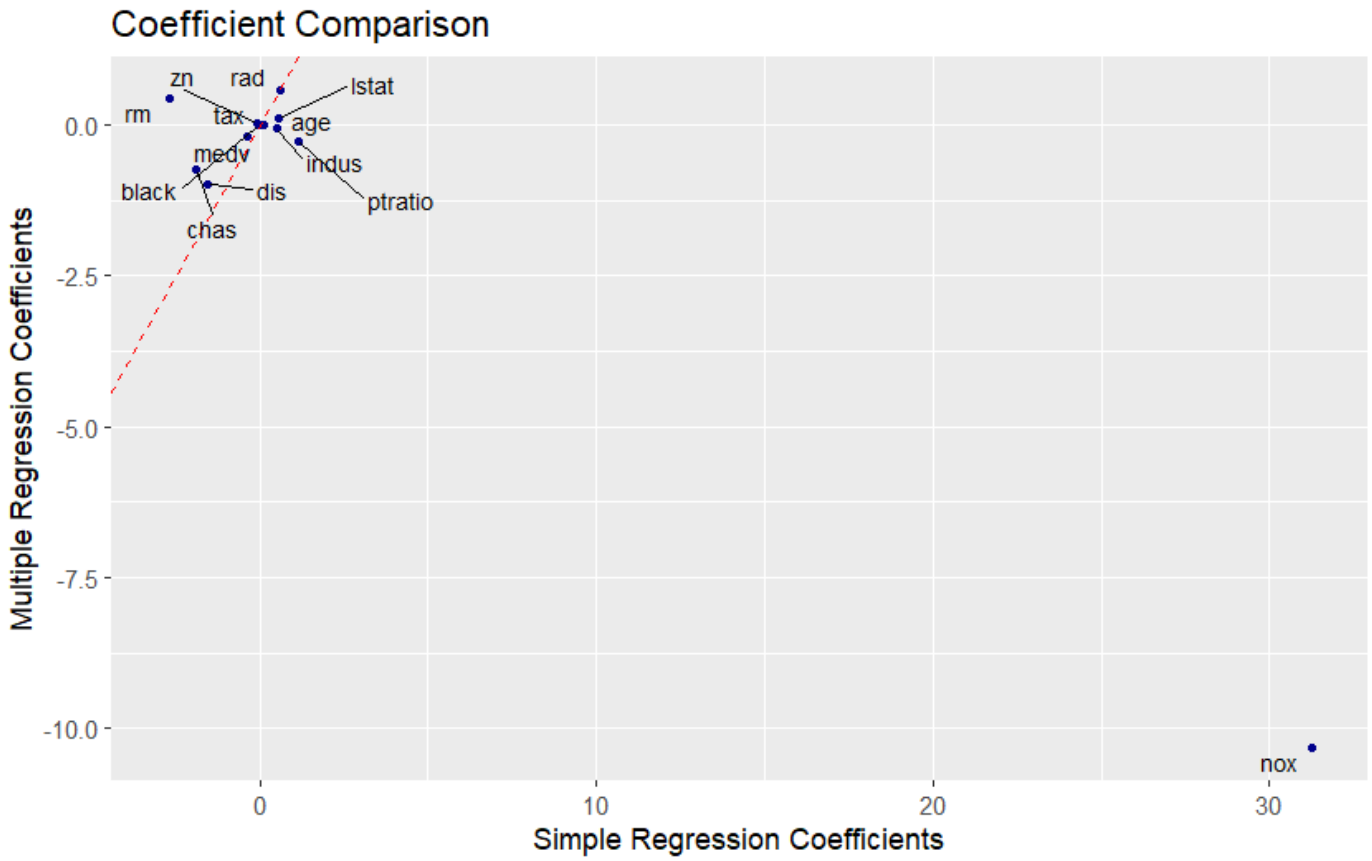
From the table, we can infer that the predictors below are statistically significant and have enough evidence to reject the null hypothesis H0 : βj = 0.

- zn
- dis
- rad
- black
- medv

The rest of the predictors do not have sufficient evidence to reject the null hypothesis.

Question (c)

How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.



- From the graph, we can observe that several predictors exhibit significant shifts or even sign reversals in their coefficients when moving from simple linear regression to multiple linear regression. This suggests the presence of multicollinearity, where predictors are correlated with one another.
- For instance, the variable nox has a simple linear regression coefficient of 31.24, but in the multiple regression model, it flips to –10.31, indicating a reversal from a positive to a negative association when controlling for other variables.
- Similarly, rm also changes direction, shifting from a negative to a positive relationship with the response variable.
- On the other hand, some predictors such as rad and dis display relatively stable coefficients across both models, suggesting their relationship with the response variable is more independent and robust, with less influence from other predictors.

Question (d)

Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form
Y = β0 + β1X + β2X2 + β3X3 + ϵ.

After fitting the model with the above expression, we can arrive with the result below.

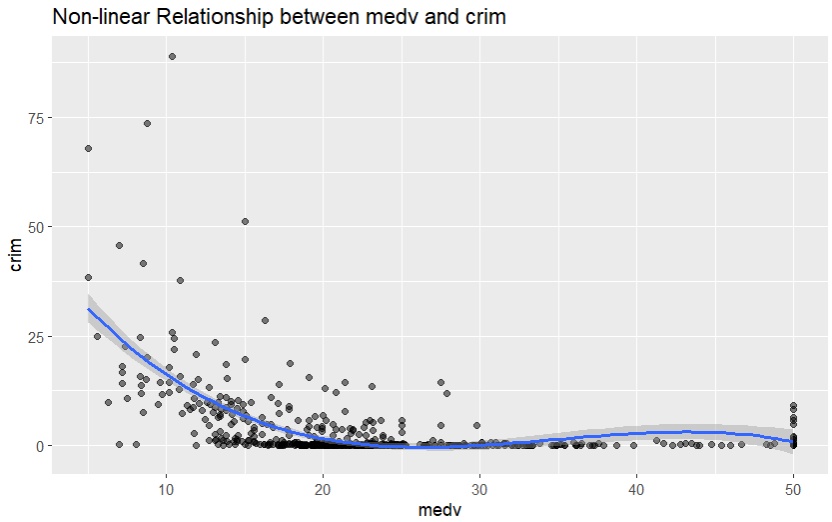
The parameters below exhibit a non-linear association between the predictor and the response.

- medv
- nox
- dis
- indus
- age
- ptratio

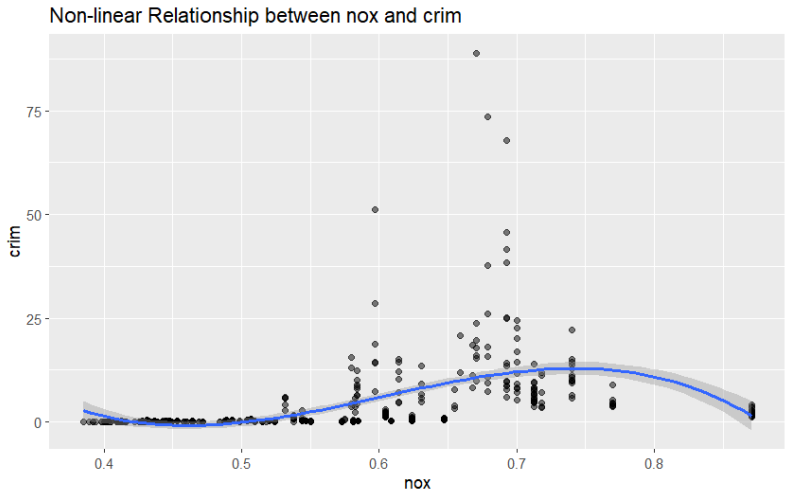
The other parameters do not exhibit non-linear relationship with the response parameter.

- For the medv parameter, we can interpret from the regression output that at low values of median home price, a one-unit increase in medv is associated with a decrease of approximately 5.09 units in the predicted crime rate.
- However, at moderate values, there is a positive correlation between medv and crime rate.
- At higher values of medv, the relationship once again becomes negative, indicating that crime decreases as medv increases.
- This alternating pattern of negative, positive, and negative associations across different ranges of medv reflects a non-linear (S-shaped) relationship, which is also evident in the graph.

```
Call:
lm(formula = crim ~ medv + I(medv^2) + I(medv^3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-24.427  -1.976  -0.437   0.439  73.655

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.1655381  3.3563105  15.840  < 2e-16 ***
medv        -5.0948305  0.4338321 -11.744  < 2e-16 ***
I(medv^2)    0.1554965  0.0171904   9.046  < 2e-16 ***
I(medv^3)   -0.0014901  0.0002038  -7.312 1.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.569 on 502 degrees of freedom
Multiple R-squared:  0.4202,    Adjusted R-squared:  0.4167
F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```



Below is another example of non-linear relationship with the crime rate.



Selection and Shrinkage Methods

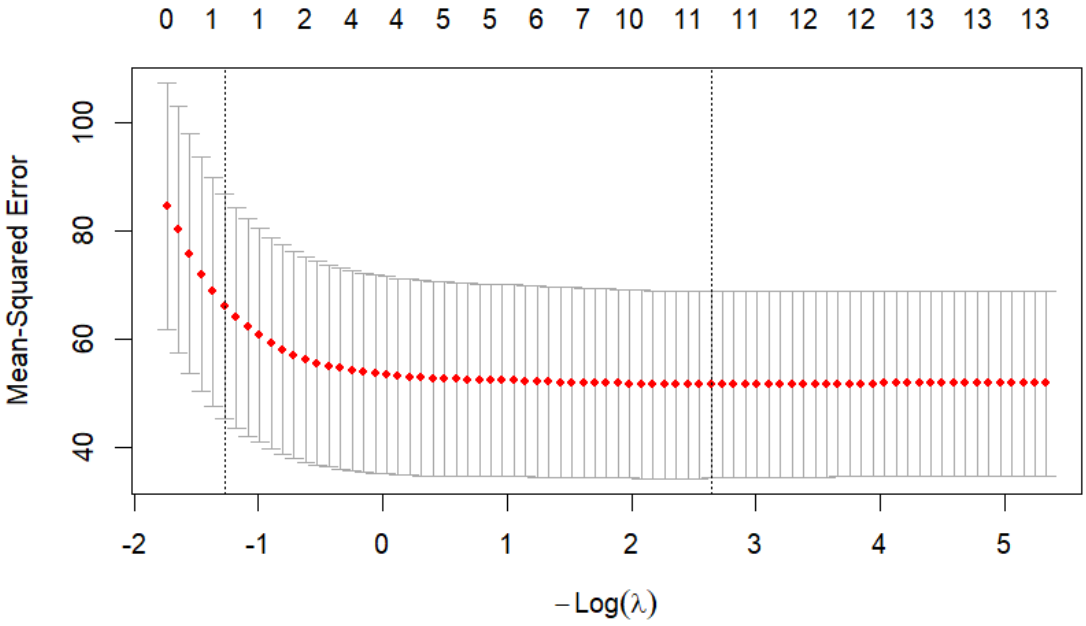
Question 11

Question(a)

We will now try to predict per capita crime rate in the Boston data set.

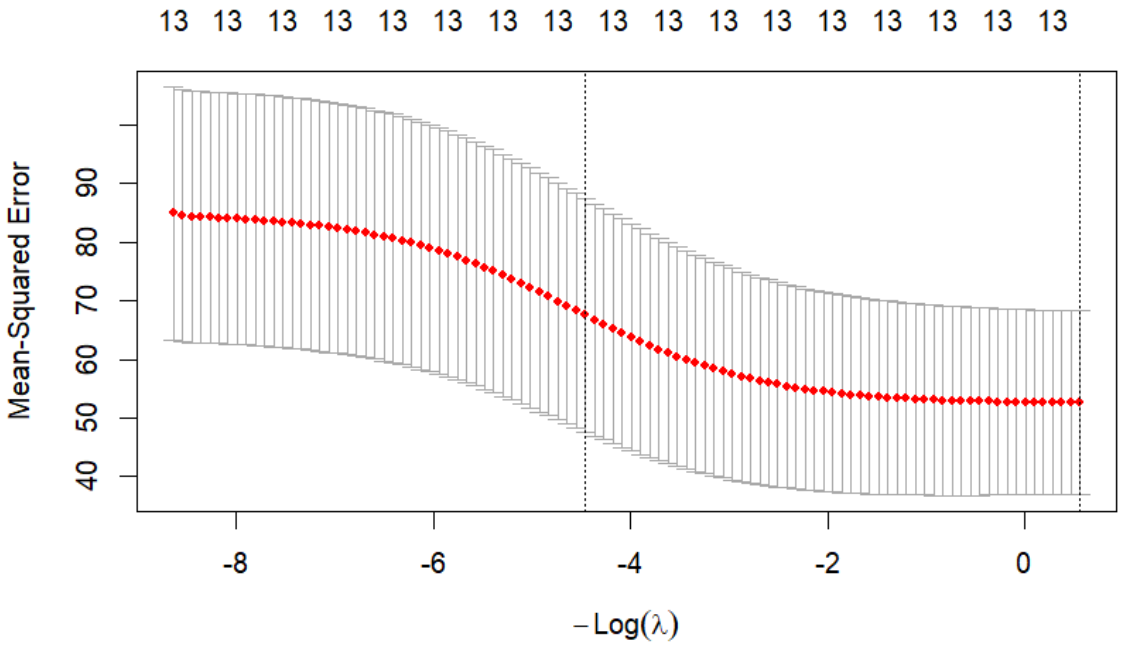
(a) Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

Lasso Regression



- When fitting Lasso regression to the Boston housing dataset using 10-fold cross-validation. From the above plot of cross validation mean-squared error against the -log(lambda) we can identify that the minimum cross-validated MSE occurred at λ = 0.072, where 11 predictors were retained in the model.
- Based on the model with λ = 0.072, a test MSE of 13.26 was achieved.

Ridge Regression



- When applying ridge regression using 10-fold cross-validation, we can identify that the minimum cross-validated MSE was present at lambda = 0.57, where all the 13 predictors were retained in the model.
- The Ridge model derived a test MSE of 13.02 at lambda = 0.57.

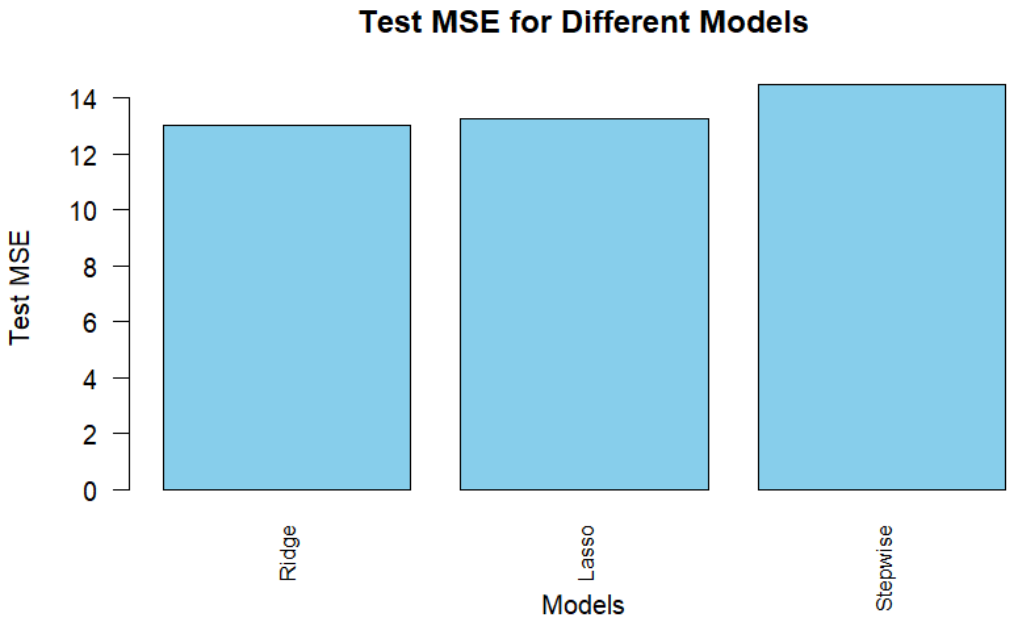
Stepwise Selection

```
> summary(boston_step)

Call:
lm(formula = crim ~ zn + nox + dis + rad + ptratio + black +
    medv, data = boston_train)

Residuals:
    Min      1Q  Median      3Q     Max
-10.869  -2.014  -0.415   1.010  74.420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.829699   7.067037   3.513 0.000493 ***
zn            0.050575   0.021732   2.327 0.020460 *
nox         -13.632761   5.869153  -2.323 0.020697 *
dis          -1.118313   0.312235  -3.582 0.000384 ***
rad           0.573786   0.060743   9.446  < 2e-16 ***
ptratio      -0.361578   0.224332  -1.612 0.107802
black        -0.008825   0.004295  -2.055 0.040557 *
medv         -0.240996   0.049768  -4.842 1.84e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.97 on 396 degrees of freedom
Multiple R-squared:  0.4375,    Adjusted R-squared:  0.4275
F-statistic:    44 on 7 and 396 DF,  p-value: < 2.2e-16
```

- From the result, the stepwise selection model has selected 7 predictors out of 13 predictors. We can interpret that the distance between the employment center, radial highway and Median value of owner-occupied homes is highly statistically significant in predicting the crime rate.
- The test MSE for the stepwise model is 14.47.

Question (b)

Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, crossvalidation, or some other reasonable alternative, as opposed to using training error.



By predicting the model using the test data, the below MSE and RMSE values were derived.

Model	Test MSE	RMSE
Ridge Regression	13.02	3.60
Lasso Regression	13.26	3.64
Stepwise Selection	14.47	3.80

Thus, we can conclude that Ridge regression has the lowest MSE and it makes accurate predictions of the crime rate when compared to the other two models on the test data.

Logistic Regression

Question 6

Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, ˆ β0 = −6, ˆ β1 = 0.05, ˆ β2 = 1.

Question (a)

Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

```
betacoeff0 <- -6
betacoeff1 <- 0.05
betacoeff2 <- 1

# Part (a) values
X1 <- 40    # hours studied
X2 <- 3.5   # GPA

# Calculate logit
logit <- betacoeff0 + betacoeff1 * X1 + betacoeff2 * X2

# Calculate probability
probability <- 1 / (1 + exp(-logit))
print(probability)
```

0.37 is the probability of the boy getting an A in the class.

Question (b)

How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

#p = 0.5
Solving the equation, 0.5 = 1/1+(e^-z)
We get the z value as 0. Since Z = 0, the provided values are substituted in the below formula.

```
X1_value <- -(betacoeff0 + betacoeff2 * X2) / betacoeff1
```

If the boy studies for 50 hours there is a 50% chance of getting an A.

Question 9

This problem has to do with odds.

Question (a)

On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

```
odds <- 0.37
probability_of_default <- odds/(1+odds)
```

0.27 is the probability of customers defaulting on their credit card payment.

Question (b)

Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?

```
probability = 0.16
odds_of_defaulting <- probability/(1-probability)
```

#0.1904 is the odds of defaulting on her credit card payment.

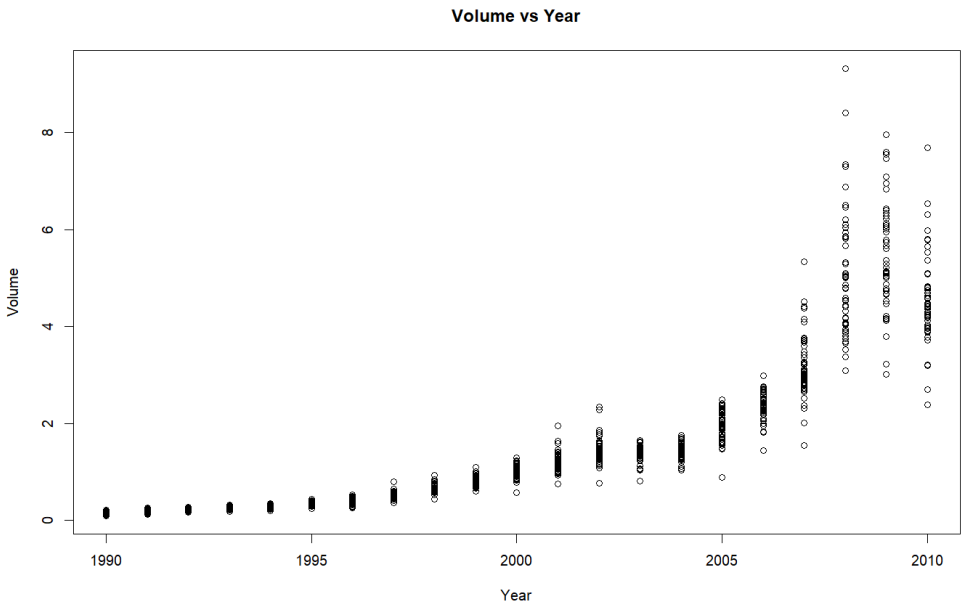
Question 13

This question should be answered using the Weekly data set, which is part of the ISLR2 package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

Question (a)

Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

- The dataset consists of percentage returns for each of the five previous days for 1089 days from the year 1990 to 2010. From the direction column we can see that the downward trend and upward trend on average was approximately equal with downward trend topping in few margins.



- As seen in the graph, we can interpret that as the Year increases the Volume (the number of shares traded on the previous day, in billions) also increases indicating a positive correlation of 0.84. The volume of shares increased steadily from 1990 to 2010.

Question (b)

Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1496.2  on 1088  degrees of freedom
Residual deviance: 1486.4  on 1082  degrees of freedom
AIC: 1500.4

Number of Fisher Scoring iterations: 4
```

- From the regression table the Lag 2 parameter is statistically significant to predict the Direction of the stocks, since it has a p-value less than 0.05.
- The Lag2 coefficient indicates that an increase in Lag2 is associated with higher odds of a positive stock market direction. The other predictors are not statistically significant since their p-values are greater than 0.05.

Question (c)

Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

From the confusion matrix, the model correctly predicted that the stock market would increase by 557 days, and it would decrease to 54 days.

```
          Actual
Predicted Down  Up
     Down   54  48
     Up    430 557
```

The overall fraction of correct predictions is as below:

(557+54)/1089 = 0.561

The logistic regression model has correctly predicted the direction of the stock market 56.1 % of the time, on the other hand it has incorrectly predicted 43.8 % of the time.

Question (d)

Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
          Actual
Predicted Down Up
     Down    9  5
     Up     34 56
```

Overall fraction of correct predictions of the testing data is as below:
(9+56)/104 = 62.5%

Overall fraction of incorrect predictions
(34+5)/104 = 37.5%

The logistic regression model has correctly predicted the direction of the stock market 62.5%

of the time, on the other hand it has incorrectly predicted 37.5% of the time.

- The model with only Lag2 as the predictor gives more accuracy when compared to the multiple logistic regression model. Also, the prediction error percentage has reduced by 6.3% and the prediction success rate has increased by 6.4% when compared to the previous multiple logistic regression model.

Regression Trees

Question 8

Use the Austin housing data from the prediction contest (austinhousing.csv) instead of the dataset in the book.
Use the following variables to generate predictions for log(latestPrice): latitude, longitude, hasAssociation, livingAreaSqFt, numOfBathrooms, numOfBedrooms. (See the description of the dataset in the individual prediciton project assignment.) When reporting your prediction errors, report them in terms of prices (not log prices).

Question (a)

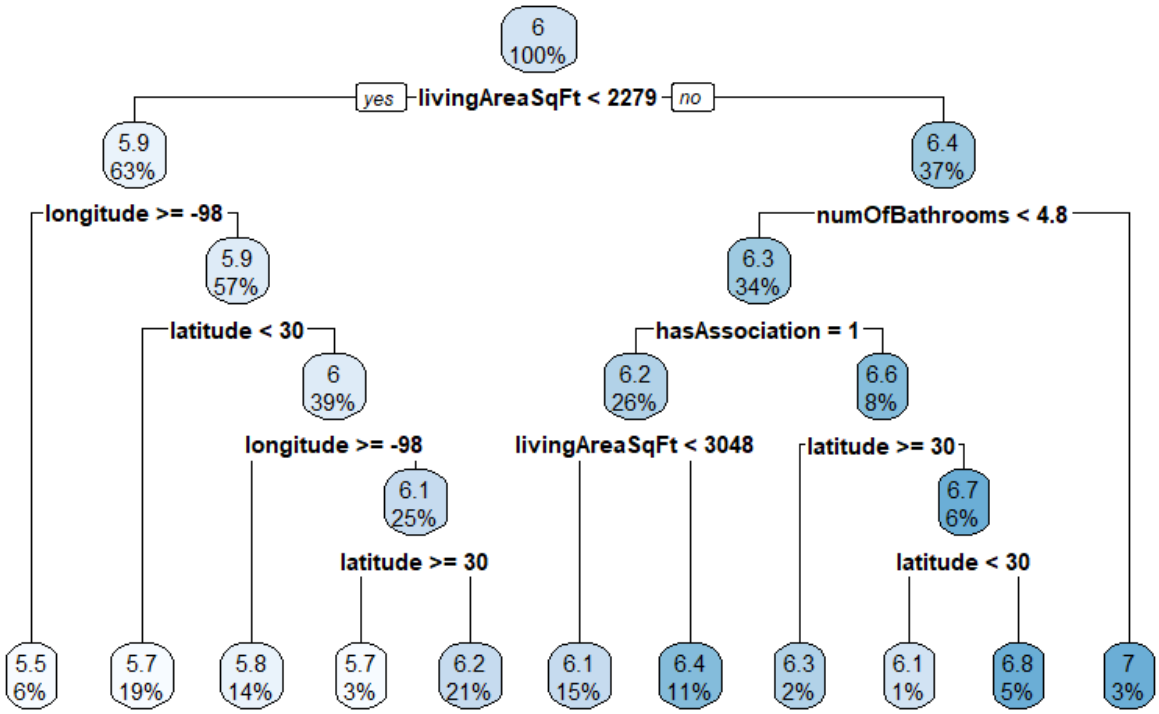
Split the data set into a training set and a test set.

```
> train_index <- sample(1:n, size = 0.8 * n)
>
> austin_train = austin_house[train_index, ]
> austin_test = austin_house[-train_index, ]
>
> dim(austin_train)
[1] 5427    7
> dim(austin_test)
[1] 1357    7
```

The Austin data set is split into training dataset (80%) and testing dataset (20%).

Question (b)

Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

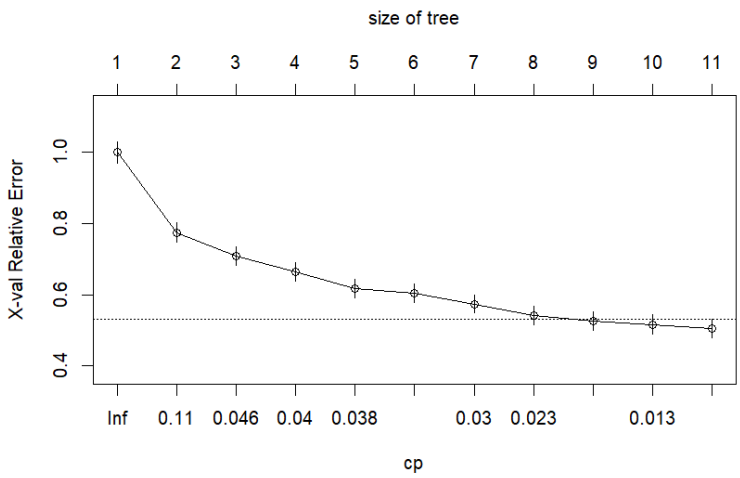



```
Variable importance
livingAreaSqFt numOfBathrooms      longitude  numOfBedrooms       latitude hasAssociation
            31              19             18             14             10              7
```

- We can interpret that the livingAreaSqFt parameter is the most important predictor for LatestPrice with 31%. Followed by other important predictors such as num of bathrooms and longitude at 19% and 18% respectively.
- The root node is based on the living area square feet where it is split into smaller homes (< 2279) and bigger homes. After that, based on the location and the number of bathrooms, the tree is split into further nodes.
- The test MSE for the regression tree is 69292.13

Question (c)

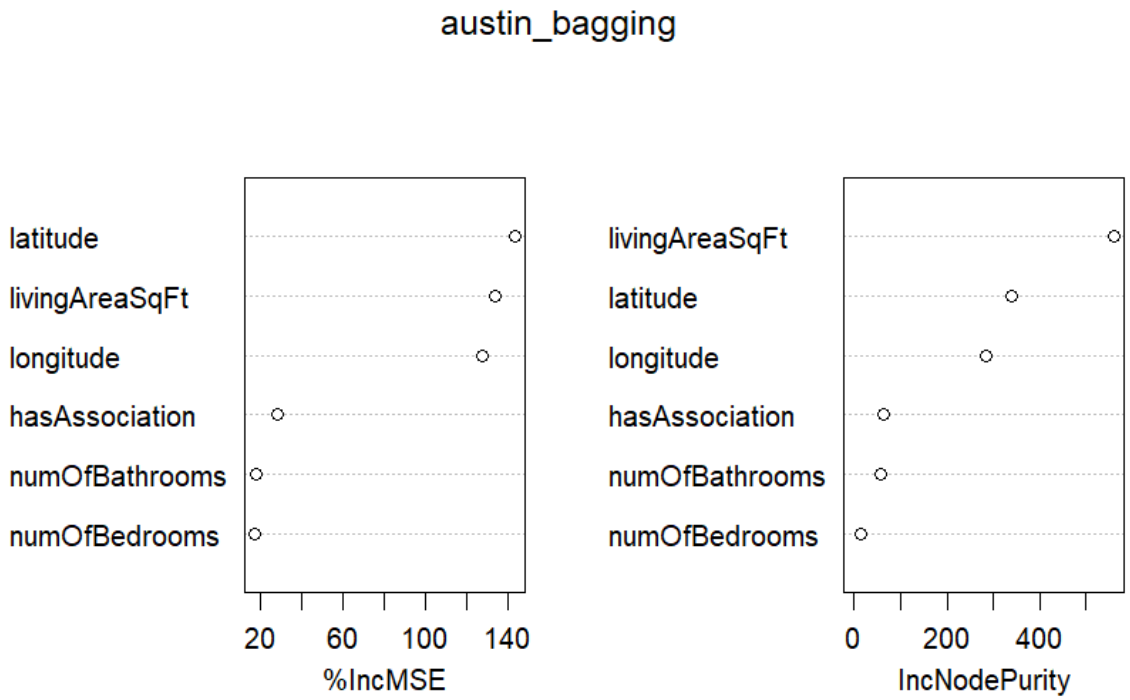
Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?



- Using cross validation, the least complexity parameter corresponded to a tree with 11 splits. However, pruning the tree yielded the same test RMSE as the Regression tree, since the Regression tree also constructed a tree with 11 splits.

Question (d)

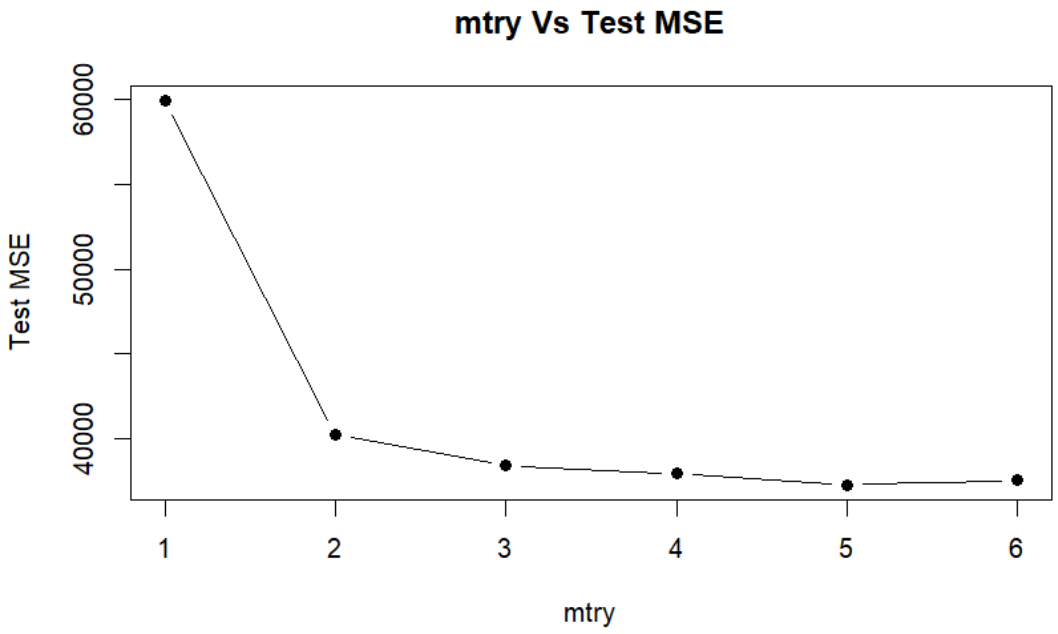
Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.



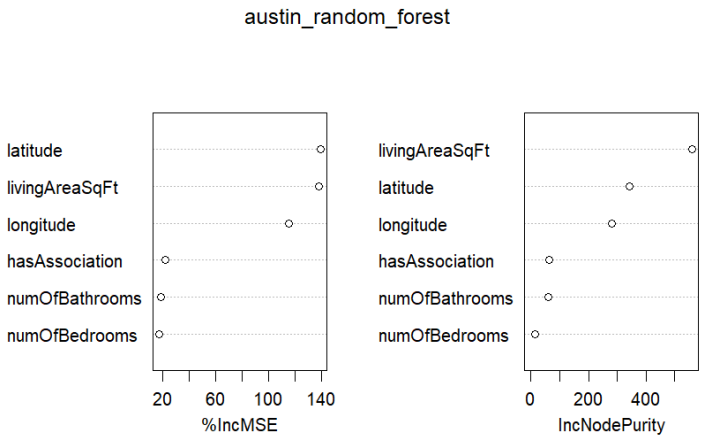
- From the importance function on the bagging model, we can infer that the most important variable is livingAreaSqFt, indicating that the size of the house plays a major role in predicting the latest Price parameter.
- The other important parameters are latitude and longitude, which infer that the location of the house is also an influential factor in deciding housing prices.
- Variables like the number of bedrooms and bathrooms showed lower importance.
- The Bagging model yielded a Test MSE of 37015.13.

Question (e)

Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.



- The importance() function indicates that the three most important predictors are livingAreaSqFt, latitude and longitude for predicting the latest house price.

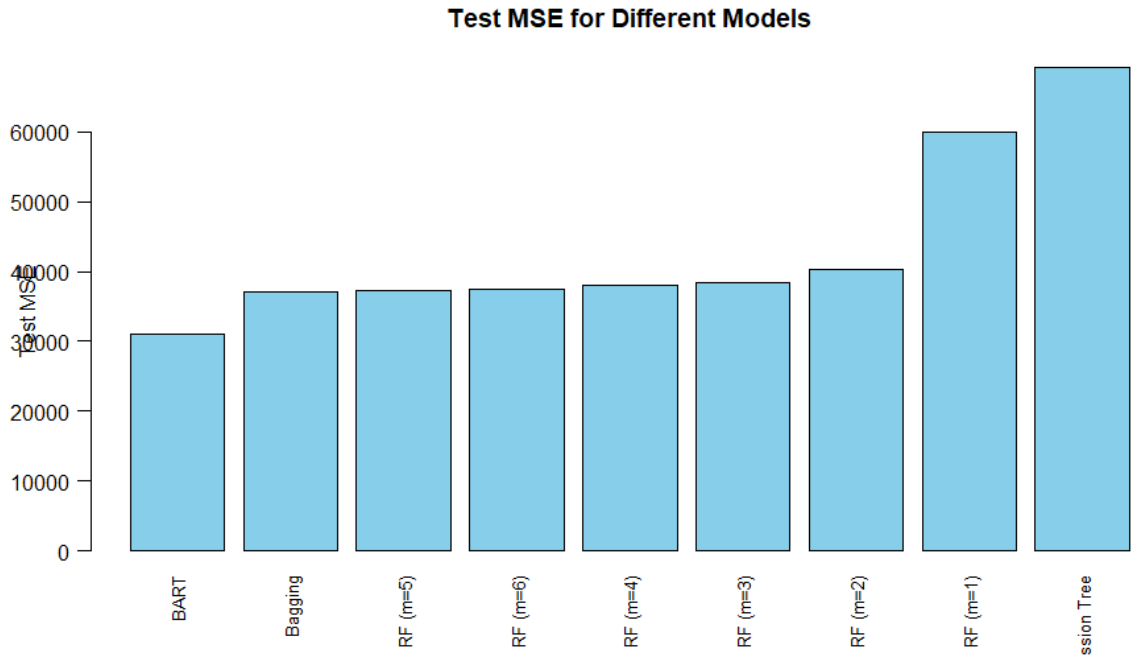


- When fitting the random forest model with different mtry values, which control the number of variables considered at each split, we observed that the test MSE value was the lowest (37,295.74) when mtry was 5.
- It decreased steadily from 59,900.99 to 37,295.74 for mtry 1 to 5. However, the test MSE slightly increased to 37,542.56 when mtry was 6.

Question (f)

Now analyze the data using BART, and report your results.

After fitting the BART model with the training and testing data we get the test MSE as below.
[1] "BART Test MSE: 31072.48"



By predicting the model using the test data, the below MSE and RMSE values were derived.

Model	Test MSE	RMSE
Regression Tree	69292.13	263.234
Bagging	37015.13	192.39
Random Forest (m = 5)	37295.74	193.12
BART	31072.48	176.27

- As seen in the above bar plot, the BART model has the least test MSE of 31, 072.48 and the Regression Tree has the highest test MSE of 69, 292. Thus, we can conclude that the BART model captures more accurate predictions of the latest housing prices.

Classification Trees

Question 11

This question uses the Caravan data set.

Question (a)

Create a training set consisting of the first 1,000 observations, and a test set consisting of the remainin g observations.

```
> training_observations <- 1:1000
> caravan_train <- Caravan[training_observations, ]
> caravan_test <- Caravan[- training_observations, ]
> dim(caravan_train)
[1] 1000   87
> dim(caravan_test)
[1] 4822   87
```

- As mentioned above in the code, the first 1000 observations of the Caravan dataset are split as the training dataset and the remaining observations are stored as the testing dataset.

Question (b)

Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most imp ortant?

```
> summary(boost_caravan)
                var     rel.inf
PPERSAUT   PPERSAUT 15.15534009
MKOOPKLA   MKOOPKLA  9.23499526
MOPLHOOG   MOPLHOOG  8.67017024
MBERMIDD   MBERMIDD  5.39403655
MGODGE       MGODGE  5.03047673
PBRAND       PBRAND  4.83740038
MINK3045   MINK3045  3.94305387
ABRAND       ABRAND  3.69692919
MOSTYPE     MOSTYPE  3.38768960
PWAPART     PWAPART  2.51970169
MGODPR       MGODPR  2.43689096
MSKC           MSKC  2.34594774
MAUT2         MAUT2  2.30973409
MFWEKIND   MFWEKIND  2.27959503
MBERARBG   MBERARBG  2.08245286
MSKA           MSKA  1.90020973
PBYSTAND   PBYSTAND  1.69481877
MGODOV       MGODOV  1.61147668
MAUT1         MAUT1  1.59879109
MBERHOOG   MBERHOOG  1.56791308
MINK7512   MINK7512  1.36255296
MSKB1         MSKB1  1.35071475
MINKGEM     MINKGEM  1.34913011
```

- The most important predictors are **PPERSAUT, MKOOPKLA, MOPLHOOG.**
- The summary output of the boosting model indicates that PPERSAUT is the most important predictor, with the highest relative influence of 15.15. This suggests that it plays a significant role in predicting the response variable.
- Following PPERSAUT, the next most influential predictors are MKOOPKLA (9.23), MOPLHOOG (8. 67) which also contribute meaningfully to the model's predictions.

Question (c)

Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20%. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compar e with the results obtained from applying KNN or logistic regression to this data set?

```
> table(Caravan_test$Purchase, boost_pred)
   boost_pred
       0    1
  0 4396  137
  1  255   34
```

The precision for the above confusion matrix is

> 34/(137 + 34)

[1] 0.1988304

- We can infer that 19.8% of the people the model predicted will purchase, actually did.

Comparing with Logistic Regression:

- When running the training data in the logistic regression we can infer that the few of the below parameters are statistically significant in predicting the Purchase parameter and has an association with it.

  MGEMOMV, MGODRK, MGODOV, MRELOV and PPERSAUT

```
> table(Caravan_test$Purchase, log_pred)
   log_pred
       0    1
  0 4183  350
  1  231   58
```

> 58/(350 + 58)
[1] 0.1421569

- The above is the confusion matrix for logistic regression using the testing data and the precision value is 0.1421, which indicates that 14.2 % of the people have been predicted to purchase and they made the purchase too.
- When comparing the precision value between the boosting model and logistic regression, we can identify that the **boosting model** has a slightly higher margin of 19.8% in identifying the actual predictors.