

Intro to ML : Take Home Assignment

Sanyam Jain (sj33448)

2023-07-28

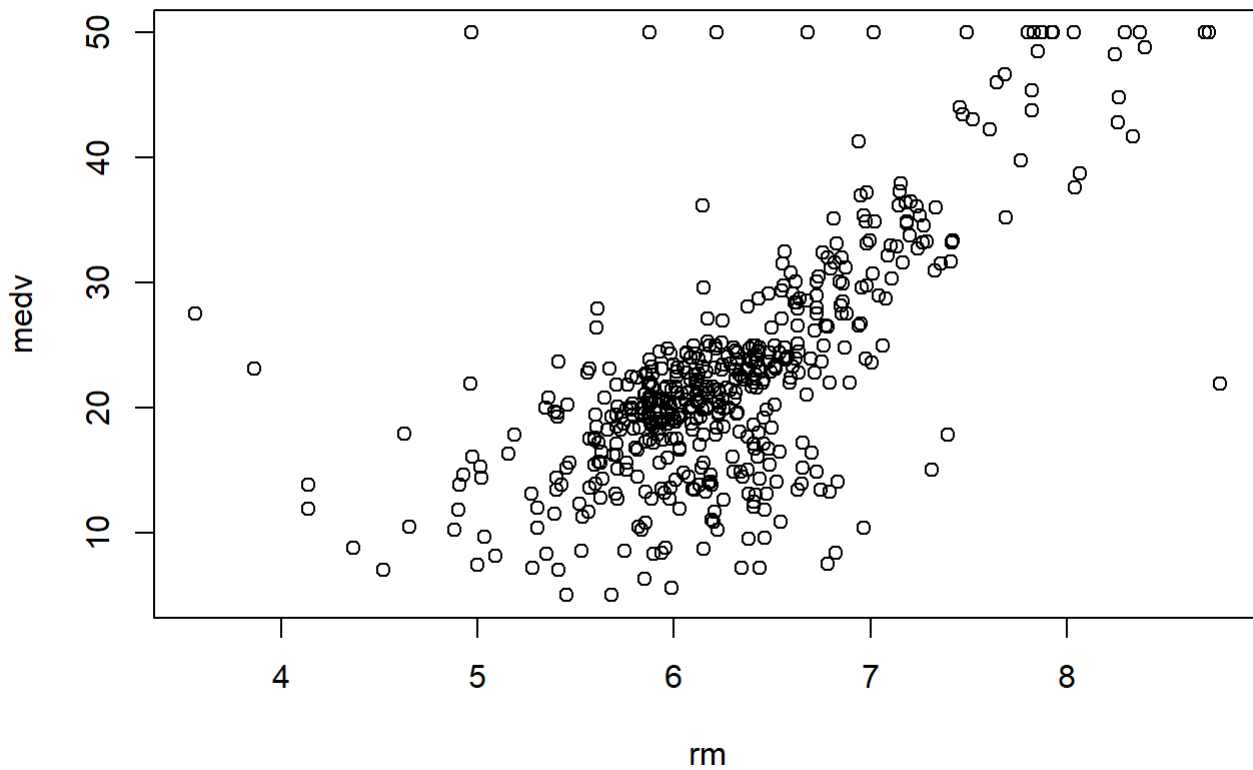
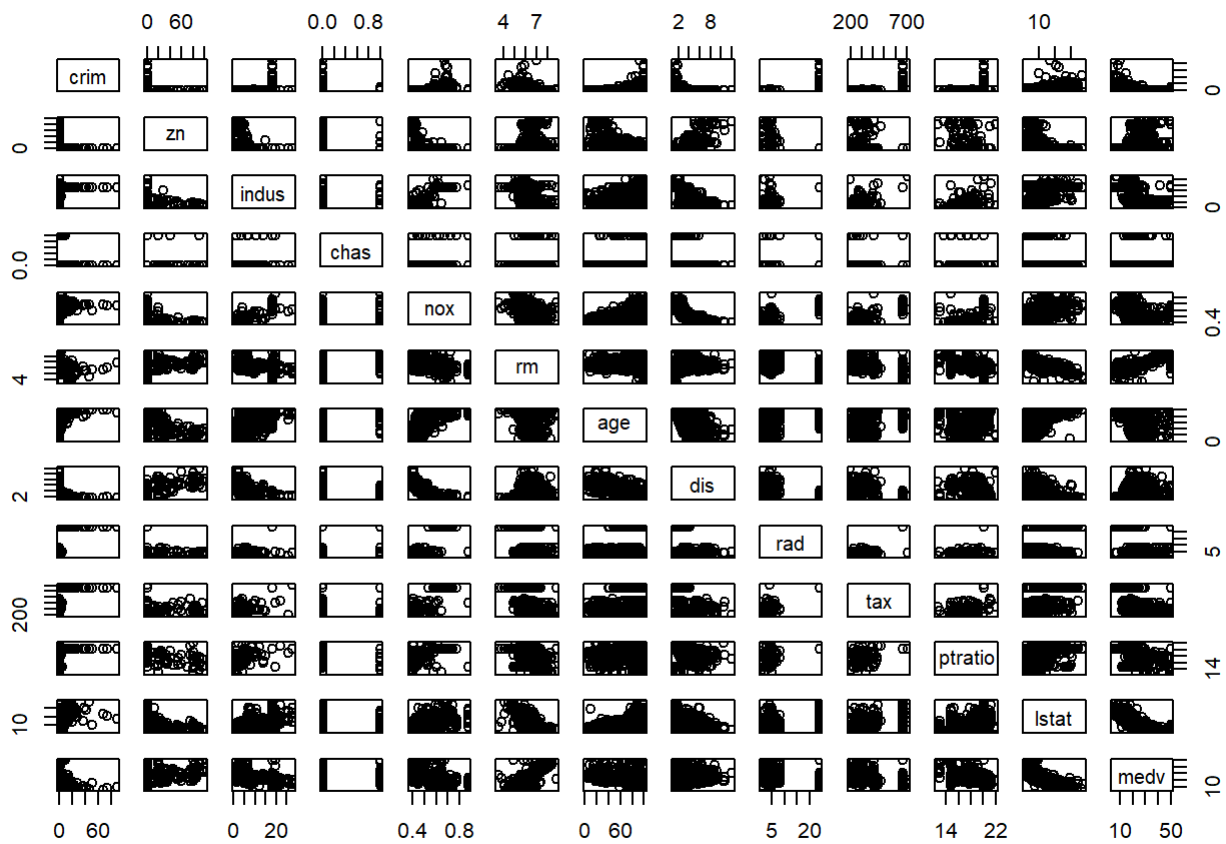
Chapter 2

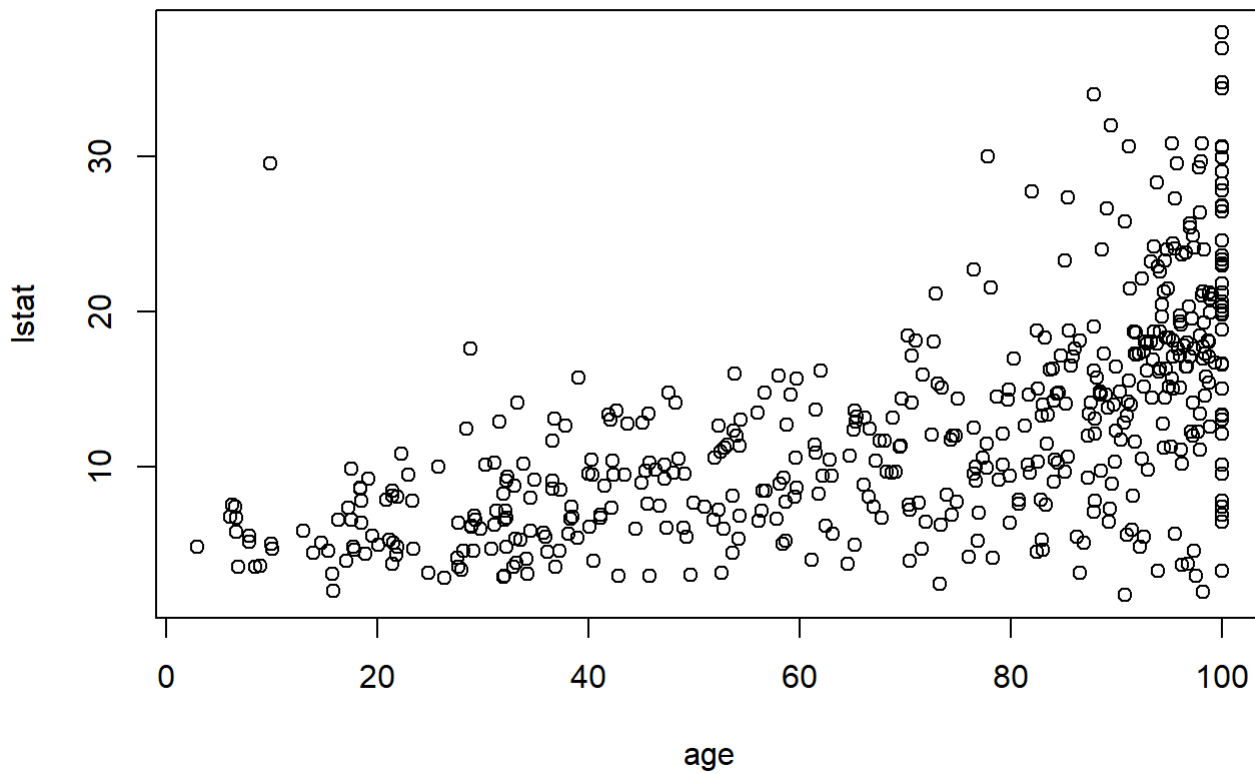
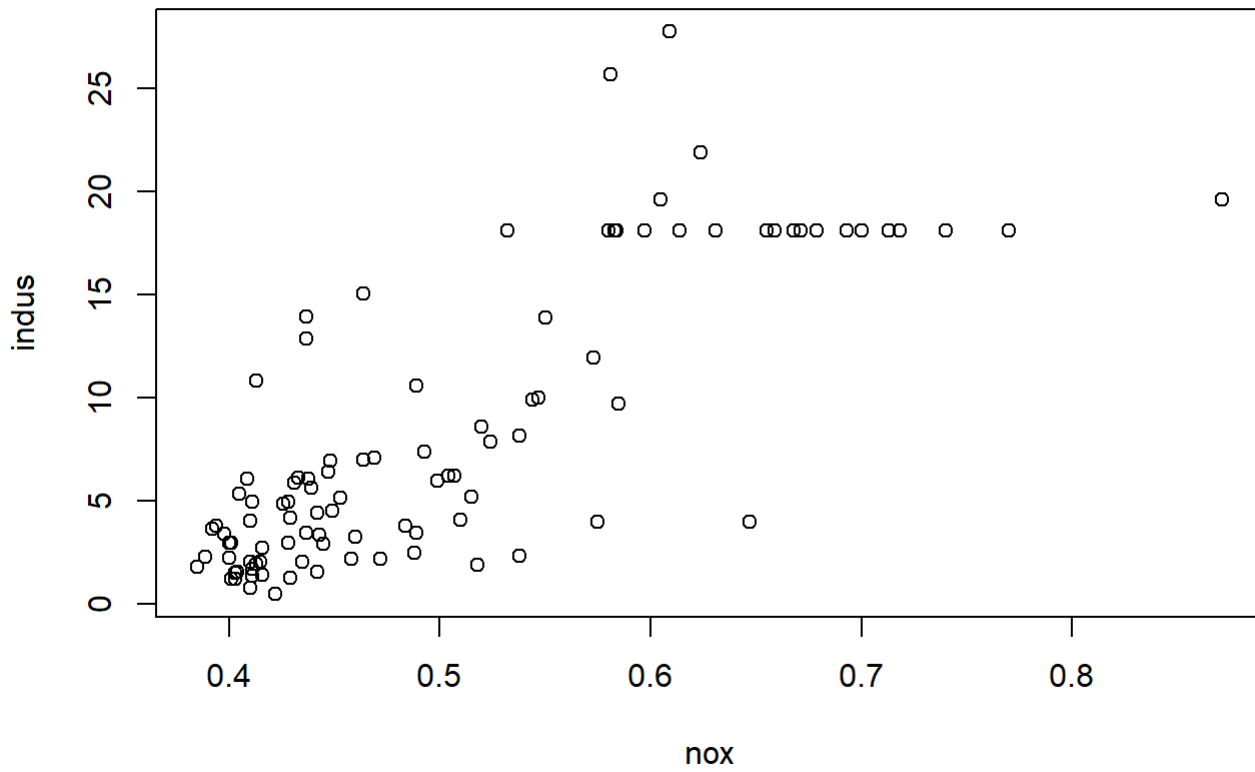
Question 10

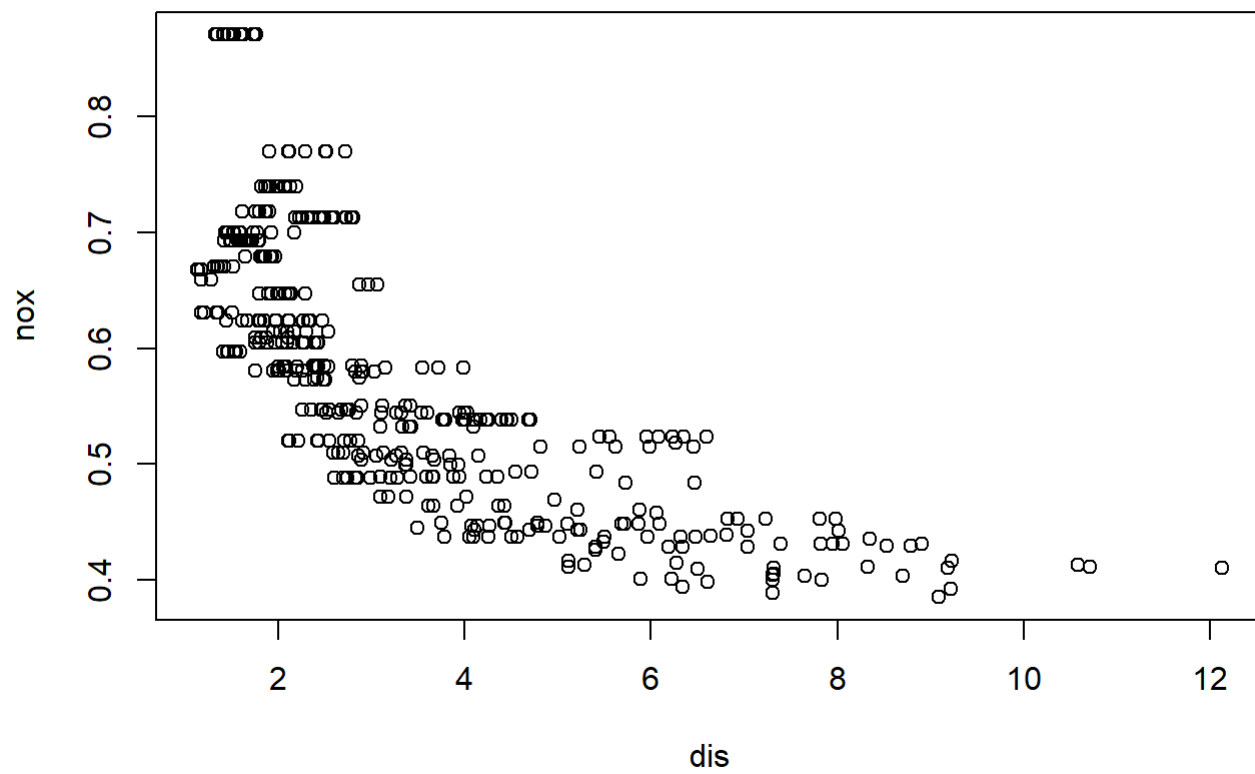
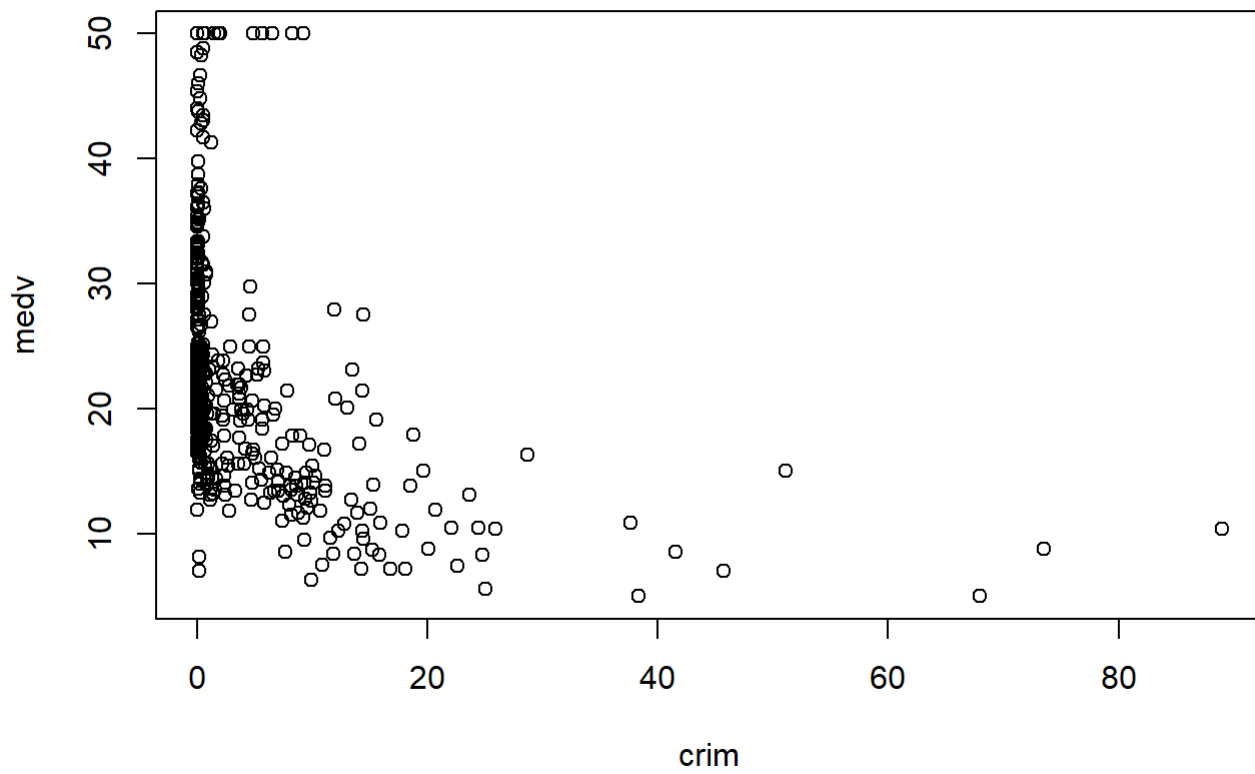
Load the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

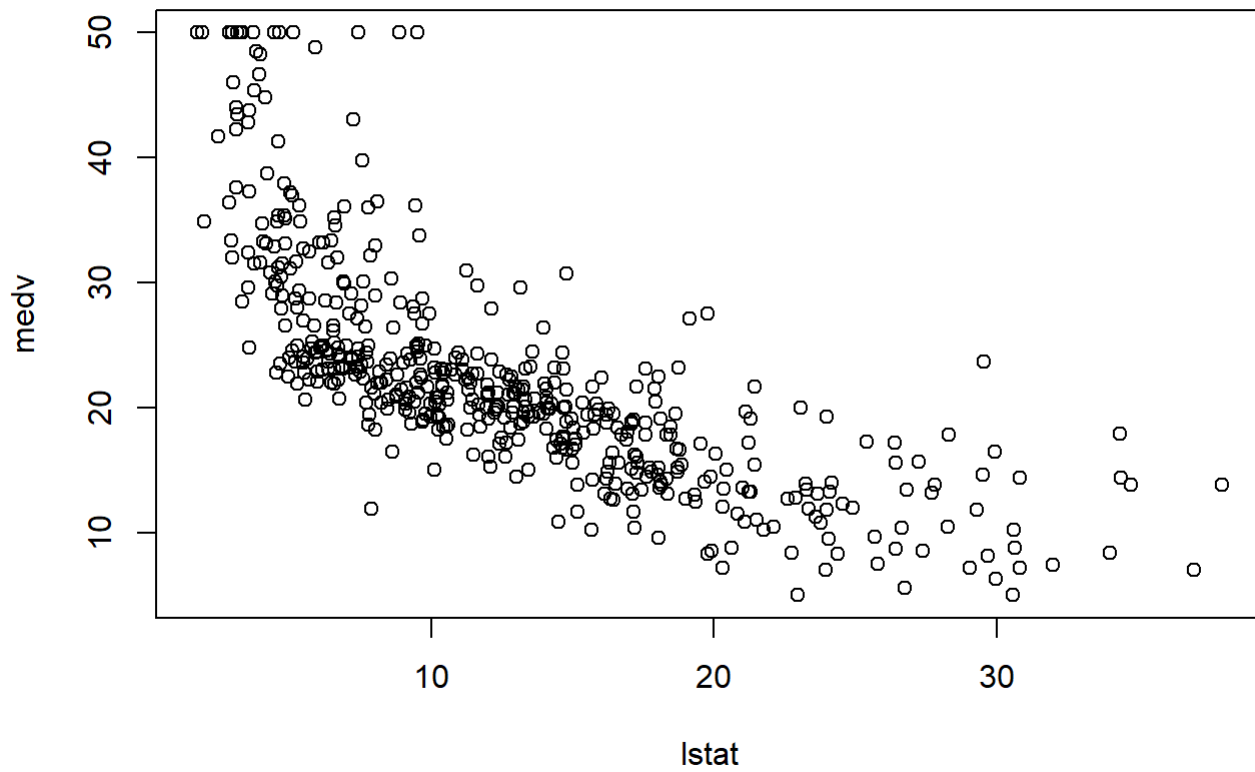
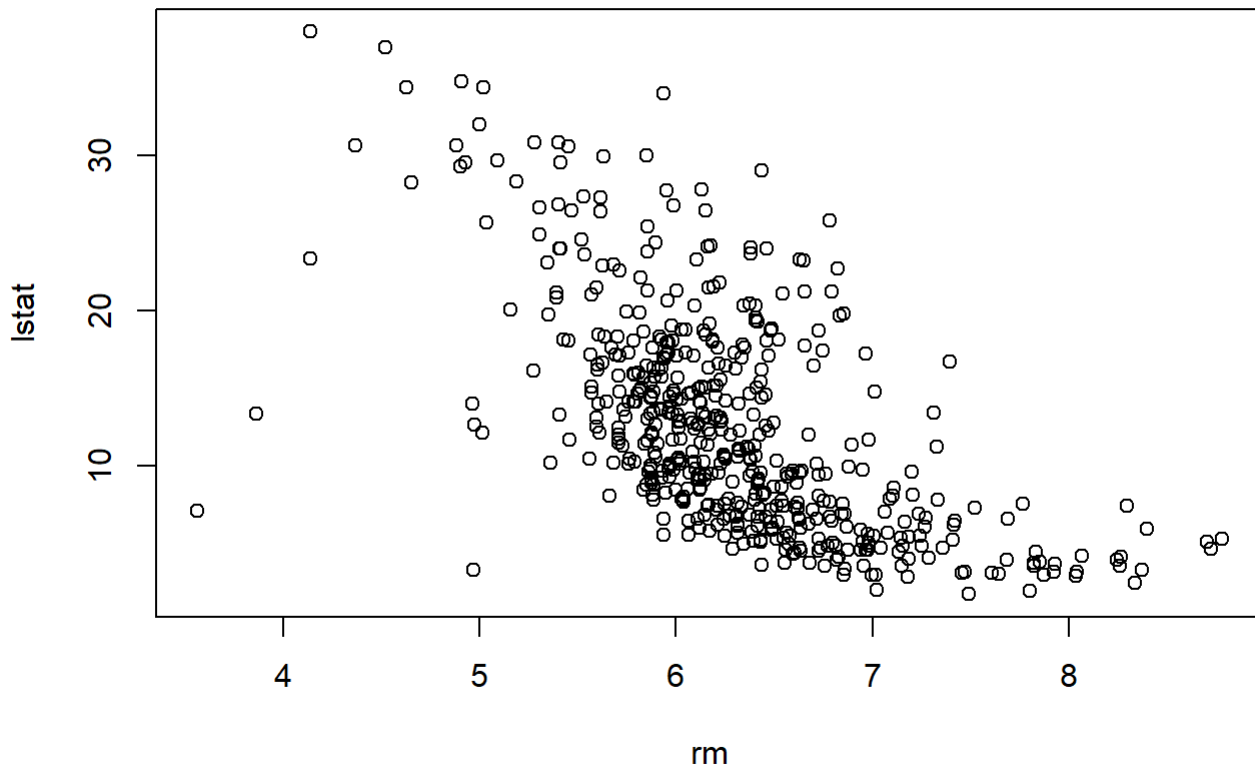
- The Boston dataset present in ISLR2 library has 13 columns and 506 rows.
- Rows represent different suburbs of Boston.
- Columns represent the various variables or features associated with census data.

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









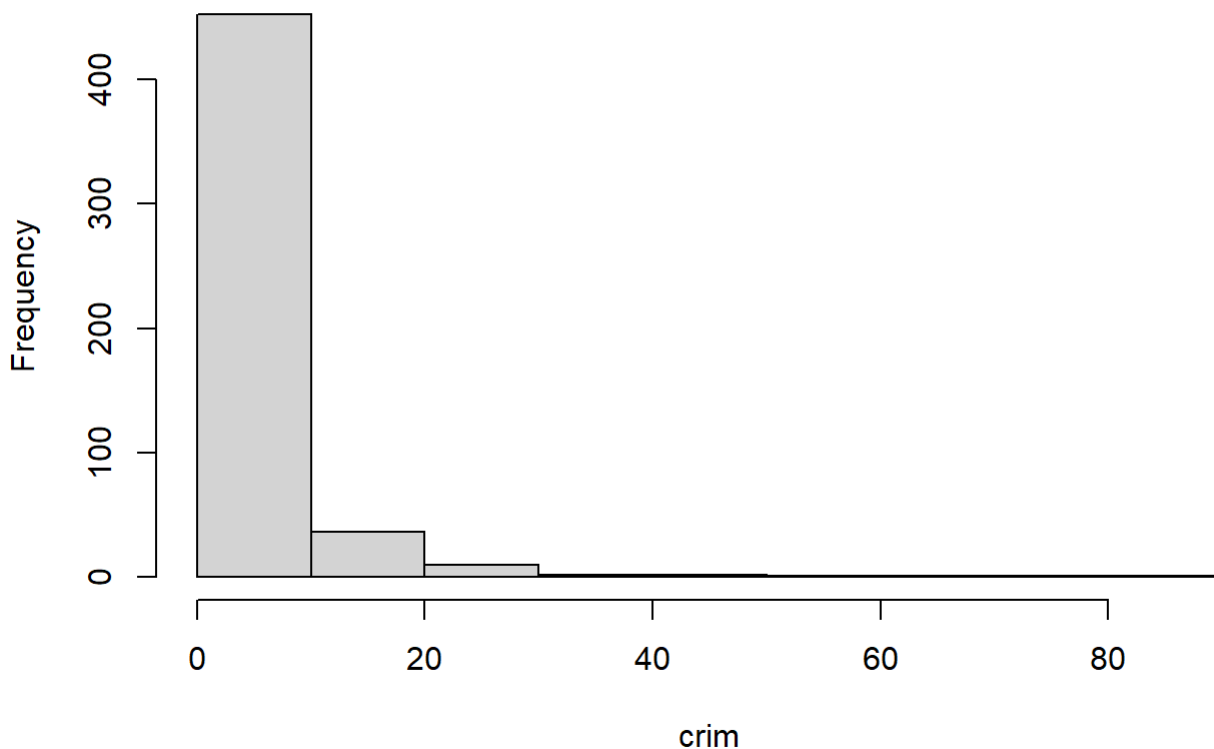
Observations

- Positive relationship pairs:
 - rm and medv: properties with higher average rooms per dwelling have higher median value
 - nox and indus : proportion of non-retail business acres per town have higher nitrogen oxide concentration.
 - age and lstat : suburbs with older units have higher percentage of lower status population.
- Negative relationship pairs:
 - medv and crim : Areas with high crime rate have low median home value
 - nox and dis : Distance to Boston employment center is negatively correlated to Nitrogen Oxide concentration
 - rm and lstat : Properties with higher average rooms per dwelling are less likely to have lower status population.
 - Lstat and Medv : Areas with higher pct of lower status population have lower median value of homes.

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00632 0.08204 0.25651 3.61352 3.67708 88.97620
```

Histogram of crim



```
## Correlation of crime rate with other variables
```

```
##          [,1]
## zn      -0.20046922
## indus    0.40658341
## chas     -0.05589158
## nox      0.42097171
## rm       -0.21924670
## age      0.35273425
## dis      -0.37967009
## rad      0.62550515
## tax      0.58276431
## ptratio  0.28994558
## lstat     0.45562148
## medv     -0.38830461
```

Observations

- CRIM Distribution is highly skewed indicating that only certain suburbs have very high crime rates.
- Positive Correlation : index , nox , age , tax , rad , lstat , ptratio.
- Negative Correlation : zn , chas , rm , dis
- Notable Trends:
 - Higher property tax rates are positively correlated to higher crime rates
 - Proximity to employment centers is observed to have relatively higher crime rates
 - High Crime rates in a suburb correlates to lower median value of propertie
 - Crime rates are higher in suburbs with more percentage of lower status population.

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 predicto

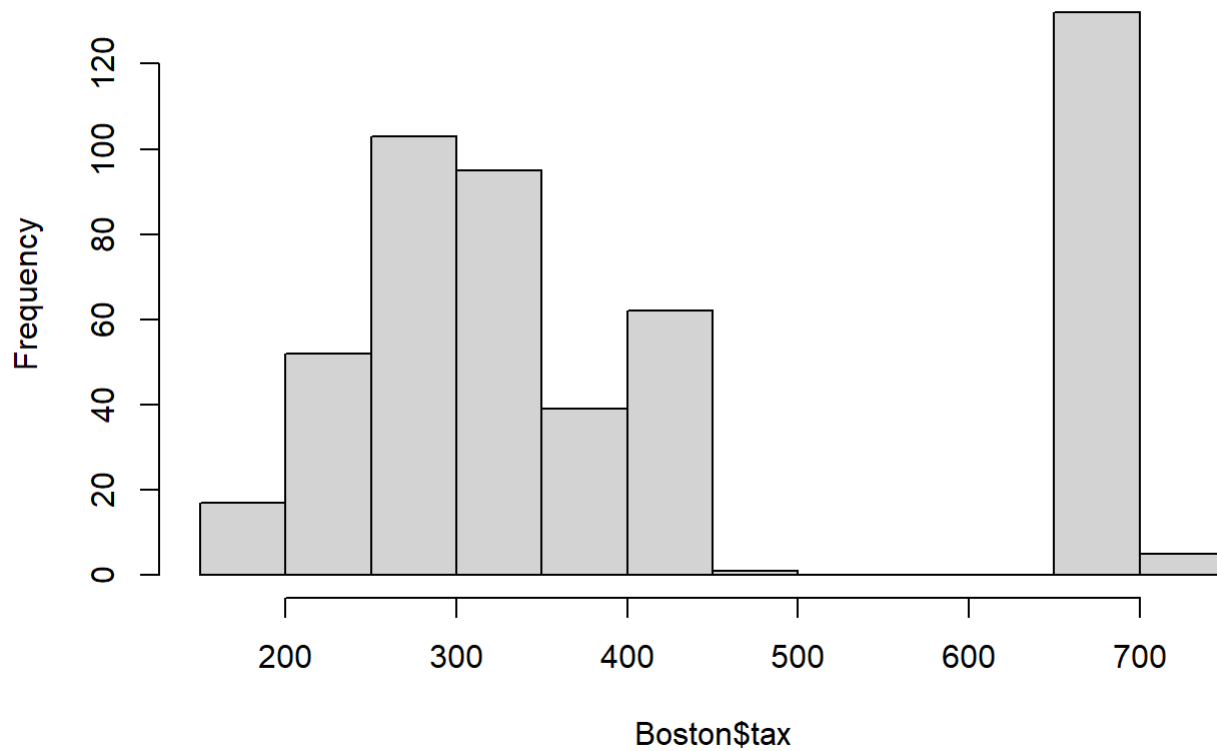
Crime Rate

```
## Range of crime rate var: 0.00632 to 88.9762
```

```
## . This shows that the variable has a very high range with a highly skewed distribution as observed in previous question. Considering very high crime rate to be 2 std dev greater than avg : 16 suburbs have very high rates
```

Tax

Histogram of Boston\$tax

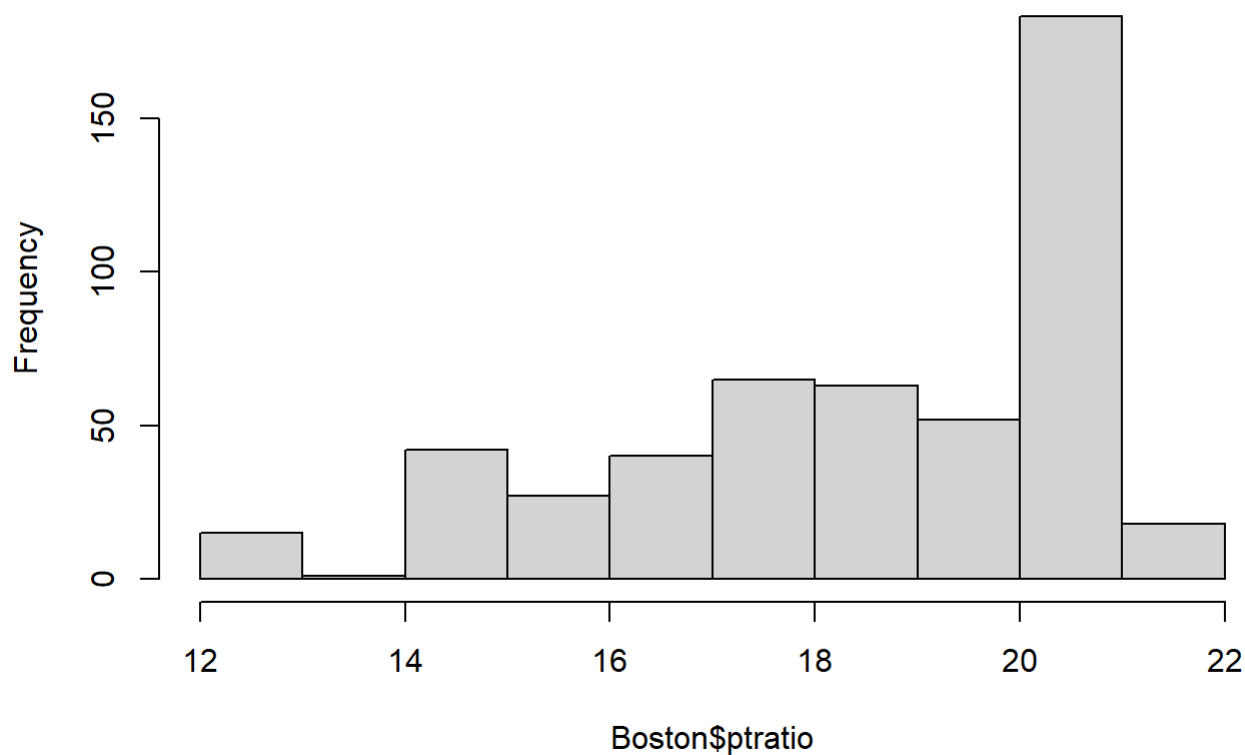


```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  187.0   279.0   330.0   408.2   666.0   711.0
```

```
## Range of Tax rate: 187 to 711
## Considering very high tax rate to be 2 std dev greater than avg , 0 suburbs can be considered to have very high tax rate relatively.
## Considering higher tax rate to be 1 std dev greater than avg , 137 suburbs can be considered to have higher tax rate relatively.
```

Pupil-Teacher Ratio

Histogram of Boston\$ptratio



```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 12.60  17.40   19.05   18.46  20.20   22.00
```

```
## Range of Pupil-Teacher Ratio: 12.6 to 22
## Considering very high ptratio to be 2 std dev greater than avg , 0 suburbs can be considered to have very ptratio relatively.
## Considering higher ptratio to be 1 std dev greater than avg , 506 suburbs can be considered to have higher ptratio rate relatively.
```

Observation:

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

```
## 35 suburbs bound the Charles River.
```

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

```
## The median pupil-teacher ratio among the towns in the data set is 19.05
```

Which census tract of Boston has 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.

```
## [1] "The suburbs with the lowest median value:"
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24

2 rows | 1-10 of 14 columns

```
## [1] "Comparing with original dataset"
```

```
##      crim      zn      indus      chas
## Min.   : 0.00632 Min.   : 0.00 Min.   : 0.46 Min.   :0.00000
## 1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000
## Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000
## Mean   : 3.61352 Mean   : 11.36 Mean   :11.14 Mean   :0.06917
## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000
## Max.   :88.97620 Max.   :100.00 Max.   :27.74 Max.   :1.00000
##      nox      rm      age      dis
## Min.   :0.3850 Min.   :3.561 Min.   : 2.90 Min.   : 1.130
## 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100
## Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207
## Mean   :0.5547 Mean   :6.285 Mean   : 68.57 Mean   : 3.795
## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188
## Max.   :0.8710 Max.   :8.780 Max.   :100.00 Max.   :12.127
##      rad      tax      ptratio      lstat
## Min.   : 1.000 Min.   :187.0 Min.   :12.60 Min.   : 1.73
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.: 6.95
## Median : 5.000 Median :330.0 Median :19.05 Median :11.36
## Mean   : 9.549 Mean   :408.2 Mean   :18.46 Mean   :12.65
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:16.95
## Max.   :24.000 Max.   :711.0 Max.   :22.00 Max.   :37.97
##      medv
## Min.   : 5.00
## 1st Qu.:17.02
## Median :21.20
## Mean   :22.53
## 3rd Qu.:25.00
## Max.   :50.00
```

Observation:

- Lstat and ptratio values for these 2 suburbs is close to maximum value in original dataset.
- Crime rate for the suburbs is very high relatively although they lie within 2 S.D. from mean value.

In this dataset, 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.

```
## 64 suburbs average more than seven rooms per dwelling.
```

```
## 13 suburbs average more than eight rooms per dwelling.
```

```
## [1] "===== Original Dataset ====="
```

```
##      crim      zn      indus      chas
## Min.   : 0.00632 Min.   : 0.00 Min.   : 0.46 Min.   : 0.00000
## 1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.: 0.00000
## Median : 0.25651 Median : 0.00 Median : 9.69 Median : 0.00000
## Mean   : 3.61352 Mean   : 11.36 Mean   : 11.14 Mean   : 0.06917
## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.: 18.10 3rd Qu.: 0.00000
## Max.   : 88.97620 Max.   : 100.00 Max.   : 27.74 Max.   : 1.00000
##      nox      rm      age      dis
## Min.   : 0.3850 Min.   : 3.561 Min.   : 2.90 Min.   : 1.130
## 1st Qu.: 0.4490 1st Qu.: 5.886 1st Qu.: 45.02 1st Qu.: 2.100
## Median : 0.5380 Median : 6.208 Median : 77.50 Median : 3.207
## Mean   : 0.5547 Mean   : 6.285 Mean   : 68.57 Mean   : 3.795
## 3rd Qu.: 0.6240 3rd Qu.: 6.623 3rd Qu.: 94.08 3rd Qu.: 5.188
## Max.   : 0.8710 Max.   : 8.780 Max.   : 100.00 Max.   : 12.127
##      rad      tax      ptratio      lstat
## Min.   : 1.000 Min.   : 187.0 Min.   : 12.60 Min.   : 1.73
## 1st Qu.: 4.000 1st Qu.: 279.0 1st Qu.: 17.40 1st Qu.: 6.95
## Median : 5.000 Median : 330.0 Median : 19.05 Median : 11.36
## Mean   : 9.549 Mean   : 408.2 Mean   : 18.46 Mean   : 12.65
## 3rd Qu.: 24.000 3rd Qu.: 666.0 3rd Qu.: 20.20 3rd Qu.: 16.95
## Max.   : 24.000 Max.   : 711.0 Max.   : 22.00 Max.   : 37.97
##      medv
## Min.   : 5.00
## 1st Qu.: 17.02
## Median : 21.20
## Mean   : 22.53
## 3rd Qu.: 25.00
## Max.   : 50.00
```

```
## [1] "===== Filtered Dataset ====="
```

```
##      crim      zn      indus      chas
## Min.   :0.02009 Min.   : 0.00 Min.   : 2.680 Min.   :0.0000
## 1st Qu.:0.33147 1st Qu.: 0.00 1st Qu.: 3.970 1st Qu.:0.0000
## Median :0.52014 Median : 0.00 Median : 6.200 Median :0.0000
## Mean   :0.71879 Mean   :13.62 Mean   : 7.078 Mean   :0.1538
## 3rd Qu.:0.57834 3rd Qu.:20.00 3rd Qu.: 6.200 3rd Qu.:0.0000
## Max.   :3.47428 Max.   :95.00 Max.   :19.580 Max.   :1.0000
##      nox      rm      age      dis
## Min.   :0.4161 Min.   :8.034 Min.   : 8.40 Min.   :1.801
## 1st Qu.:0.5040 1st Qu.:8.247 1st Qu.:70.40 1st Qu.:2.288
## Median :0.5070 Median :8.297 Median :78.30 Median :2.894
## Mean   :0.5392 Mean   :8.349 Mean   :71.54 Mean   :3.430
## 3rd Qu.:0.6050 3rd Qu.:8.398 3rd Qu.:86.50 3rd Qu.:3.652
## Max.   :0.7180 Max.   :8.780 Max.   :93.90 Max.   :8.907
##      rad      tax      ptratio      lstat      medv
## Min.   : 2.000 Min.   :224.0 Min.   :13.00 Min.   :2.47 Min.   :21.9
## 1st Qu.: 5.000 1st Qu.:264.0 1st Qu.:14.70 1st Qu.:3.32 1st Qu.:41.7
## Median : 7.000 Median :307.0 Median :17.40 Median :4.14 Median :48.3
## Mean   : 7.462 Mean   :325.1 Mean   :16.36 Mean   :4.31 Mean   :44.2
## 3rd Qu.: 8.000 3rd Qu.:307.0 3rd Qu.:17.40 3rd Qu.:5.12 3rd Qu.:50.0
## Max.   :24.000 Max.   :666.0 Max.   :20.20 Max.   :7.44 Max.   :50.0
```

Observation:

- For the suburbs $rm > 8$:
- $lstat$ and crime rate are much lower for these suburbs
- $medv$ is much higher for these suburbs

Chapter 3

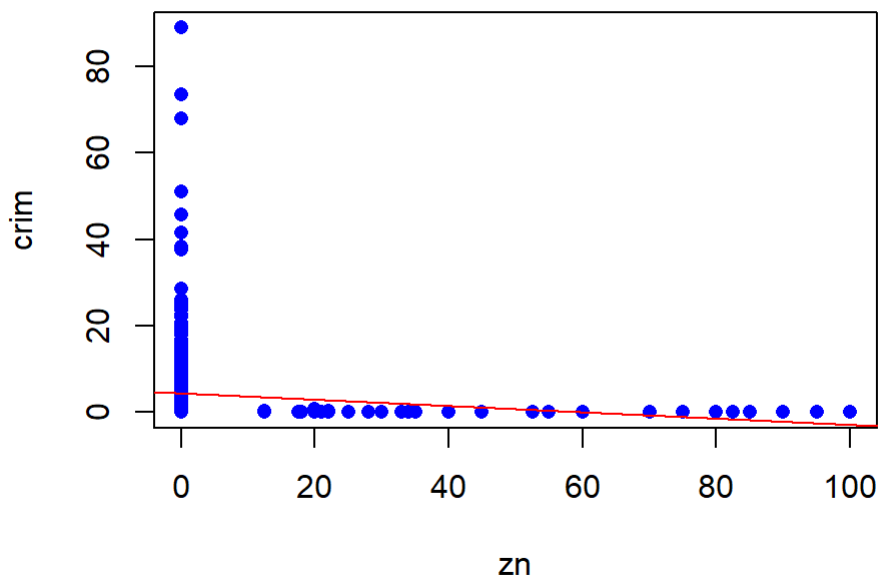
Question 15

For each predictor, fit a simple linear regression model to predict the per capita crime rate. 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.

Observations

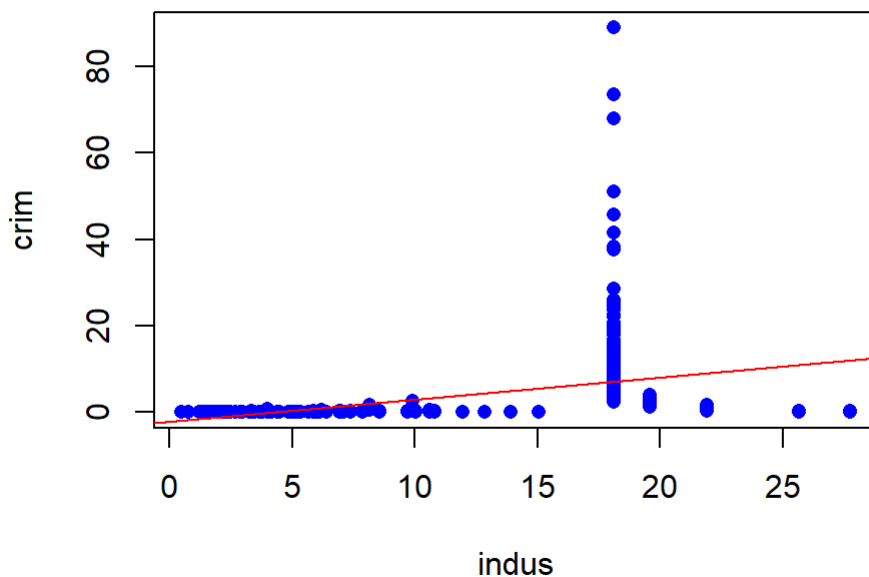
```
## regression for crim and zn
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.429 -4.222 -2.620  1.250  84.523
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675 < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic: 21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

Linear Regression Fit for zn



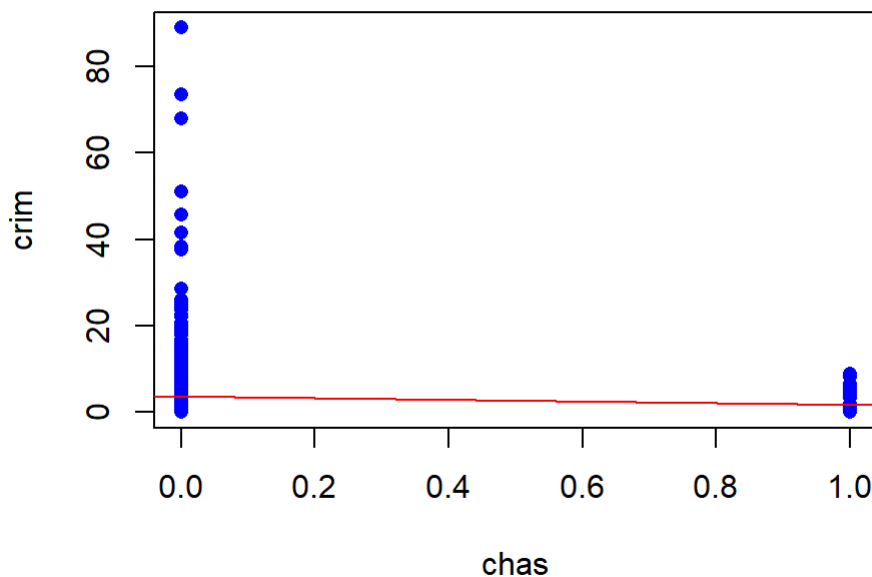
```
## regression for crim and indus
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for indus



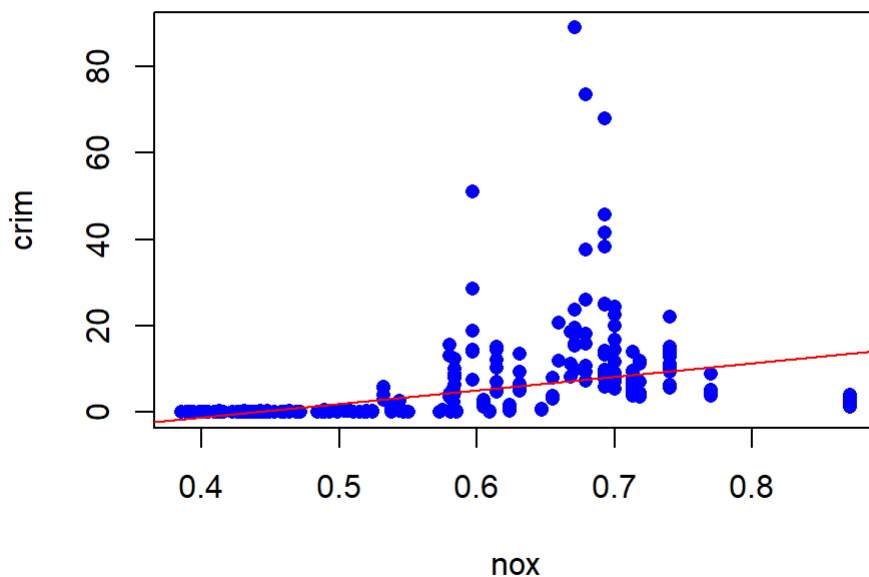
```
## regression for crim and chas
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453  <2e-16 ***
## chas         -1.8928     1.5061  -1.257   0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,    Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

Linear Regression Fit for chas



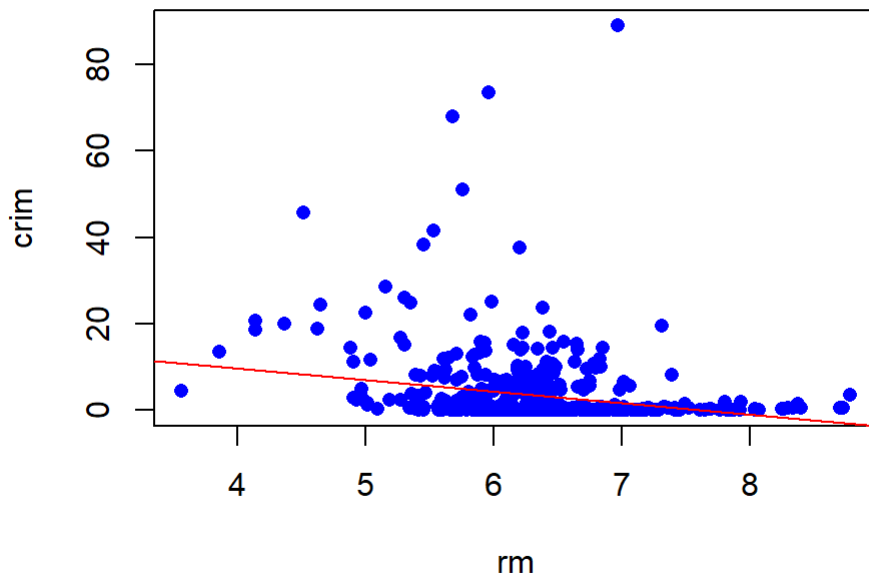
```
## regression for crim and nox
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.371  -2.738  -0.974   0.559   81.728
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699   -8.073 5.08e-15 ***
## nox           31.249      2.999   10.419 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for nox



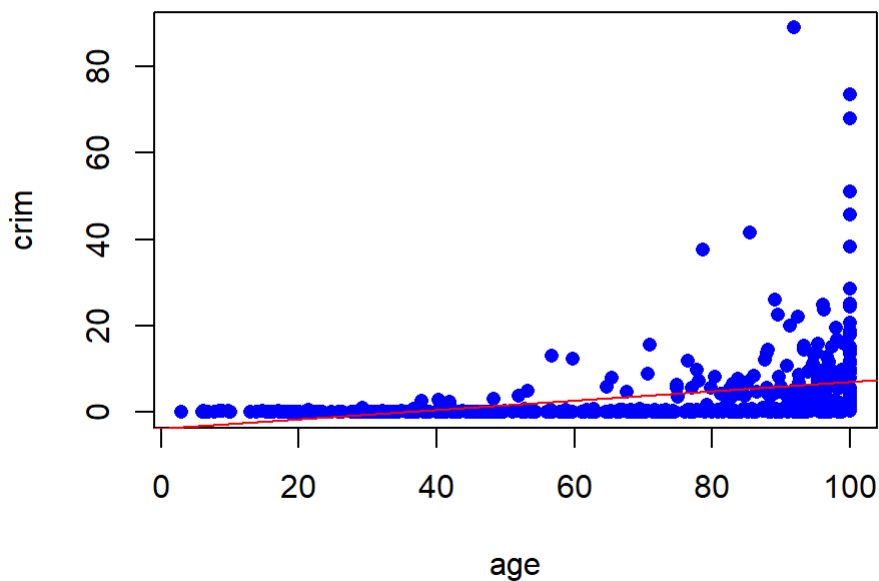

```
## regression for crim and rm
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## rm           -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

Linear Regression Fit for rm



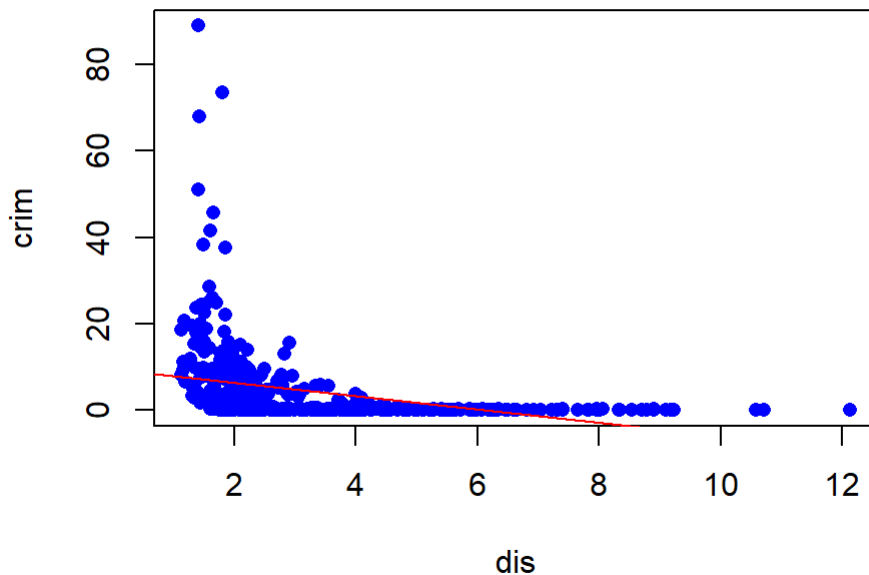
```
## regression for crim and age
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.789  -4.257  -1.230   1.527  82.849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

Linear Regression Fit for age



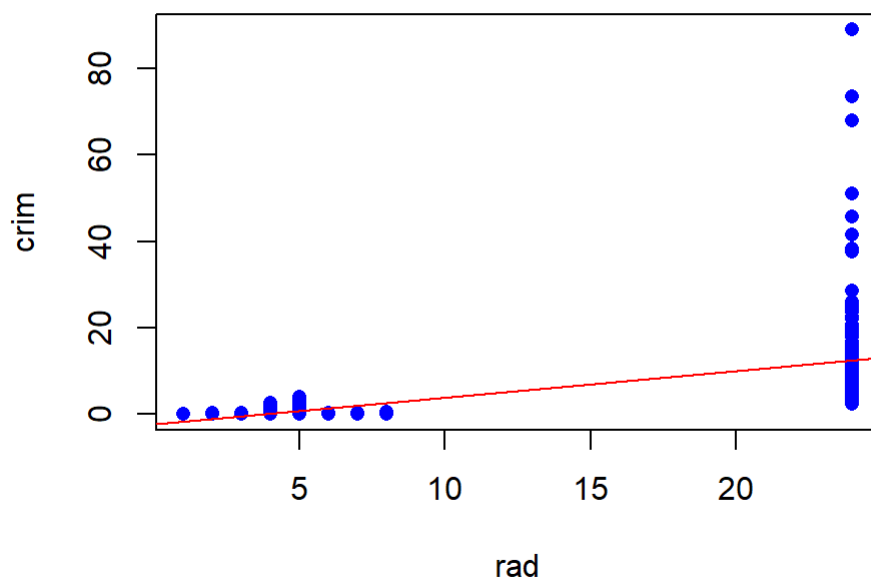
```
## regression for crim and dis
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.708 -4.134 -1.527  1.516  81.674
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4993     0.7304   13.006  <2e-16 ***
## dis          -1.5509     0.1683   -9.213  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for dis



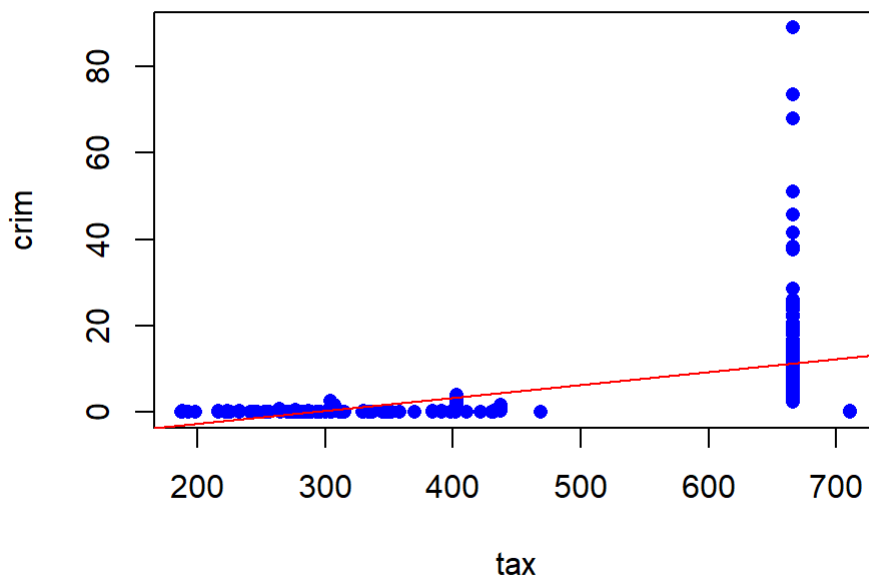
```
## regression for crim and rad
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.164  -1.381  -0.141   0.660   76.433
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716     0.44348  -5.157 3.61e-07 ***
## rad          0.61791     0.03433  17.998 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:  0.39
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for rad



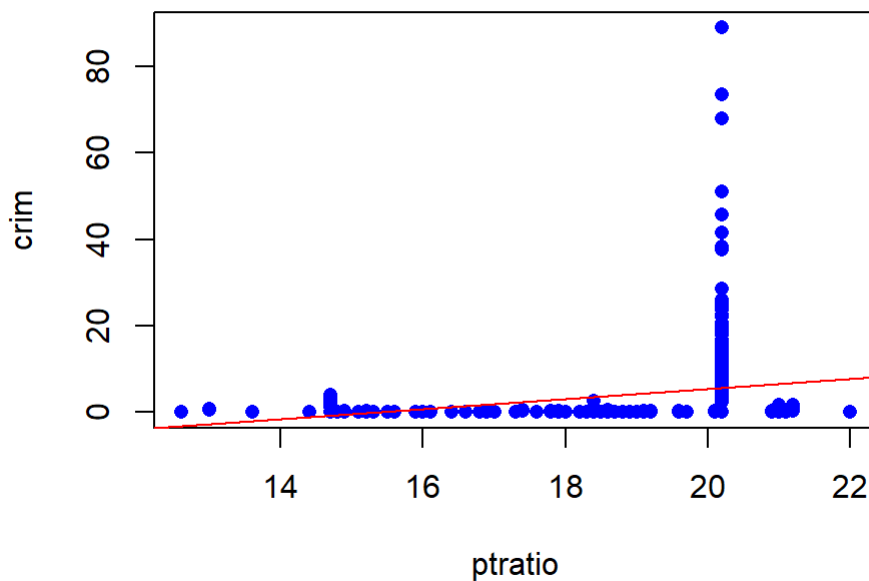
```
## regression for crim and tax
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45  <2e-16 ***
## tax          0.029742   0.001847   16.10  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for tax



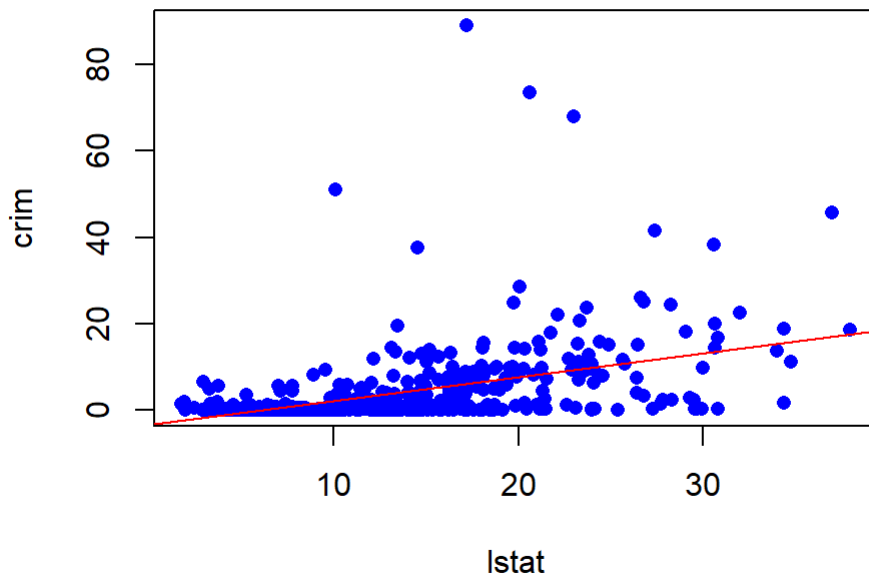
```
## regression for crim and ptratio
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio      1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

Linear Regression Fit for ptratio



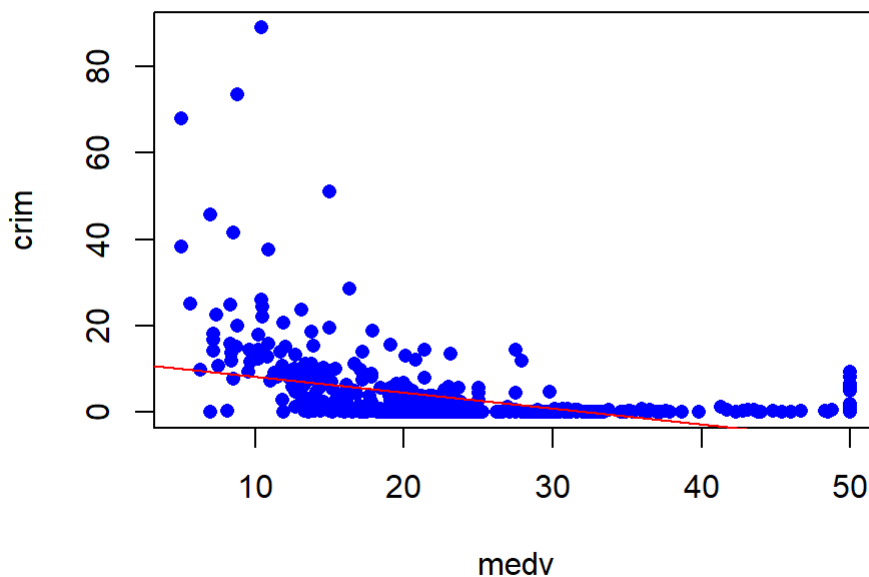
```
## regression for crim and lstat
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:  132 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for lstat



```
## regression for crim and medv
## Call:
## lm(formula = paste(response_var, "~", predictor), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.071  -4.022  -2.343   1.298   80.957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.79654    0.93419   12.63  <2e-16 ***
## medv        -0.36316    0.03839   -9.46  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

Linear Regression Fit for medv



Observations:

- Variables with Positive Relationship with crime rate - indus, nox, age, rad, tax, ptratio, lstat
- Variables with Negative Relationship with crime rate - zn, rm, chas, rm, dis, medv
- The p-value for the variable CHAS is large, indicating that the relationship between CHAS and the crime rate is not statistically significant.
- On the other hand, all the other predictor variables in the model have small p-values, indicating that they have a significant relationship with the crime rate.

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 $H_0 : \beta_j = 0$?


```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.534 -2.248 -0.348  1.087 73.923
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.7783938   7.0818258   1.946 0.052271 .
## zn          0.0457100   0.0187903   2.433 0.015344 *
## indus       -0.0583501   0.0836351  -0.698 0.485709
## chas        -0.8253776   1.1833963  -0.697 0.485841
## nox         -9.9575865   5.2898242  -1.882 0.060370 .
## rm          0.6289107   0.6070924   1.036 0.300738
## age        -0.0008483   0.0179482  -0.047 0.962323
## dis        -1.0122467   0.2824676  -3.584 0.000373 ***
## rad         0.6124653   0.0875358   6.997 8.59e-12 ***
## tax        -0.0037756   0.0051723  -0.730 0.465757
## ptratio    -0.3040728   0.1863598  -1.632 0.103393
## lstat       0.1388006   0.0757213   1.833 0.067398 .
## medv       -0.2200564   0.0598240  -3.678 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.46 on 493 degrees of freedom
## Multiple R-squared:  0.4493, Adjusted R-squared:  0.4359
## F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

Observations:

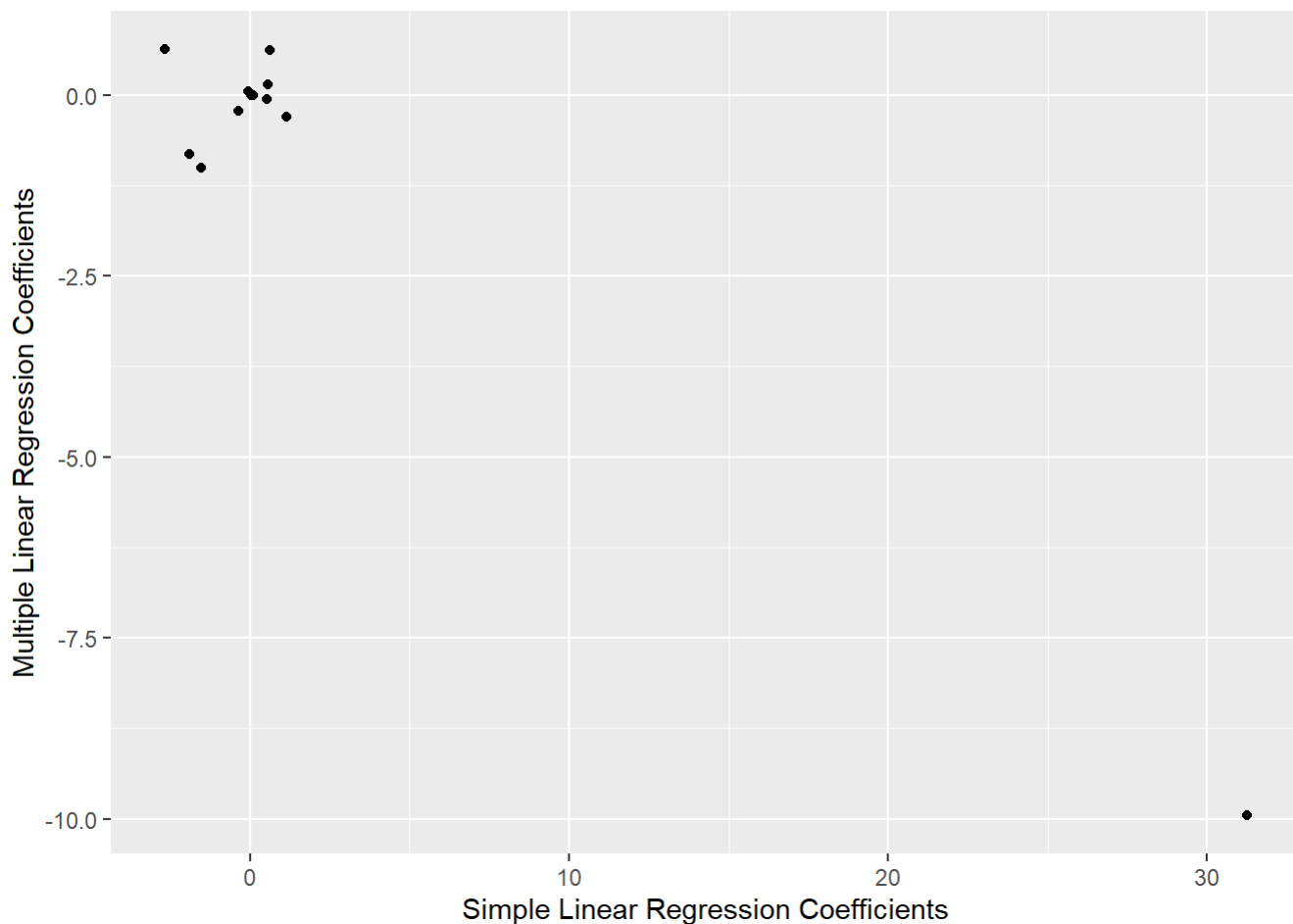
The p-values for the predictors ZN, DIS, RAD, and MEDV are all below 0.05, which suggests that these variables have a significant influence on the response variable. Therefore, we can reject the null hypothesis.

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.

Observations:

- The coefficient estimates of predictors have changed.
- The relationship of a predictor with response has also changed. For example, In simple regression ptratio had a positive relationship but in multiple regression it has a negative relationship with crime.

##	multiple_regression_coefs	simple_regression_coefs
## zn	0.0457100386	-0.07393498
## indus	-0.0583501107	0.50977633
## chas	-0.8253775522	-1.89277655
## nox	-9.9575865471	31.24853120
## rm	0.6289106622	-2.68405122
## age	-0.0008482791	0.10778623
## dis	-1.0122467382	-1.55090168
## rad	0.6124653115	0.61791093
## tax	-0.0037756465	0.02974225
## ptratio	-0.3040727572	1.15198279
## lstat	0.1388005968	0.54880478
## medv	-0.2200563590	-0.36315992



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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

```
## Predictor P_values_degree1 P_values_degree2 P_values_degree3
## 1      zn      4.697806e-06      4.420507e-03      2.295386e-01
## 2      indus    8.854243e-24      1.086057e-03      1.196405e-12
## 3      nox      2.457491e-26      7.736755e-05      6.961110e-16
## 4      rm       5.128048e-07      1.508545e-03      5.085751e-01
## 5      age      4.878803e-17      2.291156e-06      6.679915e-03
## 6      dis      1.253249e-21      7.869767e-14      1.088832e-08
## 7      rad      1.053211e-56      9.120558e-03      4.823138e-01
## 8      tax      6.976314e-49      3.665348e-06      2.438507e-01
## 9      ptratio  1.565484e-11      2.405468e-03      6.300514e-03
## 10     lstat    1.678072e-27      3.780418e-02      1.298906e-01
## 11     medv    4.930818e-27      2.928577e-35      1.046510e-12
## Coeff_values_degree1 Coeff_values_degree2 Coeff_values_degree3
## 1      -38.74984      23.93983      -10.071868
## 2      78.59082      -24.39480      -54.129763
## 3      81.37202      -28.82859      -60.361894
## 4     -42.37944      26.57677      -5.510342
## 5      68.18201      37.48447      21.353207
## 6     -73.38859      56.37304      -42.621877
## 7     120.90745      17.49230       4.698457
## 8     112.64583      32.08725      -7.996811
## 9      56.04523      24.77482     -22.279737
## 10     88.06967      15.88816     -11.574022
## 11    -75.05761      88.08621     -48.033435
```

Observations:

- Based on p-values, variables with non linear relationship with crim are:
- indus ,nox ,age ,dis ,ptratio ,medv

Chapter 6

Question 9

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

```
## Training set dimensions: 543 18
```

```
## Test set dimensions: 234 18
```

Fit a linear model using least squares on the training set, and report the test error obtained.

```
## Linear regression MSE = 1555839
```

Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

```
## Ridge regression MSE = 1555797
```

Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
## [1] 1555851
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##              s1
## (Intercept) -471.39372052
## (Intercept) .
## PrivateYes  -491.04485137
## Accept      1.57033288
## Enroll      -0.75961467
## Top10perc   48.14698892
## Top25perc   -12.84690695
## F.Undergrad 0.04149116
## P.Undergrad 0.04438973
## Outstate    -0.08328388
## Room.Board  0.14943472
## Books       0.01532293
## Personal    0.02909954
## PhD         -8.39597537
## Terminal    -3.26800340
## S.F.Ratio   14.59298267
## perc.alumni -0.04404771
## Expend      0.07712632
## Grad.Rate   8.28950241
```

```
## Mean Squared Error for Lasso: 1555851
```

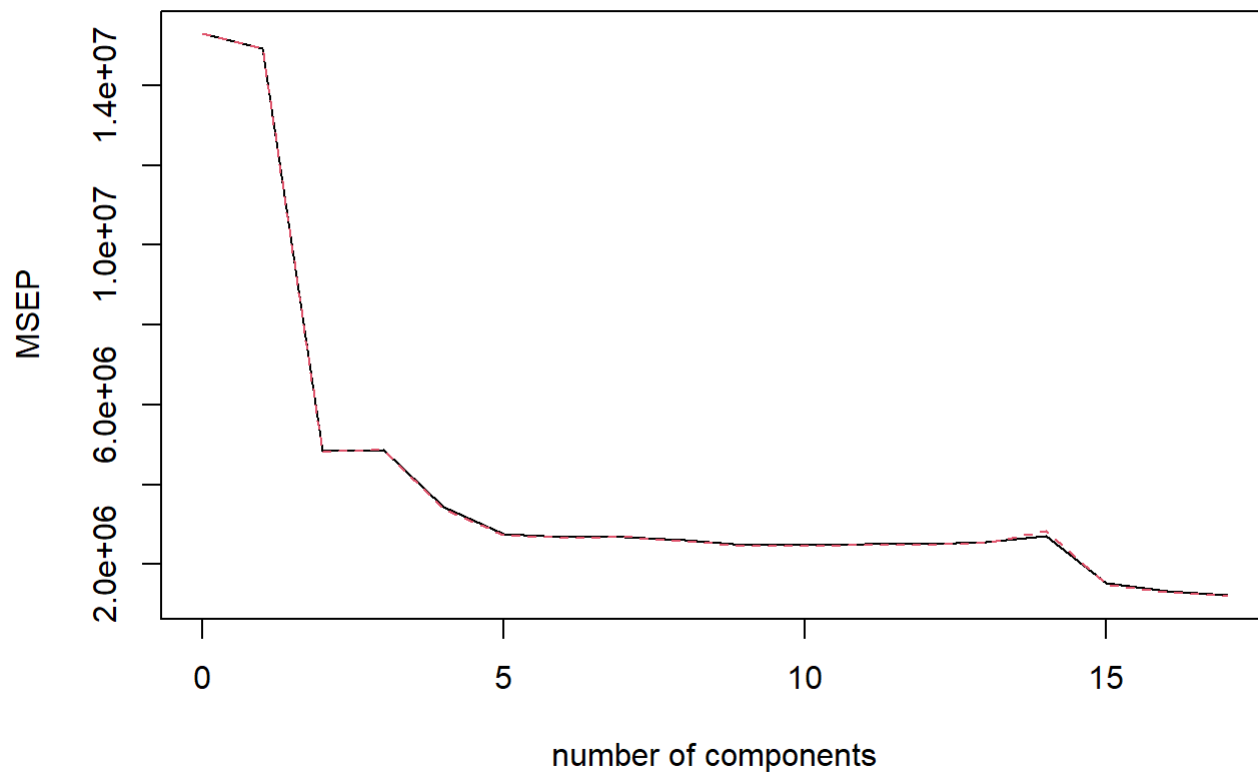
```
## Number of Non-Zero Coefficients: 18
```

Fit a PCR model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.

```
## [1] 1780624
```

```
## Mean Squared Error for PCR: 1780624
```

Apps

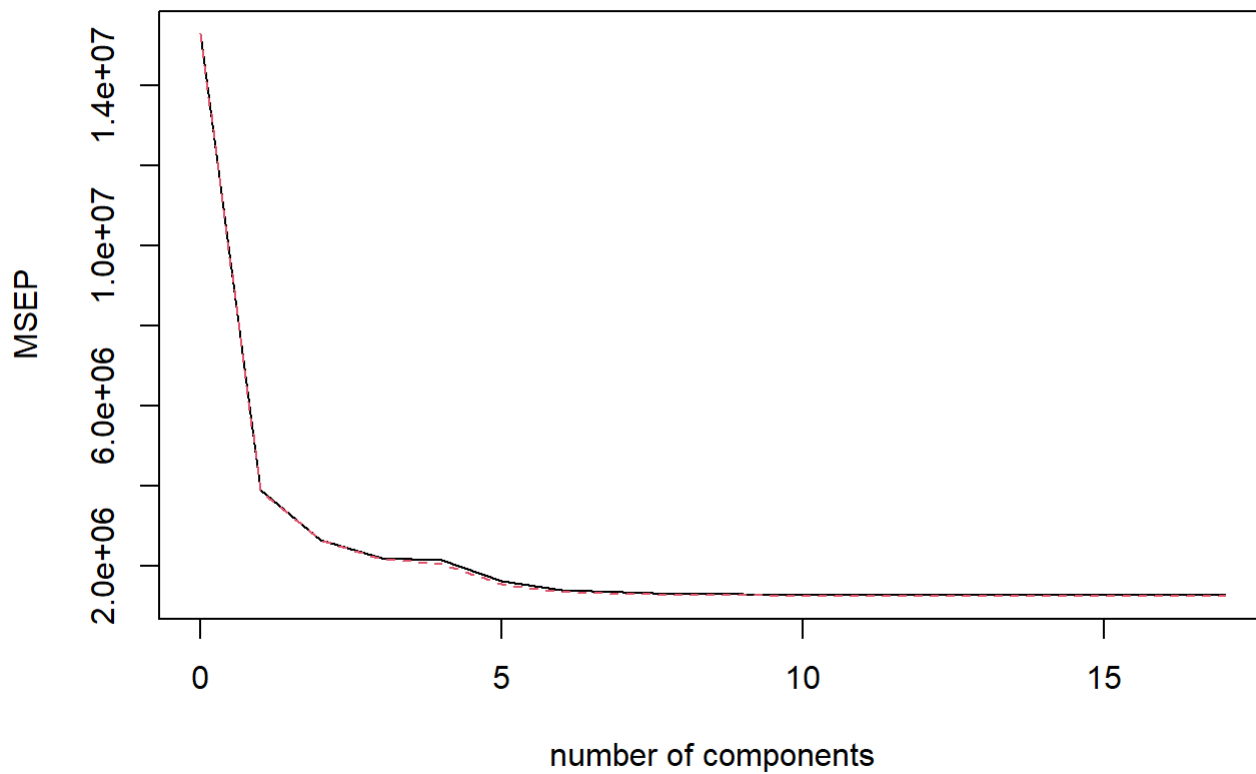


Fit a PLS model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.

```
## [1] 1558562
```

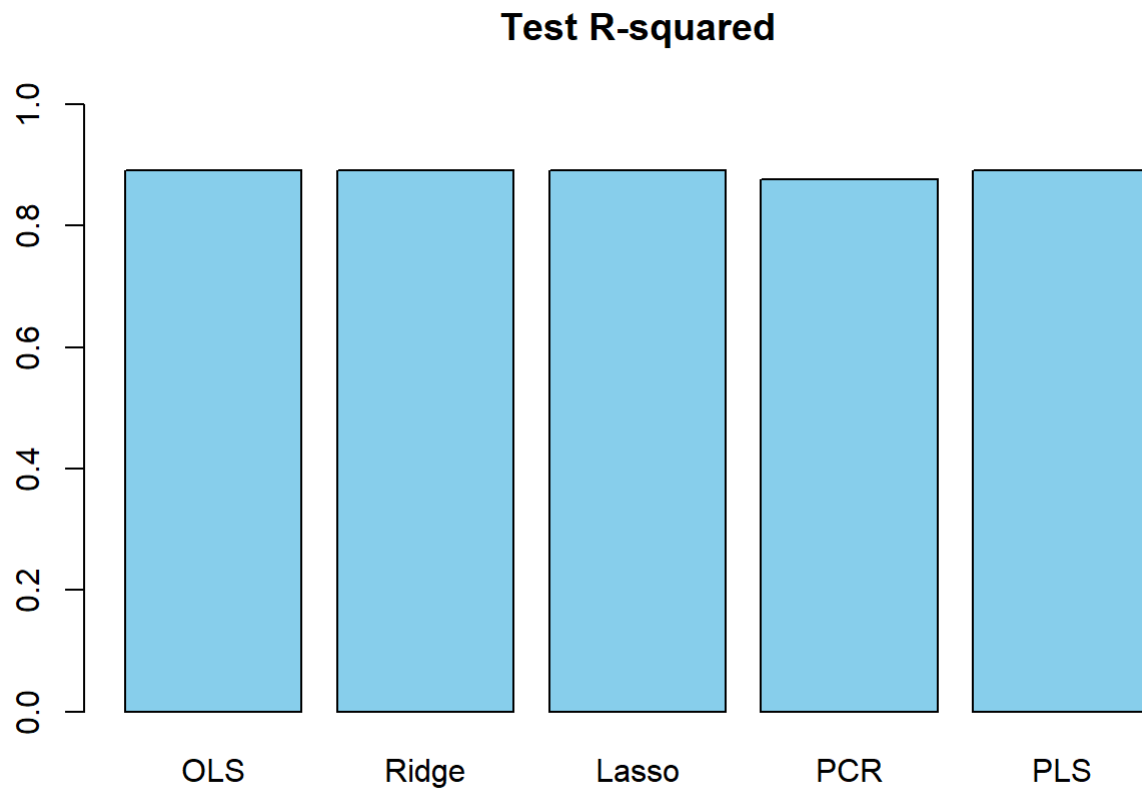
```
## Mean Squared Error for PLS: 1558562
```

Apps



Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

```
## [1] 0.8913273 0.8913302 0.8913264 0.8756264 0.8911371
```



Observations

- Lasso reduces F.Undergrad, Books and Personal variables to zero and shrinks coefficients of other variables.
- The plot shows that test R² for all models are around 0.9, with PCR having slightly lower test R² than others.

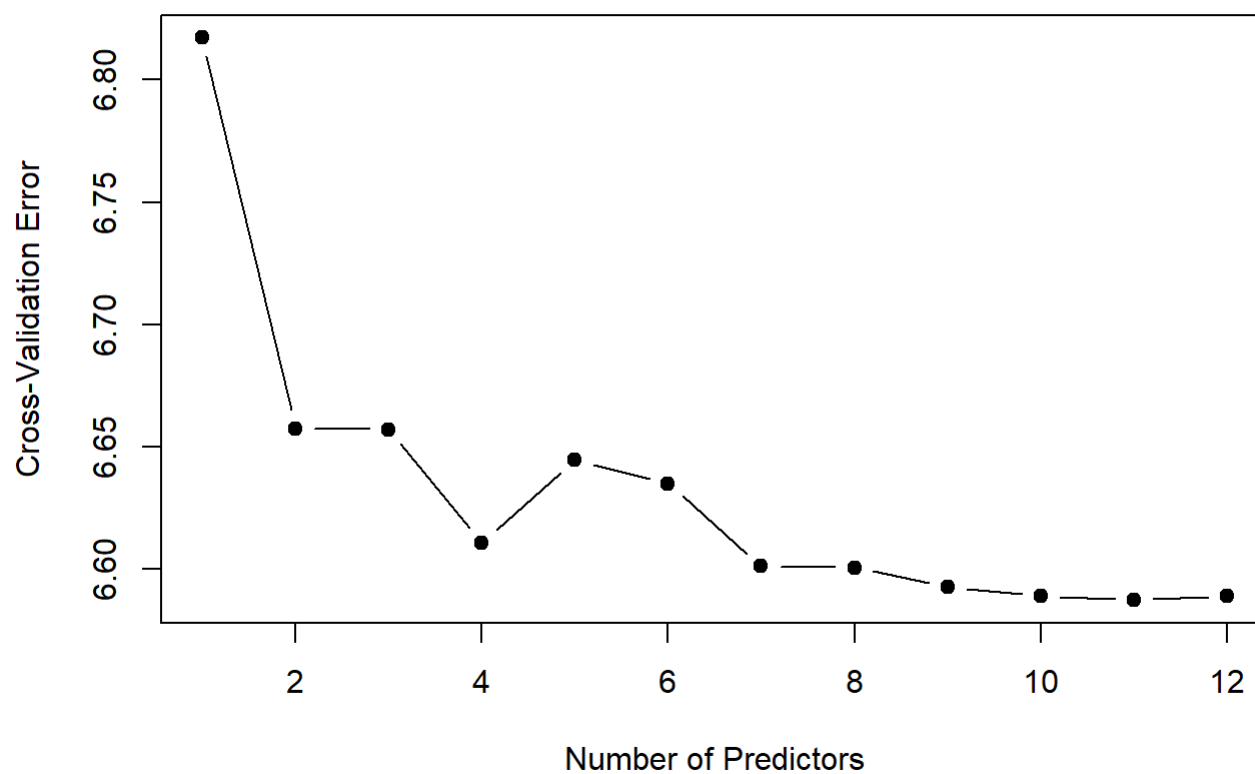
Question 11

Try out some of the regression methods on Boston data, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

Best Subset Selection

```
## [1] 12
```

Best Subset Selection



```
## [1] 6.817211 6.657478 6.657168 6.610904 6.644513 6.635045 6.601194 6.600679
## [9] 6.592704 6.589167 6.587325 6.589047
```

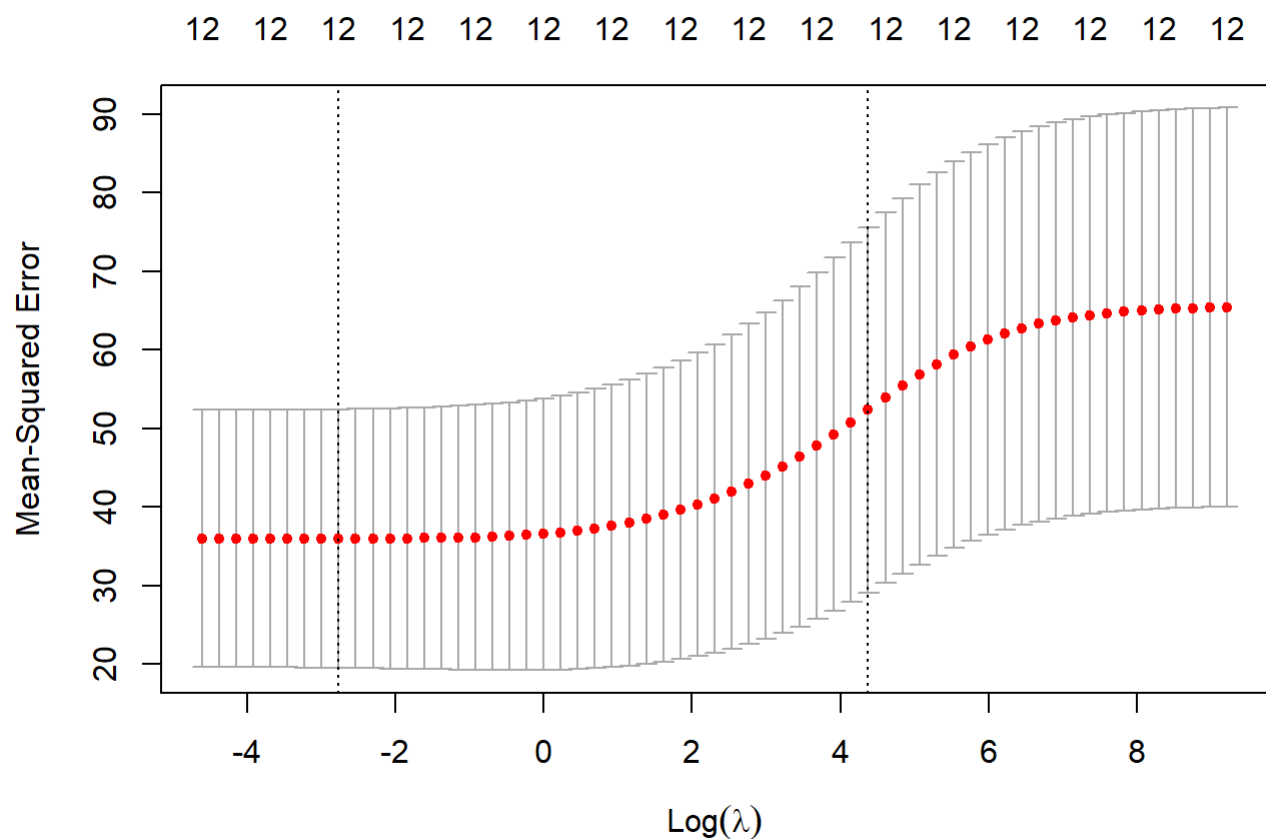
Observation:

- Cross validation error for a model with 11 variables is the lowest.

Ridge Regression

```
## Root Mean Squared Error for Ridge Regression: 7.602275
```

```
## Mean Squared Error for Ridge Regression: 57.79459
```

```
## Optimal Lambda value 0.06309573
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##              s1
## (Intercept)  0.785984530
## zn          -0.002892501
## indus        0.027226824
## chas        -0.158192556
## nox          1.827244414
## rm          -0.171419430
## age          0.005929627
## dis         -0.086550047
## rad          0.044255312
## tax          0.001991582
## ptratio      0.065220527
## lstat        0.038264057
## medv        -0.021851245
```

Lasso

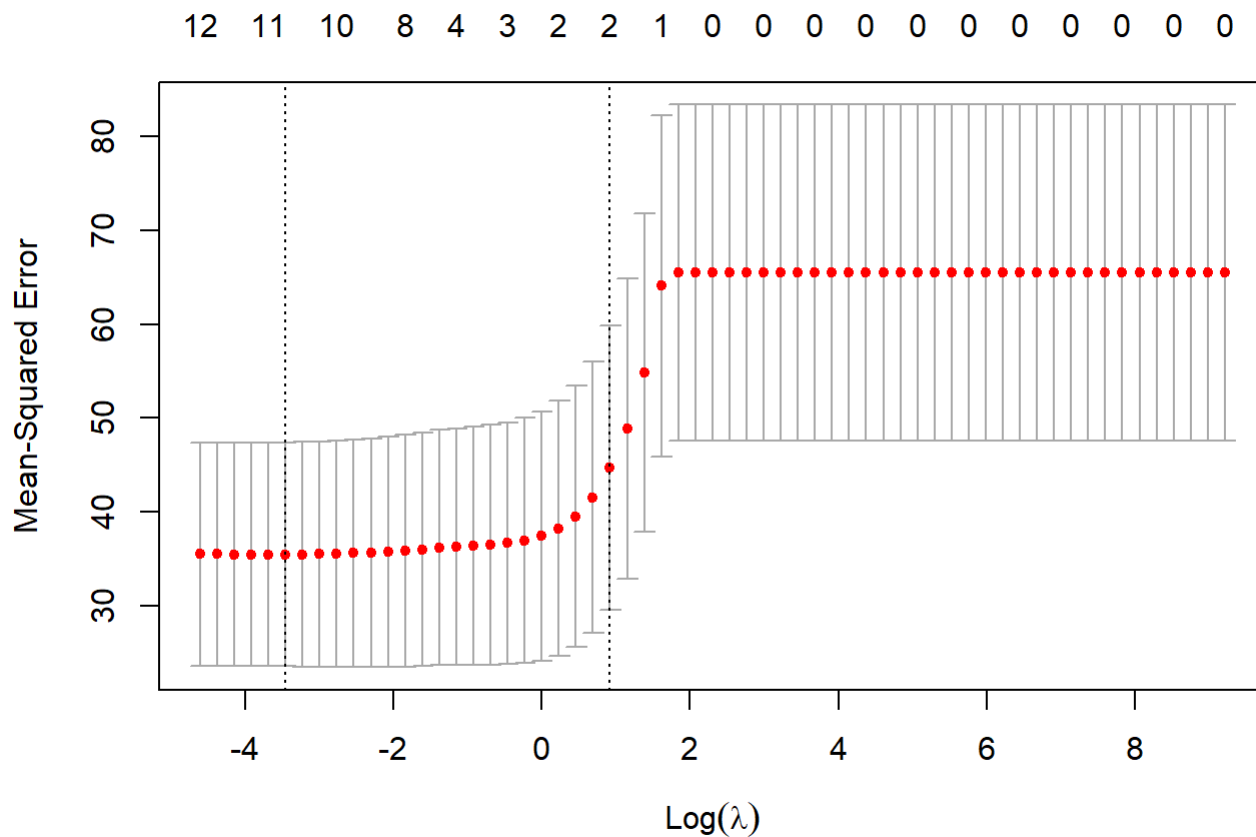
```
##
## Call: glmnet(x = model.matrix(crim ~ ., train_data)[, -1], y = train_data$crim, alpha =
1, lambda = lasso_best_lambda)
##
## Df %Dev Lambda
## 1 10 47.94 0.03162
```

```
## Root Mean Squared Error for Lasso Regression: 7.620056
```

```
## Mean Squared Error for Lasso Regression: 58.06525
```

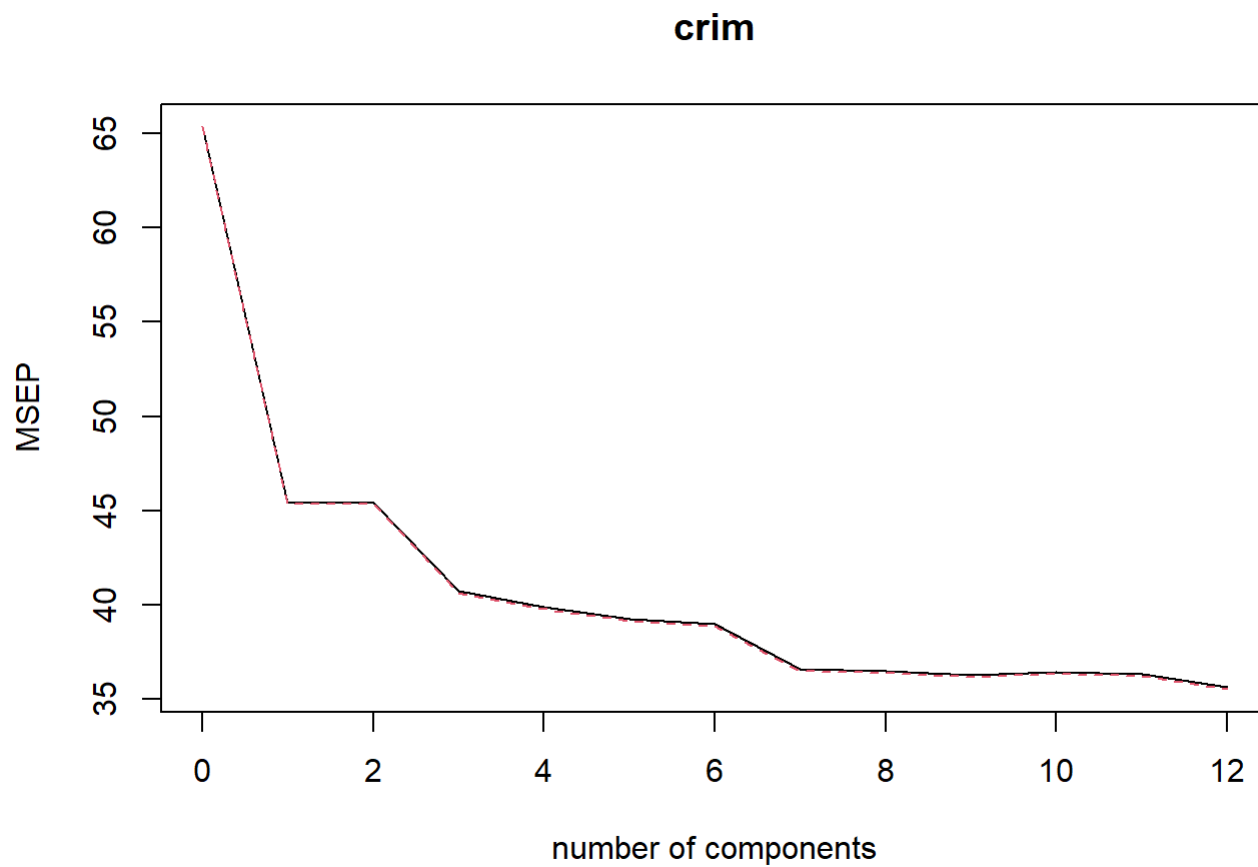
```
## Number of Non-Zero Coefficients in Lasso model: 10
```

```
## Lasso lambda best value = 0.03162278
```



```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  7.45188921
## zn          0.03198721
## indus       -0.07511476
## chas        -0.64971146
## nox         -4.86238217
## rm          0.31716023
## age         .
## dis        -0.58563461
## rad         0.52110294
## tax         .
## ptratio    -0.28952577
## lstat       0.22519281
## medv       -0.12760552
```

PCR Model



```
## Data:      X dimension: 354 12
## Y dimension: 354 1
## Fit method: svdpc
## Number of components considered: 12
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##      (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps
## CV           8.085   6.738   6.738   6.380   6.316   6.262   6.242
## adjCV        8.085   6.734   6.733   6.374   6.306   6.257   6.236
##      7 comps 8 comps 9 comps 10 comps 11 comps 12 comps
## CV          6.048   6.040   6.021   6.035   6.026   5.967
## adjCV        6.041   6.032   6.013   6.028   6.018   5.959
##
## TRAINING: % variance explained
##      1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps
## X          50.57   64.47   73.58   80.72   87.14   90.65   93.18   95.07
## crim       31.52   31.89   39.44   40.88   41.82   42.41   45.89   46.09
##      9 comps 10 comps 11 comps 12 comps
## X          96.86   98.34   99.46  100.00
## crim       46.40   46.46   46.94   48.03
```

```
## Root Mean Squared Error for PCR: 7.816138
```

```
## Mean Squared Error for PCR: 61.09201
```

12 component pcr fit has lowest CV/adjCV RMSEP.

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, cross-validation, or some other reasonable alternative, as opposed to using training error.

Observations

- MSE for different techniques are as follows:
- Lasso -> 58.06525
- Ridge Regression -> 58.251
- PCR -> 61.09201

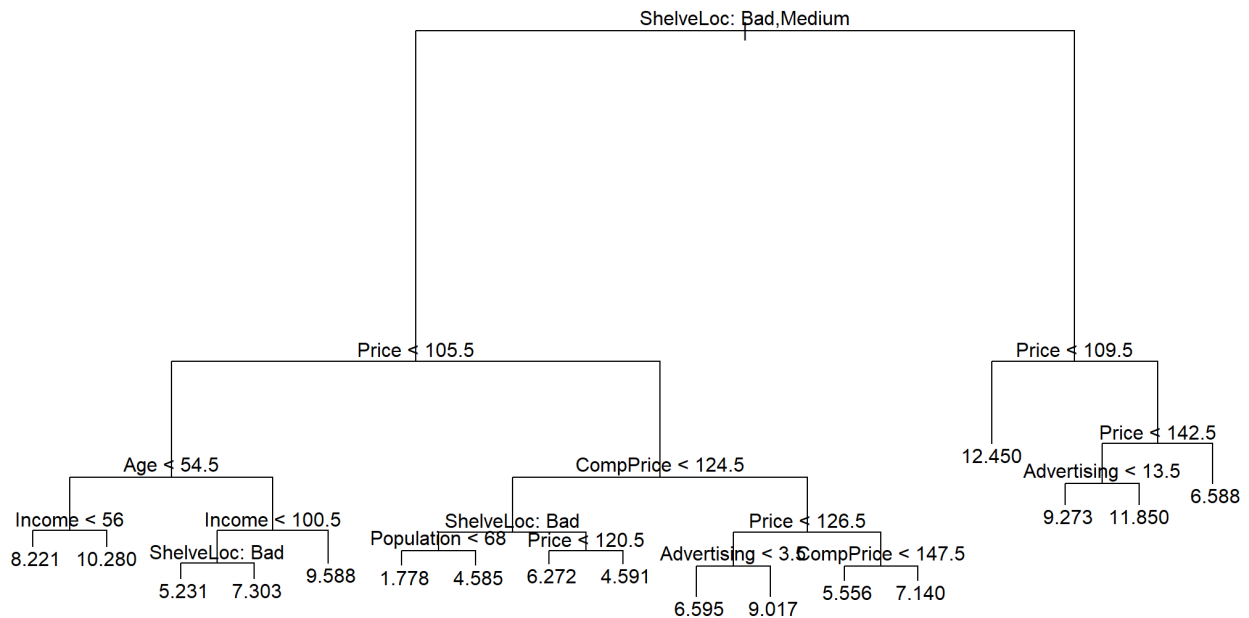
I would chose the Lasso Model because it has the lowest MSE and more interpretable comparatively since it reduces contribution of 2 variables for prediction (age and tax)

c. Does your chosen model involve all of the features in the data set? Why or why not?

Chosen Model has 10 variables as 2 variables contribution has been reduced by lasso technique.

Chapter 8: Question 8

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

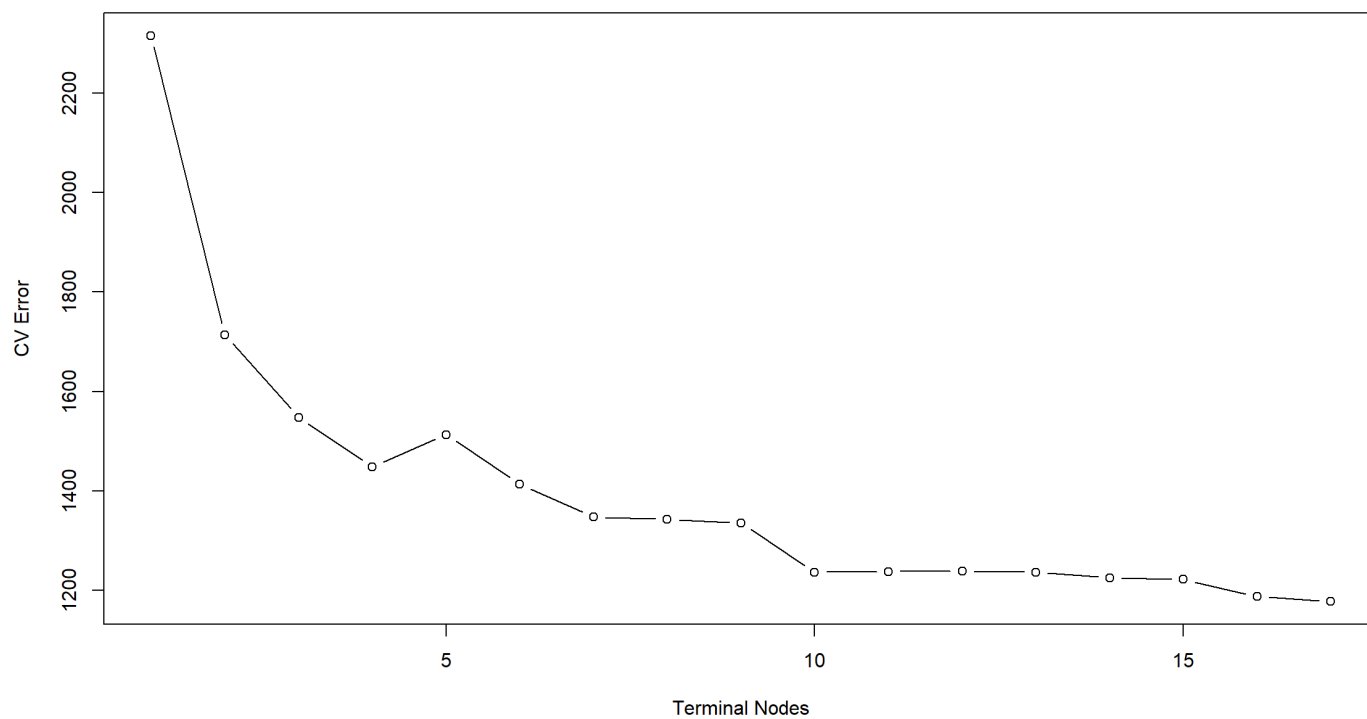


Test MSE for regression tree: 4.359207

Most important features of the tree model are **ShelveLoc** and **Price**

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

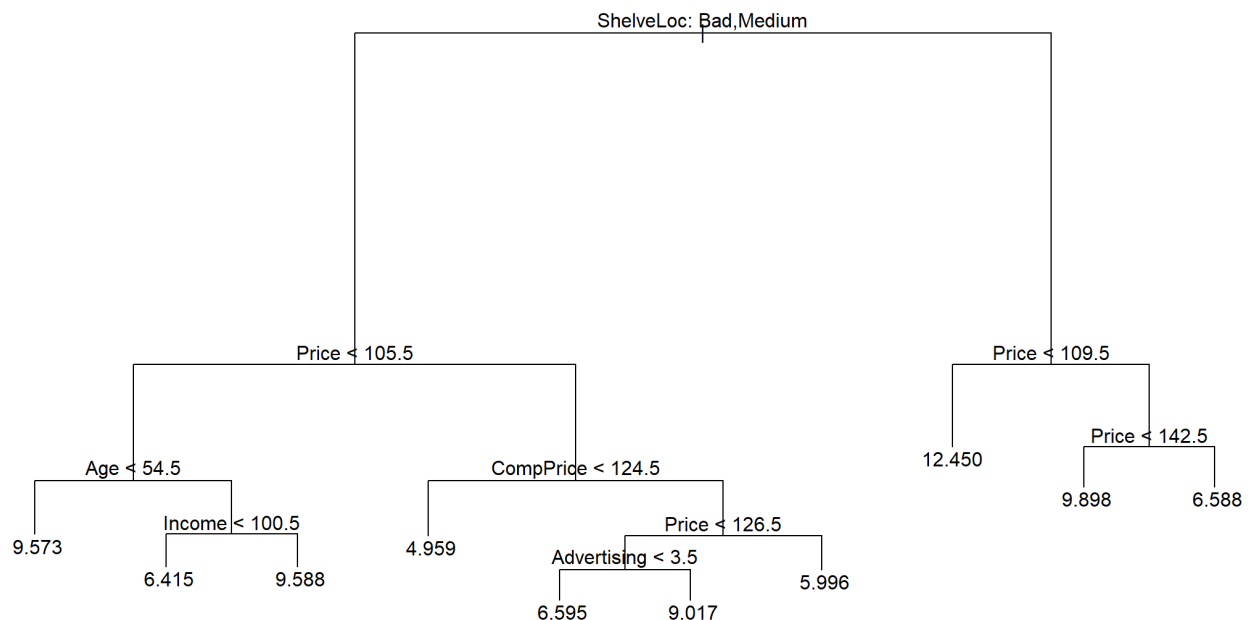
```
## $size
## [1] 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
##
## $dev
## [1] 1177.487 1187.328 1221.951 1224.846 1236.326 1238.141 1237.628 1236.377
## [9] 1335.368 1342.913 1347.593 1413.206 1513.159 1447.833 1547.726 1714.123
## [17] 2316.023
##
## $k
## [1] -Inf 25.11041 27.17616 27.80911 29.44726 30.63478 35.99601
## [8] 40.25430 58.03645 62.62515 65.54326 77.46582 101.92302 104.51745
## [15] 157.68280 223.52150 634.61364
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune" "tree.sequence"
```



```

## node), split, n, deviance, yval
##      * denotes terminal node
##
## 1) root 280 2296.00  7.603
##    2) ShelfLoc: Bad,Medium 215 1226.00  6.775
##      4) Price < 105.5 68  370.90  8.274
##        8) Age < 54.5 32  86.31  9.573 *
##        9) Age > 54.5 36  182.70  7.120
##          18) Income < 100.5 28  105.20  6.415 *
##          19) Income > 100.5 8  14.83  9.588 *
##      5) Price > 105.5 147  631.10  6.082
##        10) CompPrice < 124.5 53  187.30  4.959 *
##        11) CompPrice > 124.5 94  339.20  6.715
##          22) Price < 126.5 40  110.40  7.685
##            44) Advertising < 3.5 22  37.47  6.595 *
##            45) Advertising > 3.5 18  14.93  9.017 *
##          23) Price > 126.5 54  163.20  5.996 *
##    3) ShelfLoc: Good 65  435.90 10.340
##      6) Price < 109.5 23  62.35 12.450 *
##      7) Price > 109.5 42  215.90  9.188
##        14) Price < 142.5 33  118.00  9.898 *
##        15) Price > 142.5 9   20.40  6.588 *

```



```
## Pruned tree test MSE: 4.835634
```

Pruning the tree in this case increases the test MSE to 5.02

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.

```
##           %IncMSE  IncNodePurity
## CompPrice  37.0955879    230.468221
## Income    10.1434917    111.947577
## Advertising 22.6537826    187.171027
## Population  2.2851749     79.466849
## Price      74.6439392    645.545497
## ShelfLoc   75.1237464    713.386182
## Age        22.3025228    188.774611
## Education  6.0214758     56.339201
## Urban      0.7278781      9.285655
## US         3.6121873      7.030894
```

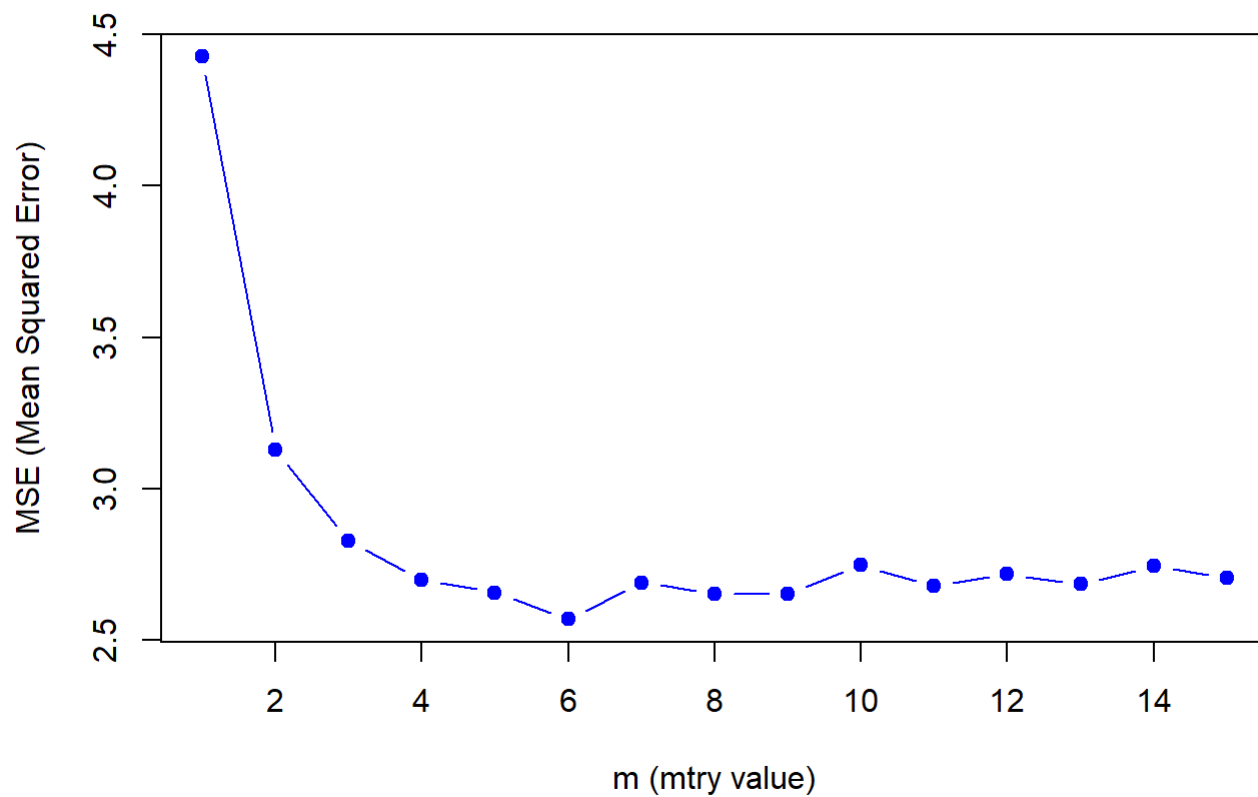
```
## Bagging tree test MSE: 2.659915
```

- With the bagging method, we observe the following:
- The most important features continue to remain Shelveloc and Price.
- Test MSE with Bagging model has decreased to the value 2.82.

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.

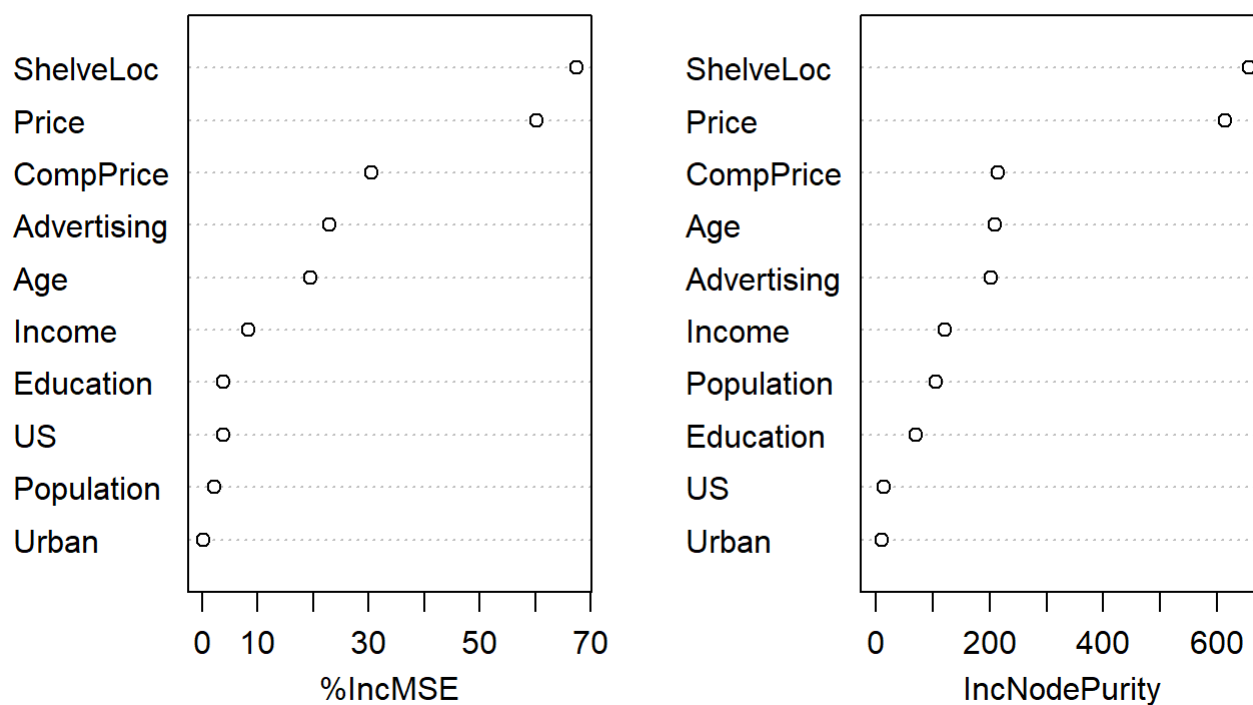

```
## [1] "M value = 1"
## Root Mean Square Error for mtry = 1 4.426806
## [1] "=====
## [1] "M value = 2"
## Root Mean Square Error for mtry = 2 3.12672
## [1] "=====
## [1] "M value = 3"
## Root Mean Square Error for mtry = 3 2.82761
## [1] "=====
## [1] "M value = 4"
## Root Mean Square Error for mtry = 4 2.698047
## [1] "=====
## [1] "M value = 5"
## Root Mean Square Error for mtry = 5 2.656558
## [1] "=====
## [1] "M value = 6"
## Root Mean Square Error for mtry = 6 2.568772
## [1] "=====
## [1] "M value = 7"
## Root Mean Square Error for mtry = 7 2.690031
## [1] "=====
## [1] "M value = 8"
## Root Mean Square Error for mtry = 8 2.653647
## [1] "=====
## [1] "M value = 9"
## Root Mean Square Error for mtry = 9 2.653374
## [1] "=====
## [1] "M value = 10"
## Root Mean Square Error for mtry = 10 2.746643
## [1] "=====
## [1] "M value = 11"
## Root Mean Square Error for mtry = 11 2.677436
## [1] "=====
## [1] "M value = 12"
## Root Mean Square Error for mtry = 12 2.718386
## [1] "=====
## [1] "M value = 13"
## Root Mean Square Error for mtry = 13 2.683535
## [1] "=====
## [1] "M value = 14"
## Root Mean Square Error for mtry = 14 2.744941
## [1] "=====
## [1] "M value = 15"
## Root Mean Square Error for mtry = 15 2.705269
## [1] "=====
```

Random Forest: mtry vs. MSE



```
## Minimum MSE: 2.568772 at m = 6
```

Variable Importance Plot (Mtry = 6)



Observations

- For the most optimal random forest model , MSE is lowest. Random forest model with $m = 9$ has lowest MSE at 2.786.
- ShelfLoc ,Price and CompPrice are the most important variables in the rf model.

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

```
## *****Calling gbart: type=1
## *****Data:
## data:n,p,np: 280, 14, 120
## y1,yn: 2.037107, -0.972893
## x1,x[n*p]: 106.000000, 0.000000
## xp1,xp[np*p]: 113.000000, 1.000000
## *****Number of Trees: 200
## *****Number of Cut Points: 67 ... 1
## *****burn,nd,thin: 100,1000,1
## *****Prior:beta,alpha,tau,nu,lambda,offset: 2,0.95,0.278247,3,0.215265,7.60289
## *****sigma: 1.051239
## *****w (weights): 1.000000 ... 1.000000
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,14,0
## *****printevery: 100
##
## MCMC
## done 0 (out of 1100)
## done 100 (out of 1100)
## done 200 (out of 1100)
## done 300 (out of 1100)
## done 400 (out of 1100)
## done 500 (out of 1100)
## done 600 (out of 1100)
## done 700 (out of 1100)
## done 800 (out of 1100)
## done 900 (out of 1100)
## done 1000 (out of 1100)
## time: 5s
## trcnt,tecnt: 1000,1000
```

```
##
## BART RMSE: 1.246607
```

##	Price	CompPrice	Age	ShelveLoc3	ShelveLoc2	US2
##	23.968	19.560	18.570	18.231	17.102	16.642
##	ShelveLoc1	Education	Urban1	Population	Urban2	Income
##	16.622	15.981	15.965	15.813	15.568	15.507
##	Advertising	US1				
##	15.493	15.410				

- BART technique produces the lowest RMSE among all the tried techniques at 1.24.
- Most important variables are Price , CompPrice , ShelveLoc1

Chapter 8: Question 11

a. Create a training set consisting of the first 1,000 observations on Caravan Dataset, and a test set consisting of the remaining observations.

```
## dimensions of training set : 1000 87
```

```
## dimensions of test set: 4822 87
```

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 important?

##	var	rel.inf
##	PPERSAUT	PPERSAUT 14.84712121
##	MKOOPKLA	MKOOPKLA 9.66048095
##	MOPLHOOG	MOPLHOOG 8.25537601
##	MBERMIDD	MBERMIDD 5.47438623
##	PBRAND	PBRAND 4.72706545
##	ABRAND	ABRAND 4.01193114
##	MGODGE	MGODGE 4.00433517
##	MOSTYPE	MOSTYPE 3.73848401
##	MINK3045	MINK3045 3.63441353
##	MSKC	MSKC 2.52724460
##	MBERARBG	MBERARBG 2.29299662
##	PWAPART	PWAPART 2.26545077
##	MINKGEM	MINKGEM 2.10526876
##	MGODPR	MGODPR 2.04483723
##	MAUT2	MAUT2 2.02550612
##	PBYSTAND	PBYSTAND 1.94412135
##	MSKA	MSKA 1.88309582
##	MBERHOOG	MBERHOOG 1.83575590
##	MAUT1	MAUT1 1.80484326
##	MSKB1	MSKB1 1.58556667
##	MINK7512	MINK7512 1.36118575
##	MGODRK	MGODRK 1.33371875
##	MHHUUR	MHHUUR 1.18051272
##	MAUT0	MAUT0 1.16464125
##	MGODOV	MGODOV 1.09171875
##	MRELOV	MRELOV 1.00599777
##	APERSAUT	APERSAUT 0.98773862
##	MFWEKIND	MFWEKIND 0.93077028
##	MFGEKIND	MFGEKIND 0.90843790
##	MINKM30	MINKM30 0.90434538
##	PMOTSCO	PMOTSCO 0.87804244
##	MOSH00FD	MOSH00FD 0.87151606
##	MRELGE	MRELGE 0.77114099
##	MBERBOER	MBERBOER 0.72600774
##	MHKOOP	MHKOOP 0.67891044
##	MINK4575	MINK4575 0.57309410
##	MINK123M	MINK123M 0.55367540
##	PLEVEN	PLEVEN 0.42876494
##	MOPLLAAG	MOPLLAAG 0.34831934
##	MRELSA	MRELSA 0.32694573
##	MGEMLEEF	MGEMLEEF 0.31862949
##	MGEMOMV	MGEMOMV 0.28151478
##	MSKB2	MSKB2 0.26920200
##	MOPLMIDD	MOPLMIDD 0.26790510
##	MZPART	MZPART 0.24092339
##	MBERARBO	MBERARBO 0.21376690
##	MFALLEEN	MFALLEEN 0.21136856
##	MZFONDS	MZFONDS 0.19384881
##	MSKD	MSKD 0.19172431
##	MBERZELF	MBERZELF 0.07544956
##	MAANTHUI	MAANTHUI 0.04190195

```

## PWABEDR    PWABEDR    0.00000000
## PWALAND    PWALAND    0.00000000
## PBESAUT    PBESAUT    0.00000000
## PVRAAUT    PVRAAUT    0.00000000
## PAANHANG   PAANHANG   0.00000000
## PTRACTOR   PTRACTOR   0.00000000
## PWERKT     PWERKT     0.00000000
## PBROM      PBROM      0.00000000
## PPERSONG   PPERSONG   0.00000000
## PGEZONG    PGEZONG    0.00000000
## PWAOREG    PWAOREG    0.00000000
## PZEILPL    PZEILPL    0.00000000
## PPLEZIER   PPLEZIER   0.00000000
## PFIETS     PFIETS     0.00000000
## PINBOED    PINBOED    0.00000000
## AWAPART    AWAPART    0.00000000
## AWABEDR    AWABEDR    0.00000000
## AWALAND    AWALAND    0.00000000
## ABESAUT    ABESAUT    0.00000000
## AMOTSCO    AMOTSCO    0.00000000
## AVRAAUT    AVRAAUT    0.00000000
## AAANHANG   AAANHANG   0.00000000
## ATRACTOR   ATRACTOR   0.00000000
## AWERKT     AWERKT     0.00000000
## ABROM      ABROM      0.00000000
## ALEVEN     ALEVEN     0.00000000
## APERSONG   APERSONG   0.00000000
## AGEZONG    AGEZONG    0.00000000
## AWAOREG    AWAOREG    0.00000000
## AZEILPL    AZEILPL    0.00000000
## APLEZIER   APLEZIER   0.00000000
## AFIETS     AFIETS     0.00000000
## AINBOED    AINBOED    0.00000000
## ABYSTAND   ABYSTAND   0.00000000

```

- The most important features for the GBM model are **PPERSAUT** , **MKOOKPLA** and **MOPLHOOG** .

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 compare with the results obtained from applying KNN or logistic regression to this data set?

Boosting

```
##
## predicted_purchase    0    1
##                No  4415  253
##                Yes   118   36
```

Logistic Regression

```
##
## log_reg_predicted_purchase    0    1
##                No  4183  231
##                Yes   350   58
```

Observations :

- Fraction of the people predicted to make a purchase do in fact make one.
- GBM : $34/(34+127) = 0.211$
- Logistic Regression : 0.142

GBM Fares better !

Chapter 10: Question 7

a. Fit a neural network to the Default data. Use a single hidden layer with 10 units, and dropout regularization. Have a look at Labs 10.9.1–10.9.2 for guidance. Compare the classification performance of your model with that of linear logistic regression.

```
## Dimensions of original dataset = 10000 4
```

```
## Dimensions of Training dataset = 7000 4
```

```
## Dimensions of Testing dataset = 3000 4
```



```
## # weights:  51
## initial  value 1373.413610
## iter   10 value 199.929221
## iter   20 value 183.149645
## iter   30 value 169.787081
## iter   40 value 166.256003
## iter   50 value 164.583964
## iter   60 value 163.453985
## iter   70 value 162.582715
## iter   80 value 161.724851
## iter   90 value 161.158318
## iter  100 value 160.849820
## final   value 160.849820
## stopped after 100 iterations
```

```
## Accuracy of Neural Network Model =  0.972
```

```
## Accuracy of Logistic Model =  0.974
```

Problem 1: Beauty Pays!

a. Using the data, estimate the effect of “beauty” into course ratings. Make sure to think about the potential many “other determinants”. Describe your analysis and your conclusions

Correlation Matrix

```
##          CourseEvals BeautyScore      female      lower nonenglish
## CourseEvals  1.00000000  0.40709120 -0.231829451 -0.24864349 -0.079891096
## BeautyScore  0.40709120  1.00000000  0.125719400  0.03257686  0.010293330
## female      -0.23182945  0.12571940  1.000000000 -0.05657933  0.003805072
## lower       -0.24864349  0.03257686 -0.056579333  1.000000000 -0.143448262
## nonenglish  -0.07989110  0.01029333  0.003805072 -0.14344826  1.000000000
## tenuretrack -0.03760944 -0.01913483 -0.074315467 -0.13663972  0.134859291
##          tenuretrack
## CourseEvals -0.03760944
## BeautyScore -0.01913483
## female      -0.07431547
## lower       -0.13663972
## nonenglish  0.13485929
## tenuretrack 1.00000000
```

- According to the correlation matrix, BeautyScore emerges as a crucial factor when estimating course ratings.

- The CourseEvals variable exhibits a positive correlation with the BeautyScore variable, suggesting that higher BeautyScore values are associated with higher CourseEval ratings.
- The impact of the nonenglish and tenuretrack variables on CourseEval appears to be relatively weak.
- Conversely, there is a negative correlation between the Female and lower variables and CourseEvals, indicating that as the values of Female and lower increase, CourseEvals tend to decrease.

Linear Regression model on BeautyScore

```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore, data = beauty_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5936 -0.3346  0.0097  0.3702  1.2321
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.71340    0.02249  165.119  <2e-16 ***
## BeautyScore    0.27148    0.02837   9.569   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4809 on 461 degrees of freedom
## Multiple R-squared:  0.1657, Adjusted R-squared:  0.1639
## F-statistic: 91.57 on 1 and 461 DF,  p-value: < 2.2e-16
```

Linear Regression Model on all variables

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = beauty_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31385 -0.30202  0.01011  0.29815  1.04929
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.06542    0.05145  79.020 < 2e-16 ***
## BeautyScore  0.30415    0.02543  11.959 < 2e-16 ***
## female      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish  -0.25808    0.08478  -3.044 0.00247 **
## tenuretrack -0.09945    0.04888  -2.035 0.04245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 457 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3399
## F-statistic: 48.58 on 5 and 457 DF,  p-value: < 2.2e-16
```

- As we introduce more predictors in multiple regression, the Adjusted R-squared increases, indicating an improved fit of the data.
- The multiple variable model takes into account all relevant factors, providing a comprehensive analysis. Thus, using the more inclusive model for predictions is recommended.
 - BeautyScore has a positive coefficient, indicating professors with higher beauty scores tend to get better course ratings.
- Female professors show a negative coefficient suggesting they tend to receive lower course ratings.
- Lower-division course instructors also have a negative coefficient, implying lower course ratings.
- Non-English speaking professors receive lower course ratings with a negative coefficient, suggesting language barriers impacting effective teaching.
- Professors on the tenure track receive lower course ratings.

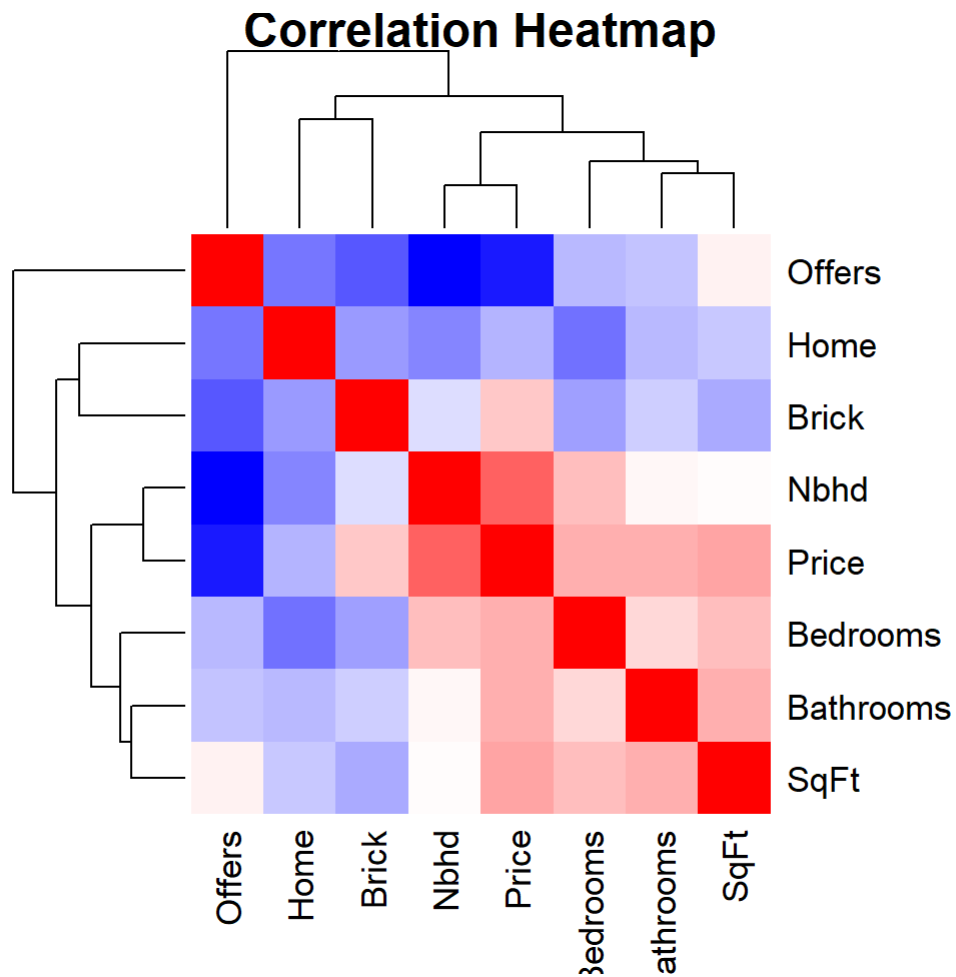
b. In his paper, Dr. Hamermesh has the following sentence: “Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible”. Using the concepts we have talked about so far, what does he mean by that?

- The statement “Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible,” highlights the challenge of understanding certain outcomes in scientific fields.
- Specifically, it questions whether a higher beauty score truly reflects a professor’s teaching ability or if it’s merely a bias on the students’ part, perceiving more conventionally attractive professors as better.
- However, this model lacks enough factors to provide a definitive answer. Since it cannot access students’ thoughts, determining the true reason behind the outcome is difficult.

Problem 2: Housing Price Structure

a. Is there a premium for brick houses everything else being equal?

Simple linear regression model with all variables



```
##
## Call:
## lm(formula = Price ~ ., data = city)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24940.6  -8383.0   430.7   7430.4  31371.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9814.663    9858.884  -0.996   0.32149
## Home           6.187       28.973   0.214   0.83128
## Nbhd          9832.281    1821.869   5.397 3.47e-07 ***
## Offers       -8351.794    1267.428  -6.590 1.24e-09 ***
## SqFt          49.811       6.769   7.359 2.53e-11 ***
## Brick       15601.818    2261.896   6.898 2.66e-10 ***
## Bedrooms     5671.911    1840.979   3.081  0.00256 **
## Bathrooms    8243.545    2449.897   3.365  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11540 on 120 degrees of freedom
## Multiple R-squared:  0.8256, Adjusted R-squared:  0.8154
## F-statistic: 81.15 on 7 and 120 DF,  p-value: < 2.2e-16
```

- The coefficient of Brick variables shows that it significantly impacts the Price of the house since its highly positive with a value of 15601.818 . This indicates a high premium for brick houses everything else being equal.
- Moreover, the confidence interval for the coefficient of Brick ranges from 13373.88702 to 21220.81203, with all values above zero, confirming that there is a positive premium associated with brick houses.

b. Is there a premium for houses in neighborhood 3?

```
##
## Call:
## lm(formula = Price ~ ., data = city[c("neigh3", "Offers", "SqFt",
##   "Brick", "Bedrooms", "Bathrooms", "Price")])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26810.5  -5953.6   -266.5   5662.9  26793.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3067.471    8746.712   0.351 0.726423
## neigh3      21937.572    2482.393   8.837 9.39e-15 ***
## Offers      -8019.003    1013.011  -7.916 1.32e-12 ***
## SqFt         52.149       5.572    9.359 5.44e-16 ***
## Brick       17058.771    1942.805   8.780 1.28e-14 ***
## Bedrooms    4070.005    1570.921   2.591 0.010751 *
## Bathrooms   7810.698    2109.060   3.703 0.000322 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9995 on 121 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8616
## F-statistic: 132.8 on 6 and 121 DF,  p-value: < 2.2e-16
```

- There is a positive correlation between the price of houses and Neighborhood 3 with coefficient of 21937.572, indicating that houses in Neighborhood 3 have a significant premium.
- The confidence interval for the coefficient of Brick ranges from 17017.713 to 26857.431, all of which are above zero and entirely positive. This strengthens the conclusion that there is a premium for houses made of brick.

c. Is there an extra premium for brick houses in neighborhood 3?

Regression Model with Brick and Nbhd3 variable

```
##
## Call:
## lm(formula = Price ~ . + Brick * neigh3, data = city[c("neigh3",
##      "Offers", "SqFt", "Brick", "Bedrooms", "Bathrooms", "Price")])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26710.2  -5797.0   -277.9   4337.6  26483.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3407.487   8559.390    0.398  0.69126
## neigh3       17709.779   2949.399    6.005 2.10e-08 ***
## Offers        -8298.791    997.359   -8.321 1.60e-13 ***
## SqFt           53.728     5.488    9.790 < 2e-16 ***
## Brick        13664.535   2327.643    5.871 3.95e-08 ***
## Bedrooms      4652.044   1554.250    2.993  0.00335 **
## Bathrooms     6407.659   2137.027    2.998  0.00330 **
## neigh3:Brick 10361.809   4100.564    2.527  0.01281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9780 on 120 degrees of freedom
## Multiple R-squared:  0.8748, Adjusted R-squared:  0.8675
## F-statistic: 119.8 on 7 and 120 DF,  p-value: < 2.2e-16
```

- Brick houses in neighborhood 3 command a substantial premium of approximately 10361.809
- The coefficient estimate for neighborhood 3 is 20681.037, with a confidence interval ranging from 625.392 to 18798.225. The positive interval reinforces the presence of a premium for houses in neighborhood 3

d. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single “older” neighborhood?

```
##
## Call:
## lm(formula = Price ~ ., data = city[c("neigh12", "Offers", "SqFt",
##   "Brick", "Bedrooms", "Bathrooms", "Price")])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26810.5  -5953.6   -266.5   5662.9  26793.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25005.043   9658.403    2.589 0.010807 *
## neigh12     -21937.572   2482.393   -8.837 9.39e-15 ***
## Offers       -8019.003   1013.011   -7.916 1.32e-12 ***
## SqFt          52.149     5.572     9.359 5.44e-16 ***
## Brick        17058.771   1942.805    8.780 1.28e-14 ***
## Bedrooms     4070.005   1570.921    2.591 0.010751 *
## Bathrooms    7810.698   2109.060    3.703 0.000322 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9995 on 121 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8616
## F-statistic: 132.8 on 6 and 121 DF,  p-value: < 2.2e-16
```

- In older neighborhood 1 and 2 the price of the house drops significantly as the coefficient is highly negative at -21937.
- Since the premium is negative for neighborhoods 1 and 2 these can be considered “old” neighborhoods.

Problem 3: What causes what??

a. Why can't I just get data from a few different cities and run the regression of “Crime” on “Police” to understand how more cops in the streets affect crime

- When using data from different cities and conducting a regression of “Crime” on “Police,” the relationship between the presence of more police officers and crime rates may not be straightforward.
- Observing a correlation between the number of police officers and crime rates in various cities does not automatically imply a direct causal relationship.
- A positive correlation might be misinterpreted as more police causing higher crime rates, but it could actually be the other way around, with higher crime rates leading to the hiring of more police officers in response to the situation.
- Data analysis and correlations can provide valuable insights, but they do not directly establish causation.

b. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the “Table 2” below.

- The researchers from UPENN utilized a natural experiment, where random assignment is possible, to ensure equal probability of participants being assigned to specific groups in the experiment.
- They focused on crime rate figures in Washington, D.C., specifically on high-alert days with an increased risk of terrorist attacks.
- City officials had no choice but to station more police on high-alert days to protect the public, regardless of the city’s crime rate.
- The results from Table 2 indicate that having more police in a city leads to a lower crime rate, as evidenced by the negative coefficient.
- Thus, it can be reasonably concluded that there is reduced crime on high-alert days when more police are present.

c. Why did they have to control for METRO ridership? What was that trying to capture?

- Data on police and crime presents challenges in determining causation: it cannot distinguish between more police leading to increased crime or vice versa.
- Observing a positive correlation between police presence and crime across cities is possible, as mayors may react to crime by hiring more police.
- Conducting a controlled experiment with random police placement is practically impossible. UPENN researchers found a natural experiment in DC, analyzing crime data on high-alert days with increased police presence.
- Controlling for ridership, high-alert days showed lower crime rates, supporting the impact of more police on crime reduction.
- Proving the definitive relationship between more police and less crime remains challenging, but the findings provide strong circumstantial evidence in favor of the connection.

d. In the next page, I am showing you “Table 4” from the research paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

- Table 4’s model considered the impact of placing more police on high-alert days in different locations across the city.
- Interaction terms, High Alert X District 1 and High Alert X Other Districts, were included to analyze the relationship between high-alert days and district locations.
- The effect of more police was most pronounced in District 1, leading to the greatest decrease in crime rate, which is logical as densely populated districts are more vulnerable to attacks.
- This aligns with the logical assumption that District 1, with more potential terrorist targets in DC, receives increased police deployment during high-alert periods, resulting in a significant crime reduction.
- Other districts also showed a negative effect, with some reduction in crime rate due to more police on high-alert days, but the effect was smaller and not statistically significant.

Problem 5: Final Project

Describe your contribution to the final group project

- Discussed about the potential problem statements to solve and presented multiple datasets to use and finalize a problem statement and ultimately arrived at the conclusion to target a healthcare problem. The original dataset we had selected was not very voluminous. So I searched Kaggle datasets which could be alternatively used for our diabetes prediction model and found a dataset which was usable and checked the data quality to kickstart our work. However we found out that it was required to be balanced as it was highly skewed toward the majority (non diabetic) class . So I applied the 'oversampling' method to balance the data while using random sampling to increased the ratio of target population from 16% to 33% in the dataset.
- I checked irregularities in the data through correlation matrix and correlation heatmap. Then , I ran multiple iterations of code process for the model development exercise to include the KNN and Boosting modeling techniques using scikit learn module while integrating feature importance , accuracy and recall value patches. I also included GridSearch CV to add hyperparameter tuning and optimized the code to make it run faster through loops , saving and reloading models through the pickle module.
- Added plots to better demonstrate and visual the code in action using matplotlib and seaborn libraries. Discussed the findings and learnings with my teammates and created my portion of the presentation which included model development and analysis of different modelling techniques. I proactively coordinated with my teammates to make sure they were on the same page with every step of the process and suggested insights to include in their parts of presentation and write-up while incorporating theirs.