
# STA 380 Take Home Exam , Book Verion -2

Viswa Tej Seela

31 July 2022

# Contents

# Chapter 2

### Question 10

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

**Q10-1 To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library.**

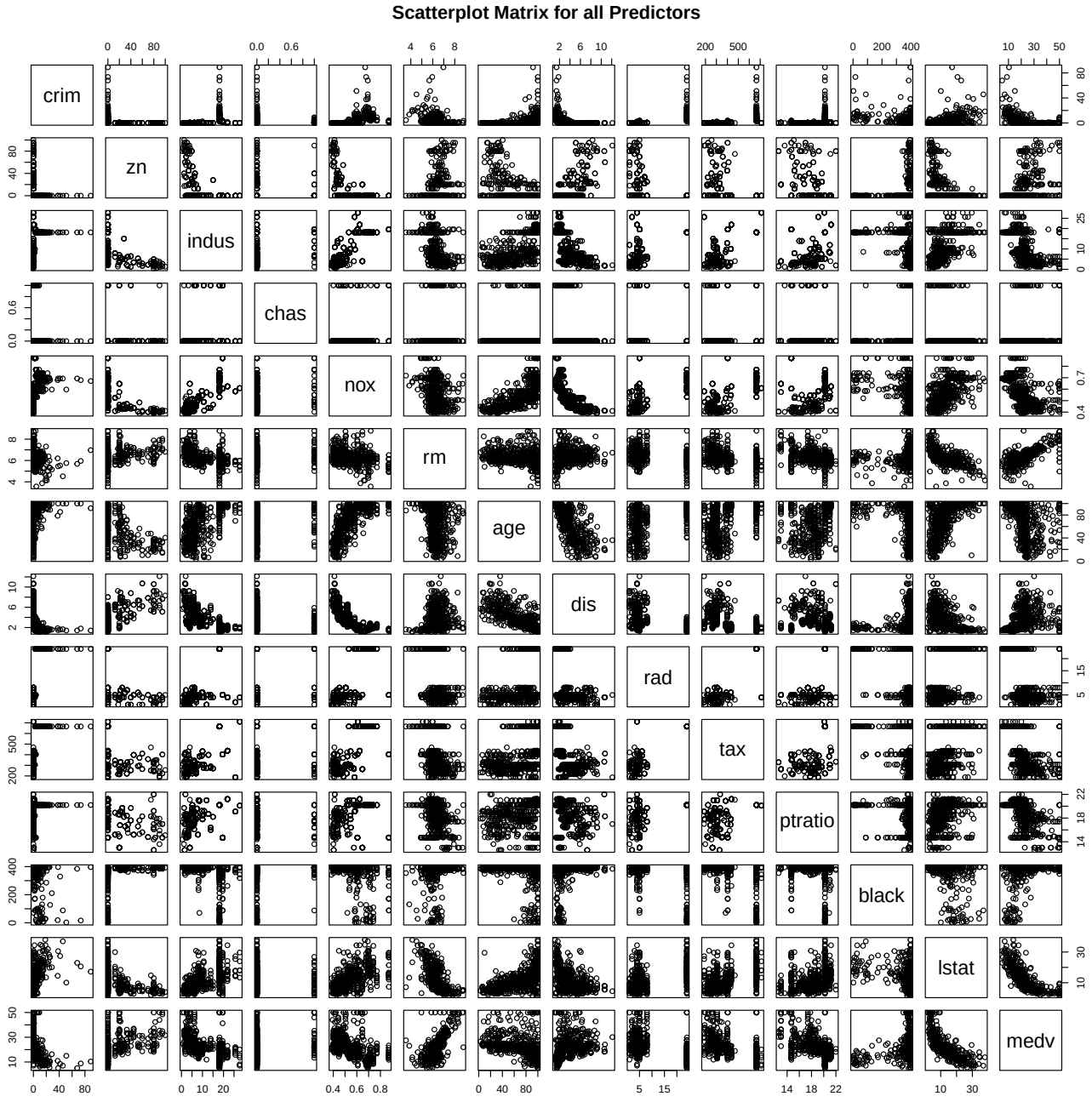
```
## [1] 506  14
```

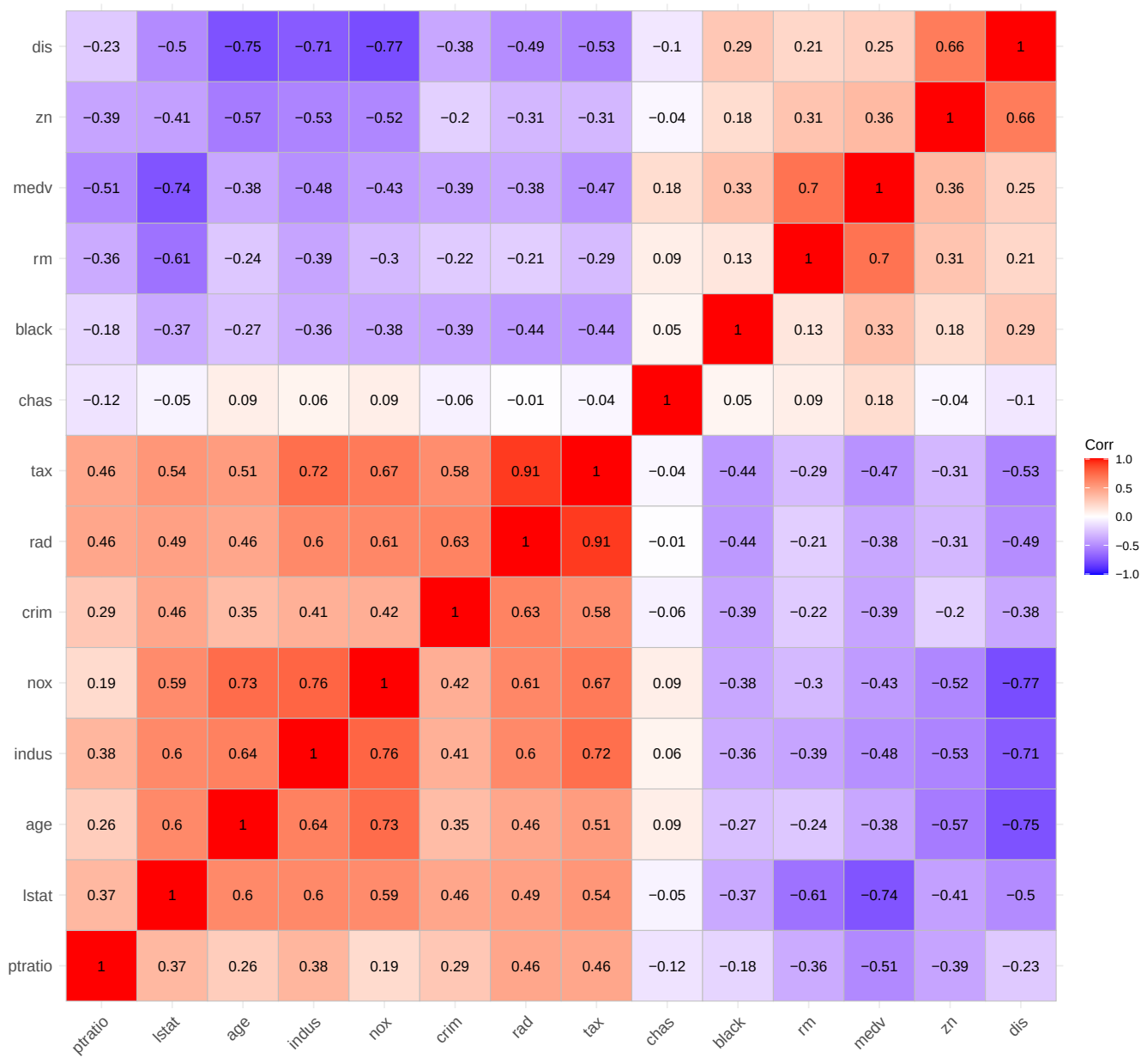
- 506 rows and 13 Columns , where row represent the set of predictor observations for a given Neighborhood in Boston and each column represents a Predictor Variable

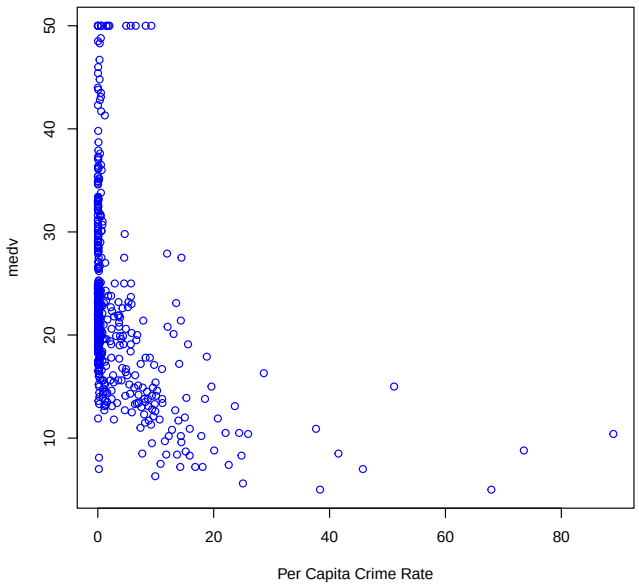
**Q10-2 Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**

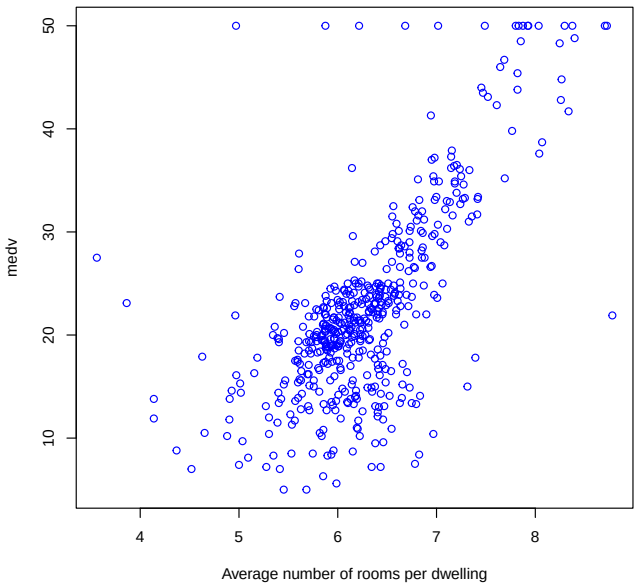
```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  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          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20
```

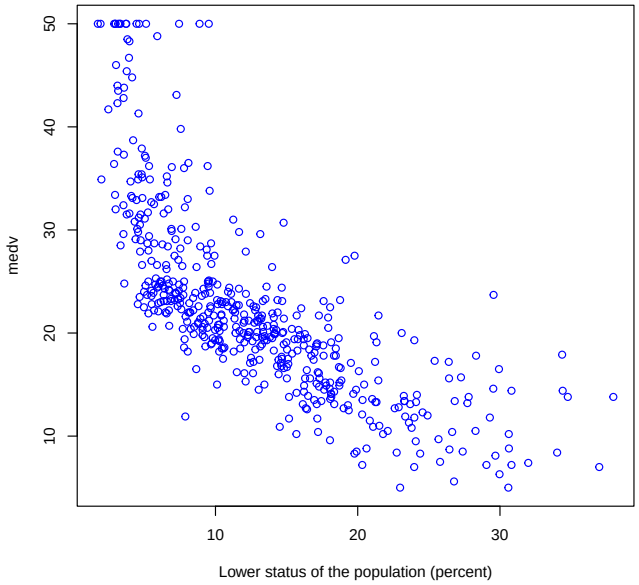
```
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

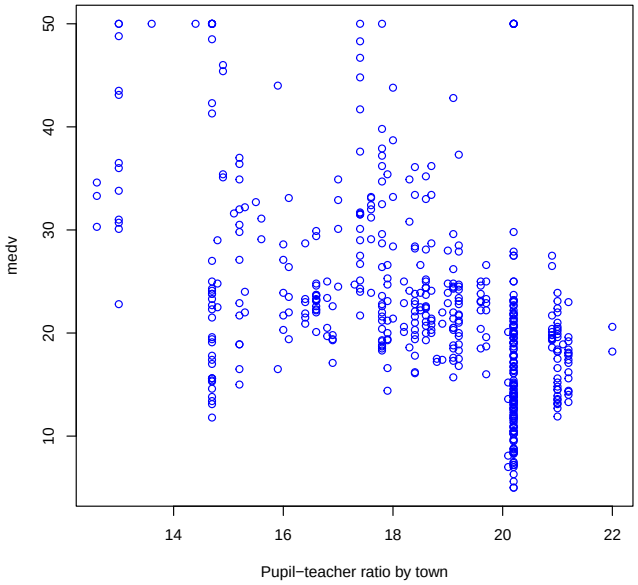



dis −0.23 −0.5 −0.75 −0.71 −0.77 −0.38 −0.49 −0.53 −0.1 0.29 0.21 0.25 0.66 1
zn −0.39 −0.41 −0.57 −0.53 −0.52 −0.2 −0.31 −0.31 −0.04 0.18 0.31 0.36 1 0.66
medv −0.51 −0.74 −0.38 −0.48 −0.43 −0.39 −0.38 −0.47 0.18 0.33 0.7 1 0.36 0.25
rm −0.36 −0.61 −0.24 −0.39 −0.3 −0.22 −0.21 −0.29 0.09 0.13 1 0.7 0.31 0.21
black −0.18 −0.37 −0.27 −0.36 −0.38 −0.39 −0.44 −0.44 0.05 1 0.13 0.33 0.18 0.29
chas −0.12 −0.05 0.09 0.06 0.09 −0.06 −0.01 −0.04 1 0.05 0.09 0.18 −0.04 −0.1
tax 0.46 0.54 0.51 0.72 0.67 0.58 0.91 1 −0.04 −0.44 −0.29 −0.47 −0.31 −0.53
rad 0.46 0.49 0.46 0.6 0.61 0.63 1 0.91 −0.01 −0.44 −0.21 −0.38 −0.31 −0.49
crim 0.29 0.46 0.35 0.41 0.42 1 0.63 0.58 −0.06 −0.39 −0.22 −0.39 −0.2 −0.38
nox 0.19 0.59 0.73 0.76 1 0.42 0.61 0.67 0.09 −0.38 −0.3 −0.43 −0.52 −0.77
indus 0.38 0.6 0.64 1 0.76 0.41 0.6 0.72 0.06 −0.36 −0.39 −0.48 −0.53 −0.71
age 0.26 0.6 1 0.64 0.73 0.35 0.46 0.51 0.09 −0.27 −0.24 −0.38 −0.57 −0.75
lstat 0.37 1 0.6 0.6 0.59 0.46 0.49 0.54 −0.05 −0.37 −0.61 −0.74 −0.41 −0.5
ptratio 1 0.37 0.26 0.38 0.19 0.29 0.46 0.46 −0.12 −0.18 −0.36 −0.51 −0.39 −0.23
ptratio lstat age indus nox crim rad tax chas black rm medv zn dis
Corr
1.0
0.5
0.0
−0.5
−1.0



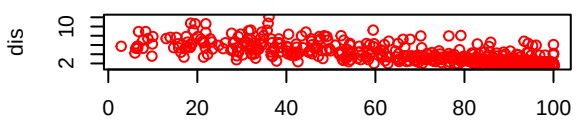



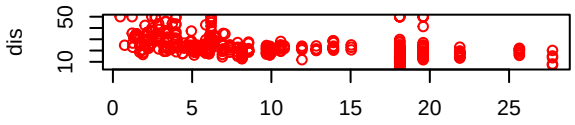

**Findings:**

- There is correlation between variables so first plotting a correlation matrix and then will deepdive into the relation between two variables
- Even though there are other varibles which have a high collinearity , i have chosen to demonstrate MEDV and DIS
- As crim increases, the medv decreases. demand for homes in more dangerous areas leads to a devaluation in the price of homes there.
- As rm increases, the medv also increases,homes with more square footage/space are valued higher than homes with less space.

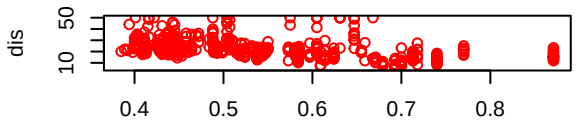
- As lstat increases, the medv decreases, s the lower status of population (percent) increases, the median value of owner-occupied homes drops.



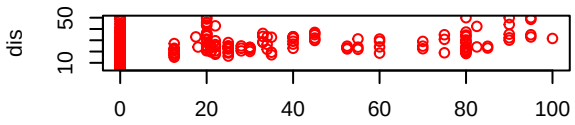

Proportion of owner−occupied units built prior to 1940



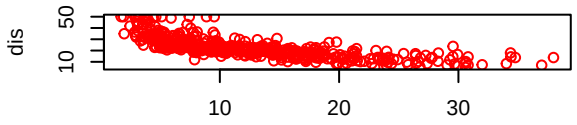

proportion of non−retail business acres per town




nitrogen oxides concentration (parts per 10 million)

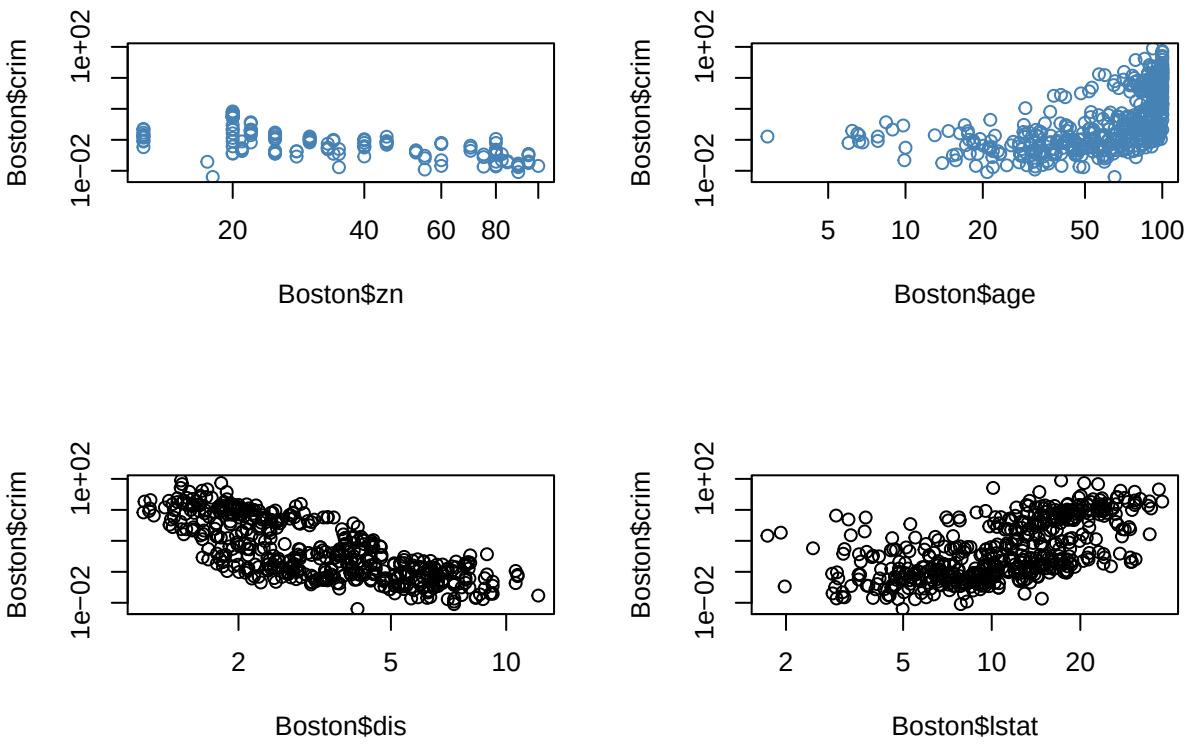



proportion of residential land zoned for lots over 25,000 sq.ft




lower status of the population (percent)

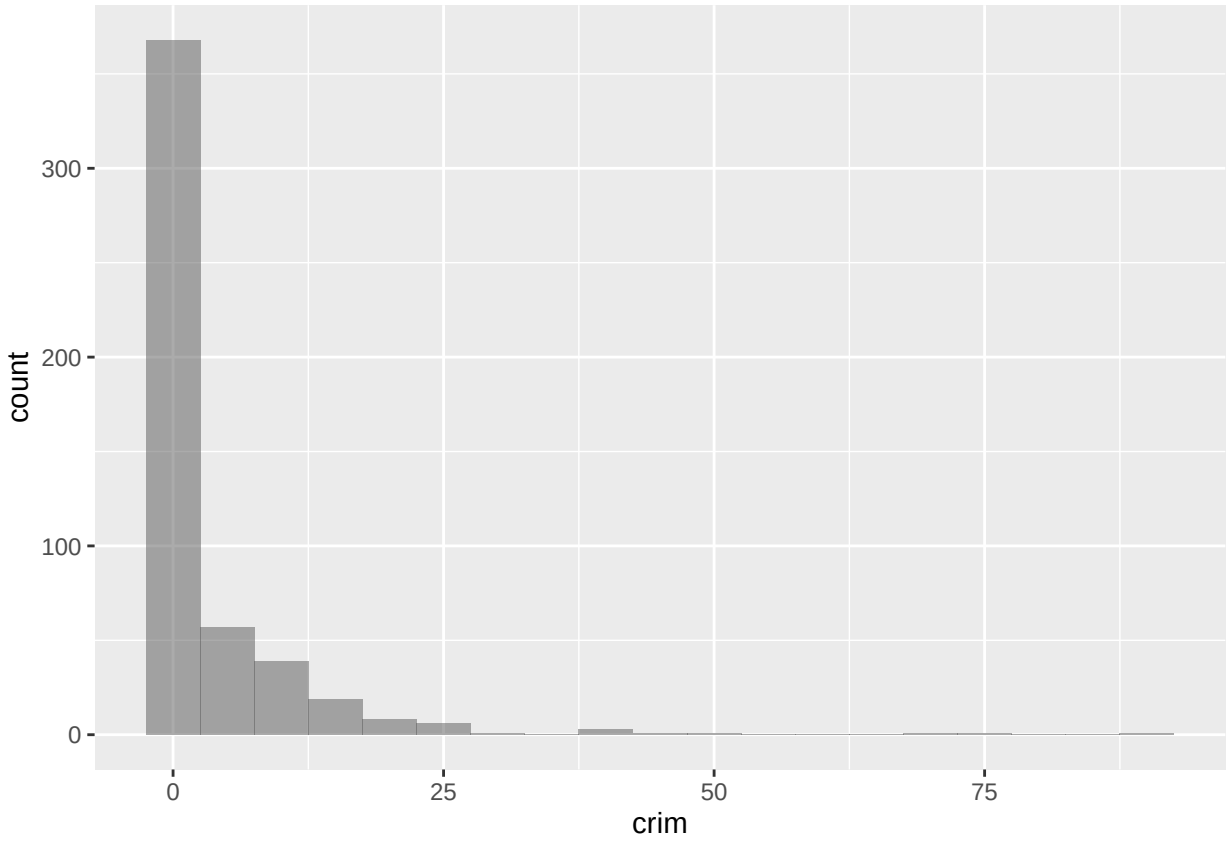
**Q10-3 Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**



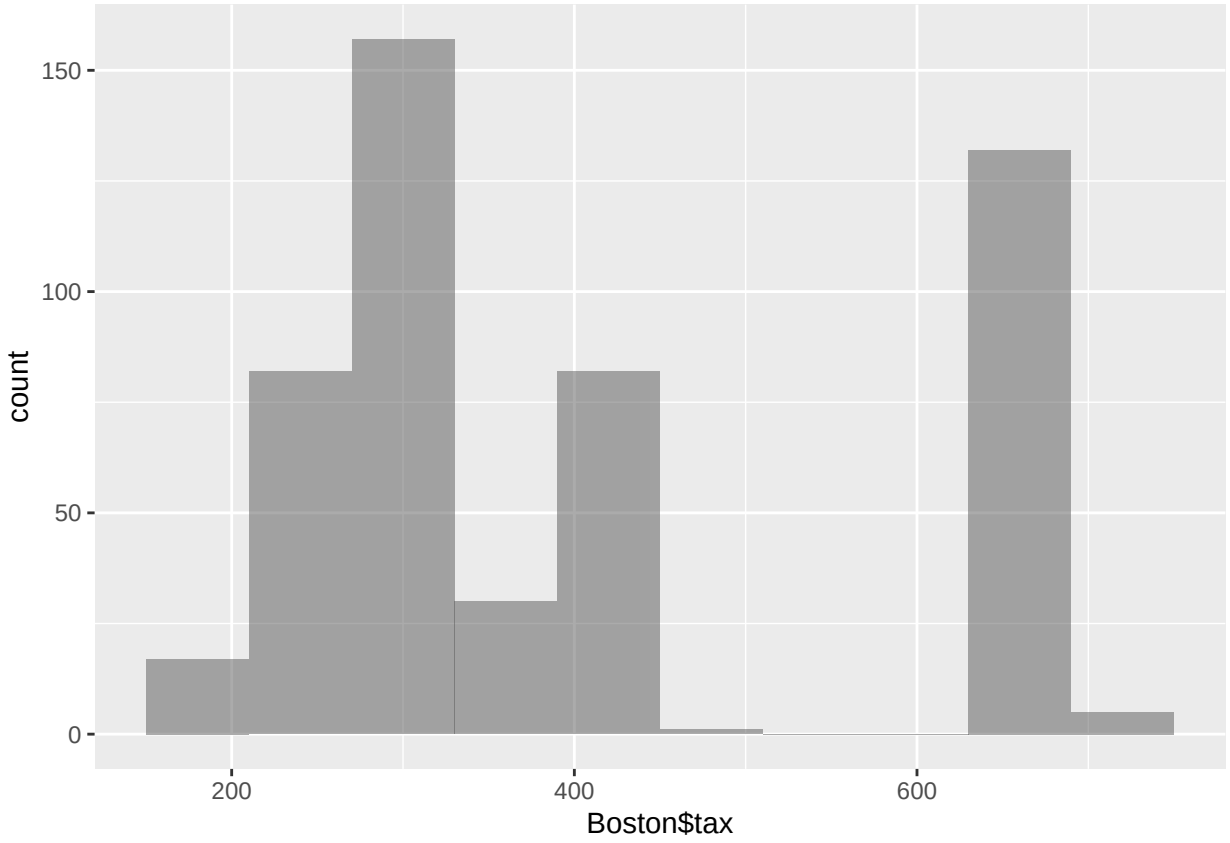

**Findings**

- As the proportion of owner-occupied units built prior to 1940 increases, the Per Capita Crime Rate increases.
- As the weighted mean of distances to five Boston employment centres increases, the Per Capita Crime Rate decreases.
- As the lower status of the population (percent) increases, the Per Capita Crime Rate increases.

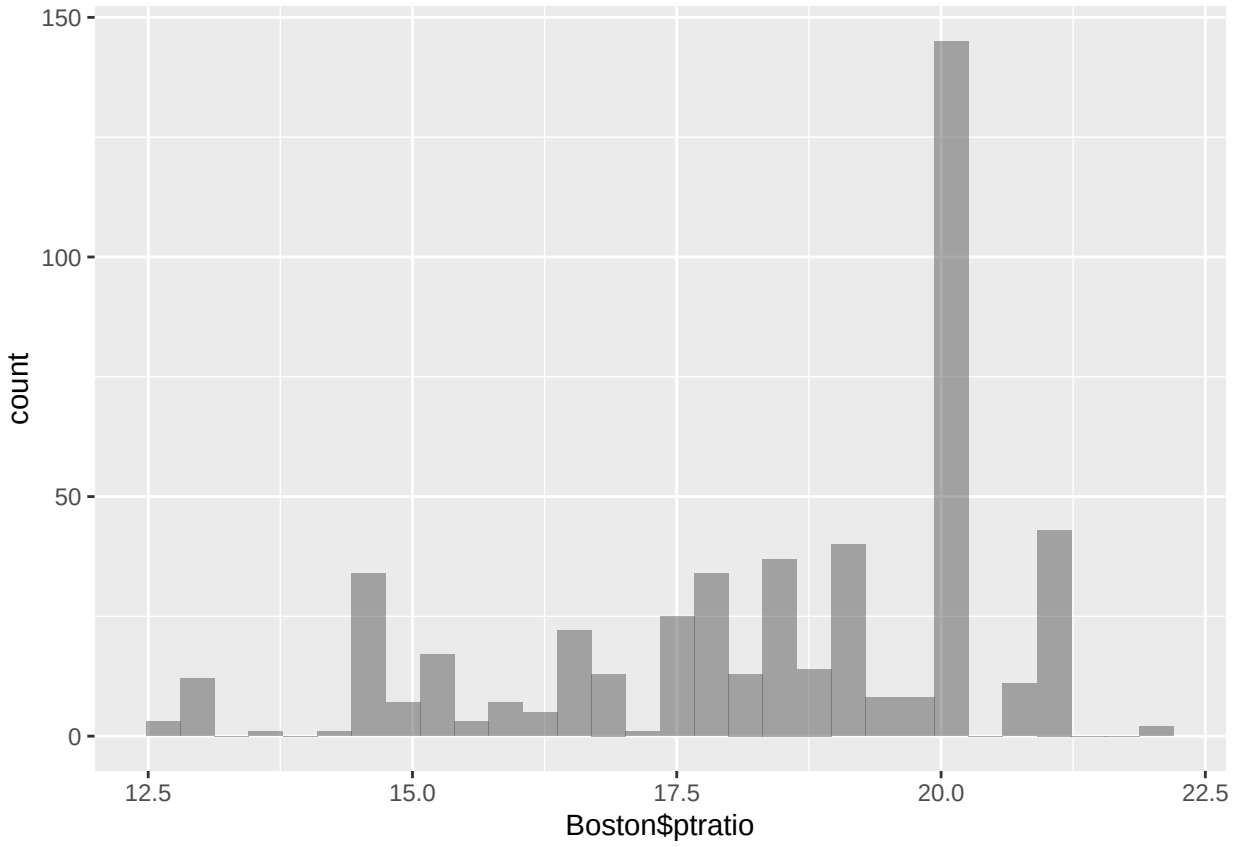
**Q10-4 Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**




- There are very few Boston suburbs with high crime rates. Majority of observations align with a zero Per Capita Crime Rate.




- High count of observations that align with a near 700 value on the x-axis.A moderate count of observations that fall within the range of 200-400 values on the x-axis.




- There is a spike approximate to 20 on the x-axis that indicates a higher frequency of observations.

**Q10-5 How many of the census tracts in this data set bound the Charles river?**

```
## [1] 35
```

**Q10-6 What is the median pupil-teacher ratio among the towns in this data set?**

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

**Q10-7 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.**

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
##     medv
## 399    5
```

```
##       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          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

- 399 suburb has lowest medv with medv 0f 5 ie. 5000$
- More Crime, Less Rooms per Dwelling, Low Status of the Population (Percent), and Low Median Value of Owner Homes.

**Q10-8 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**

```
## [1] 64
```

```
## [1] 13
```

```
##       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          black
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7
```

```
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9
##      lstat           medv
##  Min.   :2.47   Min.   :21.9
##  1st Qu.:3.32   1st Qu.:41.7
##  Median :4.14   Median :48.3
##  Mean   :4.31   Mean   :44.2
##  3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :7.44   Max.   :50.0
```

```
##           used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells 2422919 129.4    4715940 251.9         NA  4715940 251.9
## Vcells 4156283  31.8    8388608  64.0      16384  8388568  64.0
```

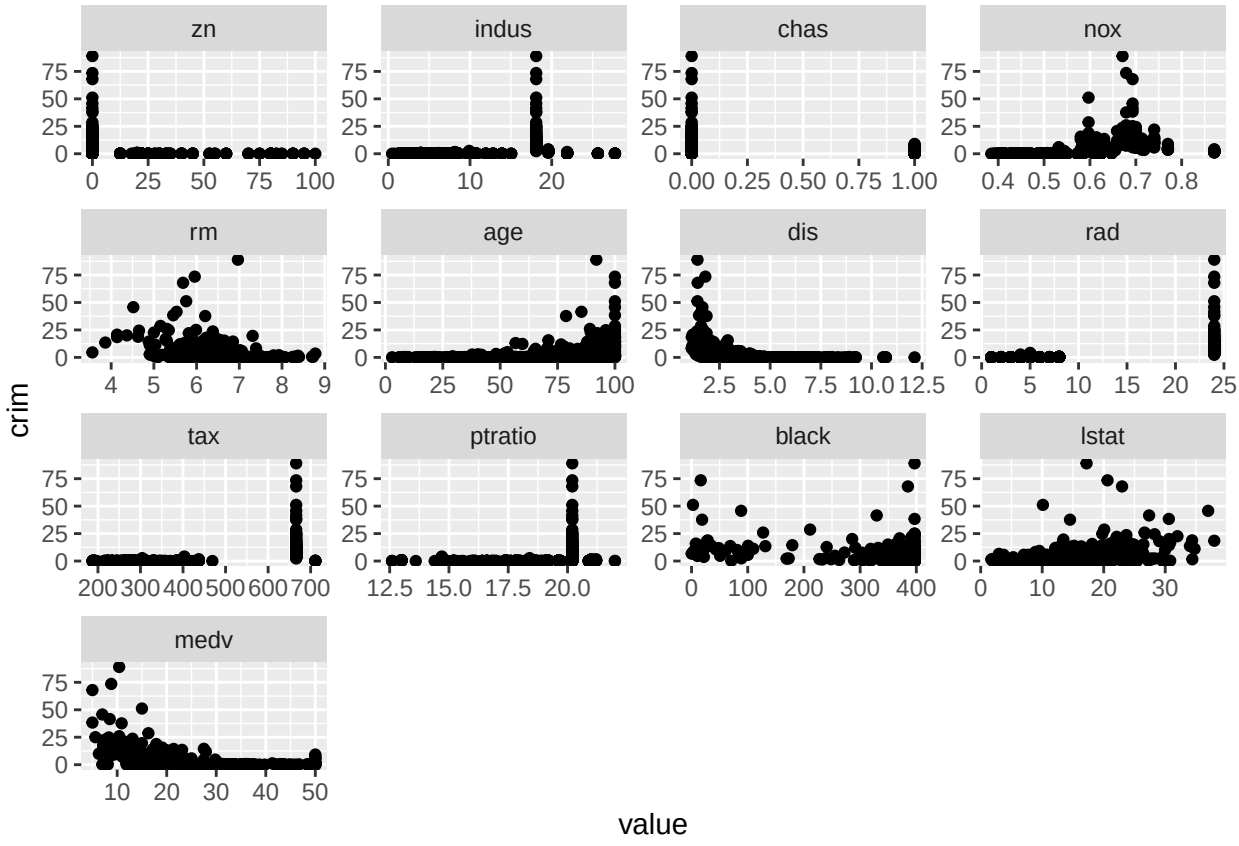
**Findings**

- 64 room have more than 7 rooms per dwelling
- 13 rooms have more than 8 dwelling
- Average Per Capita Crime Rate is quite low at 0.71879. (below Boston average)
- The typical percentage of owner-occupied housing constructed before 1940 is 71.54.
- has a typical owner-occupied home price of $44,000 on average (above the Boston average)
- has a 4.31 percent average lower status of the population (below the Boston average)

# Chapter 3

## Question 15

**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.**




```
##        crim          zn       indus        chas         nox          rm
##  1.00000000 -0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670
##         age         dis         rad         tax     ptratio       black
##  0.35273425 -0.37967009  0.62550515  0.58276431  0.28994558 -0.38506394
##       lstat        medv
##  0.45562148 -0.38830461
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston[, x], na.action = na.omit)
##
## 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 ***
## Boston[, x] -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
```

```
##        .id      X..i..
## 1       zn -0.07393498
## 2    indus  0.50977633
## 3     chas -1.89277655
## 4      nox 31.24853120
## 5       rm -2.68405122
## 6      age  0.10778623
## 7      dis -1.55090168
## 8      rad  0.61791093
## 9      tax  0.02974225
## 10 ptratio  1.15198279
## 11   black -0.03627964
## 12   lstat  0.54880478
## 13    medv -0.36315992
```

**Findings**:

- Crime rate is positively influenced by rad and tax variables
- Fitting the linear regression model for each predictor
- rad and tax predictors explain the maximum variability in the response variable with adjusted R^2 closer to 35%
- All the predictors are statistically significant in explaining the response variable based on their p-value which is less than 0.05 except for chas

**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 : j = 0?**

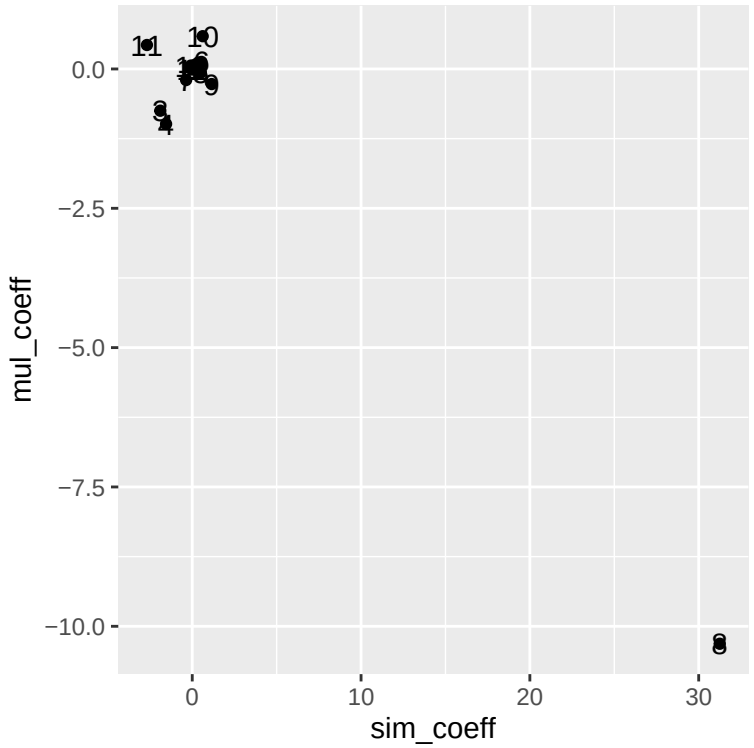
```
##
## Call:
## lm(formula = Boston$crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            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
```

**Findings**:

- All the predictors combined can successfully explain close to 45% of the variation in the response variable
- rad,dis,black,medv,zn have relatively high t-value suggesting that they have a good influence on the response variable
- Null hypothesis can be rejected for rad,dis,black,medv,zn

**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 regres- sion model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.**




- It looks like a few of the uni coefficient points change drastically. For example `nox` goes from +31 in the univariate case to -10 in the multivariate case.

**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 $Y = 0 + 1X + 2X2 + 3X3 + .$**

```
##
## Call:
## lm(formula = Boston$crim ~ Boston[, x] + I(Boston[, x]^2) + I(Boston[,
##     x]^3))
##
## 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 ***
## Boston[, x]      -5.0948305  0.4338321 -11.744  < 2e-16 ***
## I(Boston[, x]^2)  0.1554965  0.0171904   9.046  < 2e-16 ***
## I(Boston[, x]^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
```

- nox,rm,age,dis and medv have a non linear relationship with crim variable as adjusted R2 has increased by fitting a non-linear model

# Chapter 6

## Question 9

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```
## [1] 18
```

### (a) Split the data set into a training set and a test set

```
## [1] 582
```

```
## [1] 195
```

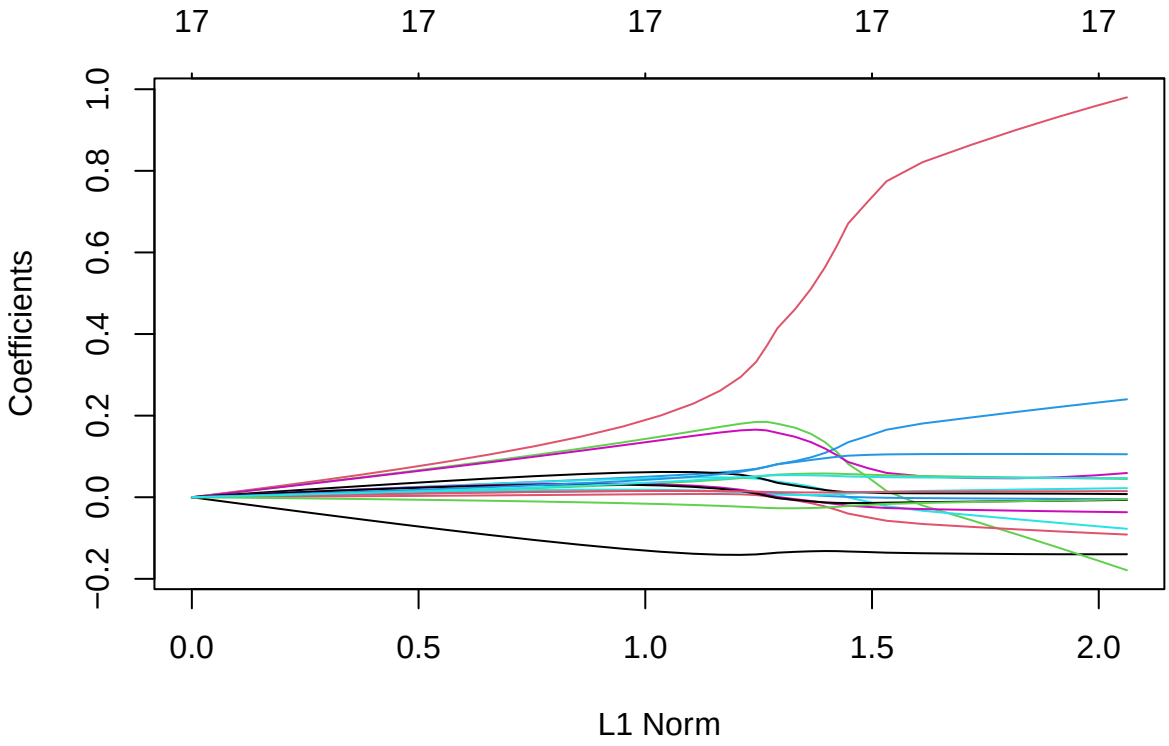
### (b) Fit a linear model using least squares on the training set, and report the test error obtained.

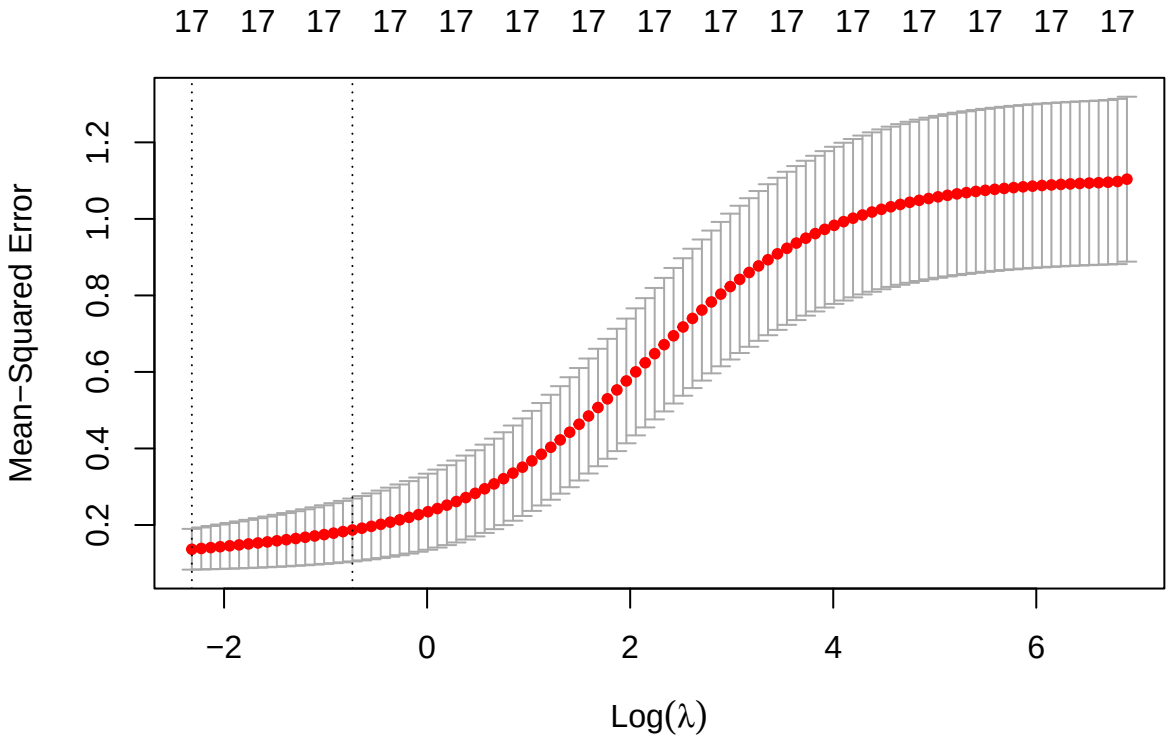
```
## [1] 0.04654965
```

- The mean squared error or MSE is 0.0465496, which is extremely large.

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

- With lambda = grid we will implement a ridge regression over a a grid ov values ranging from 10^10 to 10^-2. This way we cover the full range of scenarios from the null model containing only the intercept, to the least squares fit # When alpha = 0 we fit a ridge regression


17
17
17
17
17
Coefficients
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
2.0
L1 Norm


17 17 17 17 17 17 17 17 17 17 17 17 17 17 17
Mean−Squared Error
1.2
1.0
0.8
0.6
0.4
0.2
−2
0
2
4
6
Log(λ)

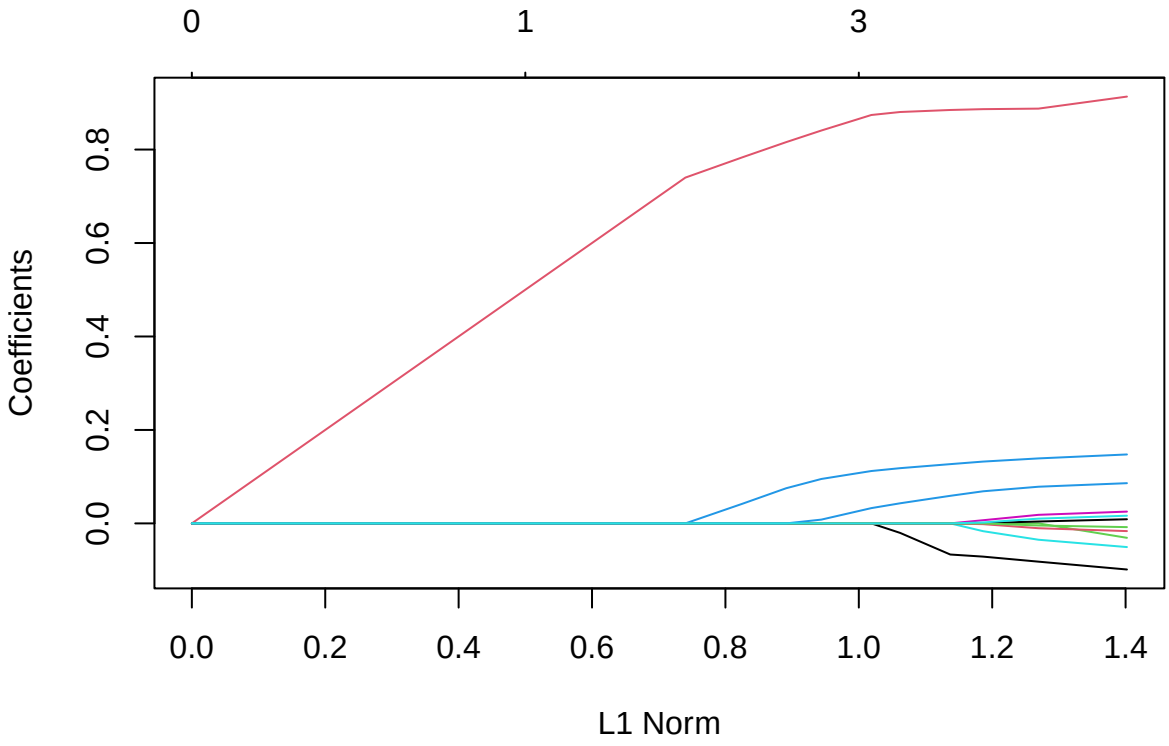
```
## [1] 0.09854043
```

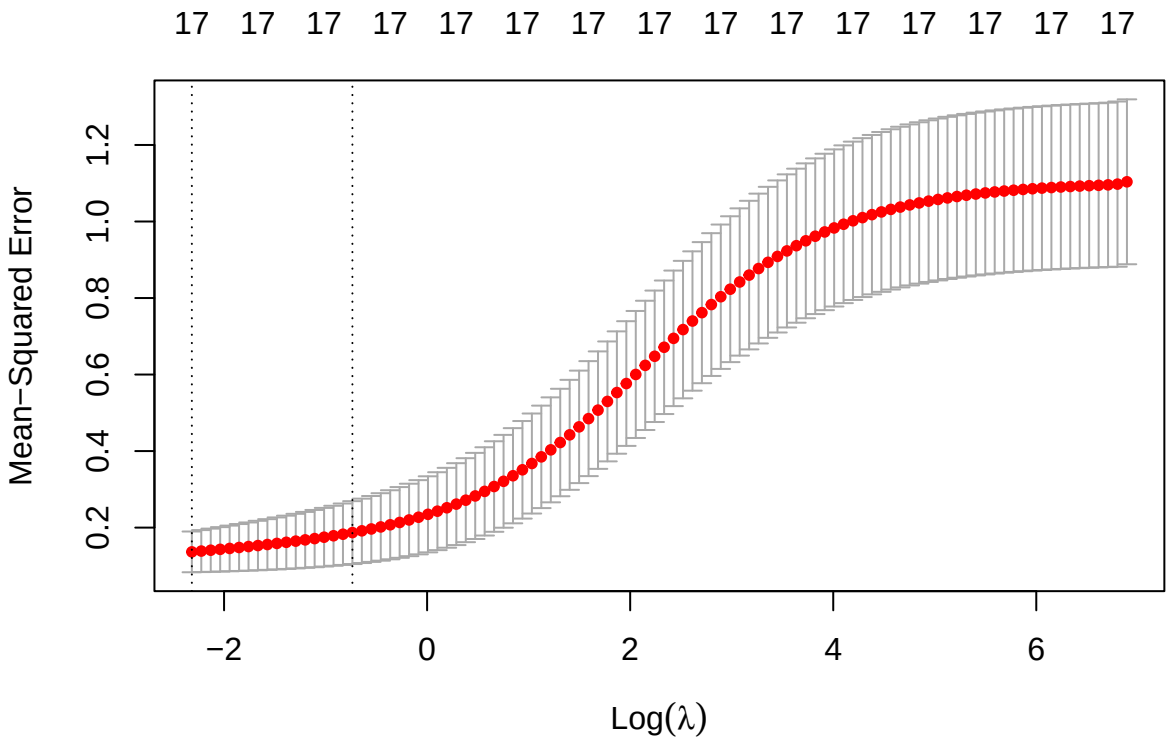
- From above, we see that the value of  that results in the smallest cross validation error for ridge regression is 0.0985404.

```
## [1] 0.0344969
```

- The test MSE is 0.0344969.

**(d) Fit a lasso model on the training set, with  chosen by cross- validation. Report the test error obtained, along with the num- ber of non-zero coefficient estimates.**





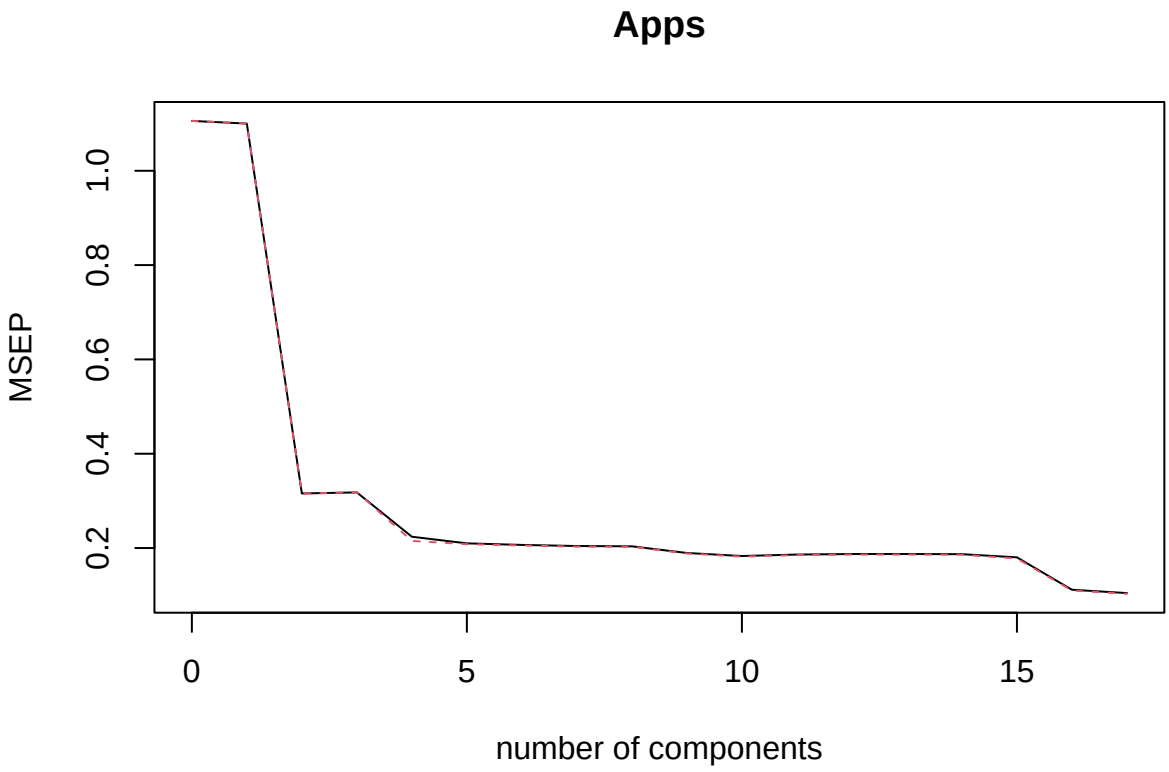

```
## [1] 0.0005258806
```

- From above, we see that the value of  that results in the smallest cross validation error for lasso is $5.2588062 \times 10^{-4}$.

```
## [1] 0.03496165
```

- The test MSE for lasso is 0.0349617.

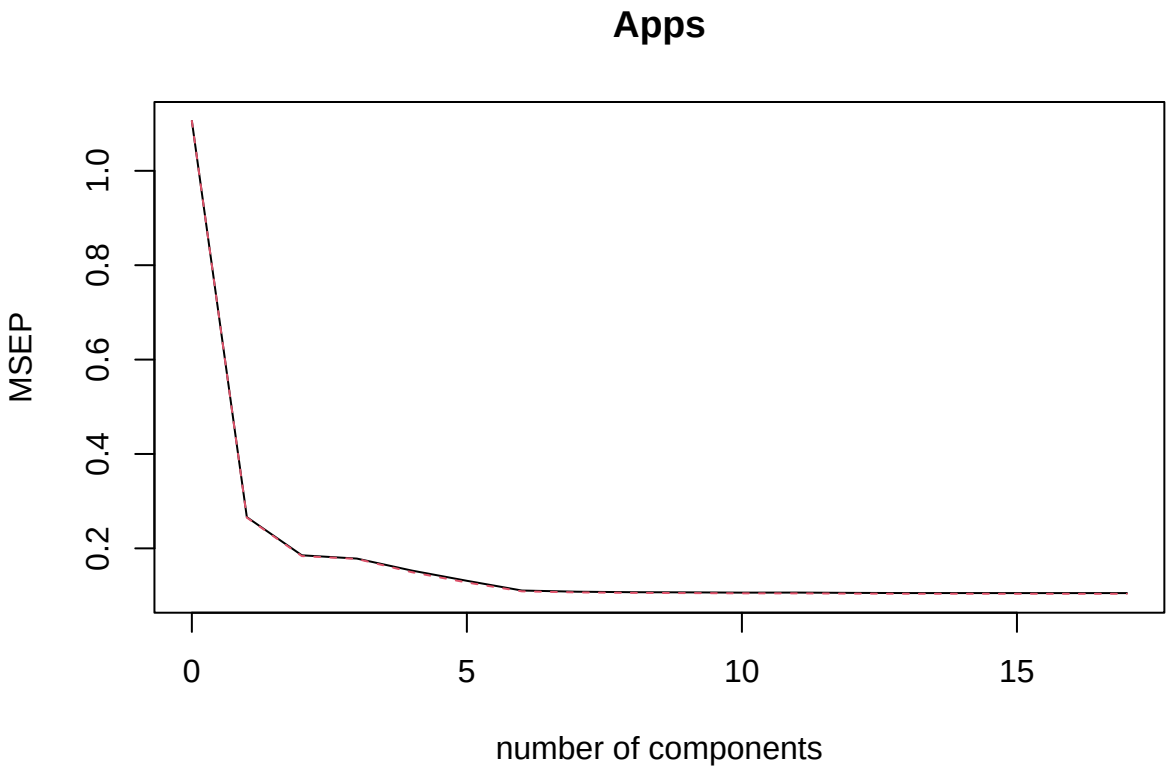
**(e) Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.**




```
## [1] NA
```

- The test MSE for PCR is $NA$ and the M values is 16 according to the graph

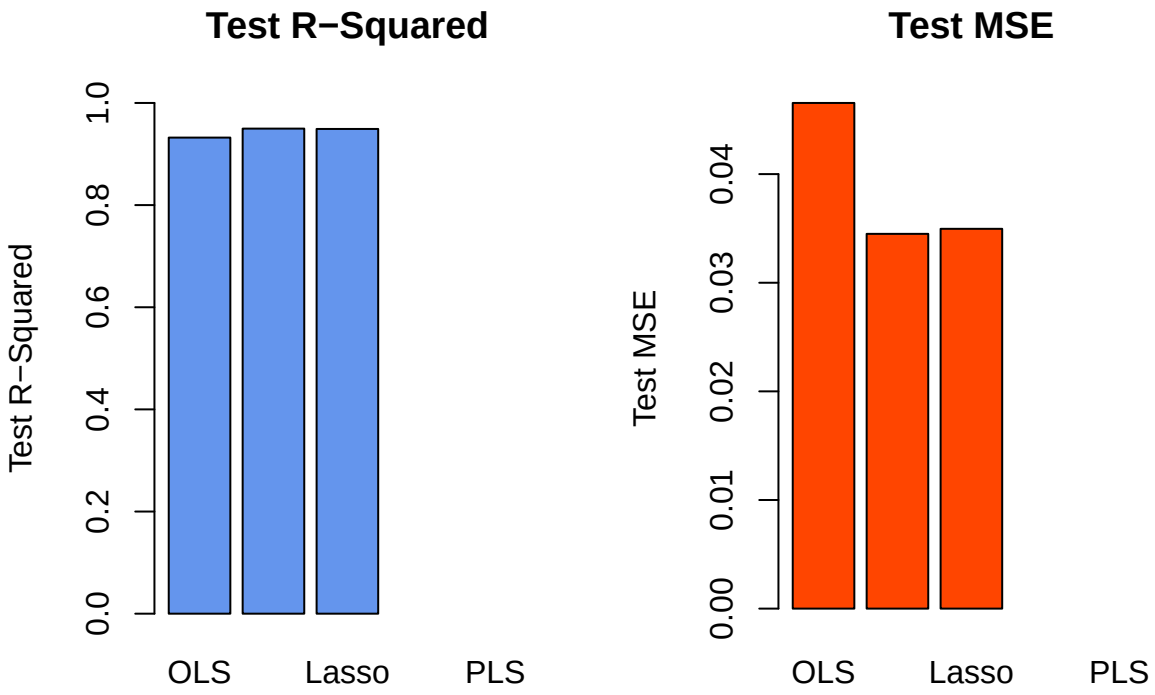
**(f) Fit a PLS model on the training set, with M chosen by cross- validation. Report the test error obtained, along with the value of M selected by cross-validation.**




```
## [1] NA
```

- The test MSE for PLS is $NA$.

**(g) 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?**




## Question 11

**(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.**

**Least squares**

```
##
## Call:
## lm(formula = crim ~ ., data = Boston, subset = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.262 -2.330 -0.452  1.065 73.765
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.835e+01  8.809e+00   2.083 0.038043 *
## zn           5.297e-02  2.463e-02   2.150 0.032242 *
## indus       -4.926e-02  1.045e-01  -0.471 0.637746
## chas        -5.041e-01  1.526e+00  -0.330 0.741398
## nox         -1.106e+01  6.734e+00  -1.643 0.101354
## rm           3.983e-01  7.073e-01   0.563 0.573711
## age          4.431e-03  2.129e-02   0.208 0.835250
```

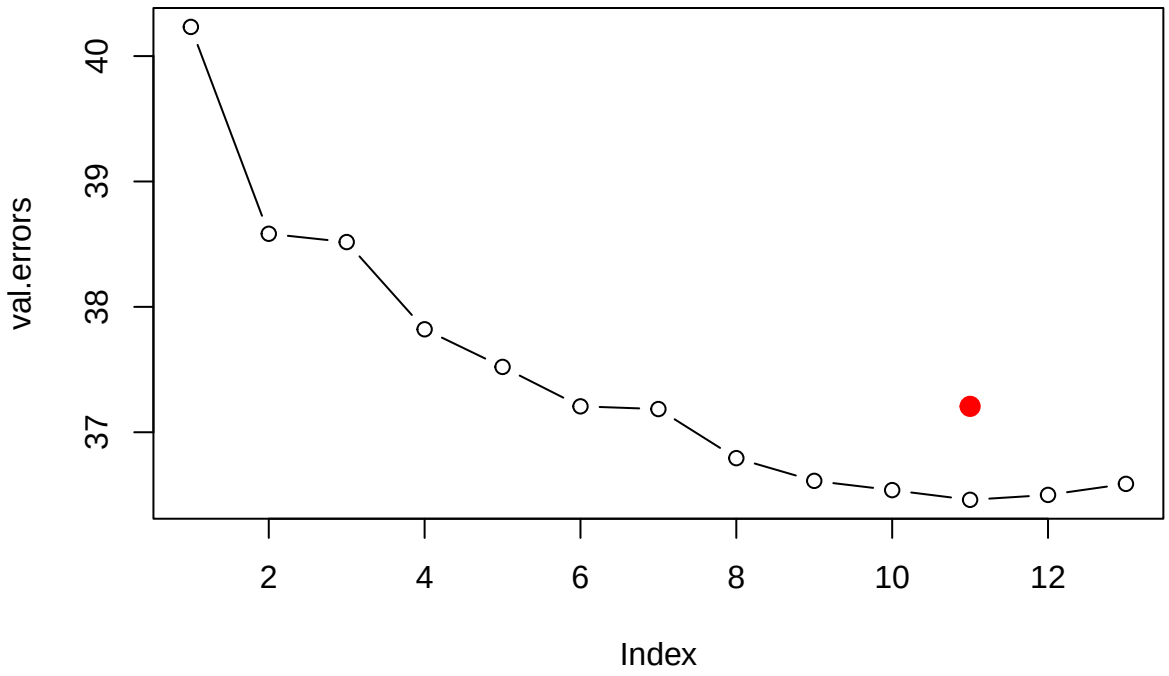
```
## dis         -1.139e+00  3.429e-01  -3.322 0.000991 ***
## rad          6.346e-01  1.112e-01   5.707 2.51e-08 ***
## tax         -4.745e-03  6.555e-03  -0.724 0.469591
## ptratio     -3.449e-01  2.375e-01  -1.453 0.147257
## black        2.775e-04  4.356e-03   0.064 0.949251
## lstat        6.032e-02  9.378e-02   0.643 0.520535
## medv        -2.537e-01  7.344e-02  -3.455 0.000620 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.66 on 340 degrees of freedom
## Multiple R-squared:  0.451,  Adjusted R-squared:  0.43
## F-statistic: 21.48 on 13 and 340 DF,  p-value: < 2.2e-16
```

- The variables that are statistically significant are zn, dis, rad, lstat, and medv.

```
## [1] 36.58822
```

**Best Subset Selection**

```
## [1] 11
```

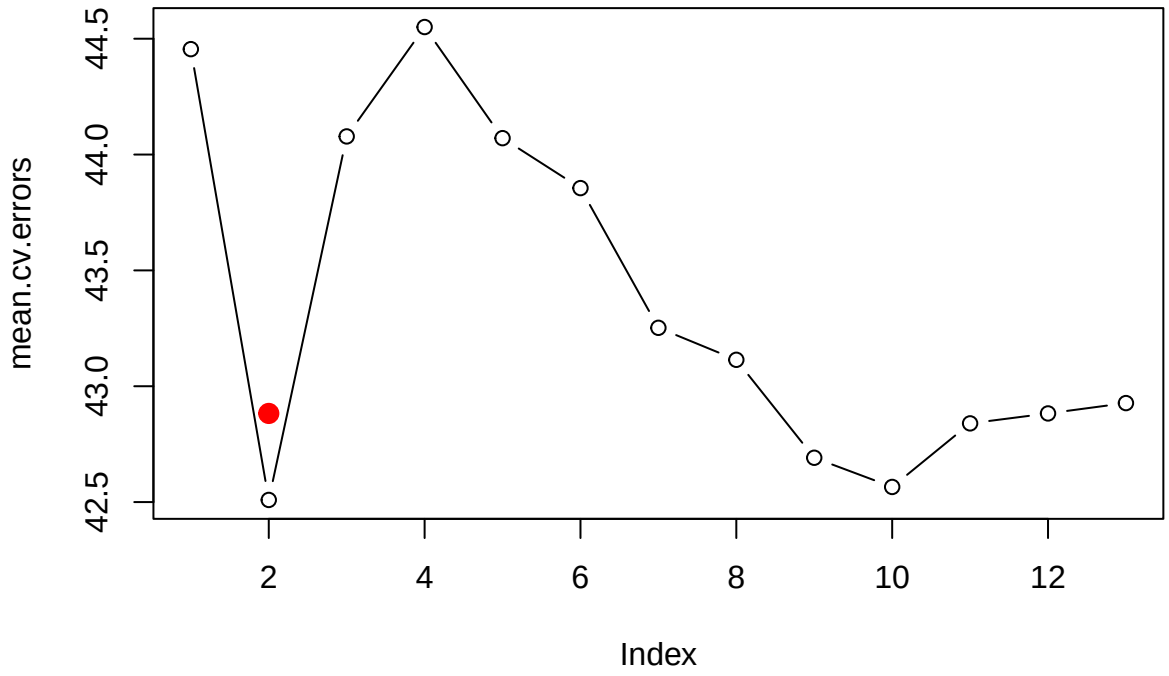



```
##  (Intercept)          zn       indus        chas         nox          rm
##  18.32940256  0.05241230 -0.05004619 -0.50032015 -10.69889459  0.41343928
##          dis         rad         tax     ptratio       lstat        medv
##  -1.15812542  0.63186884 -0.00471570 -0.33882587  0.06594473 -0.25205326
```

```
## [1] 37.20655
```

```
##  (Intercept)           zn          nox          dis          rad        black
## 14.642639407  0.053963088 -9.238768232 -0.992810697  0.499838443 -0.008710565
##         medv
## -0.195989936
```

```
##        1        2        3        4        5        6        7        8
## 44.45496 42.50923 44.07822 44.55039 44.07080 43.85521 43.25223 43.11403
##        9       10       11       12       13
## 42.69129 42.56528 42.83967 42.88256 42.92730
```

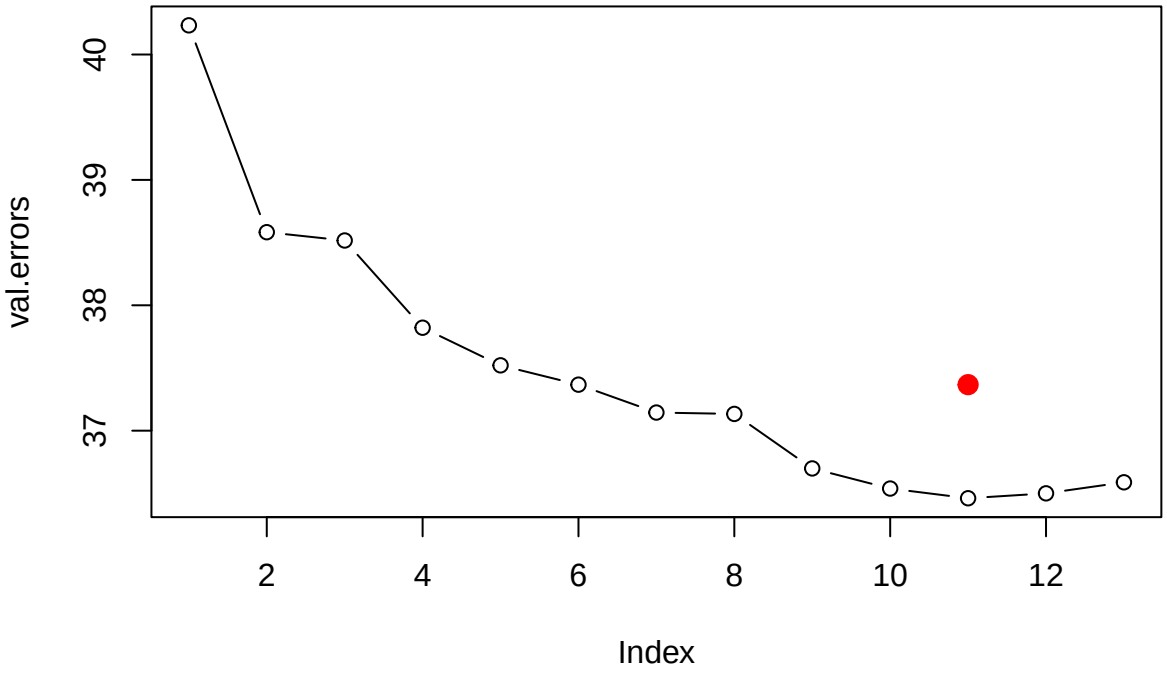



```
##   (Intercept)            zn         indus          chas           nox
##  16.985713928   0.044673247  -0.063848469  -0.744367726 -10.202169211
##            rm           dis           rad           tax       ptratio
##   0.439588002  -0.993556631   0.587660185  -0.003767546  -0.269948860
##         black         lstat          medv
##  -0.007518904   0.128120290  -0.198877768
```

```
##       12
## 42.88256
```

**Forward Stepwise**

```
## [1] 11
```



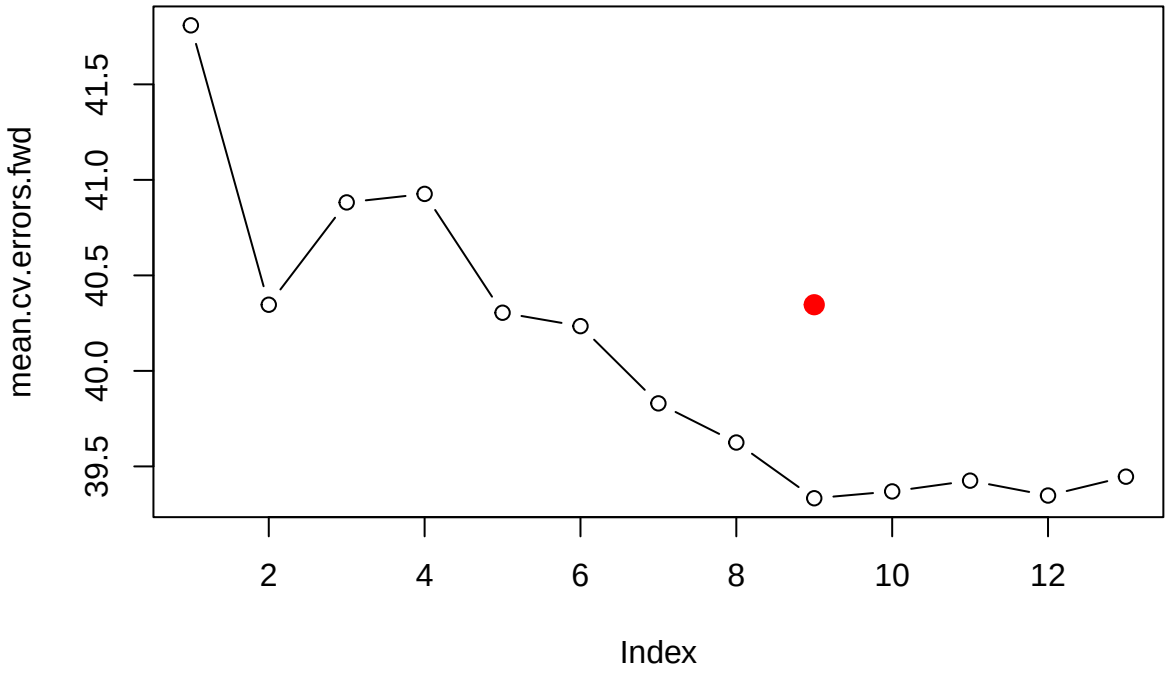

```
## [1] 37.36707
```

```
##  (Intercept)           zn          nox          dis          rad        black
## 14.642639407  0.053963088 -9.238768232 -0.992810697  0.499838443 -0.008710565
##         medv
## -0.195989936
```

- The models chosen by best subset selection, and forward selection using the validation approach chose the same variables - zn, nox, dis, rad, black, and medv. The test errors for both are the same.

```
##        1        2        3        4        5        6        7        8
## 41.80889 40.34632 40.88195 40.92660 40.30454 40.23431 39.82998 39.62513
##        9       10       11       12       13
## 39.33337 39.36912 39.42545 39.34757 39.44623
```

```
## 9
## 9
```

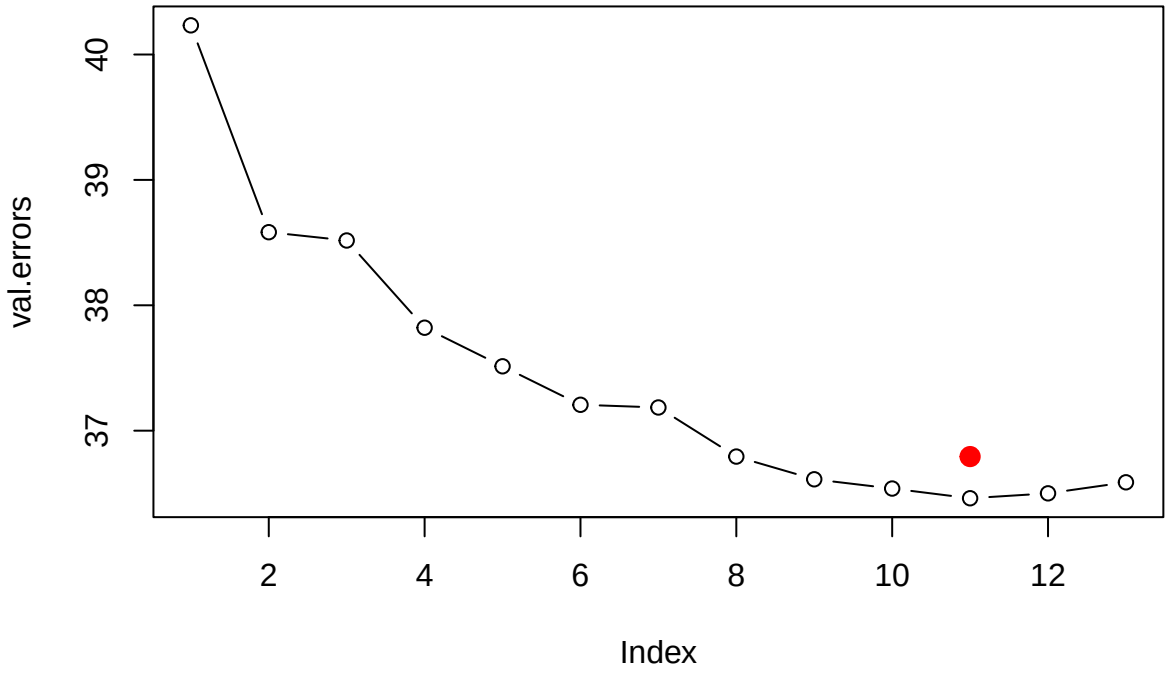



```
## (Intercept)         rad       lstat
##  -4.3814053   0.5228128   0.2372846
```

```
##        2
## 40.34632
```

**Backward Stepwise**

```
## [1] 11
```




```
## [1] 36.79335
```

```
##   (Intercept)            zn           nox           dis           rad
##  19.683127801   0.043293393 -12.753707757  -0.918318253   0.532616533
##       ptratio         black         lstat          medv
##  -0.310540942  -0.007922426   0.110173124  -0.174207166
```

```
## [1] 37.20655
```

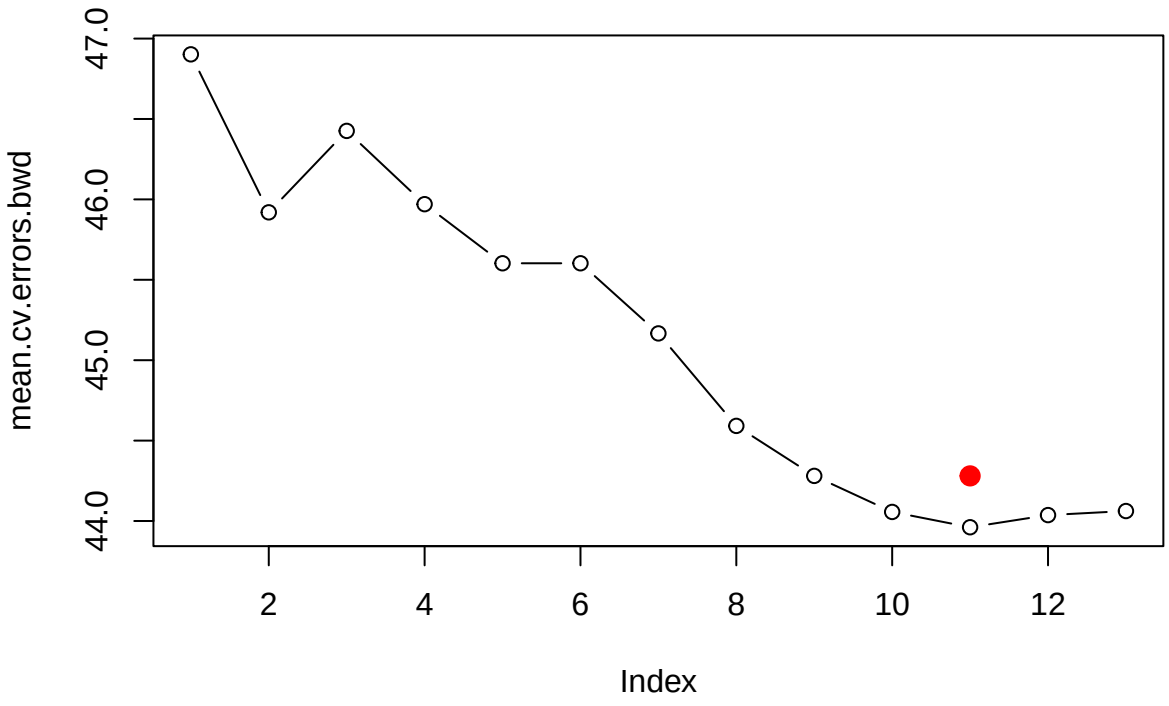
```
## [1] 37.36707
```

```
## [1] 36.79335
```

- The backward selection model has the lowest test MSE when the model is selected using the validation set approach. It uses 2 more variables than best subset and forward selection - ptratio and lstat.

```
##        1        2        3        4        5        6        7        8
## 46.90193 45.91947 46.42590 45.97024 45.60269 45.60279 45.16658 44.59142
##        9       10       11       12       13
## 44.28048 44.05640 43.96122 44.03678 44.06203
```

```
## 11
## 11
```

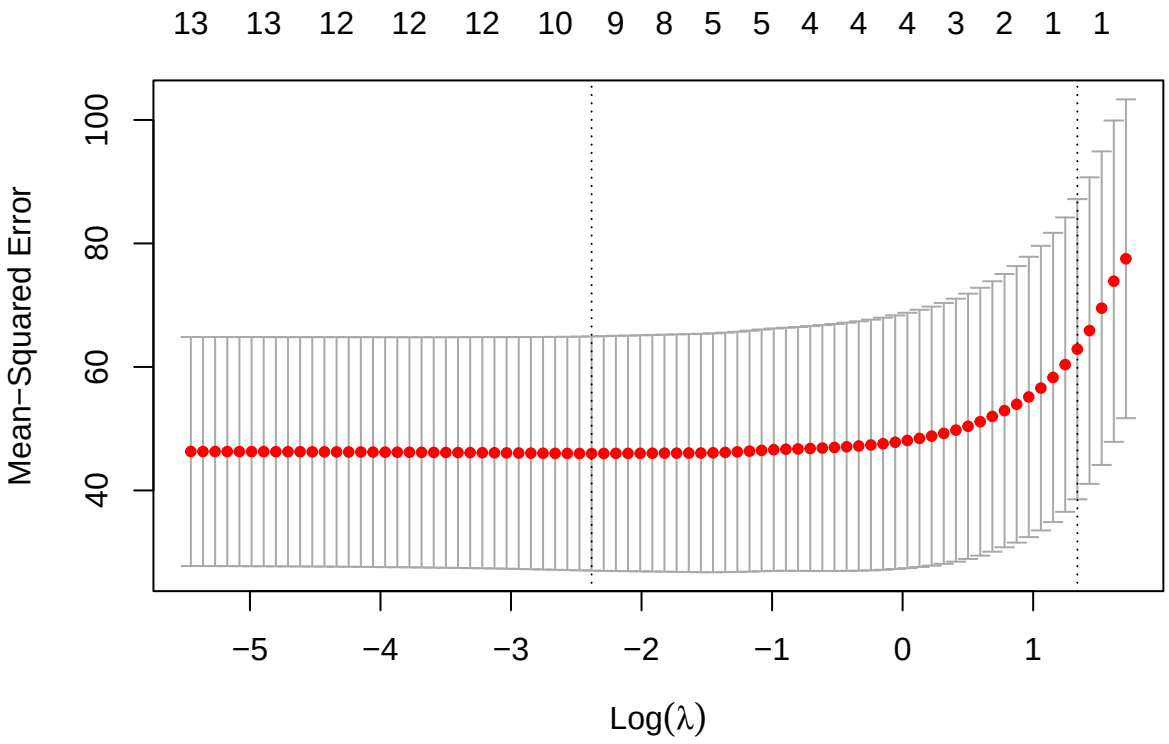



```
##   (Intercept)            zn         indus           nox           dis
##  19.124636156   0.042788127  -0.099385948 -10.466490364  -1.002597606
##           rad       ptratio         black         lstat          medv
##   0.539503547  -0.270835584  -0.008003761   0.117805932  -0.180593877
```

```
##        9
## 44.28048
```

- The backward stepwise selection model selected by cross validation has even lower test MSE.

**Lasso regression**




```
## [1] 0.09230933
```

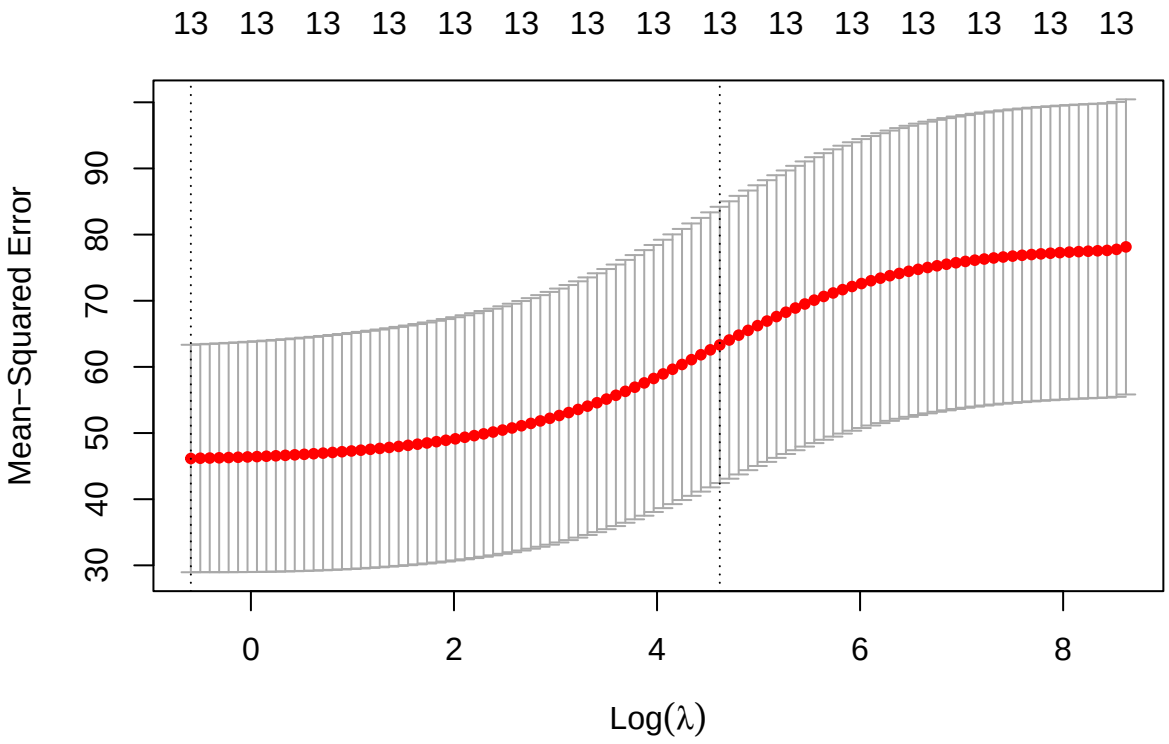
```
## [1] 36.72976
```

```
## [1] 10
```

```
## (Intercept)          zn       indus        chas         nox          rm
## 10.32161882  0.03839405 -0.05329463 -0.32855104 -3.88224471  0.05107485
##         dis         rad     ptratio       lstat
## -0.77863119  0.52502171 -0.17337006  0.05381851
```

- Lasso regression zeroed out all the variables except for the intercept.

**Ridge regression**



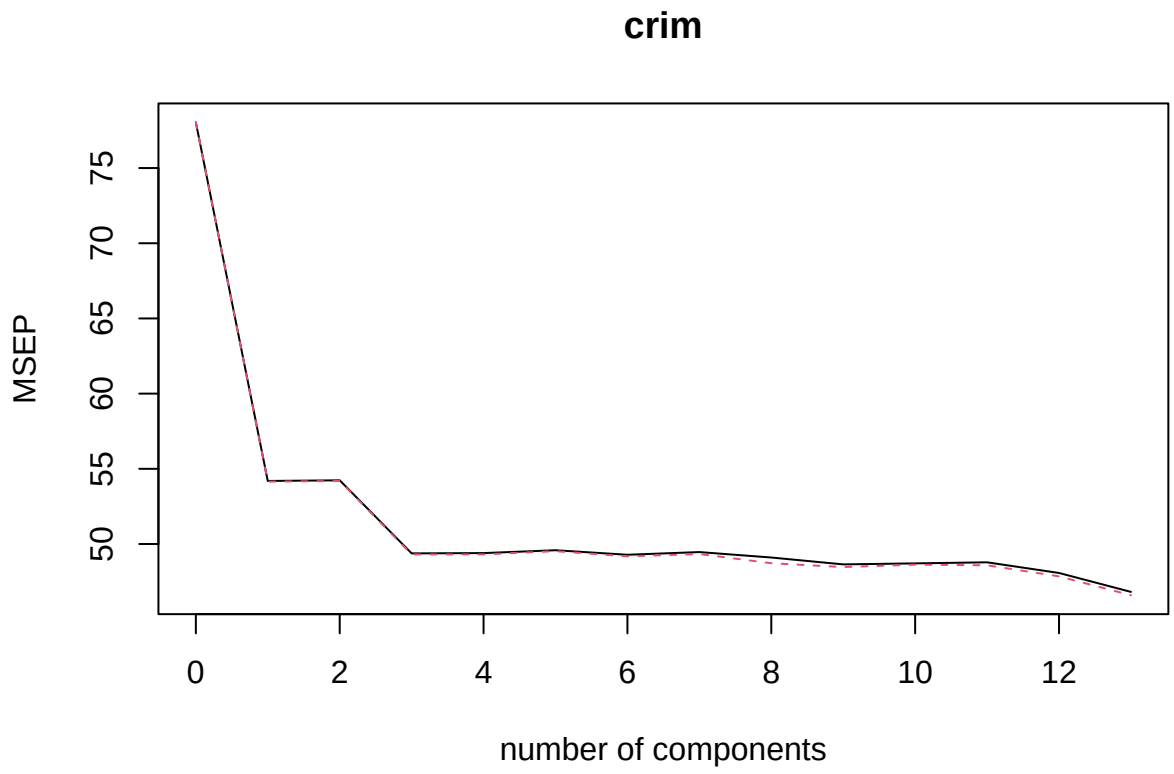

```
## [1] 0.5533799
```

```
## [1] 36.17571
```

- The ridge error is only slightly higher than the lasso

**PCR (Principal Component Regression)**

```
## Data:    X dimension: 354 13
##  Y dimension: 354 1
## Fit method: svdpc
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           8.834    7.362    7.365    7.027    7.028    7.042    7.020
## adjCV        8.834    7.358    7.361    7.022    7.022    7.037    7.013
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.033    7.007    6.975     6.979     6.984     6.934     6.842
## adjCV    7.024    6.980    6.962     6.973     6.970     6.917     6.826
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       47.99    60.42    69.63    76.41    82.83    87.86    91.23    93.43
## crim    31.27    31.40    37.67    37.79    37.79    38.97    39.32    40.69
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       95.53     97.05     98.52     99.57     100.0
## crim    41.47     41.50     41.92     43.60      45.1
```




```
## [1] 0.360531
```

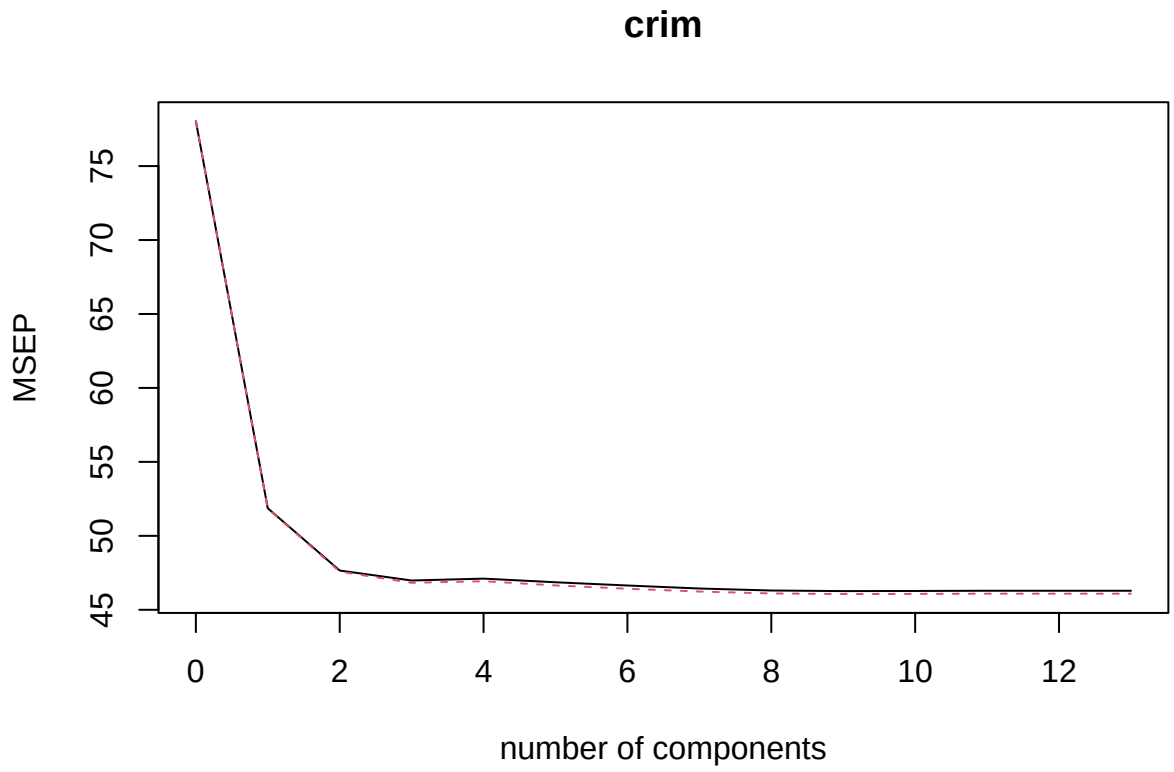
- The lowest MSE occurs when all 13 variables are present in the model, which doesn't help in dimensionality reduction. However, we can also see that between 8-13 variables, the test MSE for the 8 variable model is only higher by 0.360531.

```
## [1] 36.58822
```

**PLS (Partial Least Squares)

```
## Data:    X dimension: 354 13
##  Y dimension: 354 1
## Fit method: kernelpls
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           8.834    7.202    6.903    6.854    6.863    6.845    6.829
## adjCV        8.834    7.200    6.898    6.843    6.851    6.830    6.813
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       6.815    6.805    6.802     6.802     6.803     6.803     6.803
## adjCV    6.799    6.790    6.787     6.788     6.789     6.789     6.789
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       47.66    56.74    61.27    70.12    75.97    78.89    83.81    86.55
## crim    34.19    40.90    43.33    44.02    44.51    44.87    44.98    45.06
```

```
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       88.52     90.89     96.37     98.48     100.0
## crim    45.09     45.10     45.10     45.10      45.1
```



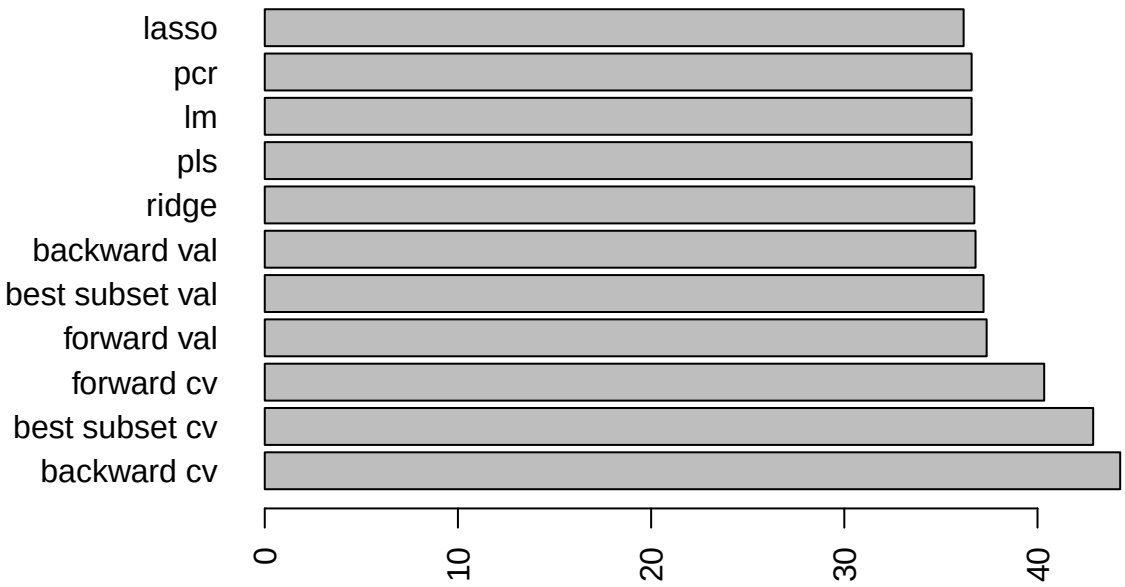

- The lowest MSE occurs when 9 variables are present.

```
## [1] 36.59429
```

**Findings:**

- Lasso regression has lower RMSE value than Ridge regression although its comparatively small

**(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.**




```
##           lasso             pcr              lm             pls           ridge
##        36.17571        36.58822        36.58822        36.59429        36.72976
##    backward val best subset val     forward val      forward cv  best subset cv
##        36.79335        37.20655        37.36707        40.34632        42.88256
##     backward cv
##        44.28048
```

**Findings**

- The best model is a 9 variable model selected using cross-validation and backward stepwise selection method.
- Ridge has all the variables considered with age and tax having a minimum effect based on their coefficient values whereas in Lasso regression, age and tax variables are eliminated due to their low effect

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

- No, it does not. It contains 9 variables: zn, indus, nox, dis, rad, ptratio, black, lstat, and medv.

# Chapter 8

## Question 8

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

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

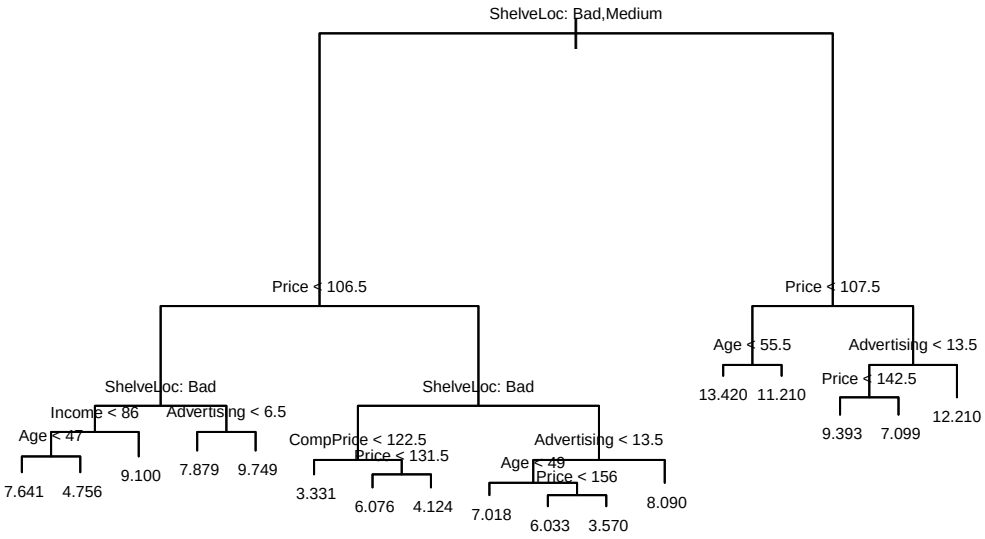
```
##     Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 395  5.35       130     58          19        366   139       Bad  33        16
## 396 12.57       138    108          17        203   128      Good  33        14
## 397  6.14       139     23           3         37   120    Medium  55        11
## 398  7.41       162     26          12        368   159    Medium  40        18
## 399  5.94       100     79           7        284    95       Bad  50        12
## 400  9.71       134     37           0         27   120      Good  49        16
##     Urban  US
## 395   Yes Yes
## 396   Yes Yes
## 397    No Yes
## 398   Yes Yes
## 399   Yes Yes
## 400   Yes Yes
```

**b) Fit a regression tree to the training set. Plot the tree, and inter- pret the results. What test MSE do you obtain?**

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats_train)
```

```
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Income"      "Age"         "Advertising"
## [6] "CompPrice"
## Number of terminal nodes:  17
## Residual mean deviance:  2.528 = 715.5 / 283
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.34800 -1.15100 -0.07915  0.00000  1.08900  4.06400
```

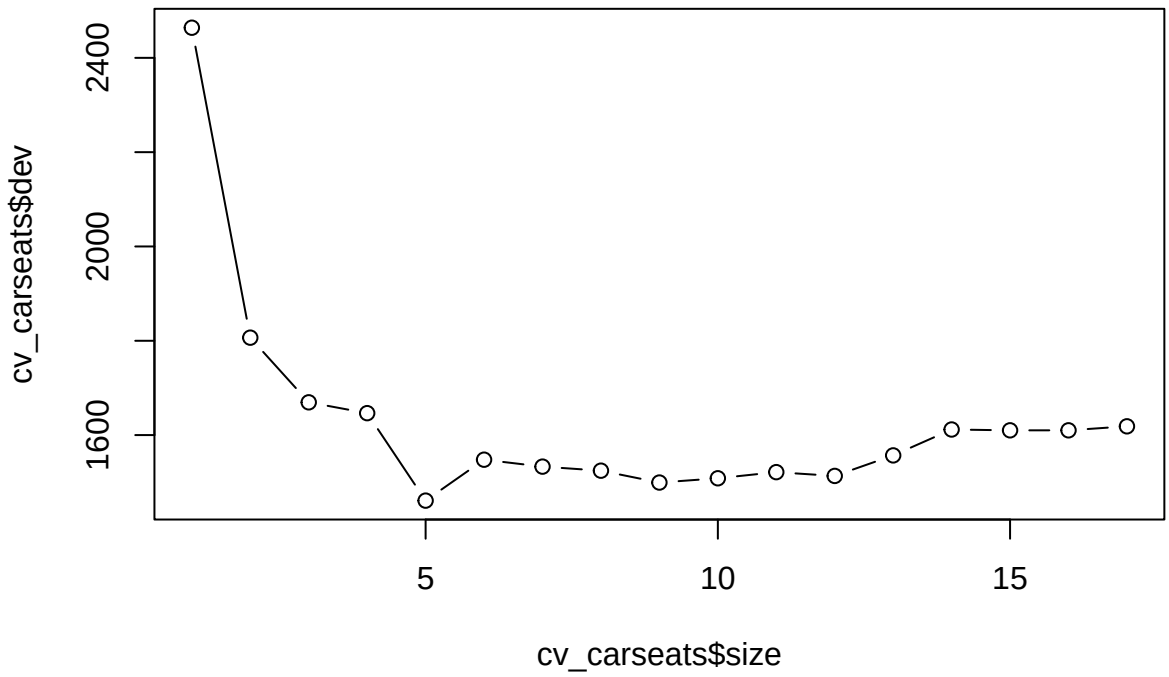
```
##            var   n         dev      yval splits.cutleft splits.cutright
## 1    ShelveLoc 300 2436.073364  7.616567            :ac              :b
## 2        Price 229 1296.153098  6.787118         <106.5          >106.5
## 4    ShelveLoc  80  373.971155  8.194250             :a              :c
## 8       Income  28  155.655611  6.976786            <86             >86
## 16         Age  19   74.504779  5.971053            <47             >47
## 32      <leaf>   8   13.309688  7.641250
## 33      <leaf>  11   22.648455  4.756364
## 17      <leaf>   9   21.360000  9.100000
## 9  Advertising  52  154.466098  8.849808           <6.5            >6.5
## 18      <leaf>  25   54.303984  7.879200
## 19      <leaf>  27   54.802741  9.748519
## 5    ShelveLoc 149  678.732413  6.031611             :a              :c
## 10   CompPrice  48  193.294392  4.665417         <122.5          >122.5
## 20      <leaf>  14   28.271571  3.331429
## 21       Price  34  129.851047  5.214706         <131.5          >131.5
## 42      <leaf>  19   57.353863  6.075789
## 43      <leaf>  15   40.564760  4.124000
## 11 Advertising 101  353.268620  6.680891           <13.5           >13.5
## 22         Age  79  247.503018  6.288481            <49             >49
## 44      <leaf>  33   94.151406  7.017576
## 45       Price  46  123.224941  5.765435           <156            >156
## 90      <leaf>  41   72.287688  6.033171
## 91      <leaf>   5   23.898600  3.570000
## 23      <leaf>  22   49.917800  8.090000
## 3        Price  71  474.221862 10.291831         <107.5          >107.5
## 6          Age  22   77.528950 12.415000          <55.5           >55.5
## 12      <leaf>  12   30.647625 13.417500
## 13      <leaf>  10   20.349160 11.212000
## 7  Advertising  49  252.993800  9.338571           <13.5           >13.5
## 14       Price  41  164.851956  8.777561         <142.5          >142.5
## 28      <leaf>  30   86.228430  9.393000
## 29      <leaf>  11   36.270691  7.099091
## 15      <leaf>   8    9.104588 12.213750
```

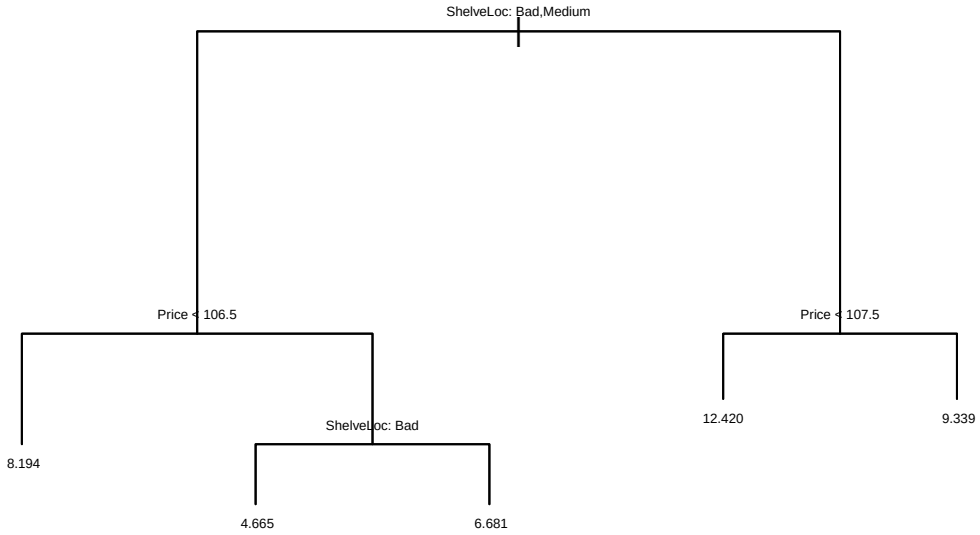



```
## [1] 5.14849
```

- The test MSE is about 4.48

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






```
## [1] 5.095198
```

- The test MSE is 4.20 which is less than the one with prunning

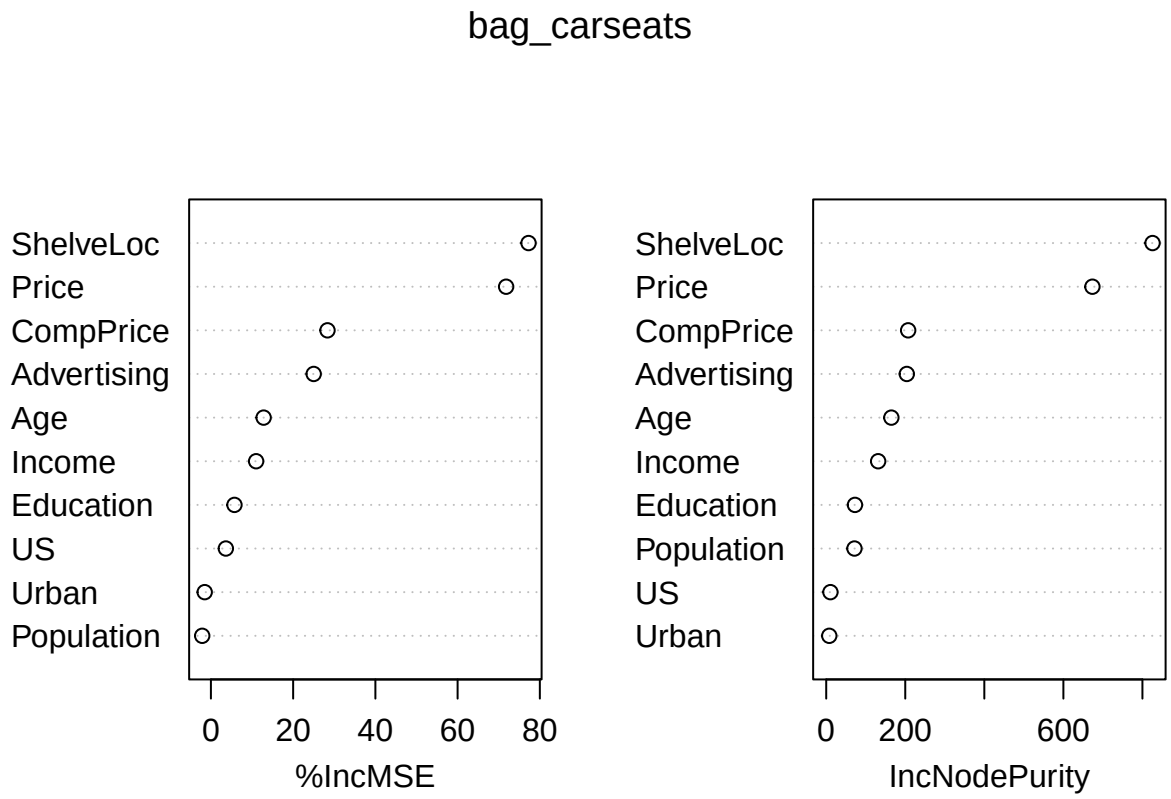
**d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to de- termine which variables are most important.**

```
##
## Bagging regression trees with 100 bootstrap replications
##
## Call: bagging.data.frame(formula = Sales ~ ., data = Carseats_train,
##     nbagg = 100, coob = TRUE, control = rpart.control(minsplit = 2,
##         cp = 0))
##
## Out-of-bag estimate of root mean squared error:  1.6486
```

```
## [1] 2.340367
```

```
##               %IncMSE IncNodePurity
## CompPrice   28.355850    207.355876
## Income      10.995942    131.465714
## Advertising 24.994722    203.908979
## Population  -2.118862     71.468063
## Price       71.801208    673.452467
## ShelveLoc   77.252876    825.577286
## Age         12.818398    164.810062
## Education    5.725214     72.726365
```

```
## Urban       -1.513533      7.746131
## US           3.629986     10.697548
```




```
## [1] 2.34352
```

- As seen above, it looks like the price of the carseat and where it is located on the shelf are the most important predictors of how a carseat will sell.
- Age, Competitor price, and advertising budget also appear to have an effect, but all other variables seem to be less important

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

```
## [1] 9
```

```
## [1] 2.313629
```

- We see that the best model uses 10 variables at each split. That model decreases test error compared to bagging.

```
##               %IncMSE IncNodePurity
## CompPrice   40.1231591     336.30816
## Income      13.7766904     174.53731
## Advertising 27.1868130     220.09826
## Population   0.9427062      98.94103
```

```
## Price       83.3416019     914.28460
## ShelveLoc   87.1747406     976.58493
## Age         27.1246025     276.24393
## Education    2.8574183      79.08633
## Urban       -1.3815670      14.38259
## US           3.4523522      15.48496
```

- ShelveLoc is the most important variable. Price, CompPrice, Advertising, and Age are also important predictors of Sale.

**Findings**:

- ShelveLoc, Price are the two most important variables
- error value forms a quadratic function while altering m value with a minimum at a certain of m

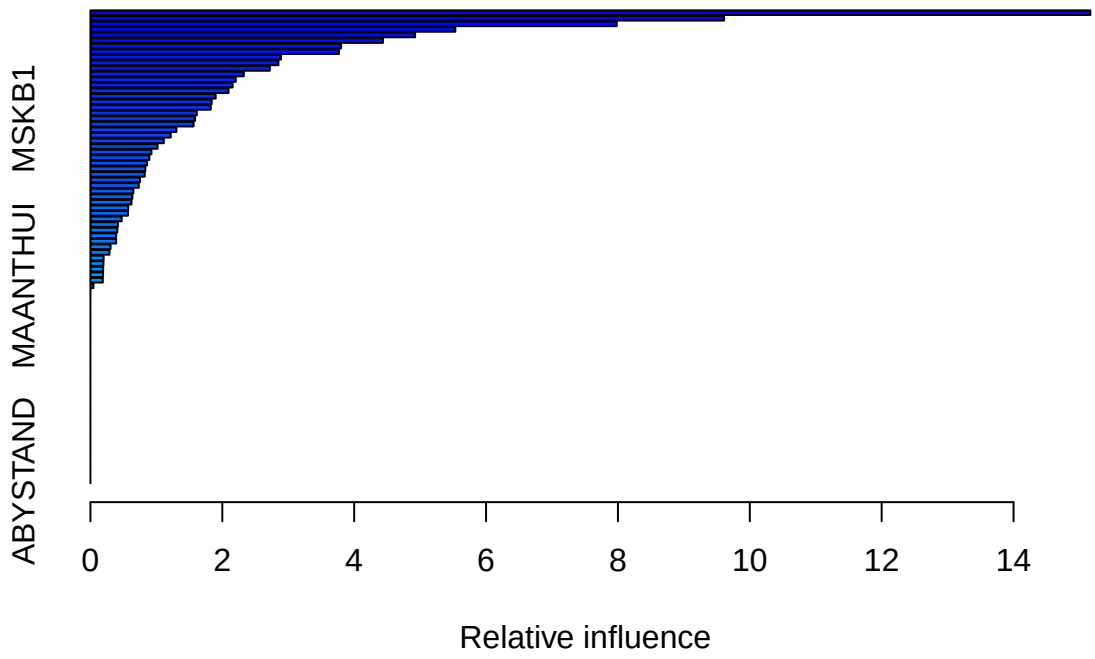
## Question 11

**a) Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.**

```
##  [1] "MOSTYPE"   "MAANTHUI" "MGEMOMV"   "MGEMLEEF" "MOSHOOFD" "MGODRK"
##  [7] "MGODPR"    "MGODOV"   "MGODGE"    "MRELGE"   "MRELSA"   "MRELOV"
## [13] "MFALLEEN" "MFGEKIND" "MFWEKIND" "MOPLHOOG" "MOPLMIDD" "MOPLLAAG"
## [19] "MBERHOOG" "MBERZELF" "MBERBOER" "MBERMIDD" "MBERARBG" "MBERARBO"
## [25] "MSKA"      "MSKB1"    "MSKB2"    "MSKC"     "MSKD"     "MHHUUR"
## [31] "MHKOOP"    "MAUT1"    "MAUT2"    "MAUT0"    "MZFONDS"  "MZPART"
## [37] "MINKM30"   "MINK3045" "MINK4575" "MINK7512" "MINK123M" "MINKGEM"
## [43] "MKOOPKLA" "PWAPART"  "PWABEDR"  "PWALAND"  "PPERSAUT" "PBESAUT"
## [49] "PMOTSCO"   "PVRAAUT"  "PAANHANG" "PTRACTOR" "PWERKT"   "PBROM"
## [55] "PLEVEN"    "PPERSONG" "PGEZONG"  "PWAOREG"  "PBRAND"   "PZEILPL"
## [61] "PPLEZIER" "PFIETS"   "PINBOED"  "PBYSTAND" "AWAPART"  "AWABEDR"
## [67] "AWALAND"   "APERSAUT" "ABESAUT"  "AMOTSCO"  "AVRAAUT"  "AAANHANG"
## [73] "ATRACTOR" "AWERKT"   "ABROM"    "ALEVEN"   "APERSONG" "AGEZONG"
## [79] "AWAOREG"   "ABRAND"   "AZEILPL"  "APLEZIER" "AFIETS"   "AINBOED"
## [85] "ABYSTAND" "Purchase"
```

**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?**

```
## gbm(formula = Purchase ~ ., distribution = "bernoulli", data = train_Caravan,
##     n.trees = 1000, shrinkage = 0.01)
## A gradient boosted model with bernoulli loss function.
## 1000 iterations were performed.
## There were 85 predictors of which 50 had non-zero influence.
```




```
##                var     rel.inf
## PPERSAUT PPERSAUT 15.16498941
## MKOOPKLA MKOOPKLA  9.61023425
## MOPLHOOG MOPLHOOG  7.98252646
## MBERMIDD MBERMIDD  5.53407671
## PBRAND     PBRAND  4.92403072
## MGODGE     MGODGE  4.43636803
## ABRAND     ABRAND  3.80183044
## MINK3045 MINK3045  3.76959522
## MAUT2       MAUT2  2.89012940
## PWAPART   PWAPART  2.85067268
## MOSTYPE   MOSTYPE  2.72053445
## MBERARBG MBERARBG  2.32508298
## MSKC         MSKC  2.20344800
## MAUT1       MAUT1  2.15637478
## MGODPR     MGODPR  2.09454762
## MGODOV     MGODOV  1.89917147
## MSKA         MSKA  1.83856589
## PBYSTAND PBYSTAND  1.82428358
## MINKGEM   MINKGEM  1.61354325
## MSKB1       MSKB1  1.58668245
## MFGEKIND MFGEKIND  1.56352878
## MFWEKIND MFWEKIND  1.30425463
## MBERHOOG MBERHOOG  1.21771296
## MGODRK     MGODRK  1.11260511
## MRELGE     MRELGE  1.01864293
## MAUT0       MAUT0  0.92496903
## MINKM30   MINKM30  0.89327927
## MRELOV     MRELOV  0.85896697
```

```
## MHHUUR     MHHUUR  0.83371895
## APERSAUT APERSAUT  0.82357657
## MSKB2       MSKB2  0.75101229
## MINK4575 MINK4575  0.73362611
## MZFONDS   MZFONDS  0.65127427
## MINK7512 MINK7512  0.63618930
## MOPLMIDD MOPLMIDD  0.62208757
## PMOTSCO   PMOTSCO  0.57312972
## PLEVEN     PLEVEN  0.56879917
## MHKOOP     MHKOOP  0.47374120
## MGEMOMV   MGEMOMV  0.41698861
## MOSHOOFD MOSHOOFD  0.40748473
## MRELSA     MRELSA  0.39062967
## MBERBOER MBERBOER  0.38995820
## MINK123M MINK123M  0.30533048
## MSKD         MSKD  0.28872748
## MZPART     MZPART  0.20007746
## MFALLEEN MFALLEEN  0.19656268
## MGEMLEEF MGEMLEEF  0.19290775
## MBERARBO MBERARBO  0.19147964
## MOPLLAAG MOPLLAAG  0.18748279
## MAANTHUI MAANTHUI  0.04456787
## MBERZELF MBERZELF  0.00000000
## 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
```

- PPERSAUT, MKOOPKLA and MOPLHOOG are three most important variables in that order.

**c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated prob- ability 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?**

```
## [1] 0.06120599 0.05308922 0.02171701 0.03998968 0.09167057 0.07937638
```

```
##    boost.pred
##        0    1
##   0 4403  130
##   1  256   33
```

```
## [1] 0.237037
```

```
##    pred
##        0    1
##   0 4183  350
##   1  231   58
```

```
## [1] 0.1421569
```

- 14% of people predicted to make purchase actually end up making one, which is less than the boosting one
- Actual positives out of predicted positives is good for knn but the total predictive is poor
- Boosting has a better prediction power with 21% positive prediction percentage

# Chapter 10

## Question 7

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
## [1] 0.95225
```

*Fitting NN*

```
## [1] 0.965
```

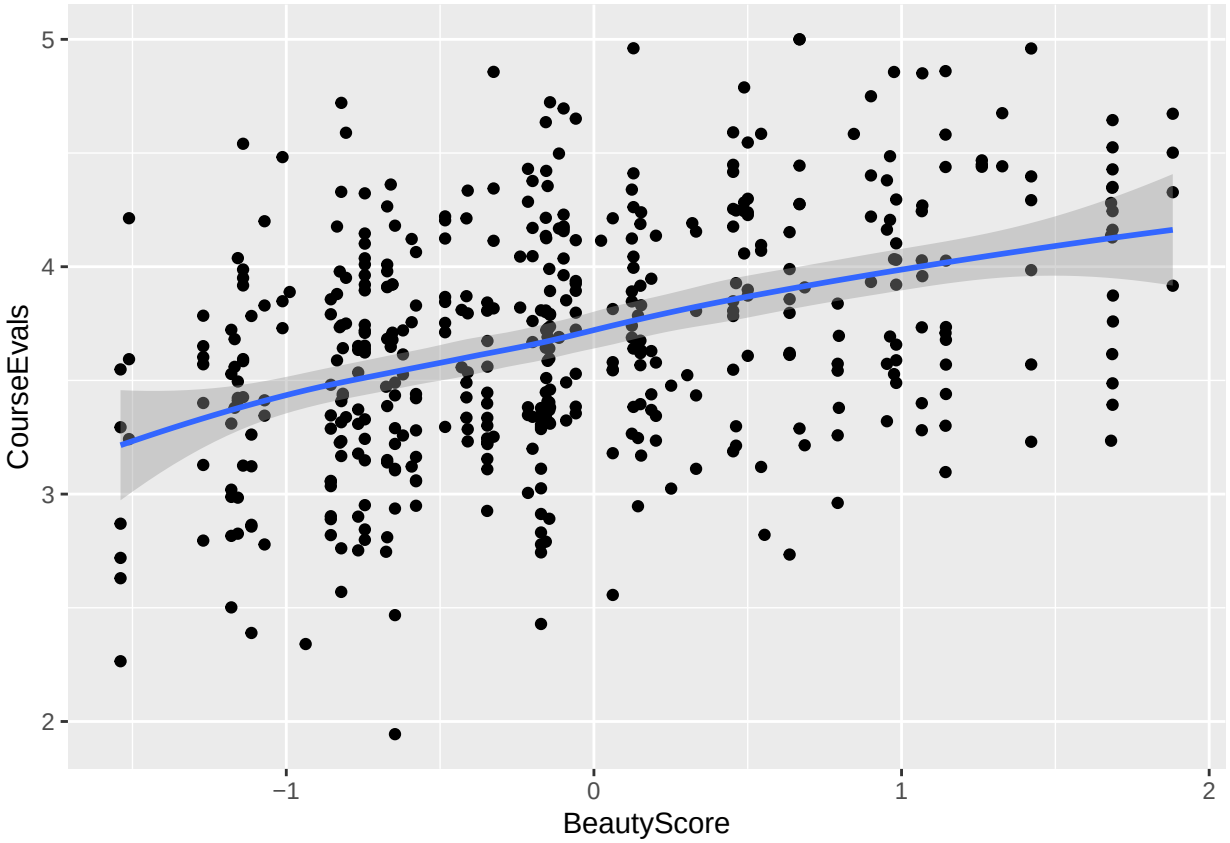
```
##           used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells 2752046 147.0    4715940 251.9         NA  4715940 251.9
## Vcells 9876612  75.4   16722626 127.6      16384 16720405 127.6
```

- As we can see, our neural network test set accuracy is about 96%. This is comparable to our linear logistic regression which had 95% accuracy.

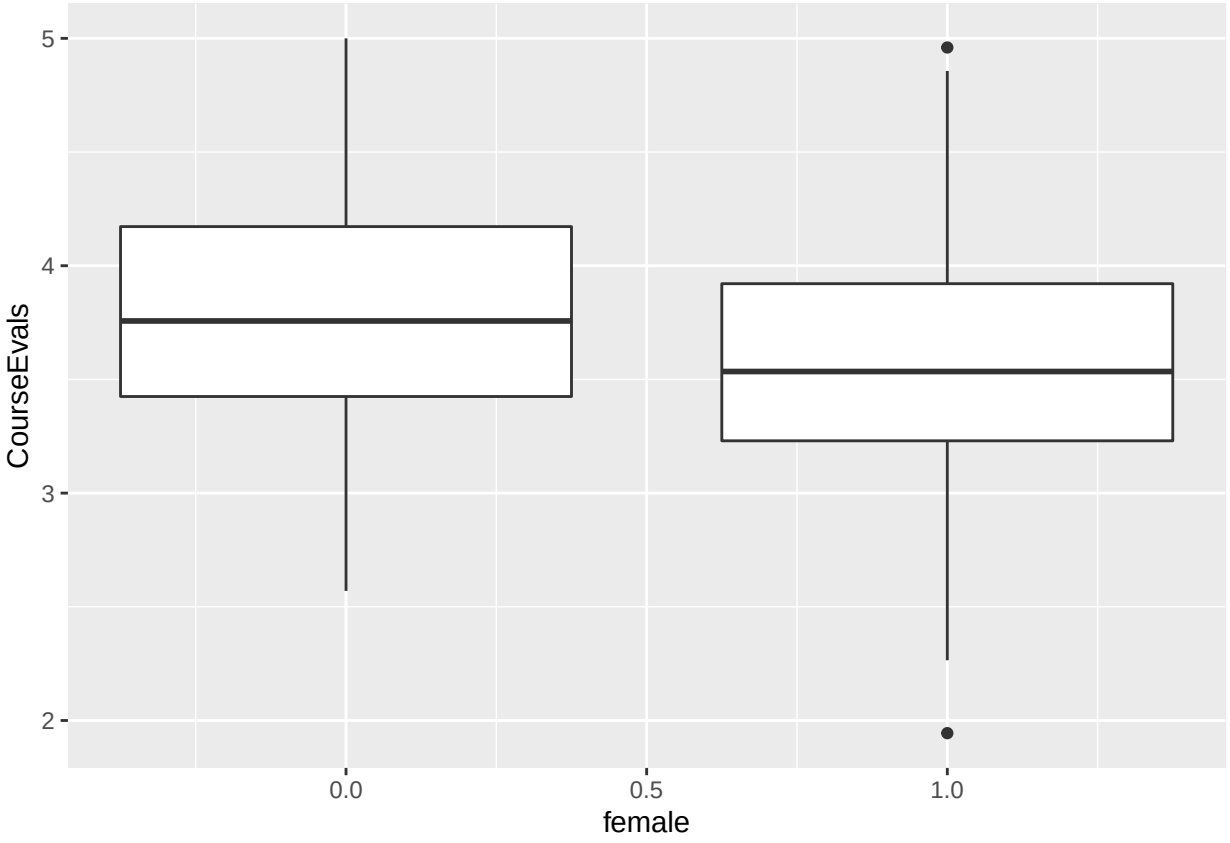
# Problem 1: Beauty Pays!

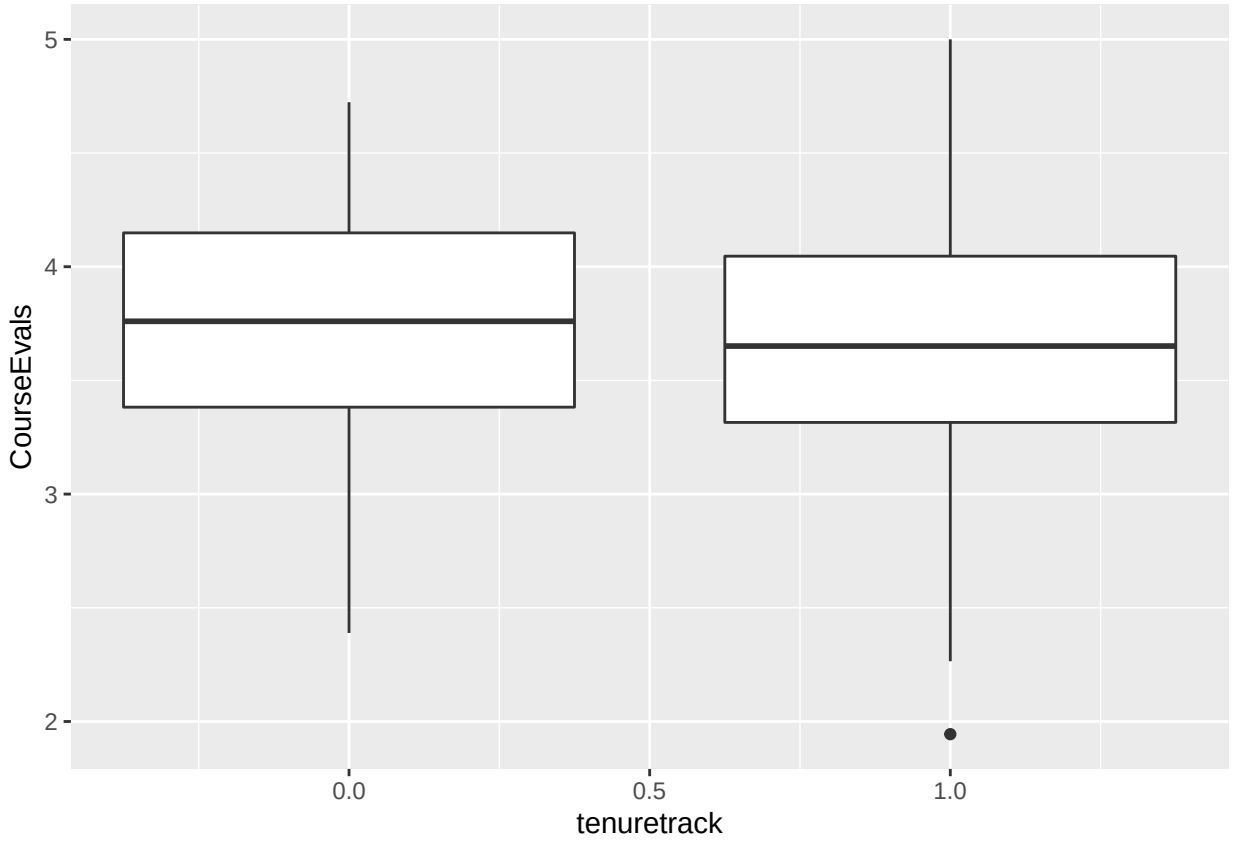
```
##   CourseEvals     BeautyScore           female           lower
##  Min.   :1.944   Min.   :-1.53884   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:3.326   1st Qu.:-0.74462   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :3.682   Median :-0.15636   Median :0.0000   Median :0.0000
##  Mean   :3.689   Mean   :-0.08835   Mean   :0.4212   Mean   :0.3391
##  3rd Qu.:4.067   3rd Qu.: 0.45725   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   : 1.88167   Max.   :1.0000   Max.   :1.0000
##    nonenglish       tenuretrack
##  Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:1.0000
##  Median :0.00000   Median :1.0000
##  Mean   :0.06048   Mean   :0.7797
##  3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000
```

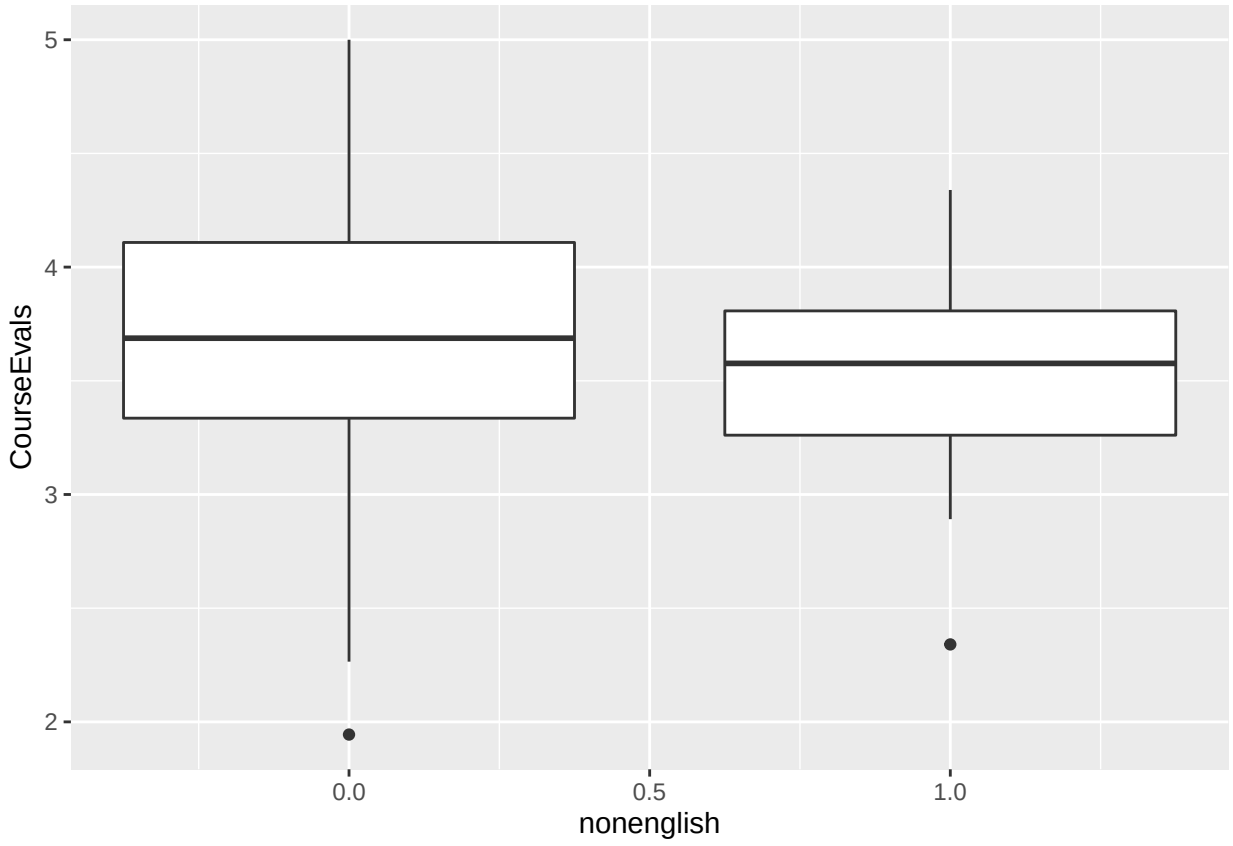
```
##
## Call:
## lm(formula = BeautyScore ~ CourseEvals, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74275 -0.54333 -0.08121  0.44552  2.04700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.34054    0.23773  -9.846   <2e-16 ***
## CourseEvals  0.61045    0.06379   9.569   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7211 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
```


5
4
3
2
CourseEvals
−1
0
1
2
BeautyScore

```
## [1] 0.4070912
```


5
4
3
2
CourseEvals
0.0
0.5
1.0
female




- Gender has an effect on the course evaluation scores

**b)**

- It is impossible to accurately identify the effect of beauty on income.Skills and strengths of different people vary at different levels.It is quite difficult to keep the other factors constant while measuring the effect of beauty on the income.As teh measurement is based on humans, there are multipe subjective choices that are made based on the situation and type of job.Sometimes it is discrinimatory to say that a person got a job just because he/she is beautiful when he/she is equally talented than any other applicant for the job.Hence it is quite challenging to distinguish between productivity and discrimation in this regard.

# Problem: 2:Housing Price Structure

```
##
## Call:
## lm(formula = midcity1$Price ~ Nbhd + Offers + SqFt + Brick +
##     Bedrooms + Bathrooms + Brick * Nbhd, data = midcity1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27225.1  -5219.0   -273.7   4297.4  27507.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3695.511   8829.382   0.419  0.67631
## Nbhd2          -1317.656   2679.849  -0.492  0.62385
## Nbhd3          16980.797   3437.529   4.940 2.60e-06 ***
## Offers         -8381.770   1068.248  -7.846 2.15e-12 ***
## SqFt              53.745      5.686   9.453 3.96e-16 ***
## BrickYes       12093.056   4082.168   2.962  0.00369 **
## Bedrooms        4777.216   1586.397   3.011  0.00318 **
## Bathrooms       6457.287   2160.867   2.988  0.00341 **
## Nbhd2:BrickYes  2668.449   5068.893   0.526  0.59957
## Nbhd3:BrickYes 11933.197   5341.027   2.234  0.02735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9847 on 118 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.8657
## F-statistic: 91.94 on 9 and 118 DF,  p-value: < 2.2e-16
```

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

- Yes there is a premium of $12093.05 if it is a brick house

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

- Yes there is a premium of $16980 if it is nbhd 3

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

- Yes, there is an extra premium of $11933

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

- Yes we can combine neighborhood 1 adn 2 as they there is not much difference in the premium and the data shows that the nieghborhood 2 is not even statistically significant

## Problem 3 : What causes what??

**1)** *Washington D.C has a terror alert system which can be used to establish the causality between cops and crime but the other cities might or might not have it.By running a regression on crime and cops in other cities, we might end up proving the correlation between them but not the causation*

**2)** *Researchers at Upenn used the terror alert system in Washington D.C to establish the causation. During the terror alert days, there are generally more cops deployed onto the streets. By comparing the correlation results between crime and cops, the researchers are able to conclude a causal relation between cops and crime.*

*From table 2,given the metro ridership is kept constant, the effect of high alert on crime is fairly constant.But we can clearly see that more the number of people on roads ,higher the crime rate.Also, introduction of metro ridership did not add much information as on high alert, it does not have a considerable effect*

**3)** *Also, there was a hypothesis that the terror alert might be causing people not to come outside and the crimes might have dropped because there were less victims to fall prey. But the researchers used the metro ridership data to check if there is any decline in ridership on a terror alert which turned out that the ridership was fairly same when compared to normal days.*

**4)** *In this model, the author was trying to show the interaction between high alert and the districts. There was a considerable between high alert and district 1 because additional police will be deployed on high alert days.But that might not be the case in other districts resulting in poor interaction between high alert and other districts*

## Problem 6 : Describe your contribution to the project

### Leadership

- I was able to guide the team towards the right approach to solve the problem. I made sure that the work that was being done was organized and structured so that there was no redundancy while performing multiple tasks.
- Apart from that, i played a key role in distributing the work among the team members to maintain the speed and made sure that we learnt from each others strengths.I encouraged the team members to ask for help when there is a roadblock

### Analysis and Modelling

- With respect to model building, i have designed the EDA steps for the project and ran iterations on models like logistic regression, random forests, KNN classification to apply the learnings from the class in the project.
- Collaboration on sharing the insights and the code on github was my idea to successfully complete the project without any obstacles.

### Presentation

- With respect to the story flow of the presentation, i suggested some key points like talking headers, simplicity/brevity of the slides to convey the right information in the right amount of time.

# References

1). Rpubs - https://rpubs.com/

2). Nueral Networks - https://rviews.rstudio.com/2020/07/20/shallow-neural-net-from-scratch-using-r-part-1/

3). ISLR Book PDF Version - https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

4). Professor Carlos Class notes - https://sites.google.com/view/predictive-modeling/home