Intro to Machine Learning (STA 380)- Take Home Exam

Ramya Madhuri Desineedi

01/08/2021

 # ISLR Book Problems

Chapter 2: 10

This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

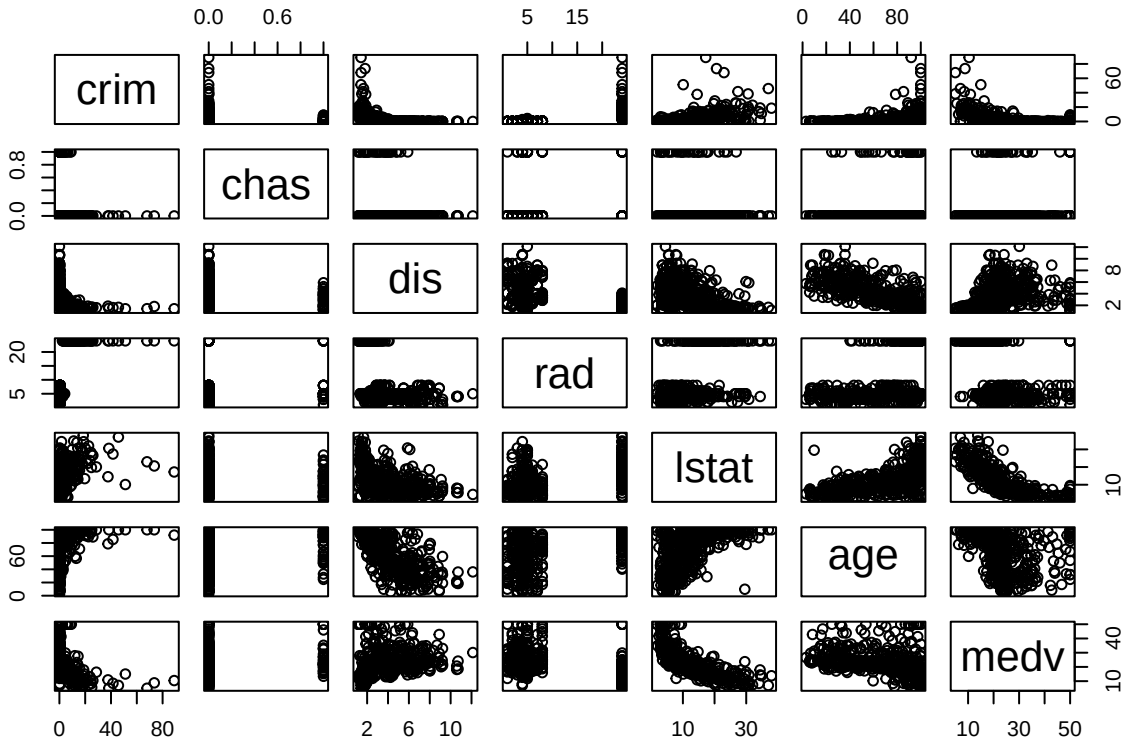
```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

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

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

- There are 506 rows and 14 columns.
- Rows represent number of data points, columns represent different parameters related to an entry.
- Here in Boston data it represents crime rate, average number of rooms per dwelling, etc... for a suburb in Boston

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



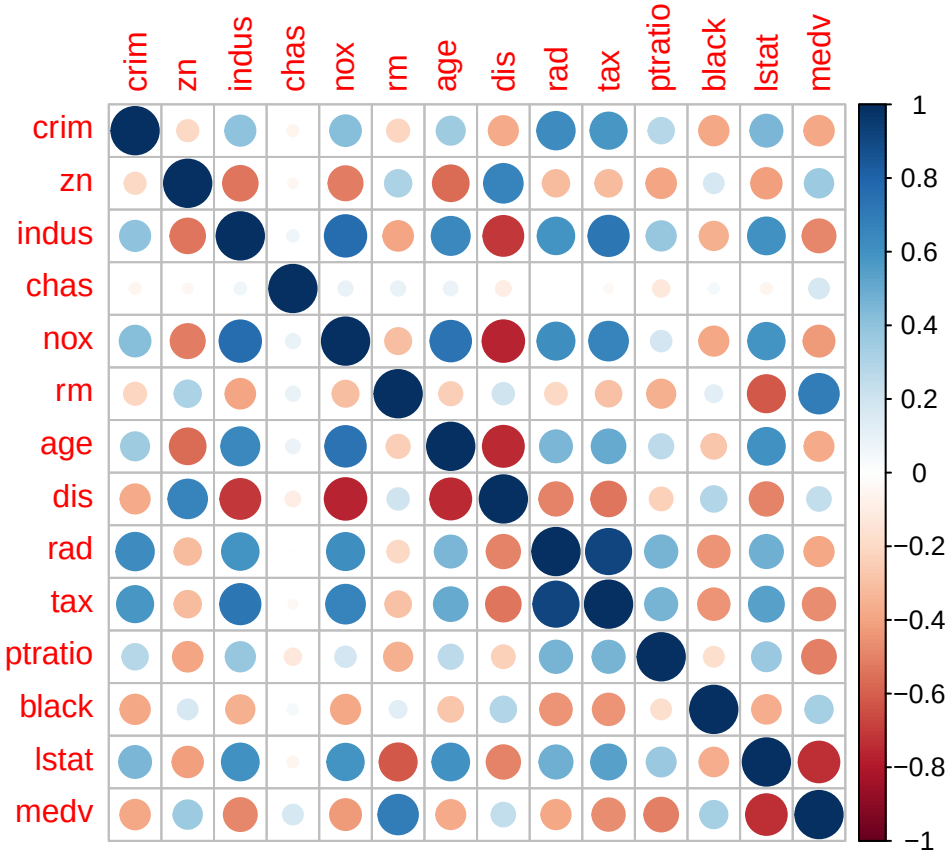
- It's hard to find correlation among the predictors from the above scatter plots. It would be better to check covariance plot instead to get the collinearity among variables/predictors.

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

```
## corrplot 0.90 loaded
```

```
##           crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
## crim      1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
## zn       -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
## indus     0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
## chas     -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
## nox       0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
## rm       -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
## age       0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
## dis      -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
## rad       0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
## tax       0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
## ptratio   0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
## black    -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18
## lstat     0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
## medv     -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
```

```
##          black lstat  medv
## crim     -0.39  0.46 -0.39
## zn        0.18 -0.41  0.36
## indus    -0.36  0.60 -0.48
## chas      0.05 -0.05  0.18
## nox      -0.38  0.59 -0.43
## rm        0.13 -0.61  0.70
## age      -0.27  0.60 -0.38
## dis       0.29 -0.50  0.25
## rad      -0.44  0.49 -0.38
## tax      -0.44  0.54 -0.47
## ptratio -0.18  0.37 -0.51
## black     1.00 -0.37  0.33
## lstat    -0.37  1.00 -0.74
## medv      0.33 -0.74  1.00
```



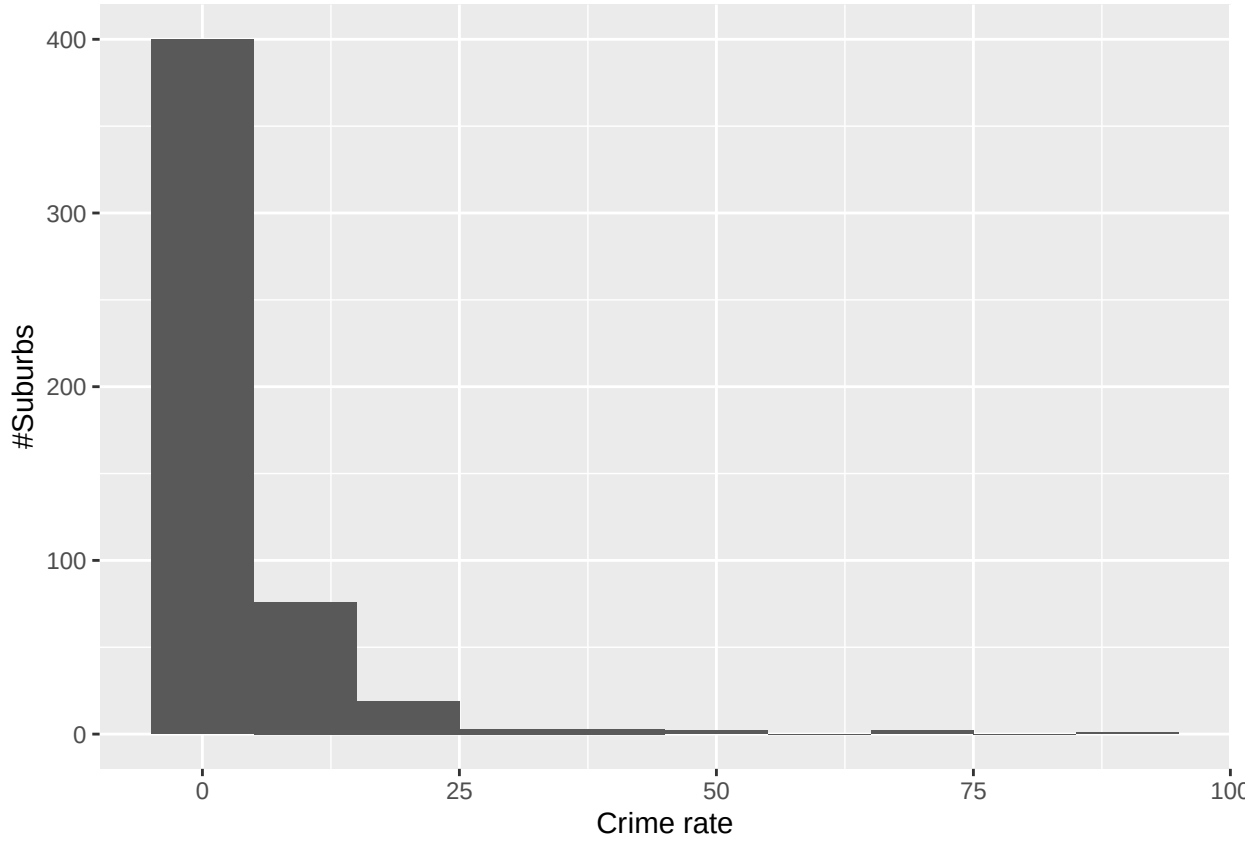
rad,tax are top 2 highly positively correlated predictors to crim followed by lstat,indus.
black, dis are top 2 highly negatively correlated predictors to crim

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor

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

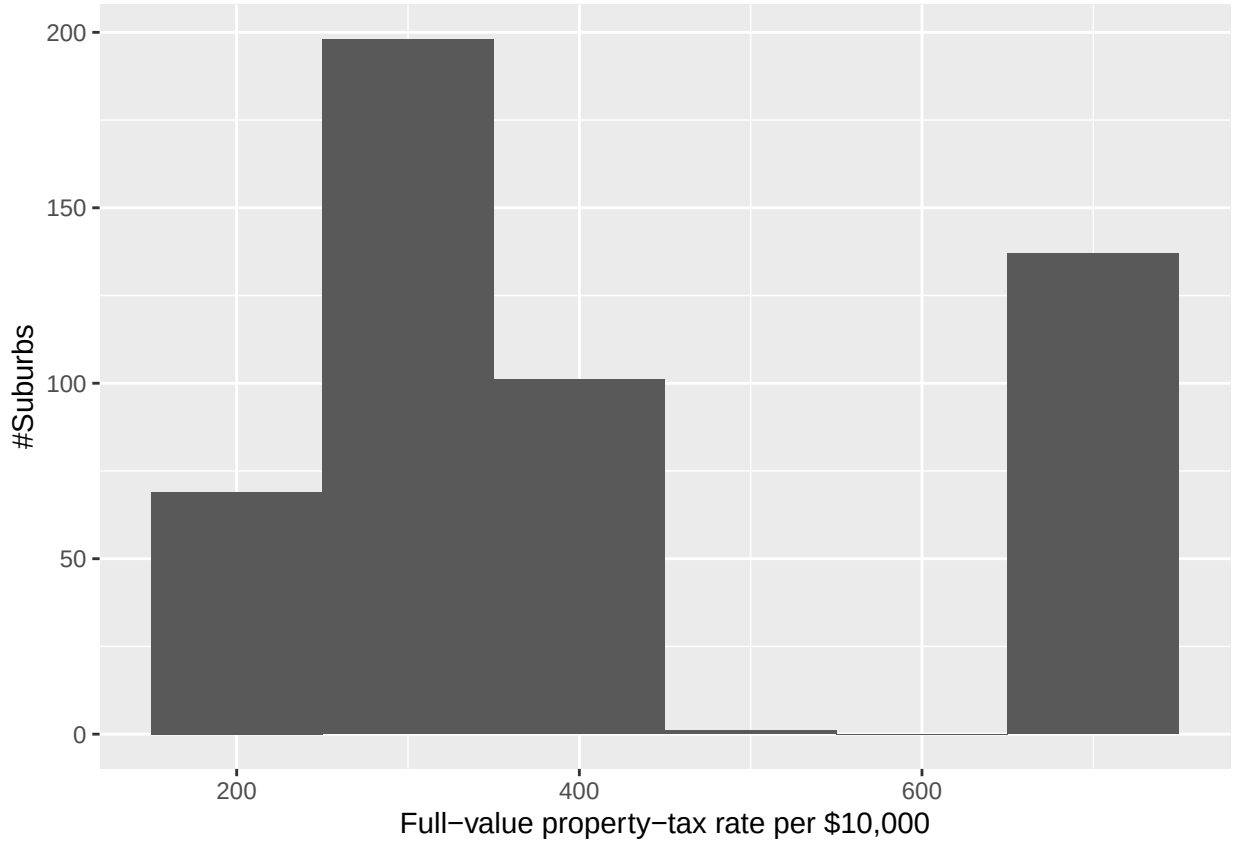
```
##          crim   zn indus chas    nox    rm   age     dis rad tax ptratio  black
```

```
## [1,]  0.00632   0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6   0.32
## [2,] 88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0 396.90
##      lstat medv
## [1,]  1.73    5
## [2,] 37.97   50
```



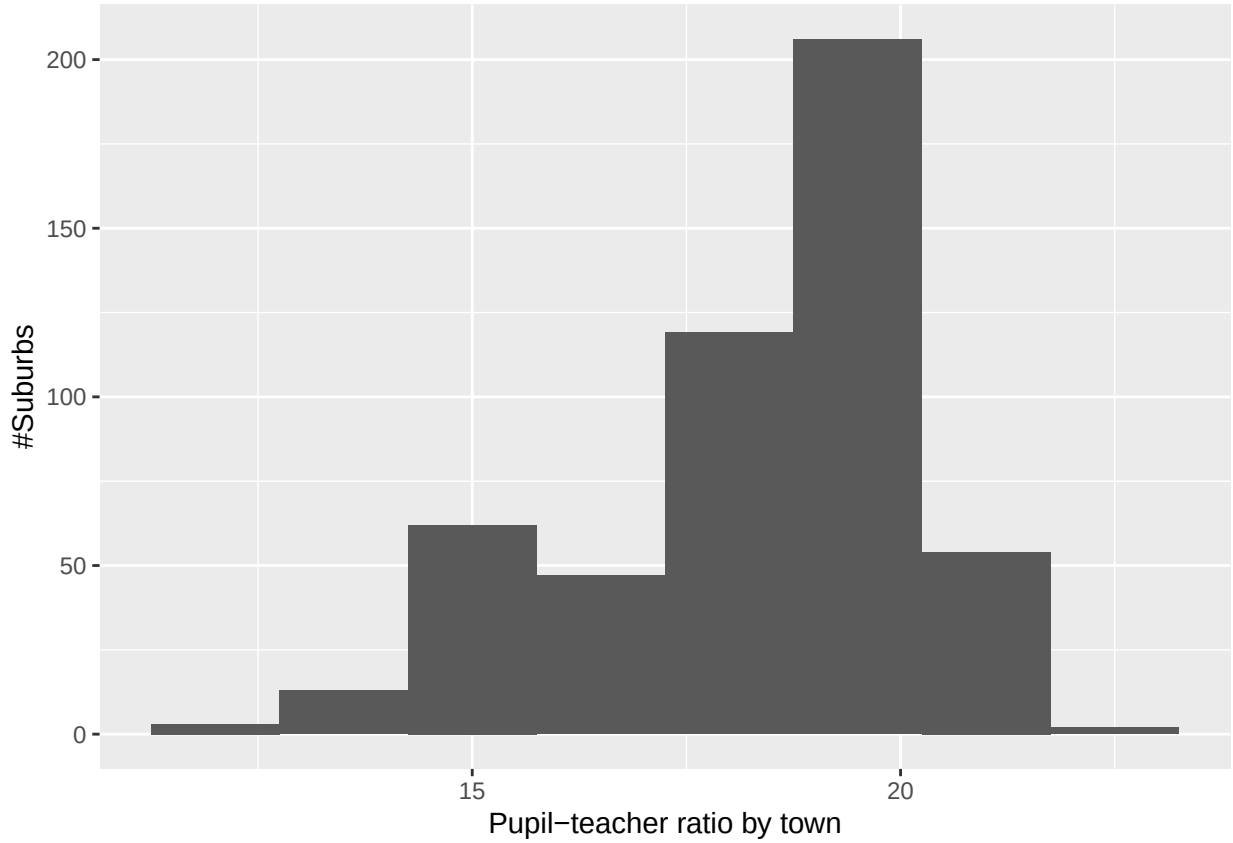
- Majority of suburbs fall below crime rate of 20.
- However there are some suburbs with crime rate above 50.Range of Crime rate is (0.006, 88.976)

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



- Majority of suburbs falls below 400 for Full-value property-tax rate per $10,000.
- However there are some suburbs with its value above 650. Range of Crime rate is (187,711)

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



- Majority of suburbs falls above 15 for Pupil-teacher ratio by town.
- However there is a small portion of suburbs with its value below 15.
- Range of Crime rate is (12.6,22)

(e) How many of the suburbs in this data set bound the Charles river?

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

- 35 suburbs bound the Charles river

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.65   17.60   17.49   18.60   20.20
```

- Median pupil-teacher ratio among the towns that bound the Charles river is 17.6

(g) Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, 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
```

Suburb 399 has lowest median value of owner- occupied homes

- Crim rate at 38.4 is very high in comparison to mean/median value
- proportion of residential land zoned for lots over 25,000 sq.ft. at 0 is equal to median value and less than mean value, which also suggests that more than half suburbs doesn't have residential land zoned for lots over 25,000 sq.ft.
- proportion of non-retail business acres per town. is very high at 18.1 is greater than mean/median value and less than mean value, and falls in 3rd Quartile
- This suburb is not bound by Charles River
- This suburb has nox of ~0.7 which is very high incomparison to other suburbs (in 4th quartile)
- This suburb has average number of rooms per dwelling of 5.5 which is less than mean/median value
- This suburb has a proportion of owner-occupied units built prior to 1940 of 100 which is max value
- This suburb has a low weighted mean of distances to five Boston employment centres of 1.49 which is quite less than mean/median value
- This suburb has index of accessibility to radial highways of 24 which is max value
- This suburb has a very high full-value property-tax rate per $10,00 of 666 which is less than mean/median value

- This suburb has a high pupil-teacher ratio by town of 20.2 which is greater than mean/median value
- This suburb has a high proportion of blacks by town of 396 which is max value

- This suburb has a high proportion of lower status population of 30.6 which is greater than mean/median value

Chapter 3: 15

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

(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.

```
##
## Call:
## lm(formula = crim ~ zn)
##
## 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
```

```
##
## Call:
## lm(formula = crim ~ indus)
##
## 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
```

```
##
## Call:
## lm(formula = crim ~ chas)
##
## 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
```

```
##
## Call:
## lm(formula = crim ~ nox)
##
## 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
```

```
##
## Call:
## lm(formula = crim ~ rm)
##
## 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

##
## Call:
## lm(formula = crim ~ age)
##
## 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

##
## Call:
## lm(formula = crim ~ dis)
##
## 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

##
## Call:
## lm(formula = crim ~ rad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.164  -1.381  -0.141   0.660  76.433
##
```

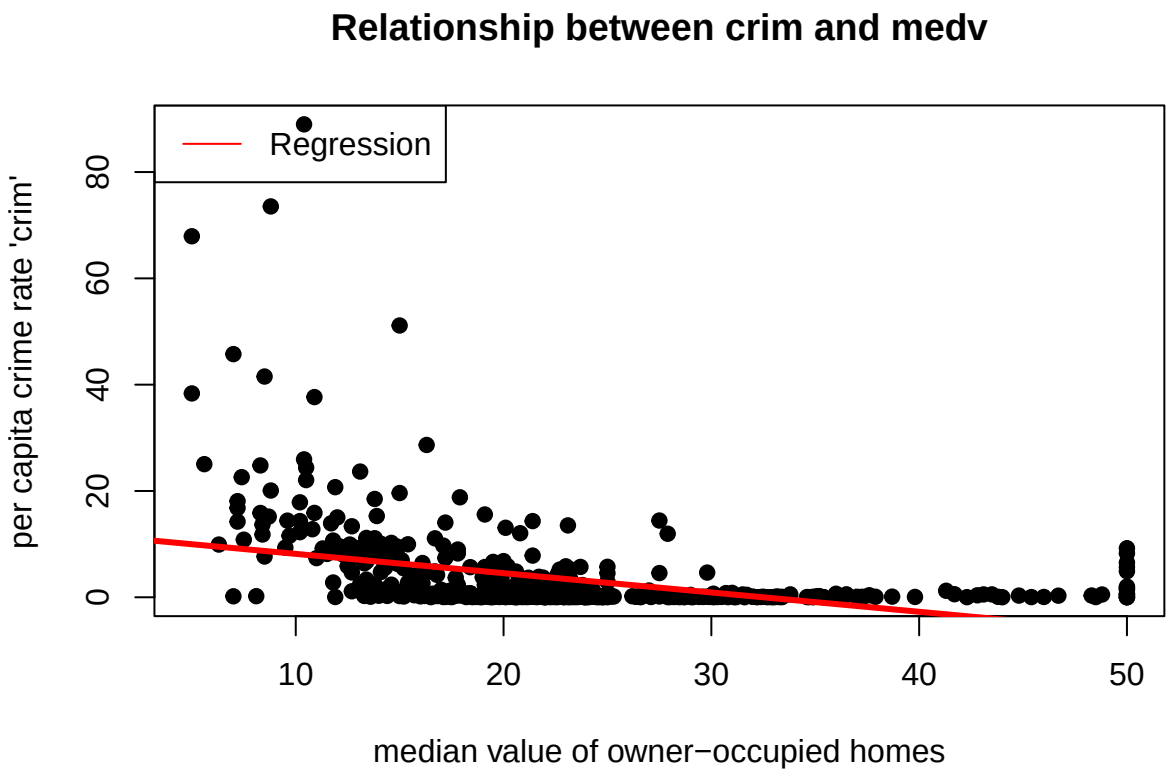
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## rad          0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 504 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
```

```
##
## Call:
## lm(formula = crim ~ tax)
##
## 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
```

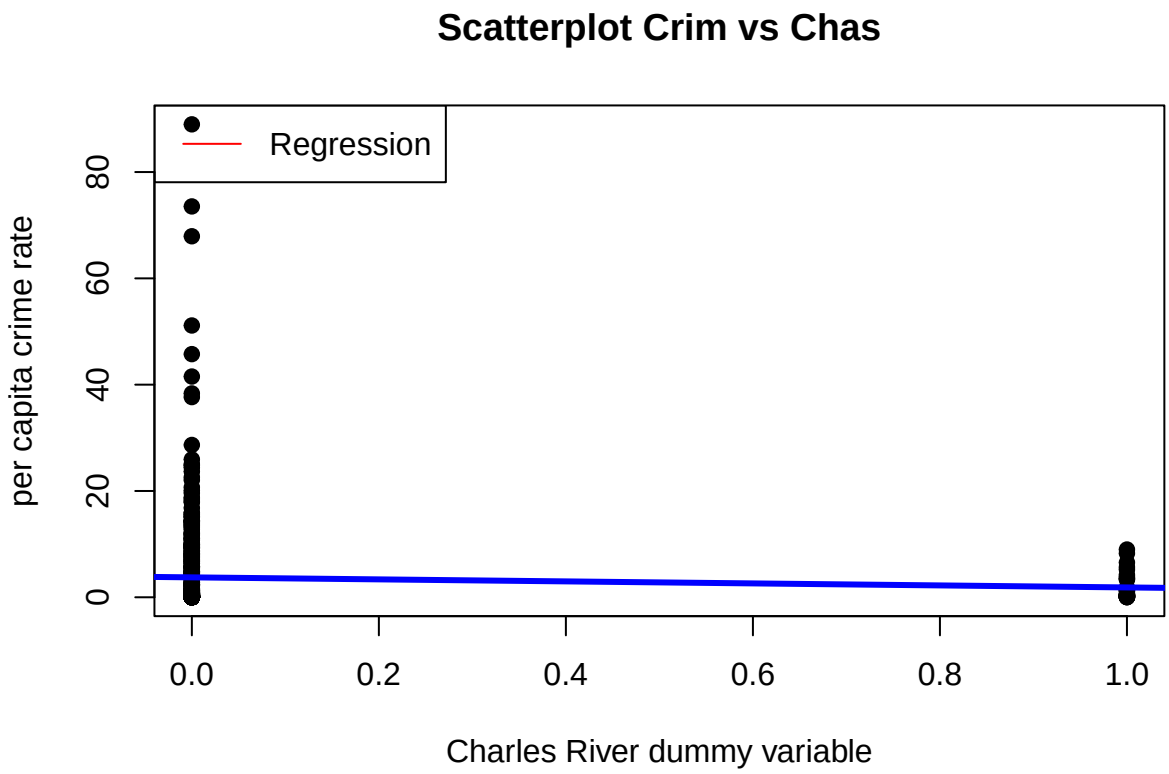
```
##
## Call:
## lm(formula = crim ~ ptratio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio       1.1520     0.1694   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
```

```
##
## Call:
## lm(formula = crim ~ black)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.756  -2.299  -2.095  -1.296  86.822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.553529   1.425903  11.609   <2e-16 ***
## black       -0.036280   0.003873  -9.367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.946 on 504 degrees of freedom
## Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
## F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = crim ~ lstat)
##
## 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


##
## Call:
## lm(formula = crim ~ medv)
##
## 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
```



- Older the building, more the crime rate



- As per univariate analysis, all predictors except "Chas" (Charles River dummy variable) turns out to be a significant variable (we can reject null hypothesis as per p-values).

- The relationship between crim and all other variables except "Chas" is statistically significant since the p value is less than 0.05.
- The low R-value signifies that the relationship between output and predictor is weak.

(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?

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv)
##
## 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
```

- We can reject null hypothesis for zn,dis,rad,black,medv

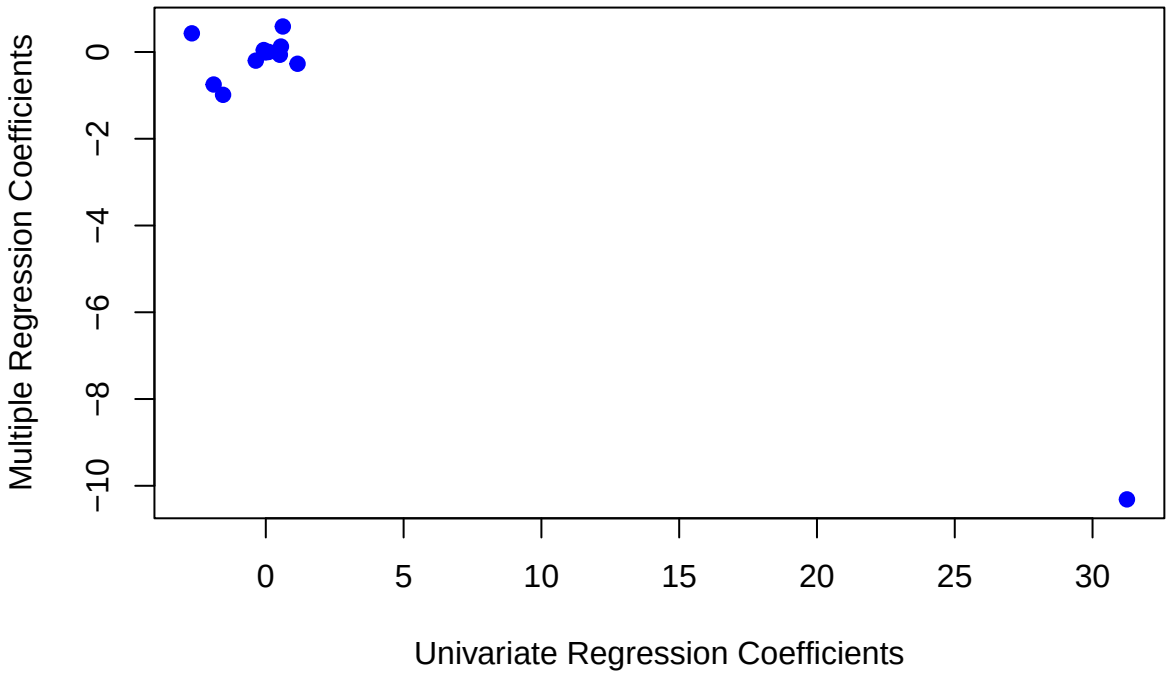
(c) How do your results from (a) compare to your results from (b)?

- In Univariate analysis all predictors except 1 came out to be very significant (at 1% level), while in multi-variate analysis only 3 variables (zn,dis,rad) turned out to be very significant (at 1% level).

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.

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv)
##
## 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
```



- Univariate and multiple linear regression coefficients are quite different for this dataset.
- In simple regression model, the fitted line represents the average effect of an increase in the predictor not taking into account the effect of other predictors.
- In multiple regression, other predictors are kept fixed, and coefficient of predictor represents its average effect on target variable with increase in that predictor.
- We can see that there are only 4-5 predictors which are important as per multiple linear regression analysis, while in univariate regression all except one predictor was important. This is because of strong correlation among predictors.

(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 .

##

```
## Call:
## lm(formula = crim ~ poly(zn, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.821 -4.614 -1.294  0.473 84.130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
## poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
## poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
## poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.372 on 502 degrees of freedom
## Multiple R-squared:  0.05824,    Adjusted R-squared:  0.05261
## F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06

##
## Call:
## lm(formula = crim ~ poly(indus, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.278 -2.514  0.054  0.764 79.713
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.614      0.330  10.950  < 2e-16 ***
## poly(indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
## poly(indus, 3)2  -24.395      7.423  -3.286  0.00109 **
## poly(indus, 3)3  -54.130      7.423  -7.292  1.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.423 on 502 degrees of freedom
## Multiple R-squared:  0.2597, Adjusted R-squared:  0.2552
## F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = crim ~ poly(nox, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.110 -2.068 -0.255  0.739 78.302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3216  11.237  < 2e-16 ***
## poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
## poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
```

```
## poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.234 on 502 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
## F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(rm, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.485  -3.468  -2.221  -0.015  87.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6135     0.3703   9.758  < 2e-16 ***
## poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
## poly(rm, 3)2  26.5768     8.3297   3.191  0.00151 **
## poly(rm, 3)3  -5.5103     8.3297  -0.662  0.50858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.33 on 502 degrees of freedom
## Multiple R-squared:  0.06779,    Adjusted R-squared:  0.06222
## F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07
```

```
##
## Call:
## lm(formula = crim ~ poly(age, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.762 -2.673 -0.516  0.019 82.842
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3485  10.368  < 2e-16 ***
## poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
## poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
## poly(age, 3)3  21.3532     7.8397   2.724  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.84 on 502 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
## F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(dis, 3))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.757  -2.588   0.031   1.267  76.378
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3259  11.087  < 2e-16 ***
## poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
## poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
## poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.331 on 502 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
## F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(rad, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.381  -0.412  -0.269   0.179  76.217
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.2971  12.164  < 2e-16 ***
## poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
## poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
## poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.682 on 502 degrees of freedom
## Multiple R-squared:    0.4,  Adjusted R-squared:  0.3965
## F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(tax, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.273  -1.389   0.046   0.536  76.950
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3047  11.860  < 2e-16 ***
## poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
## poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
## poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.854 on 502 degrees of freedom
## Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
## F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(ptratio, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.833 -4.146 -1.655  1.408 82.697
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.614      0.361  10.008  < 2e-16 ***
## poly(ptratio, 3)1    56.045      8.122   6.901 1.57e-11 ***
## poly(ptratio, 3)2    24.775      8.122   3.050  0.00241 **
## poly(ptratio, 3)3   -22.280      8.122  -2.743  0.00630 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.122 on 502 degrees of freedom
## Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085
## F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
```

```
##
## Call:
## lm(formula = crim ~ poly(black, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.096  -2.343  -2.128  -1.439  86.790
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.6135     0.3536  10.218   <2e-16 ***
## poly(black, 3)1 -74.4312     7.9546  -9.357   <2e-16 ***
## poly(black, 3)2   5.9264     7.9546   0.745    0.457
## poly(black, 3)3  -4.8346     7.9546  -0.608    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.955 on 502 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.1448
## F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(lstat, 3))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -15.234  -2.151  -0.486   0.066  83.353
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.6135     0.3392  10.654   <2e-16 ***
## poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
## poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *
## poly(lstat, 3)3 -11.5740     7.6294  -1.517   0.1299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.629 on 502 degrees of freedom
## Multiple R-squared:  0.2179, Adjusted R-squared:  0.2133
## F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = crim ~ poly(medv, 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)       3.614      0.292  12.374  < 2e-16 ***
## poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
## poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
## poly(medv, 3)3  -48.033      6.569  -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
```

Summary Table with evidence of a non-linear relationship between the per capita crime rate and the predictors

Predictor	I(Predictor)^2	I(Predictor)^3
zn	X	X
Indus	.	.
nox	.	.
rm	X	X
age	.	.
dis	.	.
rad	X	X
tax	X	X
ptratio	.	.
black	X	X
lstat	X	X
medv	.	.

Predictor	I(Predictor)^2	I(Predictor)^3
chas	NA	NA

Chapter 6: 9

In this exercise, we will predict the number of applications received using the other variables in the College data set.

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

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

```
##
## Call:
## lm(formula = Apps ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5555.2  -404.6    19.9   310.3  7577.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -630.58238  435.56266  -1.448 0.148209
## PrivateYes  -388.97393  148.87623  -2.613 0.009206 **
## Accept         1.69123    0.04433  38.153  < 2e-16 ***
## Enroll        -1.21543    0.20873  -5.823 9.41e-09 ***
## Top10perc     50.45622    5.88174   8.578  < 2e-16 ***
## Top25perc    -13.62655    4.67321  -2.916 0.003679 **
## F.Undergrad    0.08271    0.03632   2.277 0.023111 *
## P.Undergrad    0.06555    0.03367   1.947 0.052008 .
## Outstate      -0.07562    0.01987  -3.805 0.000156 ***
## Room.Board     0.14161    0.05130   2.760 0.005947 **
## Books          0.21161    0.25184   0.840 0.401102
## Personal       0.01873    0.06604   0.284 0.776803
## PhD           -9.72551    4.91228  -1.980 0.048176 *
## Terminal      -0.48690    5.43302  -0.090 0.928620
## S.F.Ratio     18.26146   13.83984   1.319 0.187508
## perc.alumni    1.39008    4.39572   0.316 0.751934
## Expend         0.05764    0.01254   4.595 5.26e-06 ***
## Grad.Rate      5.89480    3.11185   1.894 0.058662 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 993.8 on 603 degrees of freedom
## Multiple R-squared:  0.9347, Adjusted R-squared:  0.9328
## F-statistic: 507.5 on 17 and 603 DF,  p-value: < 2.2e-16
```

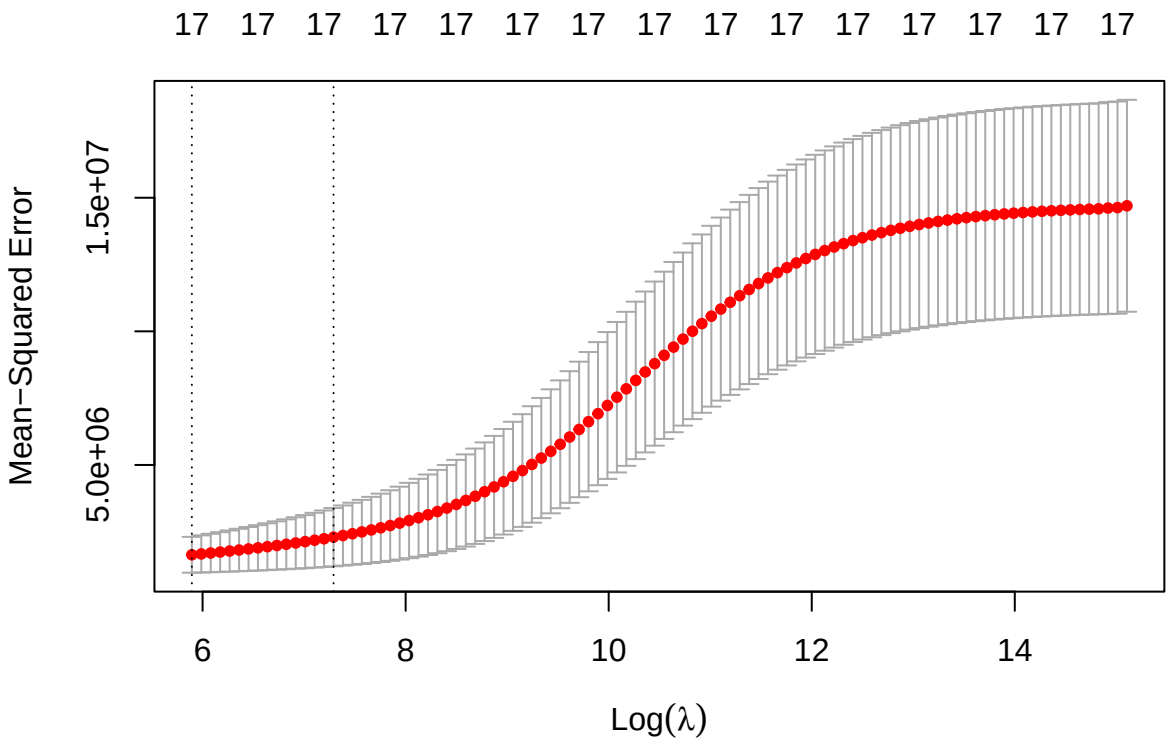
```
## [1] 1567324
```

Test MSE of linear model using least squares is 1567324

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

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-2
```



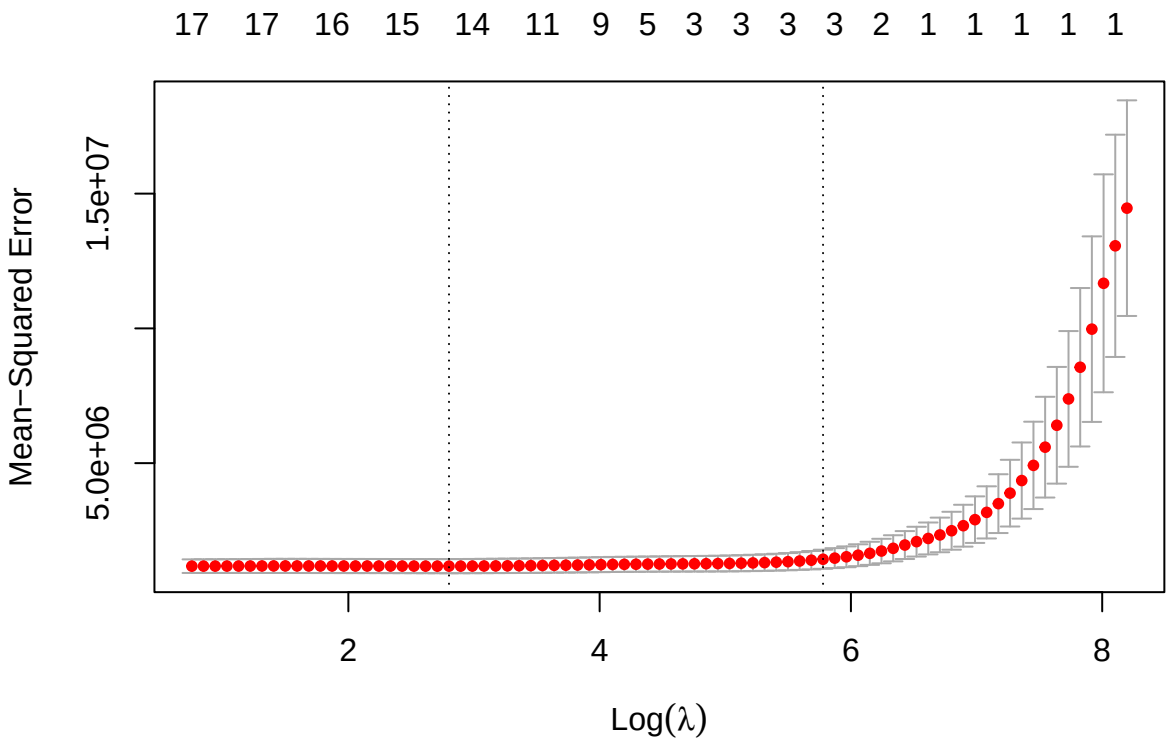

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## PrivateYes  -359.04754030
## Accept         1.06246823
## Enroll         0.38444024
## Top10perc     25.01295433
## Top25perc      1.10820539
## F.Undergrad    0.08062733
## P.Undergrad    0.03793561
## Outstate      -0.02297314
## Room.Board     0.20831052
## Books          0.21659825
## Personal      -0.02403409
## PhD           -4.82447686
## Terminal      -3.78877523
## S.F.Ratio     14.40733790
## perc.alumni   -5.81256679
## Expend         0.06094358
## Grad.Rate      8.73226951
```

```
## [1] 362.9786
```

```
## [1] 1442487
```

- Test error of ridge regression mode is 1442487
- The optimal lambda is 362.9786

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




```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## PrivateYes  -346.01504455
## Accept         1.58825724
## Enroll        -0.49445257
## Top10perc     39.13732851
## Top25perc     -5.55251034
## F.Undergrad    .
## P.Undergrad    0.05368276
## Outstate      -0.05763681
## Room.Board     0.12654679
## Books          0.14160095
## Personal       .
## PhD           -7.70609542
## Terminal       .
```

```
## S.F.Ratio      9.43931020
## perc.alumni    .
## Expend         0.05200422
## Grad.Rate      3.27270067
```

```
## [1] 16.46064
```

```
## [1] 1520014
```

- Test error of lasso model is 1520014
- optimal lambda is 16.46064

```
##   (Intercept)    PrivateYes        Accept        Enroll     Top10perc
## -697.17514891 -346.01504455    1.58825724   -0.49445257   39.13732851
##     Top25perc   P.Undergrad      Outstate    Room.Board         Books
##   -5.55251034    0.05368276   -0.05763681    0.12654679    0.14160095
##           PhD     S.F.Ratio        Expend
##   -7.70609542    9.43931020    0.05200422
```

A lot of variables' coefficients have become zero in lasso regression model

(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.

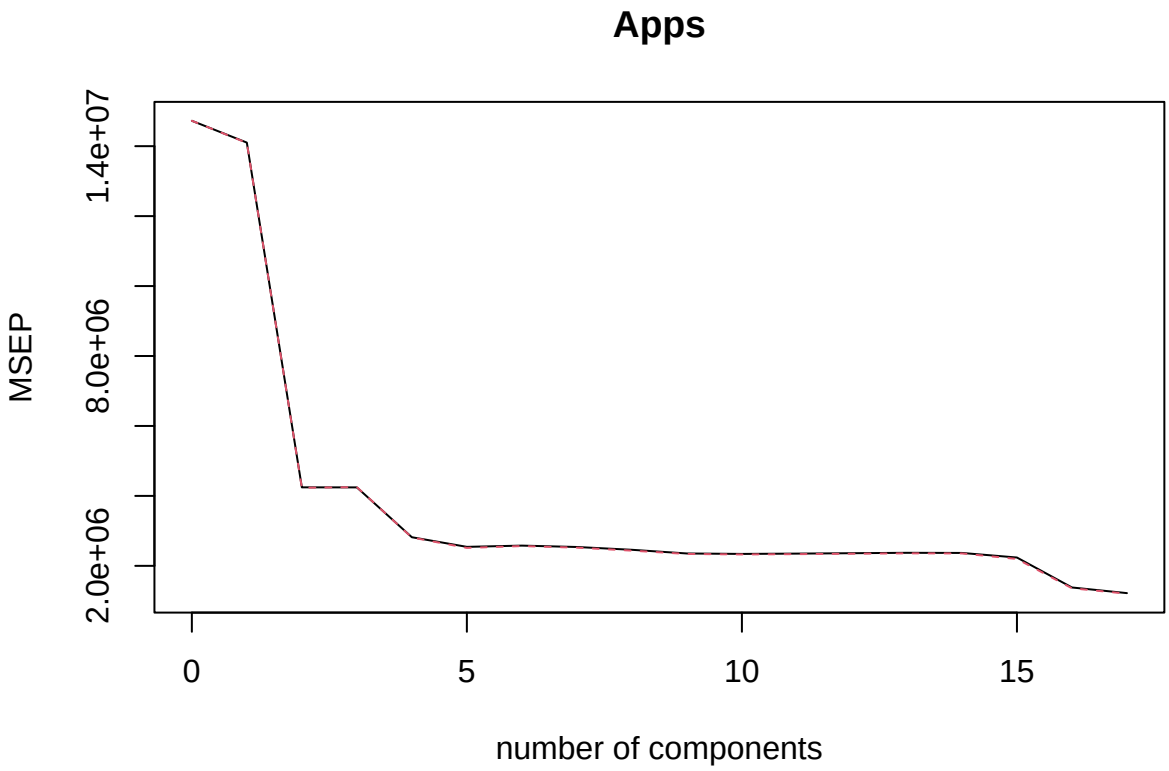
```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:corrplot':
##
##     corrplot
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
## Data:    X dimension: 621 17
##  Y dimension: 621 1
## Fit method: svdpc
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3837     3755     2060     2060     1679     1595     1606
## adjCV         3837     3754     2057     2062     1677     1585     1602
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1593     1568     1534      1530      1533      1537      1540
## adjCV     1587     1560     1530      1525      1528      1532      1535
##        14 comps  15 comps  16 comps  17 comps
## CV         1539      1496      1175      1105
## adjCV      1535      1482      1166      1097
```

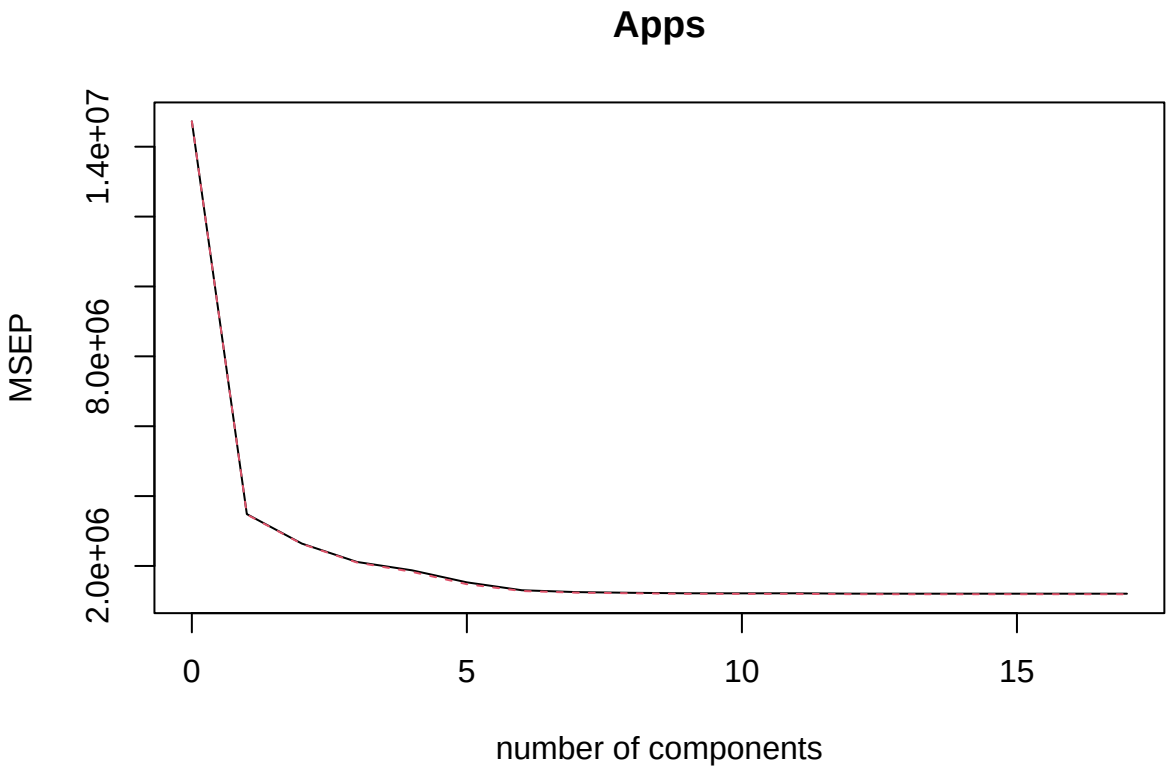
```
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X      32.003    57.06    64.13    70.03    75.36    80.38    84.09    87.44
## Apps    4.441    72.01    72.02    81.86    83.67    83.77    84.01    85.14
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       90.47     92.83     94.91     96.78     97.86     98.72     99.36
## Apps    85.42     85.76     85.76     85.76     85.89     85.95     89.93
##       16 comps  17 comps
## X        99.83    100.00
## Apps     92.89     93.47
```




```
## [1] 2579981
```

Test error is 2579981 and from the plot M can be taken as 5 components

(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] 1511731
```

Test error is 1511731 and from the graph best value for M is 6

(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 ap- proaches?

```
##      [,1]                [,2]                [,3]                [,4]  [,5]
## [1,] "OLS"               "Ridge"             "Lasso"             "PCR" "PLS"
## [2,] "0.901201944729914" "0.909071198022381" "0.904184201635825" NA    NA
```

All models have almost same R2 ~0.9-0.91. Hence there shouldn't be a lot of difference among them

Chapter 6: 11

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

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

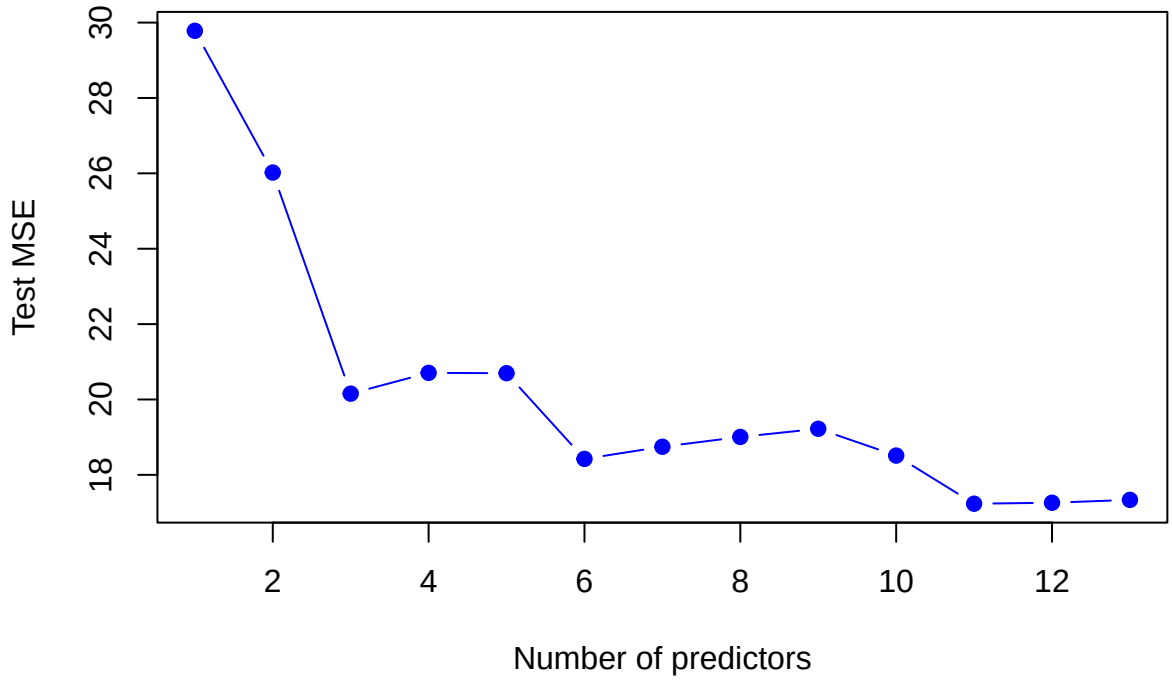
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## The following object is masked from 'package:pls':
##
##     R2

## The following objects are masked from Boston (pos = 21):
##
##     age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio, rad,
##     rm, tax, zn
```

Best subset selection

```
## Subset selection object
## Call: regsubsets.formula(medv ~ ., data = train, nbest = 1, nvmax = 13)
## 13 Variables  (and intercept)
##         Forced in Forced out
## crim        FALSE      FALSE
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
```

```
##           crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " "   " "  " " " " " " " " " " " "     " "   "*"
## 2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   "*"
## 3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "*"     " "   "*"
## 4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"     " "   "*"
## 5  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"     "*"   "*"
## 6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     " "   "*"
## 7  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"     "*"   "*"
## 8  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" "*" " " "*"     "*"   "*"
## 9  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 10  ( 1 ) " "  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 11  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```



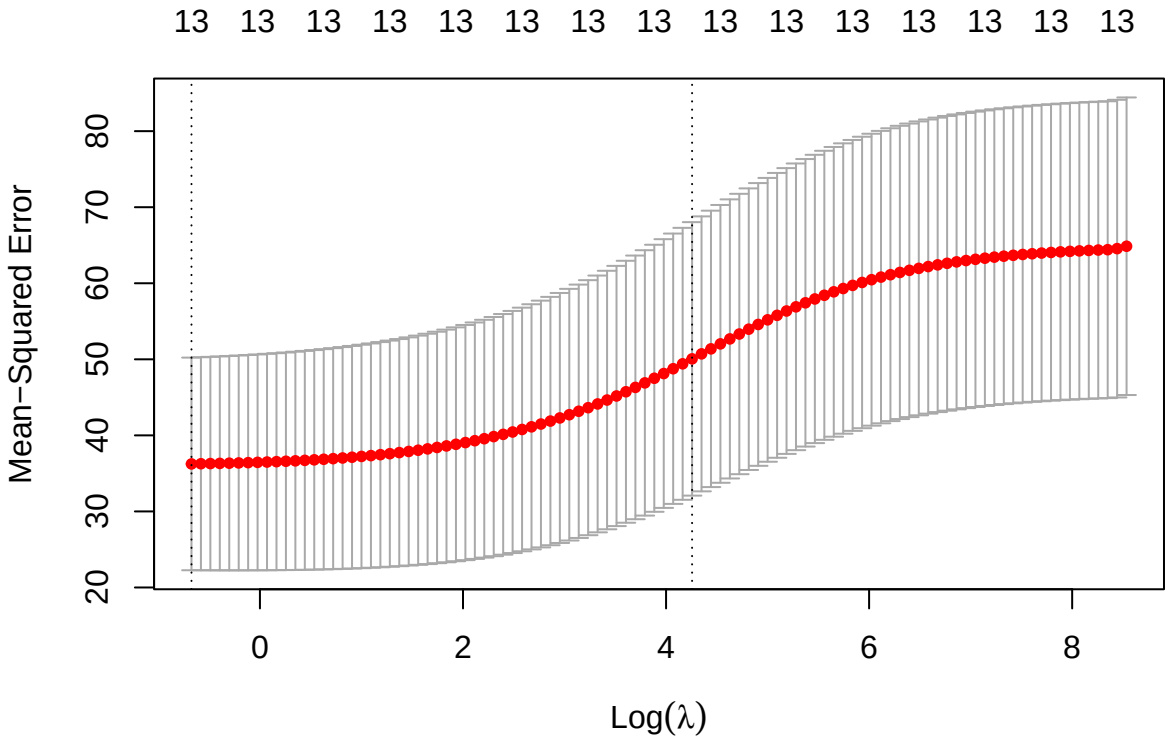

```
## [1] 11
```

```
##  (Intercept)         crim           zn         chas          nox           rm
##  32.51401751  -0.09531052   0.04027882   3.16423173 -14.84022556   4.01806843
##          dis          rad          tax      ptratio        black        lstat
##  -1.43254131   0.31633199  -0.01251988  -0.88242917   0.01047140  -0.55619049
```

```
## [1] 17.23434
```

- Best subset method selects 1 model for each number of predictor based on the minimum training set MSE

- To find best model from these 13 models, test set MSE is calculated for each model. The model with the least test set MSE is chosen as final model
- For Boston data set, final model has 11 predictor variables which is selected based on least MSE
- Test Error for Best Subset model is 17.23

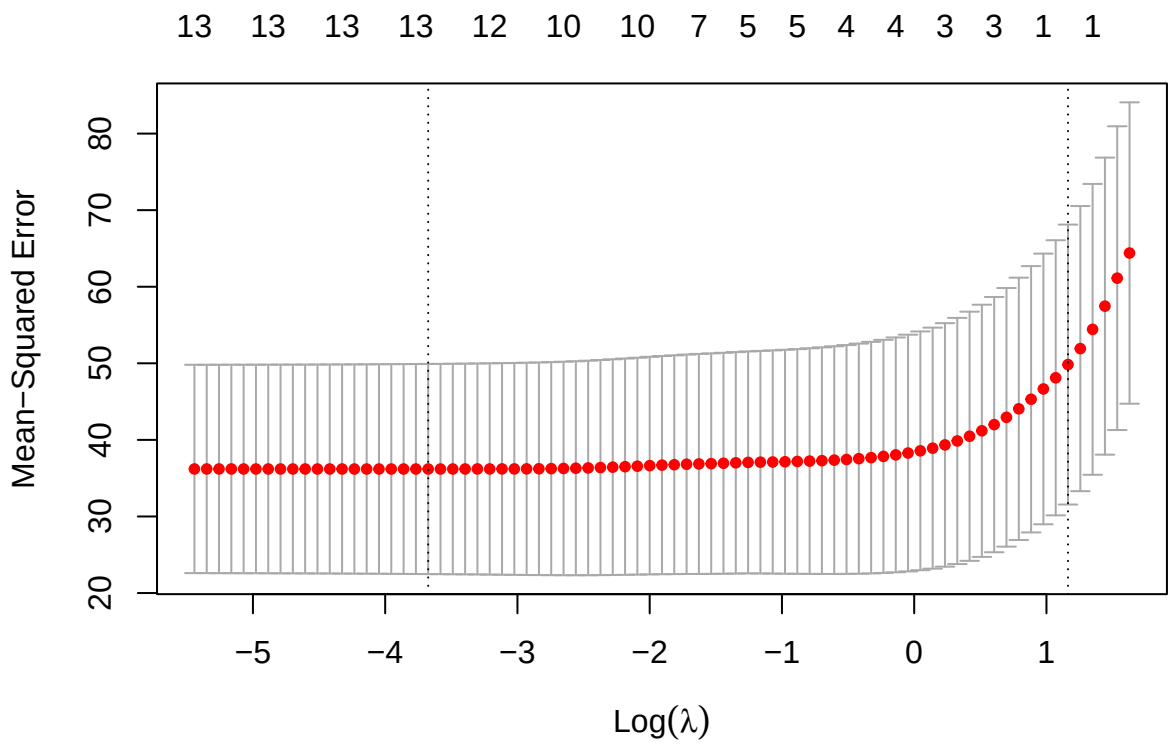


Ridge Regression

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## zn       0.027127410
## indus   -0.079905941
## chas    -0.558293904
## nox     -4.957032177
## rm      -0.114972110
## age      0.006330098
## dis     -0.560714104
## rad      0.391637626
## tax      0.002642203
## ptratio -0.158916109
## black   -0.014336192
## lstat    0.127483686
## medv    -0.094434669
```

```
## [1] 70.70553
```

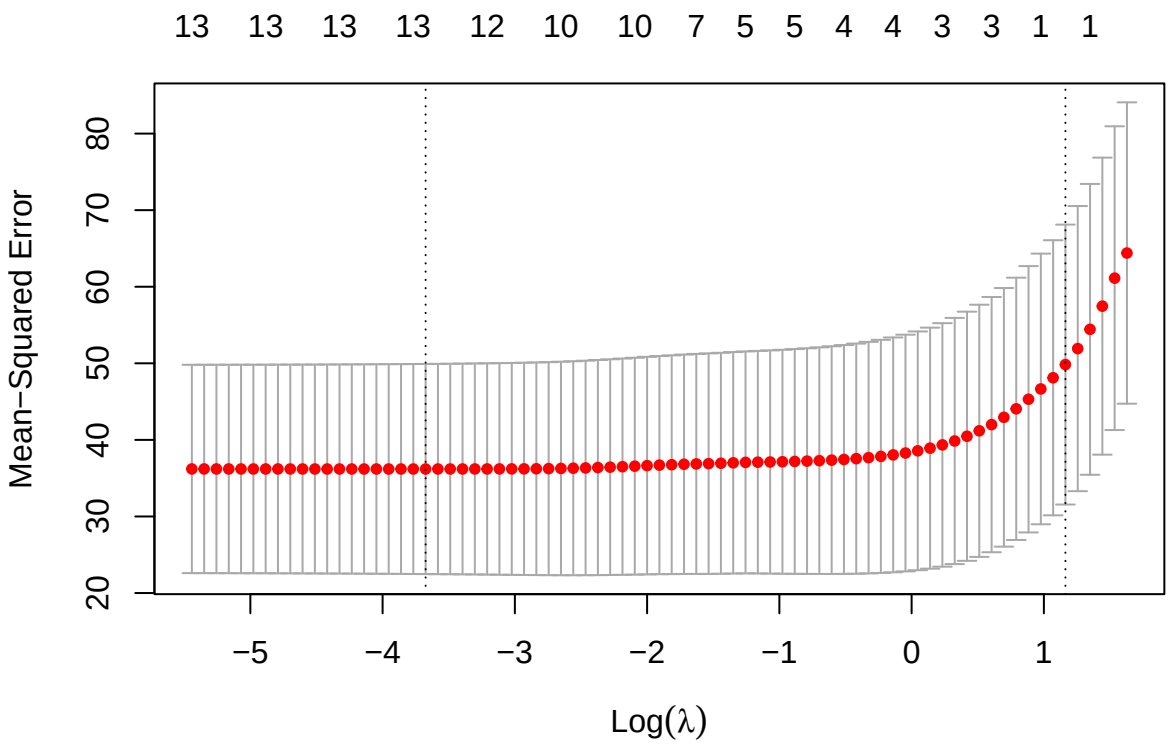
- Test Error for Lasso model is 70.71



Lasso Regression

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## zn       0.031052109
## indus   -0.069102527
## chas    -0.495047773
## nox     -7.457313888
## rm      -0.058279001
## age      0.004100168
## dis     -0.692167808
## rad      0.495473054
## tax     -0.001478808
## ptratio -0.233635980
## black   -0.013884088
## lstat    0.119943855
## medv    -0.122848974
```

```
## [1] 69.39661
```




```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## zn       0.031052109
## indus   -0.069102527
## chas    -0.495047773
## nox     -7.457313888
## rm      -0.058279001
## age      0.004100168
## dis     -0.692167808
## rad      0.495473054
## tax     -0.001478808
## ptratio -0.233635980
## black   -0.013884088
## lstat    0.119943855
## medv    -0.122848974
```

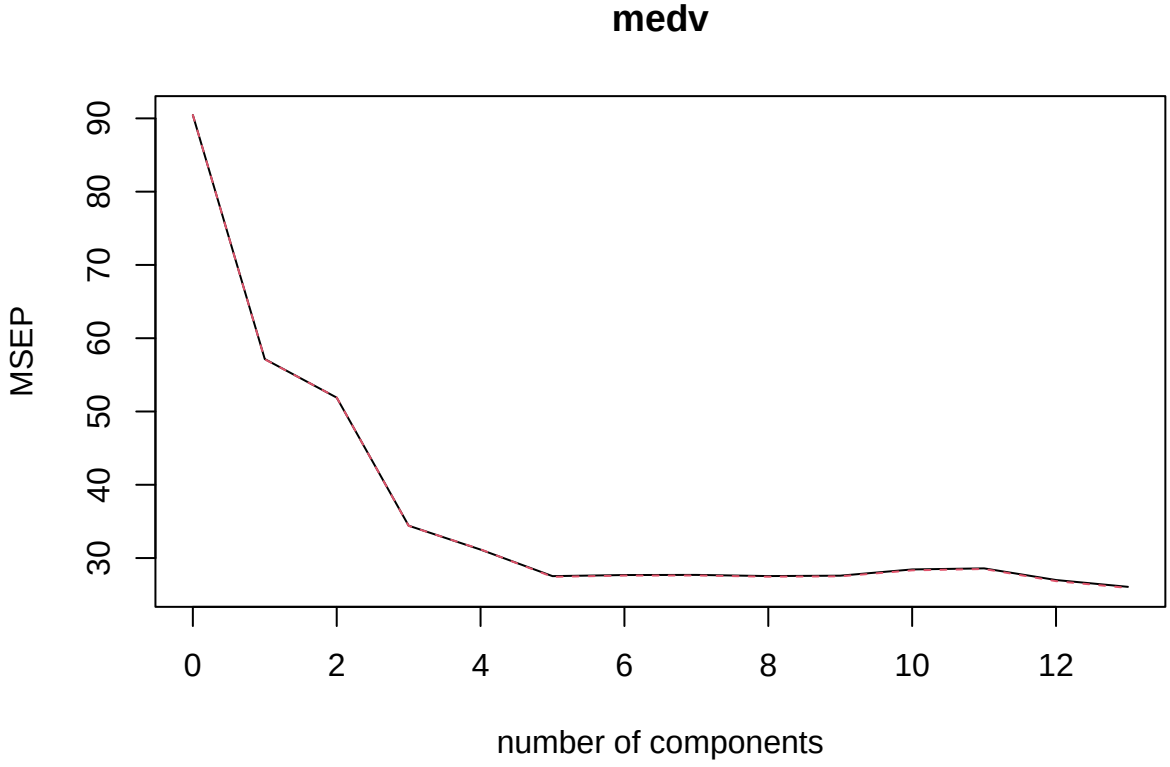
```
## [1] 69.39661
```

- Test Error for Lasso model is 69.39

Principal Component Regression

```
## Data:    X dimension: 404 13
##  Y dimension: 404 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           9.511    7.560    7.204    5.867    5.582    5.247    5.261
## adjCV        9.511    7.558    7.204    5.865    5.584    5.238    5.254
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       5.263    5.248    5.253     5.334     5.347     5.196     5.106
## adjCV    5.256    5.240    5.245     5.322     5.339     5.182     5.091
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       47.76    58.93    68.44    74.95    80.96    85.83    89.72    92.86
## medv    37.08    43.80    62.84    66.56    70.50    70.69    70.74    71.16
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       94.97     96.74     98.15     99.47    100.00
## medv    71.26     71.44     71.80     73.17     74.12
```




```
## Data:    X dimension: 404 13
##  Y dimension: 404 1
## Fit method: svdpc
## Number of components considered: 5
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps
```

```
## X       47.76    58.93    68.44    74.95    80.96
## medv    37.08    43.80    62.84    66.56    70.50

## , , 1 comps
##
##                medv
## crim    -0.588801868
## zn       0.594359188
## indus   -0.794900103
## chas     0.006755442
## nox     -0.789581241
## rm       0.482001968
## age     -0.721751861
## dis      0.738778075
## rad     -0.732505619
## tax     -0.781028476
## ptratio -0.496858010
## black    0.486748055
## lstat   -0.710440498
##
## , , 2 comps
##
##                medv
## crim    -1.26975653
## zn      -0.09509656
## indus   -0.53314409
## chas     0.93730311
## nox     -0.37519770
## rm       0.69336877
## age     -0.09176793
## dis      0.02745044
## rad     -1.29390069
## tax     -1.25754345
## ptratio -1.01351005
## black    1.08722850
## lstat   -0.82833073
##
## , , 3 comps
##
##                medv
## crim    -0.32313498
## zn       0.87384534
## indus   -0.61478210
## chas     2.12125539
## nox      0.16379693
## rm       2.76263661
## age     -0.05218162
## dis     -0.30340406
## rad     -0.30954106
## tax     -0.57549208
## ptratio -2.51153313
## black   -0.07656056
## lstat   -1.79348834
##
```

```
## , , 4 comps
##
##                medv
## crim    -0.550428741
## zn       0.439966060
## indus   -0.607885890
## chas     0.576327368
## nox      0.346441174
## rm       3.636138331
## age      0.267914868
## dis     -0.782564121
## rad     -0.361098394
## tax     -0.608362738
## ptratio -2.591044392
## black   -0.007054346
## lstat   -2.231803633
##
## , , 5 comps
##
##                medv
## crim    -0.723532547
## zn       0.027863977
## indus   -0.582899095
## chas     1.112252724
## nox      0.003375417
## rm       4.232945099
## age      0.058538387
## dis     -0.734748951
## rad      0.305679170
## tax     -0.126128915
## ptratio -1.292662974
## black    0.755318983
## lstat   -3.010694412
```

```
## [1] 21.23601
```

- Using PCR model with 5 components (which explains 80 % of variation) to predict on test data
- Test Error for PCR model is 21.23

(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.

- Minimum Test Error (Cross-validation) is for Best Subset model: 17.23, followed by PCR model: 21.23
- Best Subset Model would have performed best because it has taken different number of predictor variables and optimised MSE to get best model

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

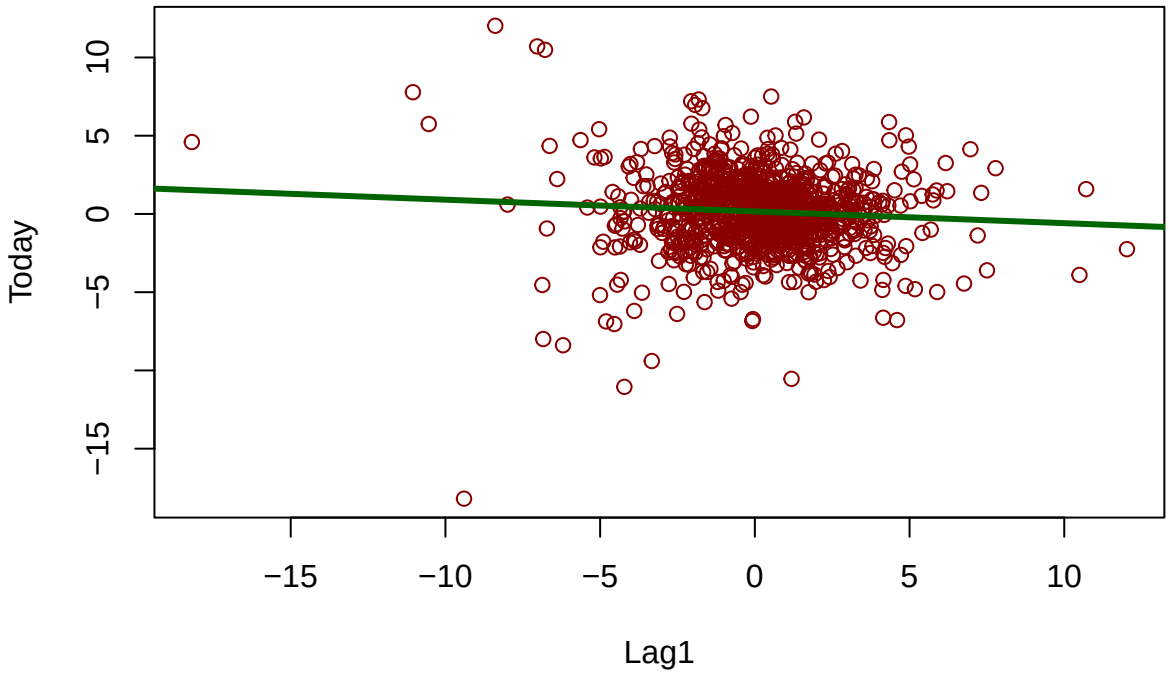
- Chosen model is Best Subset Model which uses 11/13 predictors, as majority of varianace in data can be explained by these predictors and adding more variables just makes it a model with high variance.

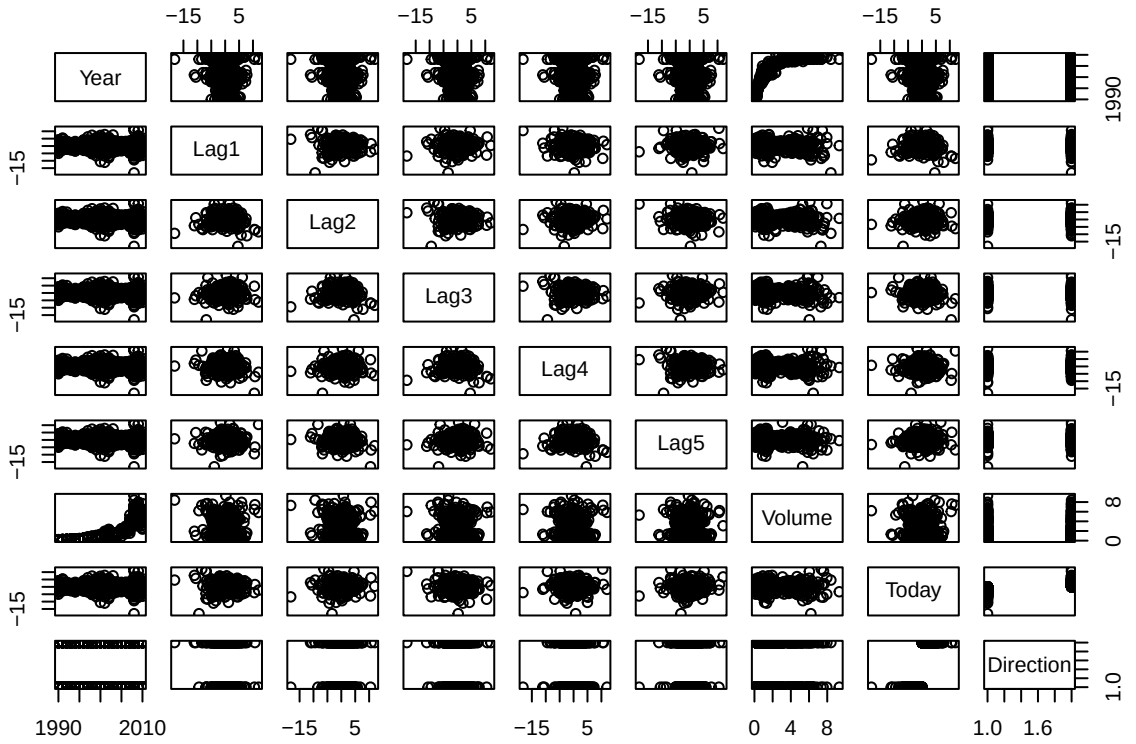
Chapter 4: #10

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

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

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```





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

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Only Lag2 is statistically significant to predict Direction.

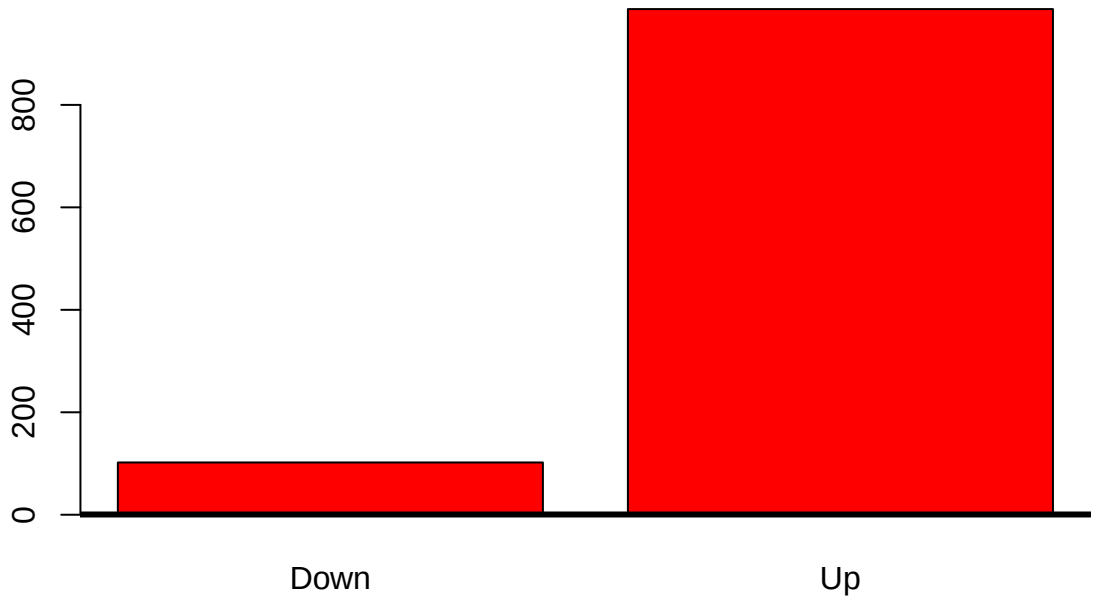
c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   54  48
##       Up    430 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.369
##
##                   Kappa : 0.035
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9207
##             Specificity : 0.1116
##          Pos Pred Value : 0.5643
##          Neg Pred Value : 0.5294
##              Prevalence : 0.5556
##          Detection Rate : 0.5115
##    Detection Prevalence : 0.9063
##       Balanced Accuracy : 0.5161
##
##        'Positive' Class : Up
##
```

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

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = "binomial", data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```




```
##       y_test
## y_pred Down Up
##   Down    9  5
##   Up     34 56
```

- Test Error Rate for Logistic Regression Model: 37.5%

(g) Repeat (d) using KNN with K = 1

```
##          y_test
## pred.knn Down Up
##        1   21 30
##        2   22 31
```

- Test Error Rate for knn Model: 50%

(h) Which of these methods appears to provide the best results on this data? Logistic Regression is a better fit for this data with 37.5% error rate

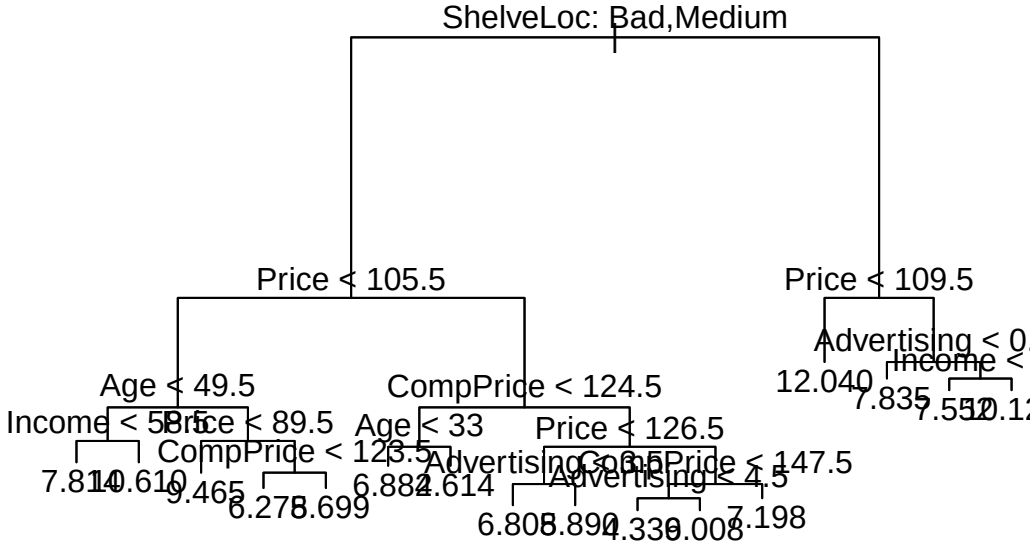
Chapter 8: #8

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

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

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test



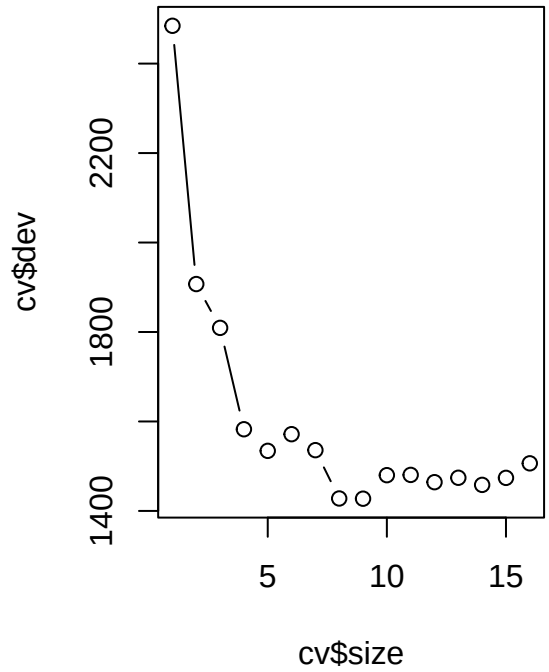
MSE do you obtain?

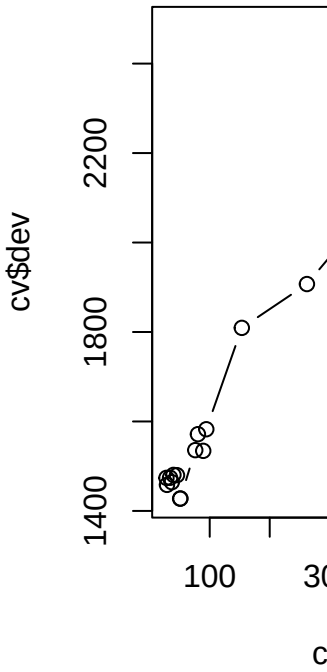
```
##
## Regression tree:
## tree(formula = Sales ~ ., data = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Income"      "CompPrice"
## [6] "Advertising"
## Number of terminal nodes:  16
## Residual mean deviance:  2.572 = 781.9 / 304
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.45400 -1.07000 -0.05544  0.00000  1.14500  4.69600
```

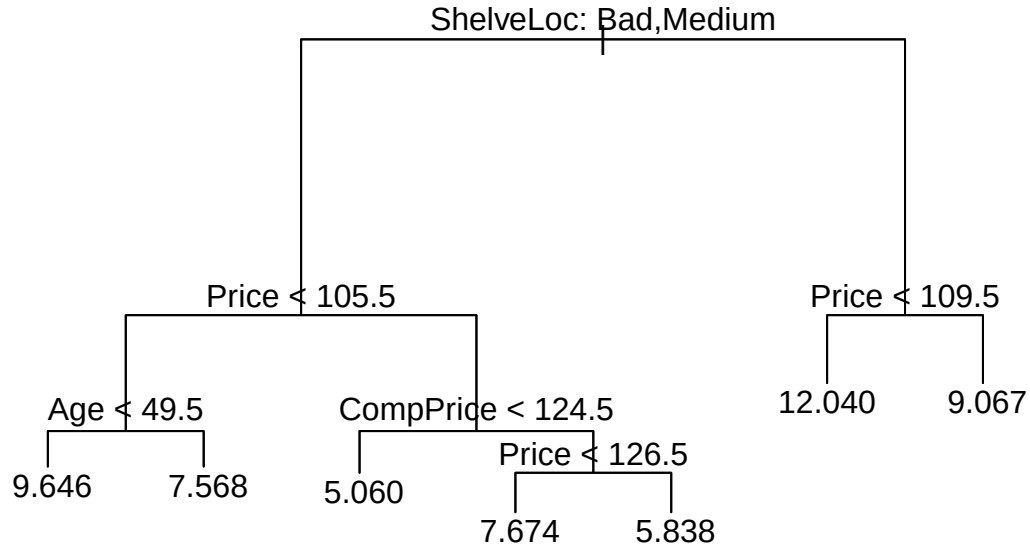
```
## [1] 4.936081
```

- Test MSE for tree model is 4.94

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








```
## [1] 5.203973
```

- Test MSE for pruned tree model is 5.2. Pruning didn't really help in reducing test-error

(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.

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

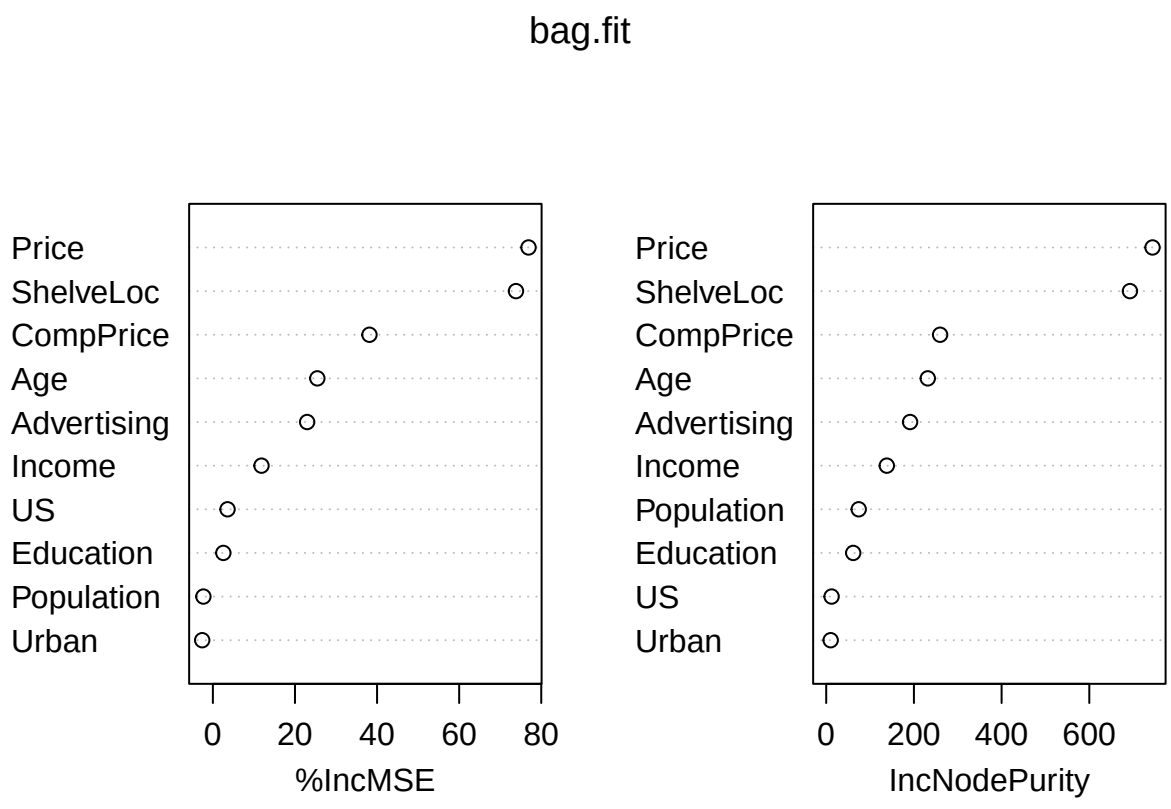
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = train, mtry = 10, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.403858
##                     % Var explained: 68.92

## [1] 2.945423
```

- Test MSE for model using the bagging approach is 2.95

```
##               %IncMSE IncNodePurity
## CompPrice   38.143176     259.77221
## Income      11.839999     138.20846
## Advertising 22.964249     191.25839
## Population  -2.309923      74.13451
## Price       76.903940     744.20064
## ShelveLoc   73.841154     692.64875
## Age         25.449768     231.66005
## Education    2.547928      61.58542
## Urban       -2.600879      10.15212
## US           3.572899      12.28877
```



- Price, ShelveLoc,CompPrice, Age are the most important predictors

(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] 5.432749
```

```
## [1] 4.045937
```

```
## [1] 3.41076
```

```
## [1] 3.190404
```

```
## [1] 3.078291
```

```
## [1] 2.911085
```

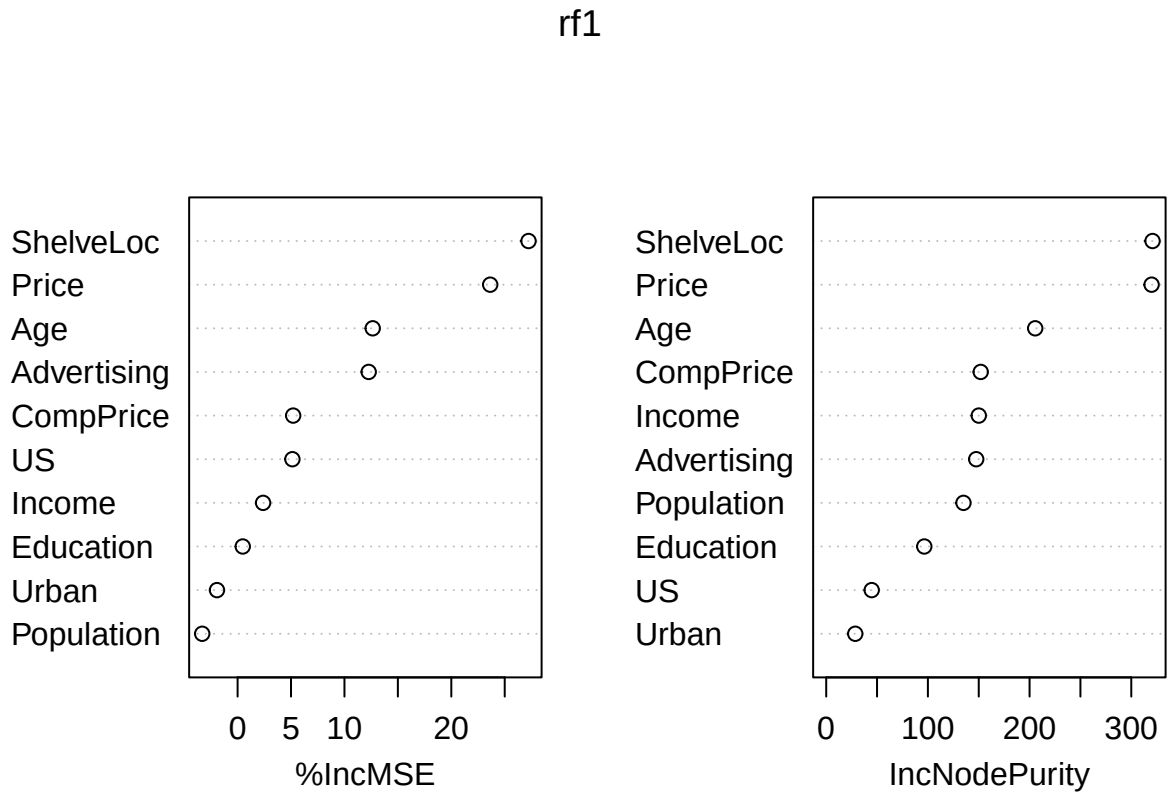
```
## [1] 2.89325
```

```
## [1] 2.916193
```

```
## [1] 2.945423
```

- Test MSE is lowest for mtry=9. MSE decreased as mtry increased from mtry=1-9 and increased a bit for mtry=10

```
##                %IncMSE IncNodePurity
## CompPrice    5.2016542     151.94693
## Income       2.3859682     150.03982
## Advertising 12.2694927     147.51034
## Population  -3.3113953     134.98232
## Price       23.6418747     319.97282
## ShelveLoc   27.2294836     320.91590
## Age         12.6451005     205.50891
## Education    0.4820841      96.46014
## Urban       -1.9290683      28.56456
## US           5.1289356      44.75617
```



- Price, ShelveLoc, Age are the most important predictors. Competitors' prices seem to have lesser impact in random forest model in comparison to bagging model, and income seems to have more impact in the random forests model than it did in the bagging model.

Chapter 8: #11

This question uses the Caravan data set.

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

(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?

```
## Loaded gbm 2.1.8
```

(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?

Gradient Boosting Model

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

	0	1
0	4410	123
1	256	33

Generalized linear Model

	0	1
0	4183	350
1	231	58

Other Problems

1.Beauty Pays!

1.1. 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.

Linear Regression Models with all predictors

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31494 -0.29771  0.01316  0.27795  1.06900
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.03717    0.05900  68.432  < 2e-16 ***
## BeautyScore  0.31522    0.02928  10.765  < 2e-16 ***
## female      -0.34492    0.04622  -7.463 6.30e-13 ***
## lower       -0.30300    0.04877  -6.212 1.43e-09 ***
## nonenglish  -0.30737    0.09656  -3.183  0.00158 **
## tenuretrack -0.07625    0.05529  -1.379  0.16871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.431 on 364 degrees of freedom
## Multiple R-squared:  0.3284, Adjusted R-squared:  0.3192
## F-statistic:  35.6 on 5 and 364 DF,  p-value: < 2.2e-16
```

From the multiple linear regression summary output with all variables, it seems that all the predictors except "tenuretrack" are important.

From the Coefficients of Regression output -

- BeautyScore seems like having a positive impact on overall score.
- Non-English has negative impact on score as people may find instructors with non-English as native language little hard to understand due to the inherent accent.
- Lower class and Gender also has negative impact on score which indicates students may rate from the inherent biases

1.2. 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? From the data, we have concluded that, keeping all other things constant, higher beauty score leads to better course ratings. However, it's hard to tell whether higher beauty leads to higher productivity, as i believe that beauty can't make them better teachers, may be it can make them more confident but it is not the only way that one can be more confident. May be they are regarded as better teachers due to the aesthetic appeal.

However it could lead to discrimination as the ratings as the Beauty is not the metric that is being judged for during course evaluations

2.Housing Price Structure

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

```
##    Home Nbhd Offers SqFt Brick Bedrooms Bathrooms  Price N1 N2 N3 N_Old
## 1     1    2      2 1790     0        2         2 114300  0  1  0     1
## 2     2    2      3 2030     0        4         2 114200  0  1  0     1
## 3     3    2      1 1740     0        3         2 114800  0  1  0     1
## 4     4    2      3 1980     0        3         2  94700  0  1  0     1
## 5     5    2      3 2130     0        3         3 119800  0  1  0     1
## 6     6    1      2 1780     0        3         2 114600  1  0  0     1
## 7     7    3      3 1830     1        3         3 151600  0  0  1     0
## 8     8    3      2 2160     0        4         2 150700  0  0  1     0
## 9     9    2      3 2110     0        4         2 119200  0  1  0     1
## 10   10    2      3 1730     0        3         3 104000  0  1  0     1
```

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

```
##
## Call:
## lm(formula = Price ~ N3 + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27246.3  -7908.8   -519.4   6919.8  26274.2
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   397.26    9812.11   0.040  0.96779
## N3          22834.16    2959.01   7.717 1.18e-11 ***
## Offers      -8078.79    1156.22  -6.987 3.83e-10 ***
## SqFt           53.39       6.30   8.475 2.98e-13 ***
## Brick       15303.21    2320.09   6.596 2.38e-09 ***
## Bedrooms     4031.36    1920.59   2.099  0.03847 *
## Bathrooms    8143.19    2623.06   3.104  0.00251 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10590 on 95 degrees of freedom
## Multiple R-squared:  0.8634, Adjusted R-squared:  0.8548
## F-statistic: 100.1 on 6 and 95 DF,  p-value: < 2.2e-16
```

```
##                    0.5 %      99.5 %
## (Intercept) -25394.59848 26189.12836
## N3           15056.17091 30612.15422
## Offers      -11118.00631 -5039.56566
## SqFt            36.82817    69.94625
## Brick         9204.67050 21401.74058
## Bedrooms     -1017.04617  9079.77194
## Bathrooms     1248.28347 15038.09632
```

- From the multiple linear regression summary output with all variables, it seems that Brick (Yes/No) and Neighborhood3(Y/N) are important variables.
- With Brick(Y/N) Regression coefficient at 15303.2 and 99% confidence interval falling in the range [9204.7,21401.7], it is evident that buyers are okay to premium for Brick, as it rejects Null Hypothesis for this variable
- With Neighborhood3(Y/N) Regression coefficient at 22834.2 and 99% confidence interval falling in the range [15056.1, 30612.1], it is evident that buyers are okay to premium for Neighborhood3, as it rejects Null Hypothesis for this variable

2.3.Is there an extra premium for brick houses in neighborhood 3? For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

```
##
## Call:
## lm(formula = Price ~ N3 * Brick + N3 + Offers + SqFt + Brick +
##     Bedrooms + Bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27177.7  -6270.5   -528.6   5313.4  25314.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   608.032   9399.382   0.065  0.94856
## N3          16595.047   3481.128   4.767 6.80e-06 ***
## Brick       10914.687   2638.150   4.137 7.65e-05 ***
## Offers      -8509.779   1116.323  -7.623 1.96e-11 ***
```

```
## SqFt           55.657      6.079   9.155 1.15e-14 ***
## Bedrooms     4515.010   1846.411   2.445  0.01634 *
## Bathrooms    6635.210   2559.694   2.592  0.01106 *
## N3:Brick    14886.148   4821.726   3.087  0.00265 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10140 on 94 degrees of freedom
## Multiple R-squared:  0.876,  Adjusted R-squared:  0.8668
## F-statistic: 94.86 on 7 and 94 DF,  p-value: < 2.2e-16
```

```
##                     0.5 %      99.5 %
## (Intercept) -24104.33197 25320.39588
## N3            7442.64622 25747.44722
## Brick         3978.60221 17850.77267
## Offers      -11444.75808 -5574.79962
## SqFt           39.67444    71.64026
## Bedrooms      -339.47774  9369.49838
## Bathrooms      -94.60294 13365.02328
## N3:Brick      2209.11863 27563.17777
```

- What we have been doing in part 1 & 2 is seeing the effect of Neighbourhood3 and Brick individually. Now, to find if buyers would be interested in paying premium for brick house in Neighborhood3, we have added interaction term N3:Brick to model and the results of multiple linear regression suggests that the variable is slightly less important in comparison to others.
- The combined effect is lower than the individual effect. Neighborhood3(Y/N): Brick(Y/N) Regression coefficient at 14907 and 99% confidence interval falling in the range [1961.7, 27852.4]. Hence, Null Hypothesis can be rejected as 0 is not part of this range. Hence we can conclude that people are ready to premium for Brick houses in Neighbourhood 3, though the spread of coefficent is high

3.What causes what??

3.1. 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? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)

- We can't directly run a regression model because we are not sure whether more crime leads to more police on ground or vice versa (classic chicken egg problem). However, they are positively correlated, but we are not sure of causal relationship between these 2 variables. Doing random experiments by changing number of cops in the street might not be a feasible experiment to do. However, to see the effect of having more cops in the streets, may be running a natural experiment like the one suggested in podcast(during high terrorist alert days) would help in knowing the actual relationship.
- The immediate reaction by the state to reducing crime would be to hire more cops to reduce crime rate*

3.2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below. Approach: Researchers from UPENN have taken data from the days when there is high alert for terrorist activity in DC. That way, there are more cops on the street and this is a natural experiment because it's imposed by state and is not related to crime in the region, hence they would be able to isolate this effect.

Results: * Low R2 for both the models indicate that there is scope in improvement for Regression model through adding more predictors.

- R2 improved by adding an extra variable metro ridership, indicates that Crime rate is dependent not only on nuber of police in the street, but also number of people on the streets. Hence we can't infer that increase in number of cops in the street alone has reduced crime rate

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

- They had to control for METRO ridership to be able to capture the impact of having more cops on crime. In the case where METRO ridership is taken as well, the coefficient of High Alert is -ve still for 99% confidence interval, hence we can reject null hypothesis for this variable
- However, we can't still conclude that high alert days have lower crime rate solely because of number of cops in the street. Due to the alert, criminals might also not come out as much as they do on normal days, due to which the crime rate could have been low.

3.4. 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?

- To test the high alert effect further interactive terms are introduced.
- The model is built to check how is crime rate effected by high alert in District1 vs all other districts.
- To test that they have taken interactive terms for High Alert and District Type(1/Others)
- Conclusion: There is effect of high alert on crime rate only in District 1

Problem 4: Neural Nets

Re-run the Boston housing data example using a single layer neural net. Cross validate for a few choices of size and decay parameters

```
## The following objects are masked from Boston (pos = 13):
##
##     age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio, rad,
##     rm, tax, zn

## The following objects are masked from Boston (pos = 32):
##
##     age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio, rad,
##     rm, tax, zn
```

Neural Net Model

```
## # weights:  91
## initial  value 454.442287
## iter  10 value 178.777522
## iter  20 value 123.242226
## iter  30 value 94.304036
```

```
## iter  40 value 83.316862
## iter  50 value 77.347629
## iter  60 value 72.246551
## iter  70 value 69.800299
## iter  80 value 68.204556
## iter  90 value 66.961741
## iter 100 value 66.584158
## final  value 66.584158
## stopped after 100 iterations
```

```
## Neural Net Test RMSE with decay:0.1 and size:6 is  0.9924577
```

Linear Regression Model

```
## Linear Regression model Test RMSE:  0.9670259
```

Neural Net Model parameter tuning

```
## Neural Net sizes:3,5,8,10,15,20
```

```
## [1] " "
```

```
## Decay Parameters:0.0001, 0.001, 0.01, 0.1, 0.3, 0.5
```

```
## Minimum RMSE obtained on Cross Validation = 0.945
```

Problem 5: Final Project

Describe your contribution to the final group project (1 page max). Group Number:3

Project Title:Job change propensity for data scientists (HR Analytics)

Team Members:

- Rushiil Deshmukh
- Ramya Madhuri Desineedi
- Jackson Hittner
- Yashpreet Kaur
- Kaushik Kumaran
- Leyang Xu

Data Set Selection: We have referred to data repository from Kaggle and Government websites to find a dataset to work for group project. In our initial group meeting, each of us have expressed our interest to work on business problems from various industries such as Healthcare, Education, Real-estate, etc. . .

After a series of discussions, we finally boiled down to a business problem of a company which is active in Big Data and Data Science and wants to hire data scientists among people who successfully pass some courses which conduct by the company. We particularly took this dataset as we wanted to understand how data analytics can be used in hiring data scientists/analysts.

Business Context: Many people signup for their training and company wants to know which of these candidates are really wants to work for the company after training or looking for a new employment because it

helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.(Source: www.Kaggle.com)

Project Objective: Predict the probability of a candidate will work for the company

Initial Division of work: We have decided to split the group into 2 sub-groups based on skillset and interests - Modelling and presentation. 4 of us have taken up EDA and modelling, 2 of us have taken up making presentation. I was part of modelling team as i wanted to try my hands-out by modelling in R.

Exploratory data analysis: Using Excel, I have done preliminary Data Exploration and data cleaning and discussed the patterns with the group, which we have used in grouping data (city) while modelling.

Models:

- I have taken up Randomforest model and run the classification model on data after cleaning.
- Calculated the variable importance and compared with other models if it is similar or not
- Used 10-fold cross-validation to find training error and compared test error with other models
- Set a threshold to classify the output into 2 classes of target (0/1) for maximum accuracy
- We finalized on GBM model which has slightly better accuracy than Logistic Regression and Random Forest Models

Presentation:

- Designed and presented slides for Data description, challenges and modelling section
- Updated modelling results in powerpoint
- Ideated on quantifying impact of our analysis
- Collaborated with others to make it consistent with other pieces in the deck