Take Home Exam - STA380 - ISLR Edition 1

Apurva Audi

7/31/2022

Book Problems :

Chapter 2 - Question 10 - Section a

```
## [1] "Columns in the Boston dataset"
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
## [1] "Dimensions in the Boston dataset :"
```

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

```
## [1] "Number of rows :"
```

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

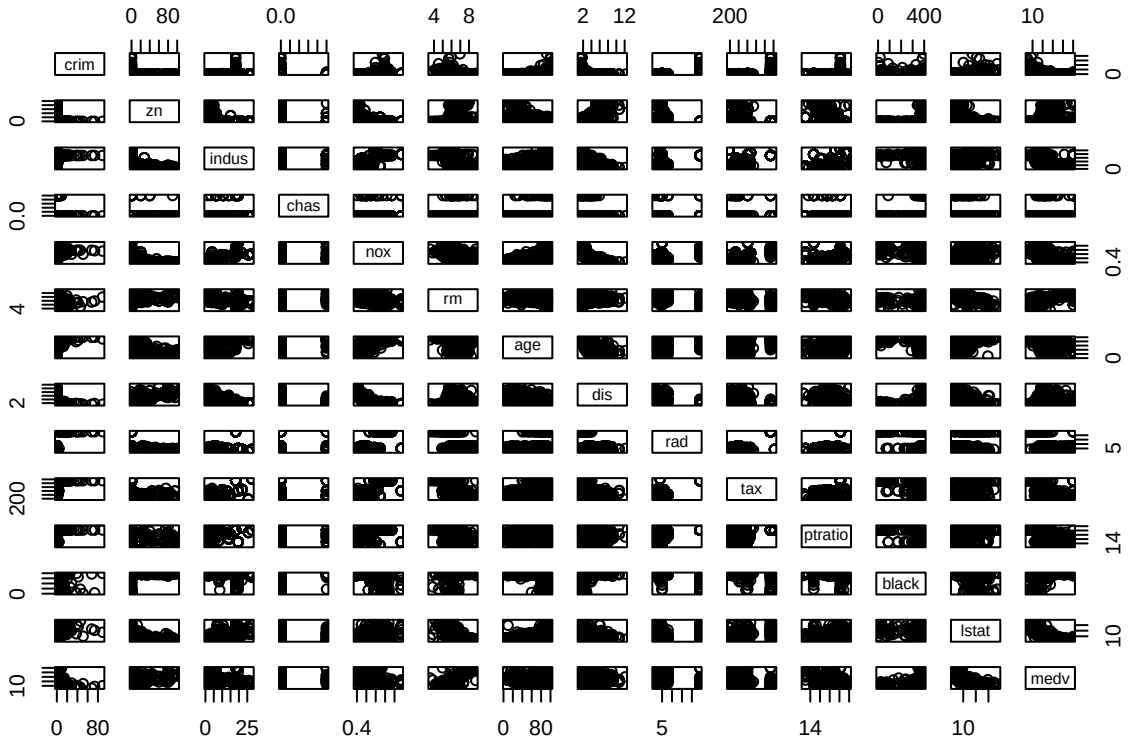
```
## [1] "Number of columns :"
```

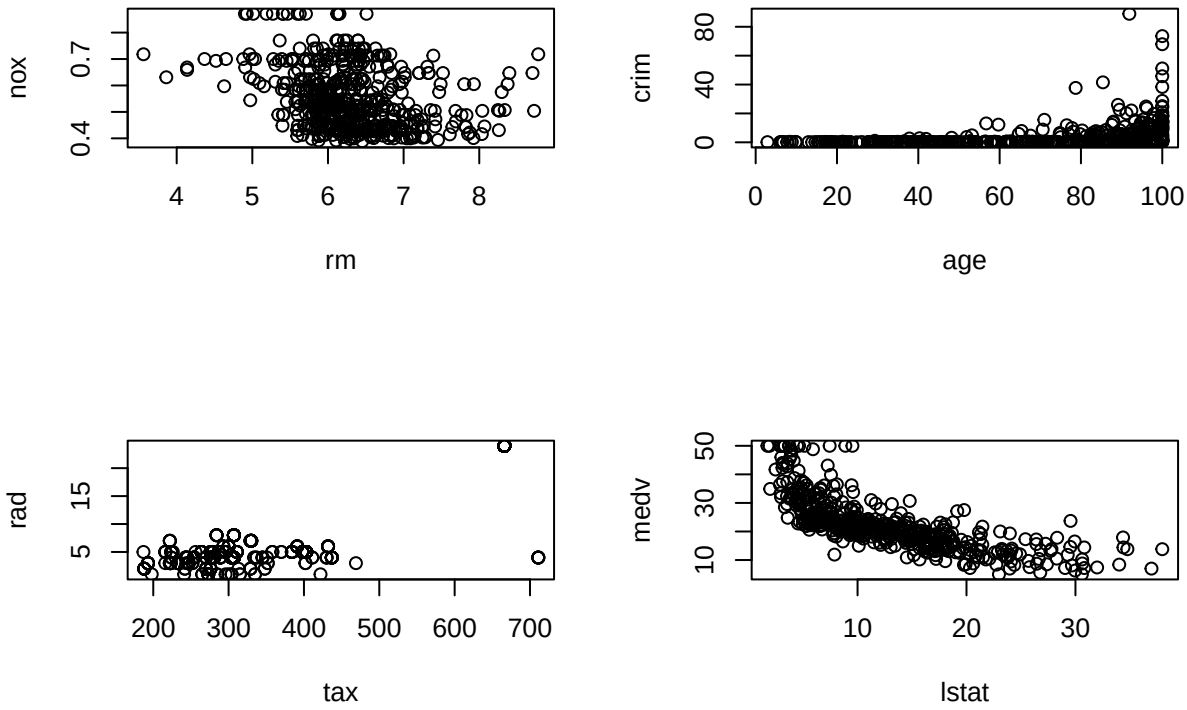
```
## [1] 14
```

There are 506 records in the data set Boston and 14 columns. The records are the different suburbs in Boston and columns are the features of the houses.

Chapter 2 - Question 10 - Section b

Pairwise plots for each predictor

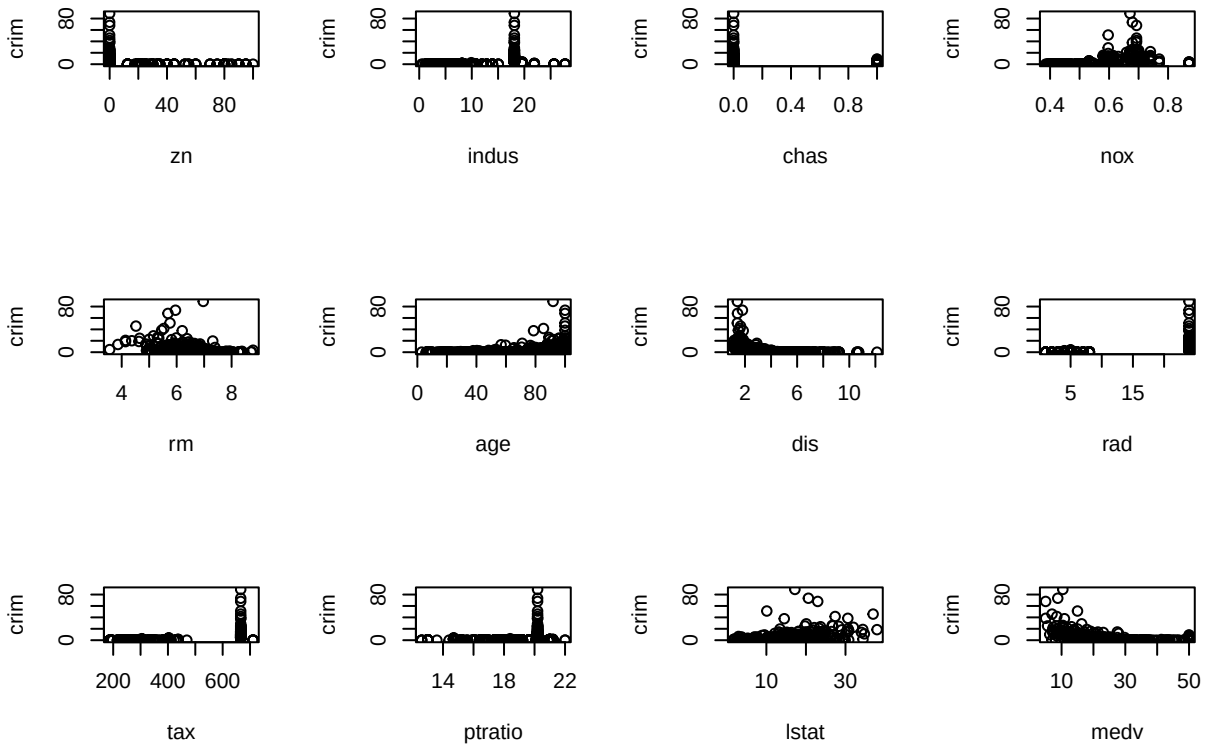

crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv



By using few of the predictors from the data set and creating some pairwise scatter plots, we can find the following :

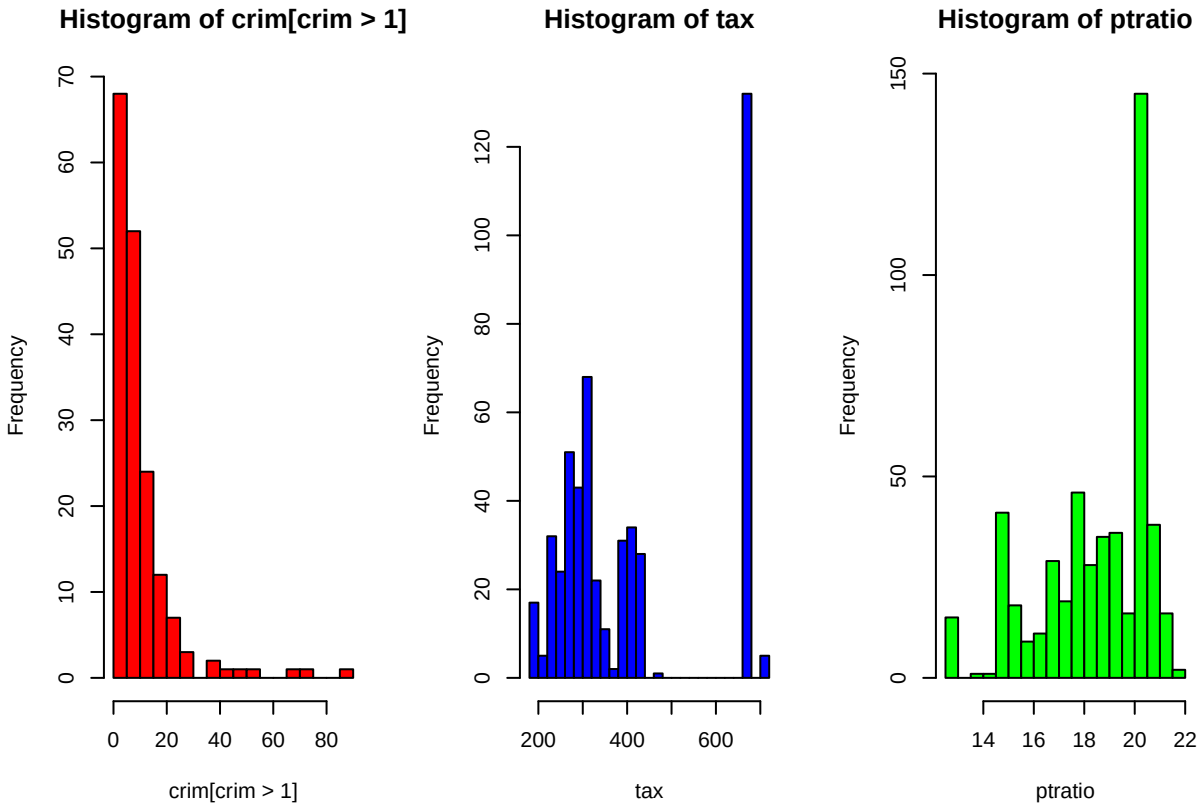
- rm & nox: no correlation.
- age & crim : positive correlation as there is exponential increase in per capita crime rate once the proportion of dwellings before 1940 rises.
- rad & tax : no clear correlation between accessibility to highways and property tax rate.
- lstat & medv : inverse relation between lower status and median value of house.

Chapter 2 - Question 10 - Section c



Best predictors are age and dis as there is a clear positive and negative correlation respectively with crim (per capita crime rate).

Chapter 2 - Question 10 - Section d



Analysis from the Histogram plots :

- Most suburbs have low crime rates with a small proportion of suburbs with crime rate > 20 and reaching above 80.
- There are two distinct groups of suburbs with low tax below 500 and high tax above 650.
- The dataset has more suburbs with high pupil to student ratio - ptratio which is a case of left skewed with highest frequency between 20 - 22.

Chapter 2 - Question 10 - Section e

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

35 suburbs are bound by Charles River as per the subset taken by this dataset.

Chapter 2 - Question 10 - Section f

```
## [1] 19.05
```

The median value for pupil-teacher ratio among the towns in this dataset is 19.05

Chapter 2 - Question 10 - Section g

```
##             399      406
## crim     38.3518  67.9208
## zn        0.0000   0.0000
## indus    18.1000  18.1000
## chas      0.0000   0.0000
## nox       0.6930   0.6930
## rm        5.4530   5.6830
## age     100.0000 100.0000
## dis       1.4896   1.4254
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## black   396.9000 384.9700
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

```
##       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
```

Record 399 and 406 have the least medv (median value of owner occupied homes). Values of other parameters for the min value of medv found for record 399 and 406.

crim - **above 3rd quartile**
zn - **minimum value**
indus - **at 3rd quartile**
chas - **minimum value - not bound by river**

nox - **above 3rd quartile**
rm - **above minimum value, below 1st quartile**
age - **maximum value**
dis - **above minimum value, below 1st quartile**
rad - **maximum value**
tax - **at 3rd quartile**
ptratio - **at 3rd quartile**
lstat - **above 3rd quartile, below maximum value**
medv - **at minimum value**

Chapter 2 - Question 10 - Section h

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

64 suburbs have average of more than 7 rooms per dwelling.

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

13 suburbs have average of more than 8 rooms per dwelling.

```
## [1] "Summary of Boston suburbs with more than eight rooms per dwelling"
```

```
##       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
```

```
## [1] "Summary of entire dataset of Boston suburbs"
```

```
##       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
```

Based on the summary of Boston suburbs and subset of Boston suburbs with average of more than 8 dwellings, we can conclude that the Boston suburbs with average more than 8 dwellings have **lower crime rate and lstat values**.

Chapter 3 - Question 15 - Section a

```
##       crim                zn              indus       chas         nox        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   N:471   Min.   :0.3850  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   Y: 35   1st Qu.:0.4490  
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710  
##        rm              age              dis              rad        
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000  
##  Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000  
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549  
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
##       tax           ptratio          black            lstat       
##  Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
```

```
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95
##  Median :330.0   Median :19.05   Median :391.44   Median :11.36
##  Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95
##  Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

As chas has 0 and 1 value based on whether the suburb is bound by river or not, we have added categories of N - not bound by river, Y - bound by river in the Boston dataset for it.

In simple linear regression models, if p-value is lesser than 0.05, it means that you can reject null hypothesis. Thus, the predictor add value to the prediction of the estimator and hence, is a significant predictor.

Linear Regression models to predict estimator - crim

```
lm_zn <- lm(crim~zn)
summary(lm_zn)
```

```
##
## 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
```

zn is a good predictor of crim based on its low p-value.

```
lm_indus <- lm(crim~indus)
summary(lm_indus)
```

```
##
## 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
```

indus is a good predictor of crim based on its low p-value.

```
lm_chas <- lm(crim~chas)
summary(lm_chas)
```

```
##
## 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
```

chas is a bad predictor of the response crim due to its low p-value and extremely low Adjusted R-squared value which signifies that this predictor is unable to explain the variability of crim.

```
lm_nox <- lm(crim~nox)
summary(lm_nox)
```

```
##
## 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
```

nox is a good predictor of crim based on its low p-value.

```
lm_rm <- lm(crim~rm)
summary(lm_rm)
```

```
##
## 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
```

rm is a good predictor of crim based on its low p-value.

```
lm_age <- lm(crim~age)
summary(lm_age)
```

```
##
## 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
```

age is a good predictor of crim based on its low p-value.

```
lm_dis <- lm(crim~dis)
summary(lm_dis)
```

```
##
## 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
```

dis is a good predictor of crim based on its low p-value.

```
lm_rad <- lm(crim~rad)
summary(lm_rad)
```

```
##
## 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
```

rad is a good predictor of crim based on its low p-value.

```
lm_tax <- lm(crim~tax)
summary(lm_tax)
```

```
##
## 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
```

tax is a good predictor of crim based on its low p-value.

```
lm_ptratio <- lm(crim~ptratio)
summary(lm_ptratio)
```

```
##
## 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
```

ptratio is a good predictor of crim based on its low p-value.

```
lm_black <- lm(crim~black)
summary(lm_black)
```

```
##
## 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
```

black is a good predictor of crim based on its low p-value.

```
lm_lstat <- lm(crim~lstat)
summary(lm_lstat)
```

```
##
## 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
```

lstat is a good predictor of crim based on its low p-value.

```
lm_medv <- lm(crim~medv)
summary(lm_medv)
```

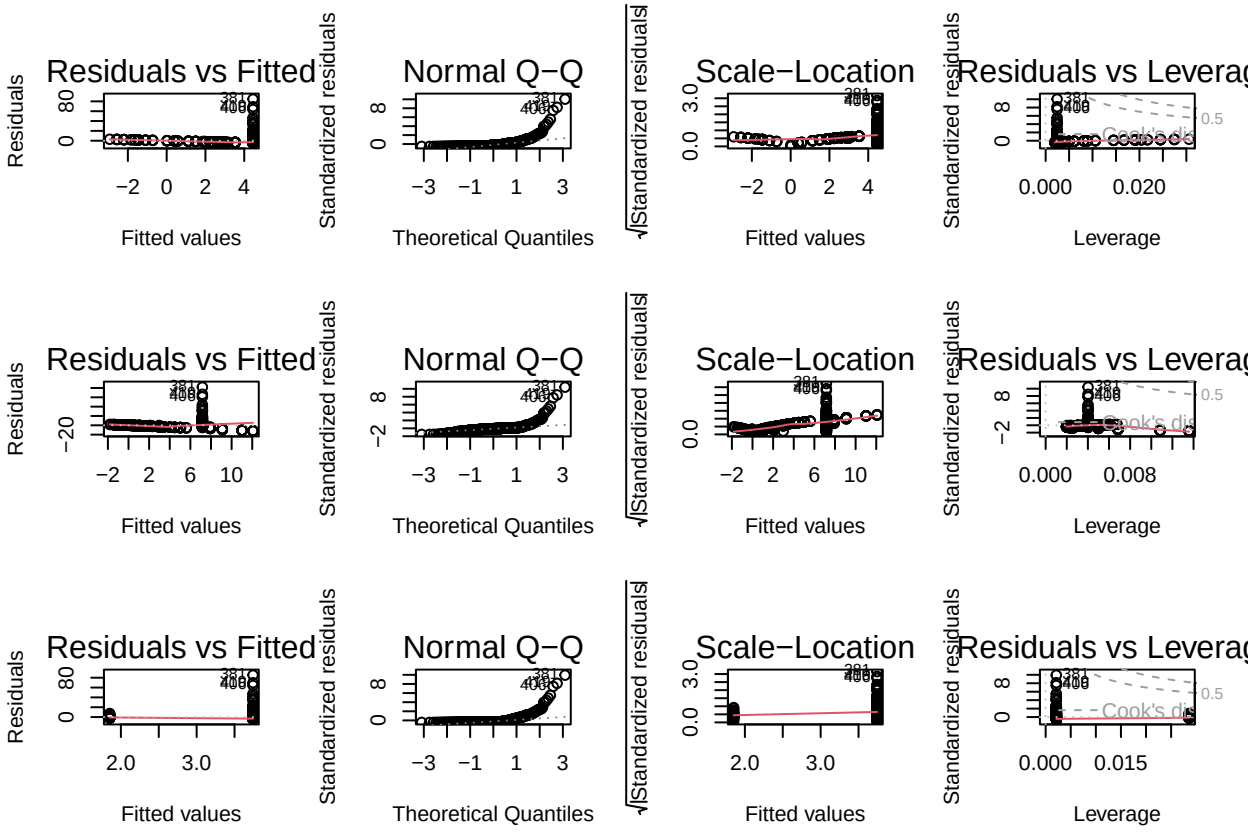
```
##
## Call:
## lm(formula = crim ~ medv)
##
```

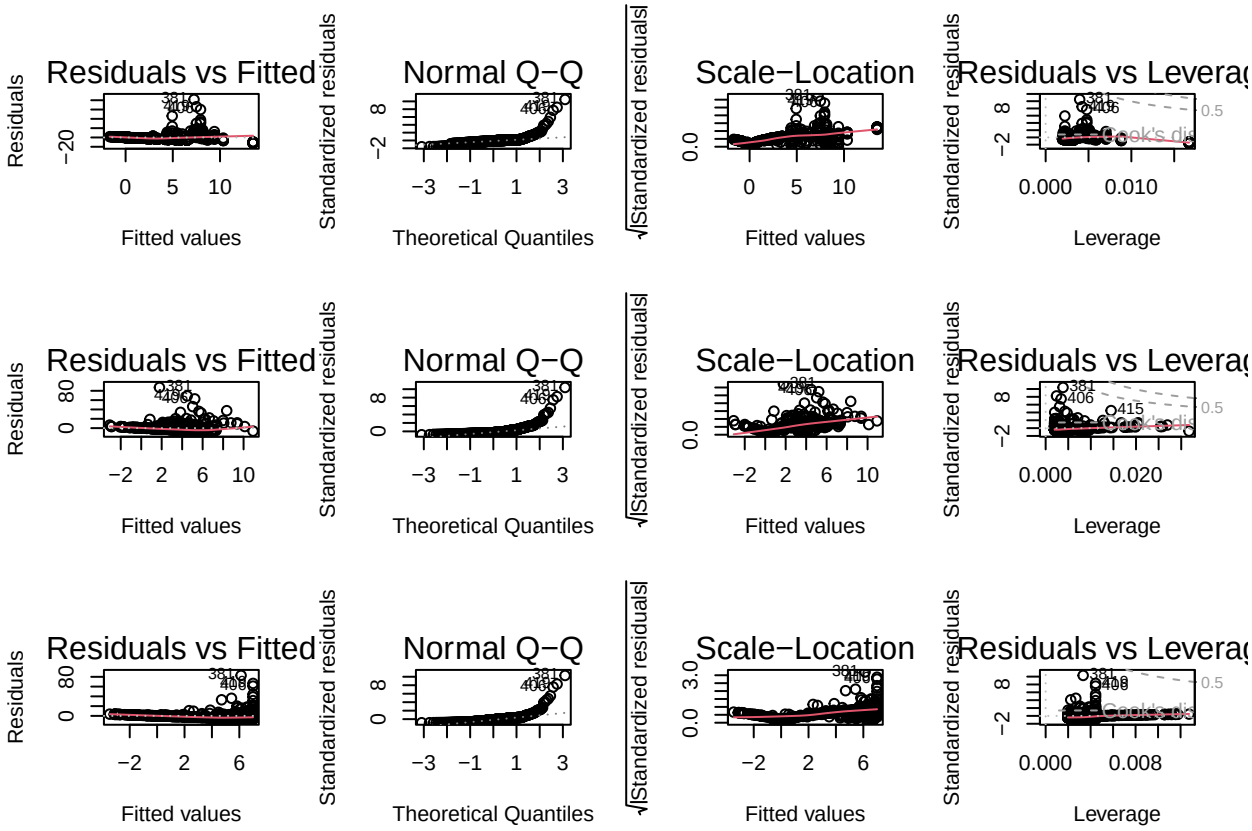
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. 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
```

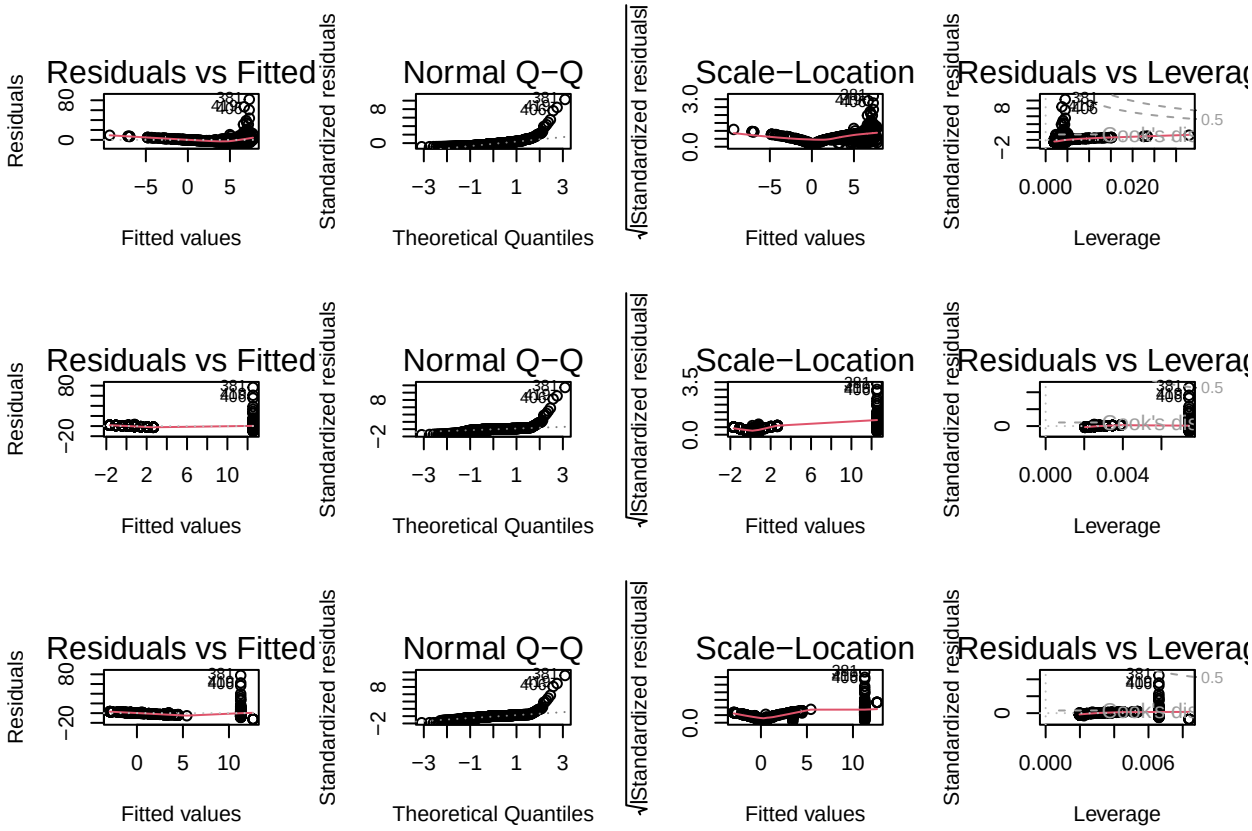
medv is a good predictor of crim based on its low p-value.

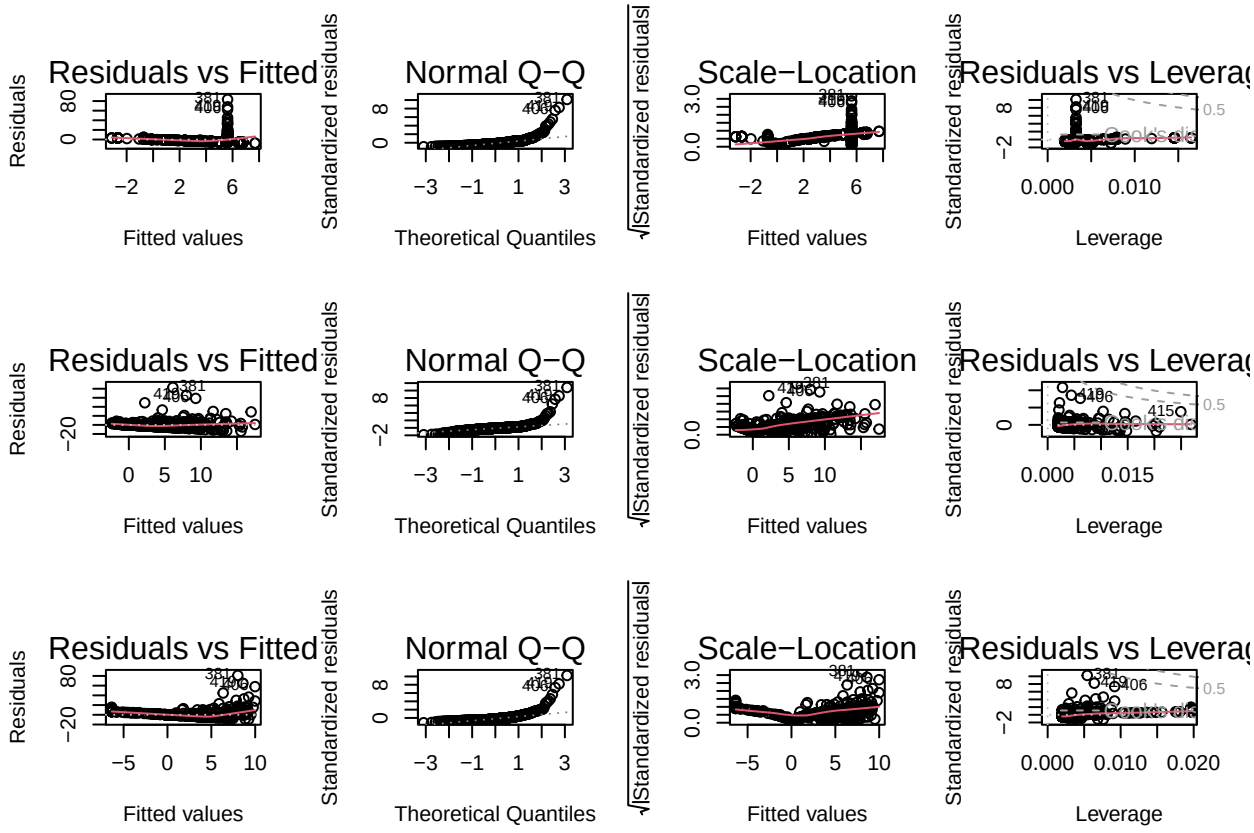
So, all predictors (except chas) have a statistically significant association with the response - crim based on the above generated linear models.

Plotting Linear Regression residuals to showcase the results of linear models.




Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverag
Standardized residuals
Leverage
Cook's dis
0.5
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverag
Standardized residuals
Leverage
Cook's dis
0.5
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverag
Standardized residuals
Leverage
Cook's dis
0.5


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Levera
Standardized residuals
Leverage
Cook's dis
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Levera
Standardized residuals
Leverage
Cook's dis
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Levera
Standardized residuals
Leverage
Cook's dis



Chapter 3 - Question 15 - Section b

```
lm_all_predict <- lm(crim~., data=Boston)
summary(lm_all_predict)
```

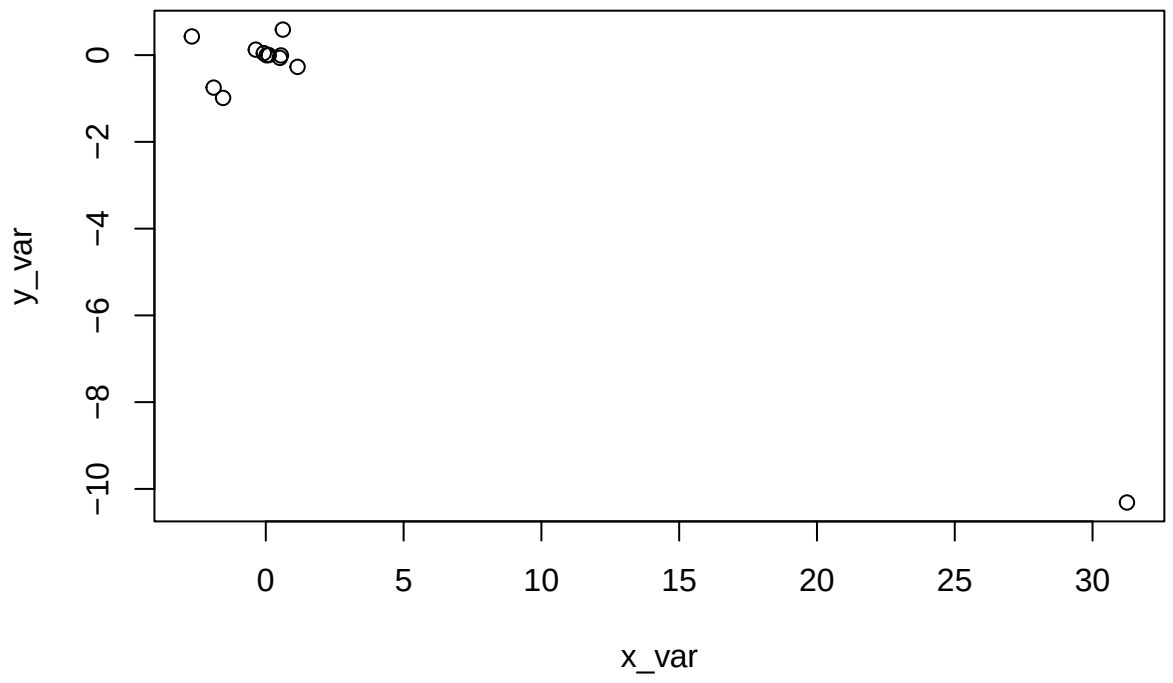
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chasY        -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
```

To reject the null hypothesis, we need to check the Pr(>|t|) corresponding to each of the predictors in the multiple regression model.

For following predictors the Pr(>|t|) is less than 0.05 meaning that it is significant predictor : **zn , dis , rad , black & medv**

Chapter 3 - Question 15 - Section c



The coefficients of all the respective univariate regression models (on X-axis) and the multiple regression models (on Y-axis) are plotted for all predictors.

All the coefficients pairs for predictors in their univariate and multiple regression model are closer to that of other predictors. However, the variable coefficient for nox is the farthest with approximately 31 in univariate and -10 in multiple regression model.

Chapter 3 - Question 15 - Section d

To understand if there is any non-linear association between any of the predictors and the response, each predictor is fit in polynomial form with degree 3.

We compare the R-squared value of polynomial fit and to corresponding linear models. If the R-squared value for polynomial model is higher than linear model, then polynomial model explains the variability of response better than linear models.

As chas is a qualitative predictor and thus, isn't included in this non-linear model.

```
lm_zn <- lm(crim~poly(zn,3))
summary(lm_zn)
```

```
##
## 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
```

The R-squared value for linear model of zn is 0.03828 but for polynomial fit, it is 0.05261. Thus, non-linear association of zn with response is a better fit due to higher R- squared value.

```
lm_indus <- lm(crim~poly(indus,3))
summary(lm_indus)
```

```
##
## 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
```

The R-squared value for linear model of indus is 0.1637 but for polynomial fit, it is 0.2552. Thus, non-linear association of indus with response is a better fit due to higher R- squared value.

```
lm_nox <- lm(crim~poly(nox,3))
summary(lm_nox)
```

```
##
## 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
```

The R-squared value for linear model of nox is 0.1756 but for polynomial fit, it is 0.2928. Thus, non-linear association of nox with response is a better fit due to higher R- squared value.

```
lm_rm <- lm(crim~poly(rm,3))
summary(lm_rm)
```

```
##
## 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
```

The R-squared value for linear model of rm is 0.04618 but for polynomial fit, it is 0.06222. Thus, non-linear association of rm with response is a better fit due to higher R- squared value.

```
lm_age <- lm(crim~poly(age,3))
summary(lm_age)
```

```
##
## 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
```

The R-squared value for linear model of age is 0.1227 but for polynomial fit, it is 0.1693. Thus, non-linear association of age with response is a better fit due to higher R- squared value.

```
lm_dis <- lm(crim~poly(dis,3))
summary(lm_dis)
```

```
##
## 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
```

The R-squared value for linear model of dis is 0.1425 but for polynomial fit, it is 0.2735. Thus, non-linear association of dis with response is a better fit due to higher R- squared value.

```
lm_rad <- lm(crim~poly(rad,3))
summary(lm_rad)
```

```
##
## 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
```

The R-squared value for linear model of rad is 0.39 but for polynomial fit, it is 0.3965. Thus, non-linear association of rad with response is a better fit due to higher R- squared value.

```
lm_tax <- lm(crim~poly(tax,3))
summary(lm_tax)
```

```
##
## 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
```

The R-squared value for linear model of tax is 0.3383 but for polynomial fit, it is 0.3651. Thus, non-linear association of tax with response is a better fit due to higher R- squared value.

```
lm_ptratio <- lm(crim~poly(ptratio,3))
summary(lm_ptratio)
```

```
##
## 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
```

The R-squared value for linear model of ptratio is 0.08225 but for polynomial fit, it is 0.1085. Thus, non-linear association of ptratio with response is a better fit due to higher R- squared value.

```
lm_black <- lm(crim~poly(black,3))
summary(lm_black)
```

```
##
## 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
```

The R-squared value for linear model of black is 0.1466 but for polynomial fit, it is 0.1448. Thus, linear association of black with response is a better fit due to higher R- squared value.

```
lm_lstat <- lm(crim~poly(lstat,3))
summary(lm_lstat)
```

```
##
## 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
```

The R-squared value for linear model of lstat is 0.206 but for polynomial fit, it is 0.2133. Thus, non-linear association of lstat with response is a better fit due to higher R- squared value.

```
lm_medv <- lm(crim~poly(medv,3))
summary(lm_medv)
```

```
##
```

```
## 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
```

The R-squared value for linear model of medv is 0.1491 but for polynomial fit, it is 0.4167. Thus, non-linear association of medv with response is a better fit due to higher R- squared value.

So, all predictors except chas and black show a non-linear association with the response crim.

Chapter 6 - Question 9 - Section a

Loading dataset College from ISLR library. No missing values were found in the dataset.

```
## [1] 0
```

Splitting the data set into training and testing set.

```
## [1] "Training set dimensions : "
```

```
## [1] 388  18
```

```
## [1] "Testing set dimensions : "
```

```
## [1] 389  18
```

Chapter 6 - Question 9 - Section b

- Using the lm(), multiple regression model is build on the training set.
- The training set is used to predict the number of applications (Apps) and finally, MSE is computed.

```
## [1] 1387559
```

This is the test MSE for linear model.

Chapter 6 - Question 9 - Section c

- Creating a Ridge regression model with the College dataset and finding the lambda value with the minimum error during cross validation.

```
## [1] 37.64936
```

This is the best lambda value which we can use in our prediction and compute the test MSE for ridge model.

```
## [1] 1545689
```

The test MSE for Ridge model is slightly higher than that of Linear model with least squares.

Chapter 6 - Question 9 - Section d

- Creating a Lasso model with the College dataset and finding the lambda value with the minimum error during cross validation.

```
## [1] 3.053856
```

The best lambda value for Lasso model is obtained through cross validation against the training set.

```
## [1] 1385378
```

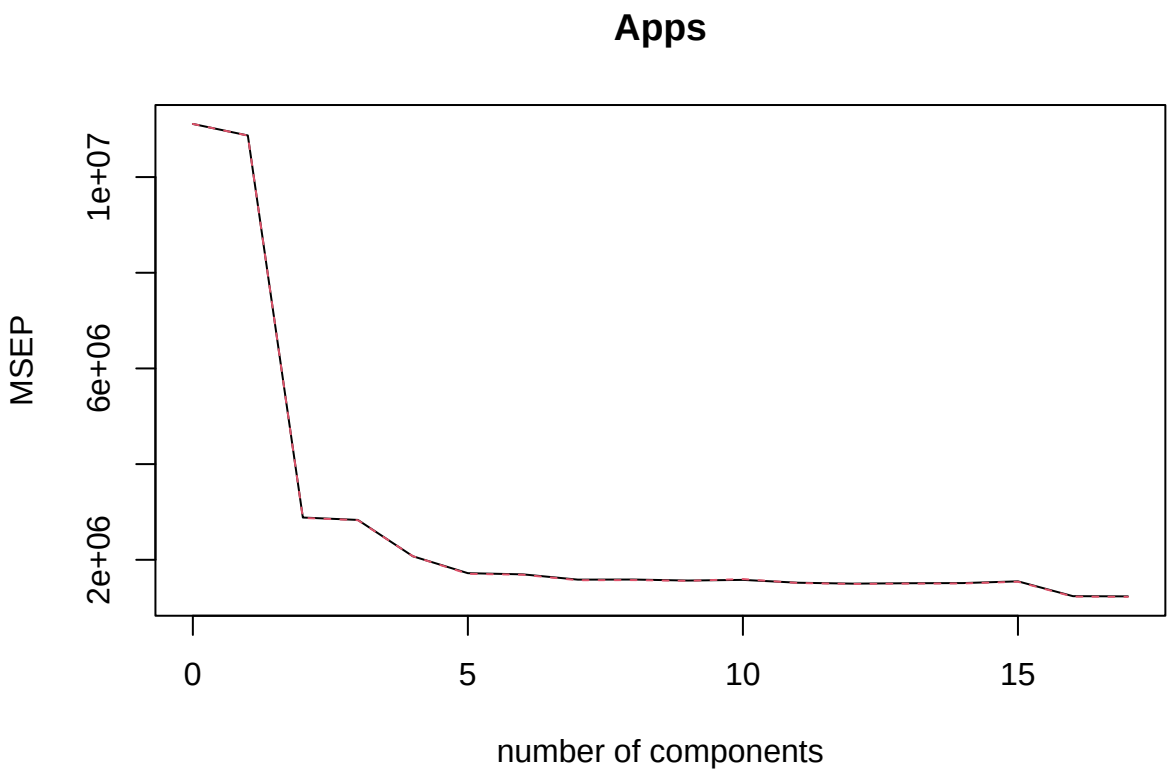
This test MSE is slightly higher than linear model but lower than Ridge model.

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept) -478.64330143
## (Intercept)    .
## PrivateYes  -489.92446319
## Accept         1.56498323
## Enroll        -0.71663618
## Top10perc     47.47015419
## Top25perc    -12.32720583
## F.Undergrad    0.03582449
## P.Undergrad    0.04420581
## Outstate      -0.08224532
## Room.Board     0.14848102
## Books          0.01291890
## Personal       0.02819927
## PhD           -8.28800099
## Terminal      -3.22400745
## S.F.Ratio     14.19347049
## perc.alumni   -0.10834739
## Expend         0.07676428
## Grad.Rate      8.13014741
```

As observed in the output, almost all predictors have non-zero coefficients. When any predictor has a zero coefficient based on the Lasso model, we can remove these predictors from our analysis leading to shrinkage and dimension reduction.

Chapter 6 - Question 9 - Section e

- Using pls library to build a PCR model and validate the total number of predictors that will provide the least error.



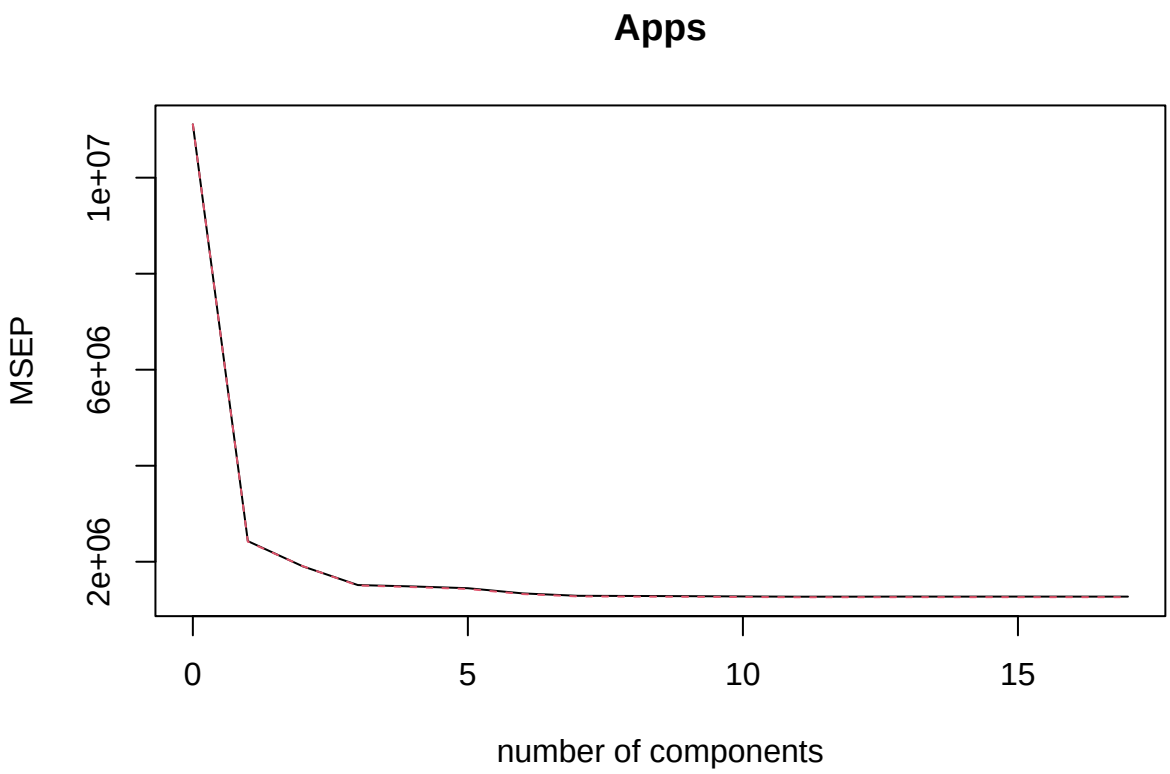
The MSEP exponentially reduces until ncomp = 10 after which there is hardly any change. Thus, we can use 10 as the number of components for our model.

```
## [1] 2996155
```

The test MSE for PCR model is higher than linear model.

Chapter 6 - Question 9 - Section f

- Fitting a PLS model by using M through cross validation.

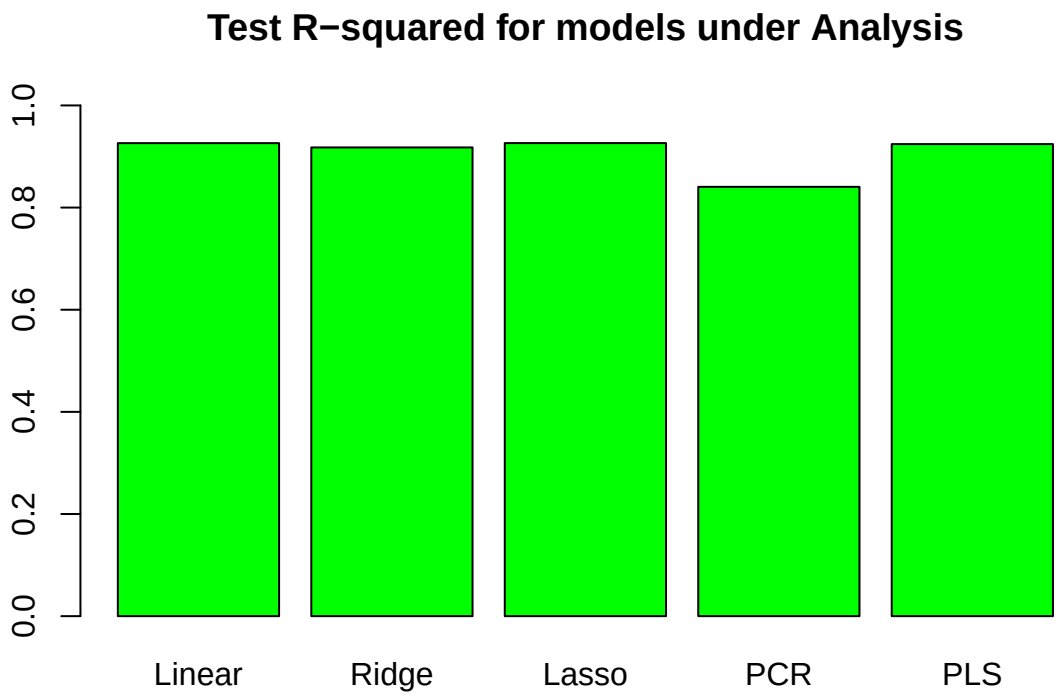



```
## [1] 1423073
```

The test MSE for PSL model is comparable to that of linear model with least squares with just a slight difference.

Chapter 6 - Question 9 - Section g

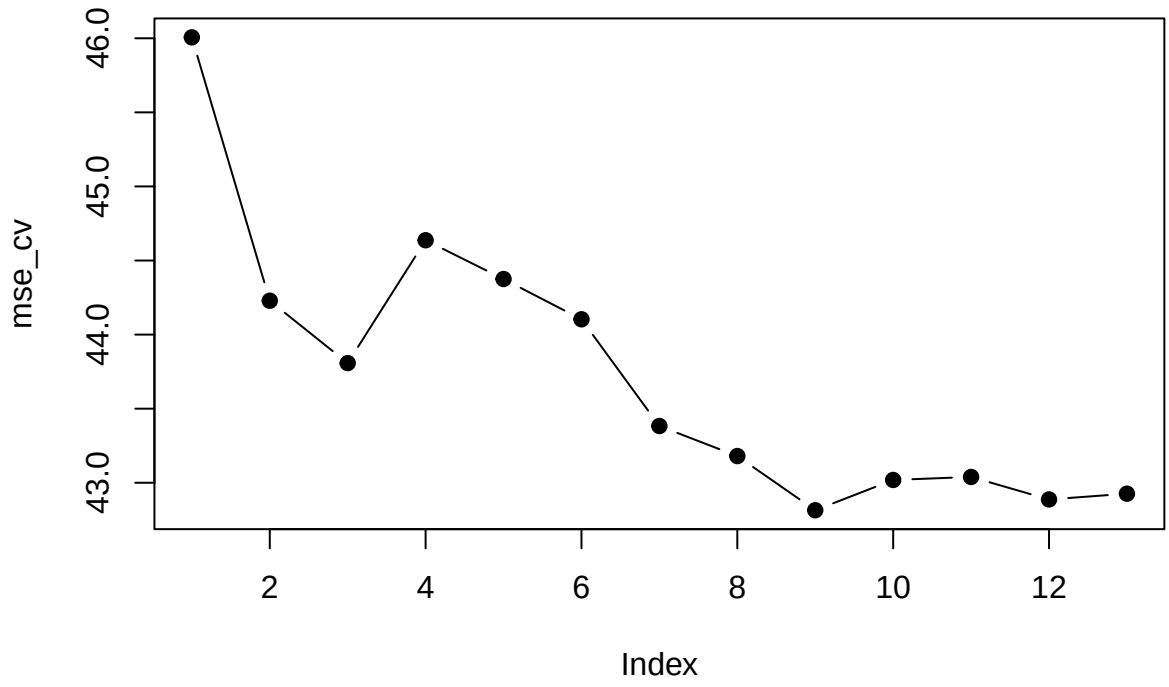
- The results of Linear model with least squares, Ridge and Lasso model have similar results. Based on Lasso model, the coefficients of variables are shrunk and it reduces the F.Undergrad and Books variables to zero.
- Comparing the test R squared values for all models : Linear, Ridge, Lasso, PCR and PLS models.



This bar plot shows that all models except for PCR have a test R squared value of 0.9, although PLS model has a slightly higher R squared. Any of the models except PCR can be used to predict the number of applications (Apps) with about 90% accuracy.

Chapter 6 - Question 11 - Section a

- Best Selection Method used with the data set Boston to predict per capita crime rate



This plot shows that for the best index at which the MSE value is least.

```
## [1] "Index for the lowest MSE : "
```

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

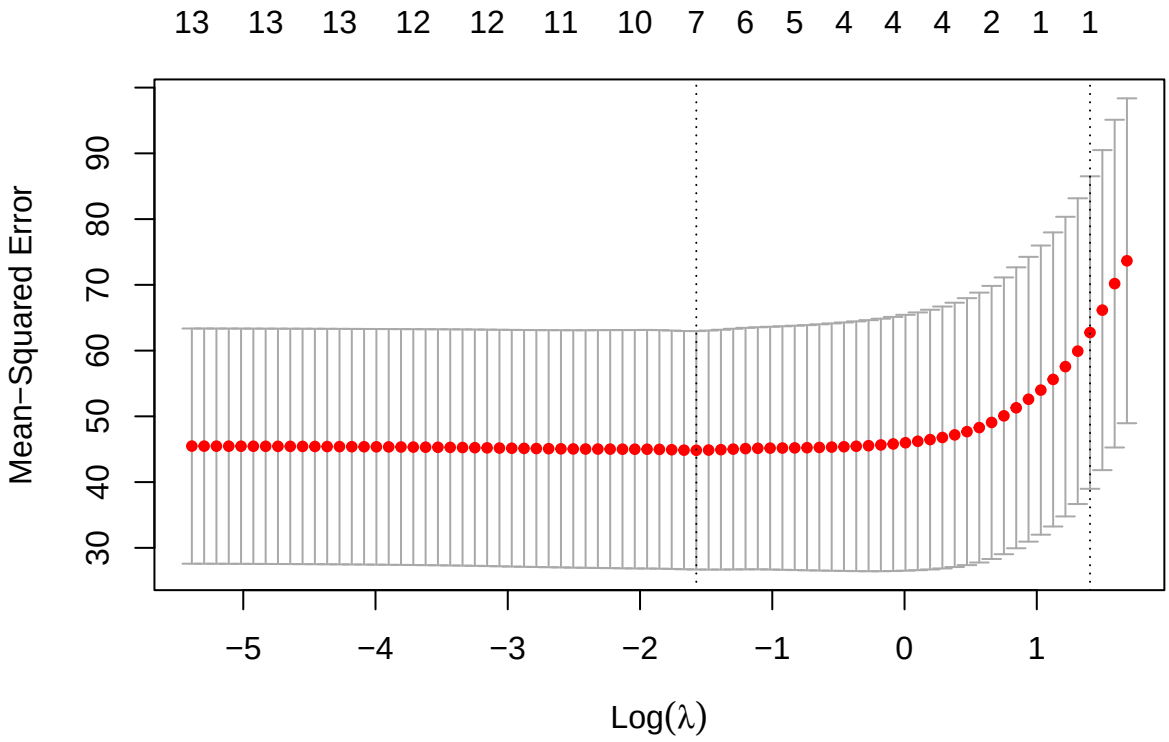
```
## [1] "Minimum MSE value :"
```

```
## [1] 42.81453
```

This shows the lowest MSE value obtained via best selection method.

Using Lasso Regression to predict per capita crime rate

Plot for cross validation in Lasso model



Coefficients for Lasso model cross validation

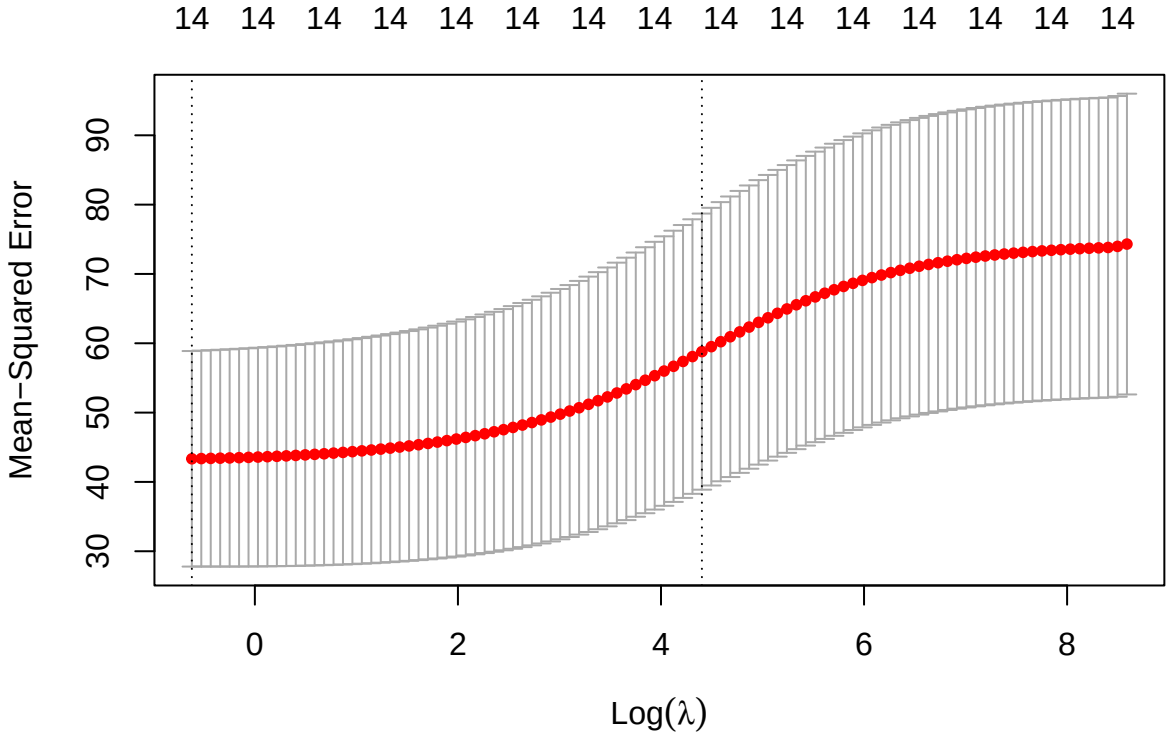
```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) 2.176491
## zn          .
## indus       .
## chasN       .
## chasY       .
## nox         .
## rm          .
## age         .
## dis         .
## rad         0.150484
## tax         .
## ptratio     .
## black       .
## lstat       .
## medv        .
```

```
## [1] "Test MSE for Lasso model : "
```

```
## [1] 62.74783
```

This shows the MSE value for Lasso model which is slightly higher than the best selection method.

Using Ridge Regression to predict per capita crime rate




```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  1.378868104
## zn          -0.002955708
## indus        0.029308357
## chasN        0.152157898
## chasY       -0.152154852
## nox          1.877361697
## rm          -0.142466331
## age          0.006217963
## dis         -0.094695187
## rad          0.045930738
## tax          0.002085959
## ptratio      0.071079829
## black       -0.002603532
## lstat        0.035722766
## medv        -0.023418669
```

According to Ridge regression model, all predictors need to be considered for analysis.

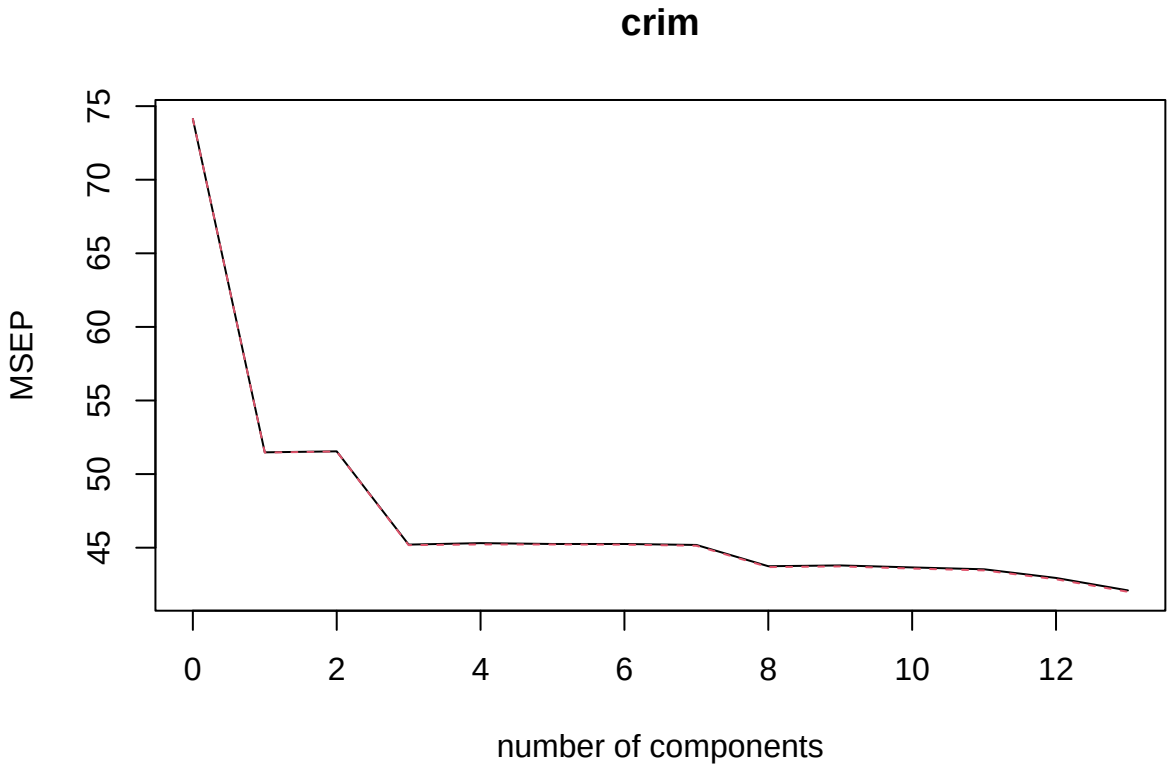
```
## [1] "Test MSE for Ridge model : "
```

```
## [1] 58.79457
```

The MSE value for Ridge model is 57.54 which is comparable to Lasso model.

Using PCR Model to predict per capita crime rate

```
## Data:    X dimension: 506 13
##  Y dimension: 506 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            8.61    7.175    7.180    6.724    6.731    6.727    6.727
## adjCV         8.61    7.174    7.179    6.721    6.725    6.724    6.724
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       6.722    6.614    6.618     6.607     6.598     6.553     6.488
## adjCV    6.718    6.609    6.613     6.602     6.592     6.546     6.481
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       47.70    60.36    69.67    76.45    82.99    88.00    91.14    93.45
## crim    30.69    30.87    39.27    39.61    39.61    39.86    40.14    42.47
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       95.40     97.04     98.46     99.52     100.0
## crim    42.55     42.78     43.04     44.13      45.4
```



PCR fit for 13th component has lowest CV and adj CV, so no dimension reduction is required. MSE value is around 44.

The validation plot for PCR fit shows that the MSE value is least when the number of predictors considered are 13.

Chapter 6 - Question 11 - Section b

The best subset selection method provides the model with the lowest cross validation error.

Chapter 6 - Question 11 - Section c

The nine parameter model chosen by best subset selection method has the best cross-validated MSE even though the 13 component PCR model is the simplest to build.

Chapter 8 - Question 8 - Section a

Splitting the Carseats data set into test and train with a split of 50-50%

```
## [1] "Carseats dataset :"
```

```
## [1] 400  11
```

```
## [1] "Training Dataset dimensions :"
```

```
## [1] 200  11
```

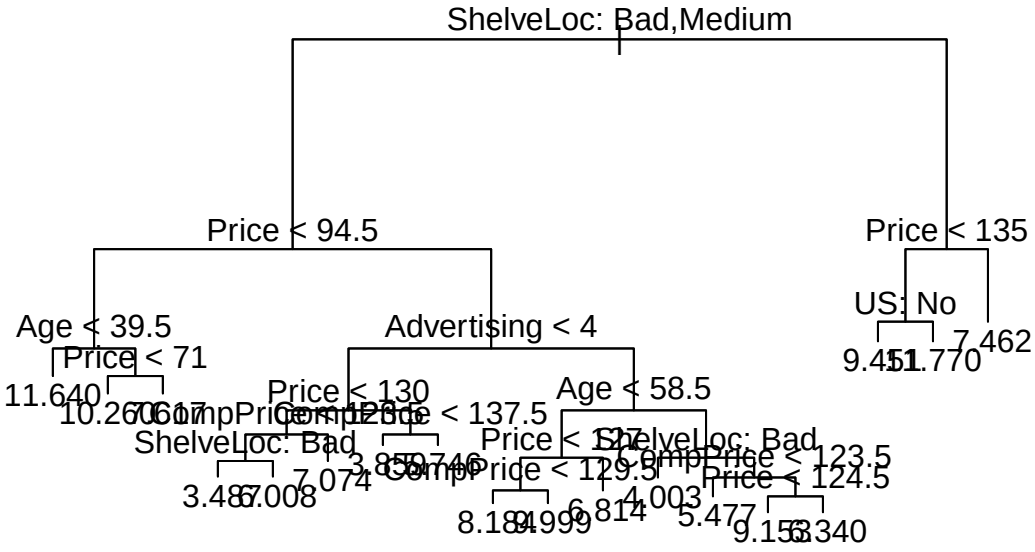
```
## [1] "Testing Dataset dimensions :"
```

```
## [1] 200  11
```

Fitting the regression tree using the training data set and calculating the test error MSE.

Chapter 8 - Question 8 - Section b

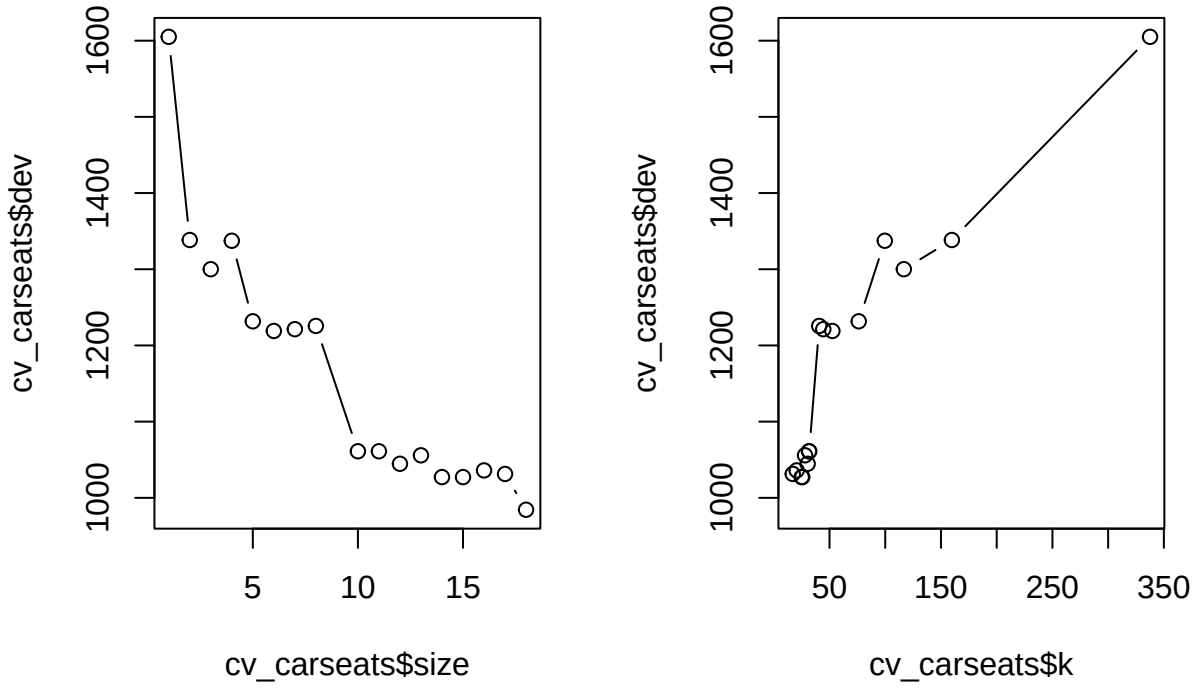
```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats_train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Advertising" "CompPrice"
## [6] "US"
## Number of terminal nodes:  18
## Residual mean deviance:  2.167 = 394.3 / 182
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.88200 -0.88200 -0.08712  0.00000  0.89590  4.09900
```




```
## [1] 4.922039
```

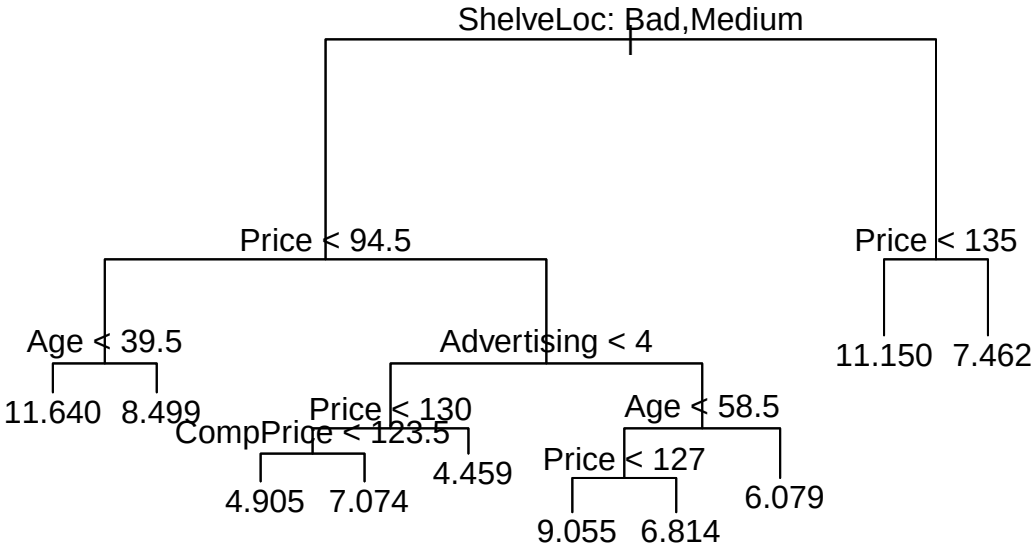
The test MSE is 4.92 for the regression tree.

Chapter 8 - Question 8 - Section c



Based on cross validation plots, the lowest size is obtained at 9 as the slope is almost zero after this point.

Pruning the tree which was formed above.




```
## [1] 4.918134
```

Pruning the tree increases the MSE value and thus, the pruned tree isn't very useful as compared to the previously formed regression tree.

Chapter 8 - Question 8 - Section d

```
## randomForest 4.7-1.1
```

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

```
## [1] 2.588149
```

Bagging reduced the TEST MSE significantly.

Finding **variable importance** through bagging of trees

```
##                %IncMSE IncNodePurity
## CompPrice   25.6142224    170.664874
## Income       4.5915046     91.047400
## Advertising 12.9597786     98.653166
## Population  -0.7823257     57.653647
## Price       55.2302897    510.311641
## ShelveLoc   48.5480002    381.630885
```

```
## Age         16.7267043    158.261715
## Education    0.9105144     43.460189
## Urban        0.2260420      9.347446
## US           5.8455044     18.261630
```

Based on the output of variable importance and incremental node purity, the predictors **Price, ShelveLoc**, **CompPrice** & **Age** are the most important predictors of the Sales in Carseats.

Chapter 8 - Question 8 - Section e

```
## [1] 2.710806
```

For Random Forest, the test MSE is higher than that for bagging.

Computing the importance of variables using Random Forest model

```
##                %IncMSE IncNodePurity
## CompPrice   18.3350929     163.77398
## Income       4.7546303     115.94202
## Advertising 10.4995404      98.99790
## Population  -0.8865995      78.91872
## Price       45.2942175     451.27503
## ShelveLoc   40.8250417     333.45728
## Age         13.6552105     165.03566
## Education    0.9386481      56.40921
## Urban       -0.8796633      11.21732
## US           6.1388918      25.33755
```

Based on the values for m in the random forest, the test MSE changes from **2.5 to 3**. The important variables are - **Price, ShelvLoc**, **CompPrice** and **Age** in the same order.

Chapter 8 - Question 11 - Section a

- Purchase is a qualitative parameter and thus, it is converted into numerical value of 0 and 1 for modeling.
- Training set is created using the first 1000 observations and remaining observations are in the testing dataset.

```
## [1] "Dimensions in Caravan dataset :"
```

```
## [1] 5822   86
```

```
## [1] "Dimensions in Caravan training dataset :"
```

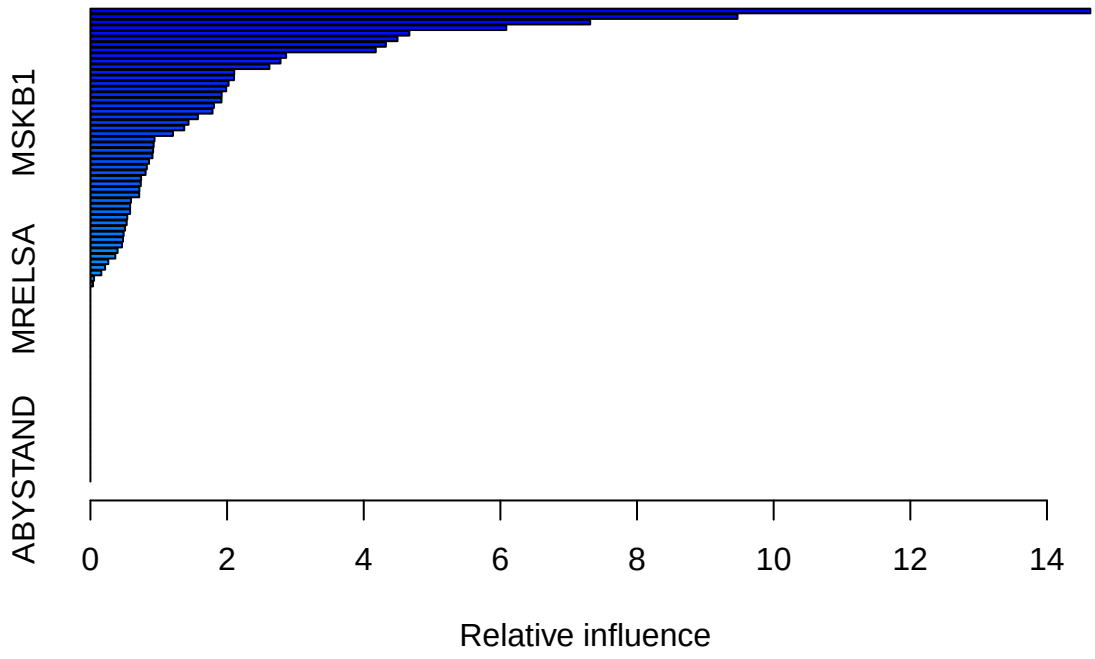
```
## [1] 1000   86
```

```
## [1] "Dimensions in Caravan testing dataset :"
```

```
## [1] 4822   86
```

Chapter 8 - Question 11 - Section b

Building a boosting model with 1000 trees and shrinkage parameter 0.01




```
##                var     rel.inf
## PPERSAUT PPERSAUT 14.63504779
## MKOOPKLA MKOOPKLA  9.47091649
## MOPLHOOG MOPLHOOG  7.31457416
## MBERMIDD MBERMIDD  6.08651965
## PBRAND     PBRAND  4.66766122
## MGODGE     MGODGE  4.49463264
## ABRAND     ABRAND  4.32427755
## MINK3045 MINK3045  4.17590619
## MOSTYPE   MOSTYPE  2.86402583
## PWAPART   PWAPART  2.78191075
## MAUT1       MAUT1  2.61929152
## MBERARBG MBERARBG  2.10480508
## MSKA         MSKA  2.10185152
## MAUT2       MAUT2  2.02172510
## MSKC         MSKC  1.98684345
## MINKGEM   MINKGEM  1.92122708
## MGODPR     MGODPR  1.91777542
## MBERHOOG MBERHOOG  1.80710618
## MGODOV     MGODOV  1.78693913
## PBYSTAND PBYSTAND  1.57279593
## MSKB1       MSKB1  1.43551401
```

```
## MFWEKIND MFWEKIND  1.37264255
## MRELGE     MRELGE  1.20805179
## MOPLMIDD MOPLMIDD  0.93791970
## MINK7512 MINK7512  0.92590720
## MINK4575 MINK4575  0.91745993
## MGODRK     MGODRK  0.90765539
## MFGEKIND MFGEKIND  0.85745374
## MZPART     MZPART  0.82531066
## MRELOV     MRELOV  0.80731252
## MINKM30   MINKM30  0.74126812
## MHKOOP     MHKOOP  0.73690793
## MZFONDS   MZFONDS  0.71638323
## MAUT0       MAUT0  0.71388052
## MHHUUR     MHHUUR  0.59287247
## APERSAUT APERSAUT  0.58056986
## MOSHOOFD MOSHOOFD  0.58029563
## MSKB2       MSKB2  0.53885275
## PLEVEN     PLEVEN  0.53052444
## MINK123M MINK123M  0.50660603
## MBERARBO MBERARBO  0.48596479
## MGEMOMV   MGEMOMV  0.47614792
## PMOTSCO   PMOTSCO  0.46163590
## MSKD         MSKD  0.39735297
## MBERBOER MBERBOER  0.36417546
## MGEMLEEF MGEMLEEF  0.26166240
## MFALLEEN MFALLEEN  0.21448118
## MBERZELF MBERZELF  0.15906143
## MOPLLAAG MOPLLAAG  0.05263665
## MAANTHUI MAANTHUI  0.03766014
## MRELSA     MRELSA  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
```

The table shows the relative influence of all predictors in the boosting model which is plotted in the graph as well.

PPERSAUT , MKOOPKLA and **MOPLHOOG** are three important variables in the same order.

Chapter 8 - Question 11 - Section c

```
##    boost_pred
##        0    1
##   0 4410  123
##   1  256   33
```

Confusion matrix generated based on the testing and prediction data set.

The fraction of people predicted to be make a purchase and made one is.

```
## [1] 0.2115385
```

Based on the Boosting model, about **21%** people from the predicted purchasers actually purchased.

Applying Logistic Regression to the training data

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

Based on Logistic Regression, the fraction of people predicted to be make a purchase and made one.

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

About **14.2%** people from the predicted Purchasers actually purchased. The performance of Logistic Regression is lower than Boosting.

Chapter 10 - Question 7

- Using the dist() function, the Euclidean distance is computed for all records.
- Correlation is calculated based on cor() function

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

This is the Correlation Matrix for the data set.

```
##             Murder   Assault  UrbanPop      Rape
## Murder   0.0000000 0.1981267 0.9304274 0.4364212
## Assault  0.1981267 0.0000000 0.7411283 0.3347588
## UrbanPop 0.9304274 0.7411283 0.0000000 0.5886588
## Rape     0.4364212 0.3347588 0.5886588 0.0000000
```

This is the matrix which shows the Euclidean distances.

```
##            Murder  Assault UrbanPop
## Assault  19.41642
## UrbanPop 91.18188 72.63057
## Rape     42.76927 32.80636 57.68856
```

As the two correlation based and euclidean distances are computed, one can observe that they are approximately same ratio. For example for Assault and Murder features, value from correlation matrix is 0.1981267 and from Euclidean matrix is 19.41642 which is a ration of 1:100. This trend is maintained for other features as well.Thus,the proportionality holds true.

Problem - Beauty Pays!

Problem - Beauty Pays! - Section 1

```
## [1] "Dimesnions of the Beauty dataset :"
```

```
## [1] 463   6
```

```
## [1] "Column names in the dataset : "
```

```
## [1] "CourseEvals" "BeautyScore" "female"      "lower"       "nonenglish"
## [6] "tenuretrack"
```

Running multiple regression model to understand the effect of all predictors on CourseEvals

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = Beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31385 -0.30202  0.01011  0.29815  1.04929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.06542    0.05145  79.020  < 2e-16 ***
## BeautyScore  0.30415    0.02543  11.959  < 2e-16 ***
## female      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish  -0.25808    0.08478  -3.044  0.00247 **
## tenuretrack -0.09945    0.04888  -2.035  0.04245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 457 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3399
## F-statistic: 48.58 on 5 and 457 DF,  p-value: < 2.2e-16
```

Even after removing tenuretrack variable from the regression analysis, the R-squared for the regression model remains same.

```
##
## Call:
## lm(formula = CourseEvals ~ . - tenuretrack, data = Beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33068 -0.29470  0.01019  0.28716  1.02593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.98262    0.03159 126.080  < 2e-16 ***
## BeautyScore  0.30447    0.02552  11.931  < 2e-16 ***
## female      -0.32523    0.04076  -7.980 1.19e-14 ***
## lower       -0.33174    0.04264  -7.781 4.84e-14 ***
## nonenglish  -0.27839    0.08448  -3.295  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 458 degrees of freedom
## Multiple R-squared:  0.3411, Adjusted R-squared:  0.3354
## F-statistic: 59.28 on 4 and 458 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore + female + lower + nonenglish,
##     data = Beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33068 -0.29470  0.01019  0.28716  1.02593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.98262    0.03159 126.080  < 2e-16 ***
## BeautyScore  0.30447    0.02552  11.931  < 2e-16 ***
## female      -0.32523    0.04076  -7.980 1.19e-14 ***
## lower       -0.33174    0.04264  -7.781 4.84e-14 ***
```

```
## nonenglish  -0.27839    0.08448  -3.295  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 458 degrees of freedom
## Multiple R-squared:  0.3411, Adjusted R-squared:  0.3354
## F-statistic: 59.28 on 4 and 458 DF,  p-value: < 2.2e-16
```

The regression model results show that the predictor BeautyScore has positive correlation with CourseEvals when the model incorporates all features to predict course ratings. The BeautyScore predictor has a statistically significant coefficient which means that if we keep all other determinants or predictors constant, BeautyScore will have a direct positive impact on course ratings.

As p value is lesser than 0.05 for BeautyScore,female, nonenglish and lower, null hypothesis can be rejected and these predictors highly effect the estimator - CourseEvals

Based on the coefficients it can be deduced that : + For female instructors, the data set shows a negative bias which means that they receive lesser ratings as compared to the other male instructors. + For lower divisions, students taking lower classes may or may not find the classes important as they are mandatory and may not be interested. + For non-english speakers, inability to fluently communicate in English might hamper in teaching and end up in lesser course rating. + For BeautyScore, higher the value, higher the course ratings.

Problem - Beauty Pays! - Section 2

The regression analysis conducted using the data set in Dr. Hamermesh's study clearly suggsests that BeautyScore is the best predictor of CourseEvals with the highest positive correlation and statistically significant coefficient.

This seems like a discrimination as we aren't basing the prediction of course rating on factors like productivity and skills of the instructor. Hence, using BeautyScore is not a great conclusion and shouldn't be generalized or standardized as a means to predict evaluation of instructors for their courses.

Problem - Housing Price Structure

Problem - Housing Price Structure - Section 1

- The parameters in midcity data set like Brick is qualitative and Nbhd has values of 1,2,3 based on the types of neighbourhood.
- Dummy variable is created to furnish one-hot encoding for brick,N1,N2 and N3.

```
##       Home             Nbhd           Offers           SqFt
##  Min.   :  1.00   Min.   :1.000   Min.   :1.000   Min.   :1450
##  1st Qu.: 32.75   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1880
##  Median : 64.50   Median :2.000   Median :3.000   Median :2000
##  Mean   : 64.50   Mean   :1.961   Mean   :2.578   Mean   :2001
##  3rd Qu.: 96.25   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2140
##  Max.   :128.00   Max.   :3.000   Max.   :6.000   Max.   :2590
##     Brick              Bedrooms       Bathrooms         Price
##  Length:128         Min.   :2.000   Min.   :2.000   Min.   : 69100
##  Class :character   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:111325
##  Mode  :character   Median :3.000   Median :2.000   Median :125950
##                     Mean   :3.023   Mean   :2.445   Mean   :130427
```

```
##                    3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:148250
##                    Max.   :5.000   Max.   :4.000   Max.   :211200
```

```
## [1] "Summary of paramters after creating dummy variables : "
```

```
##       Home             Nbhd           Offers           SqFt
##  Min.   :  1.00   Min.   :1.000   Min.   :1.000   Min.   :1450
##  1st Qu.: 32.75   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1880
##  Median : 64.50   Median :2.000   Median :3.000   Median :2000
##  Mean   : 64.50   Mean   :1.961   Mean   :2.578   Mean   :2001
##  3rd Qu.: 96.25   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2140
##  Max.   :128.00   Max.   :3.000   Max.   :6.000   Max.   :2590
##      Brick           Bedrooms       Bathrooms         Price
##  Min.   :0.0000   Min.   :2.000   Min.   :2.000   Min.   : 69100
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:111325
##  Median :0.0000   Median :3.000   Median :2.000   Median :125950
##  Mean   :0.3281   Mean   :3.023   Mean   :2.445   Mean   :130427
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:148250
##  Max.   :1.0000   Max.   :5.000   Max.   :4.000   Max.   :211200
##        N1               N2              N3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3438   Mean   :0.3516   Mean   :0.3047
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

Fitting the regression model

```
##
## Call:
## lm(formula = Price ~ ., data = midcity1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27897.8  -6074.8    -48.7   5551.8  27536.4
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22572.432  10287.606   2.194 0.030168 *
## Home           -11.456     25.387  -0.451 0.652616
## Offers       -8350.128   1103.693  -7.566 8.96e-12 ***
## SqFt            53.634      5.926   9.051 3.30e-15 ***
## Brick        17313.540   1988.548   8.707 2.12e-14 ***
## Bedrooms      4136.461   1621.775   2.551 0.012023 *
## Bathrooms     7975.157   2133.831   3.737 0.000287 ***
## N1          -20534.706   3176.051  -6.465 2.33e-09 ***
## N2          -22264.319   2540.699  -8.763 1.56e-14 ***
## N3                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10050 on 119 degrees of freedom
```

```
## Multiple R-squared:  0.8688, Adjusted R-squared:   0.86
## F-statistic: 98.54 on 8 and 119 DF,  p-value: < 2.2e-16
```

As coefficients for N3 are NA, it can be inferred that N3 parameter is highly correlated to N1 and N2. Thus, as the rule of one-hot encoding goes, we can remove N1 or N2 as N1 and N2 are both traditional houses and N3 is for modern houses.

```
##
## Call:
## lm(formula = Price ~ ., data = midcity2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27897.8  -6074.8    -48.7   5551.8  27536.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2037.726   8911.501   0.229 0.819524
## Home          -11.456     25.387  -0.451 0.652616
## Offers      -8350.128   1103.693  -7.566 8.96e-12 ***
## SqFt           53.634      5.926   9.051 3.30e-15 ***
## Brick       17313.540   1988.548   8.707 2.12e-14 ***
## Bedrooms     4136.461   1621.775   2.551 0.012023 *
## Bathrooms    7975.157   2133.831   3.737 0.000287 ***
## N2          -1729.613   2433.756  -0.711 0.478675
## N3          20534.706   3176.051   6.465 2.33e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10050 on 119 degrees of freedom
## Multiple R-squared:  0.8688, Adjusted R-squared:   0.86
## F-statistic: 98.54 on 8 and 119 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept) -15607.93640 19683.38913
## Home           -61.72532    38.81249
## Offers      -10535.54898 -6164.70615
## SqFt            41.90048    65.36778
## Brick        13376.01617 21251.06453
## Bedrooms       925.18504  7347.73765
## Bathrooms     3749.95923 12200.35499
## N2           -6548.69199  3089.46637
## N3           14245.80700 26823.60505
```

Additionally, using 95% confidence level, the confidence interval for Brick houses is [13376.01617,21251.06453] which doesn't include zero. Thus, this is another evidence that brick is an important factor to predict price of houses. (Checking the null hypthesis to ascertain if the Brick is important factor given all other factors remain constant.)

Problem - Housing Price Structure - Section 2

To validate if there is any premium for houses in the N3 neighbourhood, test the hypothesis whether coefficient of N3 is null within the confidence interval at 95% confidence level. The previous output of confidence

interval shows that the range of coefficients is [17233.48078,27295.15689] which doesn't include zero. Thus, there is a premium for N3 houses keeping all other predictors constant.

Problem - Housing Price Structure - Section 3

```
##
## Call:
## lm(formula = Price ~ Home + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms + N2 + N3 * Brick, data = midcity2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27515.9  -5681.0   -459.6   4451.0  26695.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2885.512   8738.695   0.330  0.74183
## Home          -11.799     24.875  -0.474  0.63613
## Offers      -8486.348   1082.875  -7.837 2.26e-12 ***
## SqFt           54.726      5.823   9.397 5.36e-16 ***
## Brick       13839.320   2413.580   5.734 7.69e-08 ***
## Bedrooms     4605.046   1600.639   2.877  0.00477 **
## Bathrooms    6556.432   2170.200   3.021  0.00309 **
## N2           -846.146   2412.025  -0.351  0.72636
## N3          17086.915   3417.999   4.999 2.02e-06 ***
## Brick:N3    10192.783   4178.971   2.439  0.01621 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9850 on 118 degrees of freedom
## Multiple R-squared:  0.8751, Adjusted R-squared:  0.8656
## F-statistic:  91.9 on 9 and 118 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept) -14419.48380 20190.50773
## Home           -61.05960    37.46078
## Offers      -10630.73590 -6341.96077
## SqFt            43.19341    66.25763
## Brick         9059.77458 18618.86626
## Bedrooms      1435.34506  7774.74785
## Bathrooms     2258.84419 10854.01891
## N2           -5622.61257  3930.32100
## N3           10318.34591 23855.48395
## Brick:N3      1917.28263 18468.28393
```

By interacting the paramters N3 and Brick, we can observe that the corresponding coefficient is 10192.783. Analyzing this for different houses in N1 and N3 with brick houses.

- For a brick house in traditional neighbourhood, the premium is corresponding to N1 and Brick which is : 846.146 + 13839.320 = 14,685.466
- For a brick house in modern neighbourhood, the premium is corresponding to N3, Brick and interaction of N3 : Brick which is : 13839.320 + 17933.061 + 10192.783 = 41,965.164

Additionally,using 95% confidence level, the confidence interval for Brick houses in N3 neighbourhood is [1917.28263,18468.28393] which doesn't include zero. Thus, this is another evidence that brick is an important factor to predict price of houses. (Checking the null hypthesis to ascertain if the Brick is important factor given all other factors remain constant.)

Problem - Housing Price Structure - Section 4

```
##                    2.5 %      97.5 %
## (Intercept) -15607.93640 19683.38913
## Home           -61.72532    38.81249
## Offers      -10535.54898 -6164.70615
## SqFt           41.90048    65.36778
## Brick       13376.01617 21251.06453
## Bedrooms      925.18504  7347.73765
## Bathrooms    3749.95923 12200.35499
## N2          -6548.69199  3089.46637
## N3          14245.80700 26823.60505
```

Checking the confidence interval against the original dummy variables that were created while fitting the model.

As observed in this, house in neighbourhood N2 with a confidence level of 95% range in [-6548.69199,3089.46637] which includes zero. Thus, the null hypothesis stands true and coefficient of N2 houses can be zero. This would be mean that there is no premium of N2 houses and N2 and N1 can be combined together.

Problem - What causes what??

Problem - What causes what?? - Section 1

The podcast discusses about the study wherein the data set contains information about crime and Police. To run a regression analysis, it is important to understand the causality here. It isn't clear if the number of cops decreases the crime in the cities or more cops are deployed because of rise of crime in the city.

Even if we fetch more data from different cities, this causal relationship between crime and number of cops isn't defined. Additionally, we need to control for other variables and observe if there is any change correlation in the two factors in question.

Problem - What causes what?? - Section 2

The researchers at UPENN were able to collect data regarding number of cops and crime rate using a natural experiment wherein these variables were isolated :

- During the high alert days (due to terrorist threats) in DC, the number of cops deployed increases by the mayor. This decision to deploy more cops is independent of crim and hence, this is a classic example to isolate number of cops and its effect on crime to observe a relationship.
- From Table 2 - Effect of Police on Crime , it is evident that on high alert days, the crime rate is reduced as the coefficient corresponding to it is negative (while controlling for ridership).

Problem - What causes what?? - Section 3

With reference to the UPENN study, the researchers accounted for controlling the variable METRO ridership. As previously mentioned, on high alert days, the number of cops deployed increased and the observed crime rate reduced. It is possible that the crime rate reduced because the number of people who were on the streets reduced because of the high alert on those days.

However, if we check the number of riders on these high alert days, we can ensure that this apparent decrease in the crime rate is not solely because the number of people on the streets were less. Also, there is possibility that the people engaged in criminal activities limited their mobility because of high alert.

Once, the observed METRO ridership is constant, a clear relationship between number of cops and crime rate can be deduced. In this research this hypothesis was tested and it was found that the number of victims didn't reduce leading to lesser crime rate, rather the riderhsip remained constant on high alert and normal days.Hence, controlling for METRO ridership, they inferred that increasing the number of cops has an influence in reducing the crime rate.

Problem - What causes what?? - Section 4

In Table 4, the model is estimating the effect of high alert days on crime on district 1 and other areas in the town. This analysis shows that there is a clear correlation relationship between high alert days and regions is only evident in district 1 as compared to other districts. This inference is apt as most potential targets for terrorist attacks are in DC in district 1 and hence, more cops are deployed in this area as compared to other towns. We can conclude that the impacts of other districts are small, and according to the standard error, they can be zero.

Problem - Final Project

Contribution to the project :

For our final group project, we decided to work on an interesting data set in Kaggle which holds a collection of global startups with varied features, in an attempt to classify their status as Success or Failure. This data set has about 44k records for startups from 120+ countries and across varied markets like Entertainment, News, AI, etc. The major obstacle in the analysis and modeling observed through exploratory data analysis were as follows: + Raw data set was quite unclean and diligent pre-processing was required to ensure smooth analysis + Imbalances in the classes of Success and Failure - About 700 records were corresponding to Failure and rest were for Success which is quite a contrast to the real-world scenario.

To combat these issues, first, I limited our scope of analysis to predict success of startups based out of USA in top 8 startup regions - Austin, Atlanta, Boston, Chicago, LA, New York, Seattle & SF Bay Area. The subset of this dataset was taken via coding in R. Again the market types in the dataset had about 600+ markets which were broadly categorized as Technology and Non-Technology to achieve some insights through the type of market predictor.

After this, the dataset was further cleaned and transformed : + removed records with missing values + converted strings into numerical values especially for funding amounts along with removal of ',' from funding columns + converted dates like funding start and end dates into difference in years to include provide numerical inferential value + handled categorical data by creating bins and using one-hot encoding to easily use the dataset in the models

This was done for all parameters with vast range of values like funding rounds, funding amount, venture, seed, funding in each round A to H, market type and estimator - status of the startup.

Once, I cleaned the data set, I observed a significant imbalance in classes for the estimator status of startups with about 700 records for status label - 0 which is failure of startup and about a whooping 12k+ records for success. Thus, it will be inaccurate to use this data to train our models and so, I explored two possible

options of under sampling and oversampling. Under sampling would have drastically reduced our data set and we would lose valuable insights. Hence, in our case, I have used oversampling via the package ROSE in R, to replicate the minority class and to have a balance in both classes.

Learning Outcomes :

- Hands-on for data cleaning and transformation
- Learnt various methods for dealing with imbalances in data and their impact on classification models as it can easily over fit the data
- Thinking out of the box to understand rationally how various predictors can correlate through EDA and regression models without judgement bias