STA 380: Intro to Machine Learning Take Home Exam: ISLR Edition 1

Muskaan Singhania

7/31/2022

Chapter 2 - Question 10

(a)

```
##      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
```

Boston DataSet loaded from the MASS Library

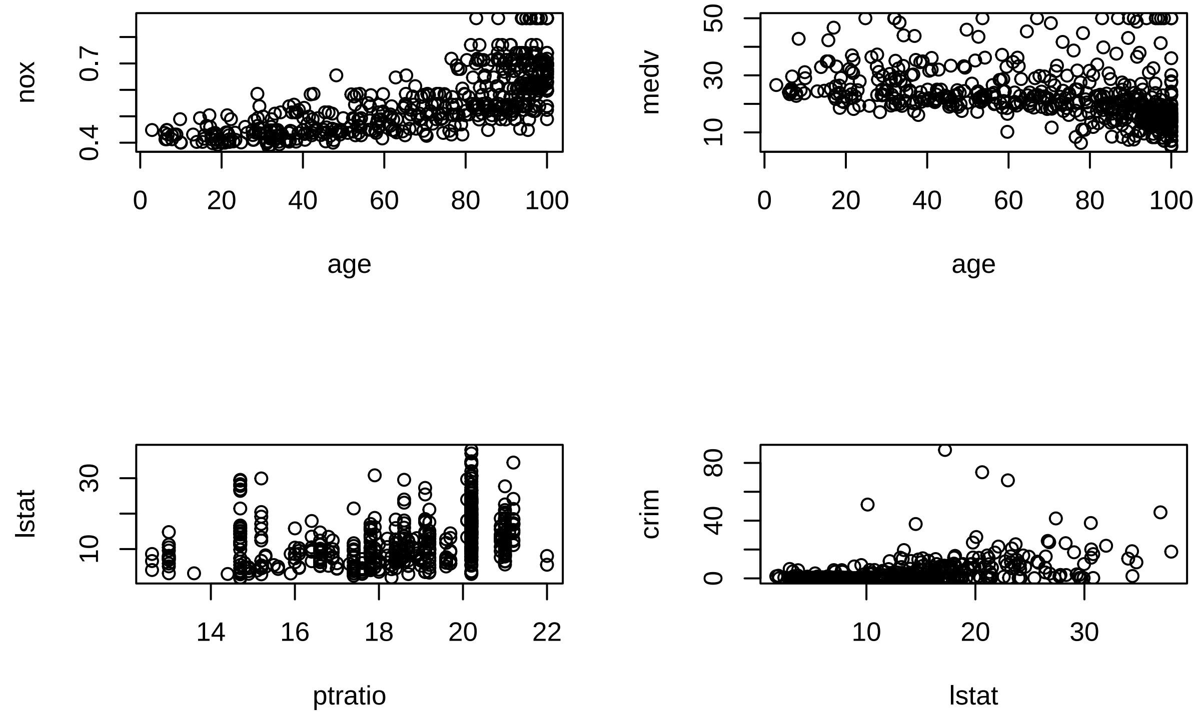
```
## [1] 506
```

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

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

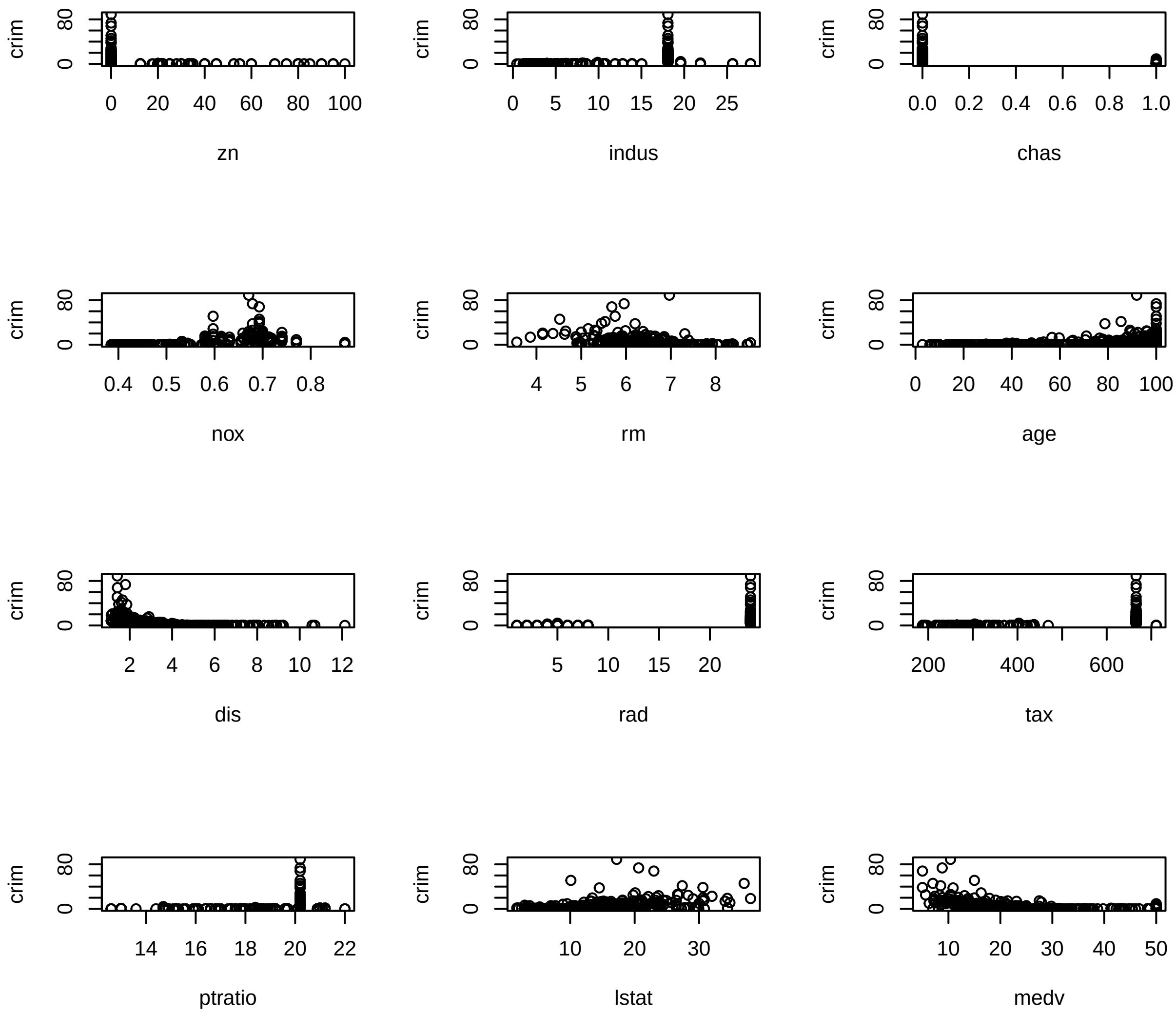
The numbers of rows here are **506** and the columns are **13**

(b)



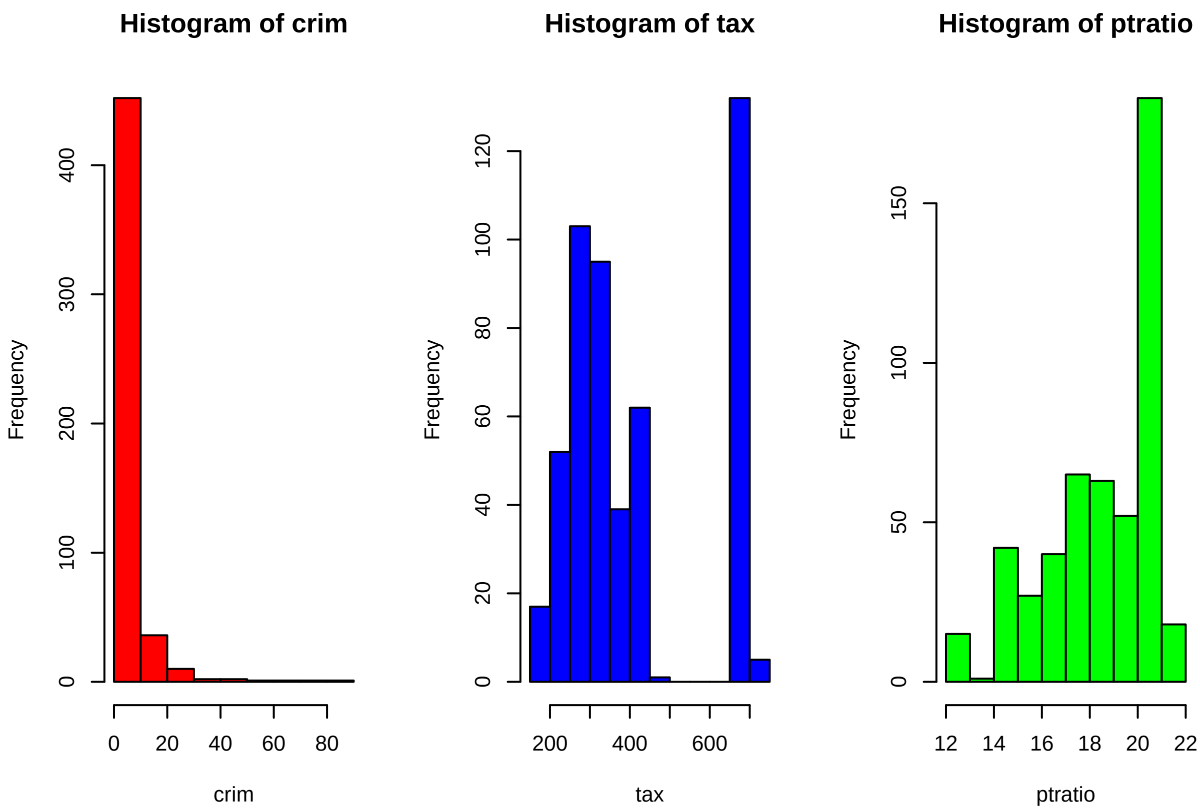
Findings: From the above plots
a. nox and age are positively correlated
b. medv and age are not that correlated
c. lstat and pratio are not correlated
d. lsat and crim are positively correlated

(c)



Findings:
From the above plots it's clear that the variable **crim** is correlated with **nox, rm, age, dis, lstat, medv**

(d)



Findings:
(1) A majority of suburbs have a crime rate lower than 30.
(2) There is a large proportion of suburbs having a tax rate greater than 600 and rest are spread out between 200 to 450.
(3) There is a large number of suburbs having a ptratio greater than 20 and the rest are spread out between 12 to 20

(e)

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

35 suburbs bound by **Charles River**

(f)

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

The **Median ptratio** is **19.05**

(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
```

As seen from the output, the suburb with the lowest median value of owner-occupied homes are **399** and **406** As for the values for all the other predictors for these suburb, please refer to the output above

Comparison of the values of all the predictors with the general trend:
(1) crim - Higher crime 75% of other suburbs zn - Lowest Zn indus - Indus at
(2) 3rd quartile chas - Lowest chass nox - Higher nox than 75% of the other (3) surburbs rm - Highest age dis - Lower dis than 25% of the other suburbs
(4) rad - Highest rad tax - tax at 3rd quartile ptratio - ptratio at 3rd quartile
(5) lstat - Higher lstat than 75% of the other suburbs medv -
(6) Lowest medv

(h)

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

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

```
##       crim               zn             indus             chas       
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age              dis       
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
```

```
## Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2
## 3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7
## Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9
##     lstat           medv
## Min.   :2.47   Min.   :21.9
## 1st Qu.:3.32   1st Qu.:41.7
## Median :4.14   Median :48.3
## Mean   :4.31   Mean   :44.2
## 3rd Qu.:5.12   3rd Qu.:50.0
## Max.   :7.44   Max.   :50.0
```

64 suburbs average more than **7** rooms per dwellings
13 suburbs average more thean **8** rooms per dwelling

Findings:
On comparing the summary of suburbs with dwellings more than 8 and summary of the entire dataset, we observe that the suburbs with dwellings more than 8 have lower crime and lower lstat

Chapter 3 - Q15

(a)

```
##      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
```

```
## [1] N N N N N N
## Levels: N Y
```

Encoded the variable **chas** as 0 & 1 so as to be able to use it in our analysis further on

```
##
## 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
```

Based on the low p-value, we can say that zn has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that indus has a statistically significant association with crim

```
##
## 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
```

Based on the high p-value, we can say that indus doesn't have a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that nox has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that rm has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that age has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that dis has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that rad has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that tax has a statistically significant association with crim

```
##
## 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
```

Based on the low p-value, we can say that ptratio has a statistically significant association with crim

```
##
## 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
```

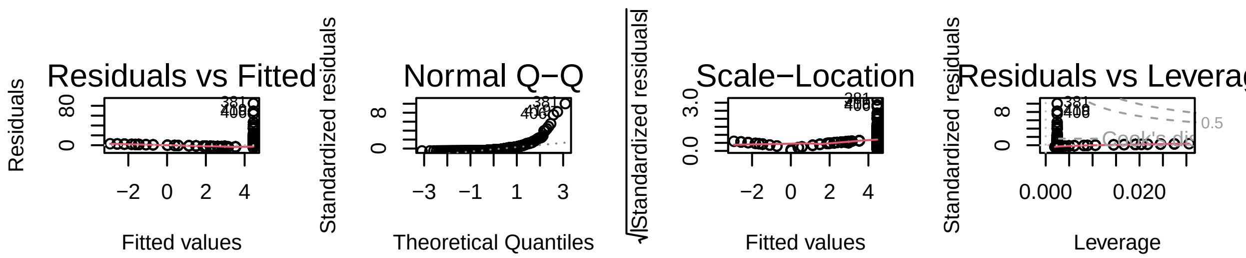
Based on the low p-value, we can say that black has a statistically significant association with crim

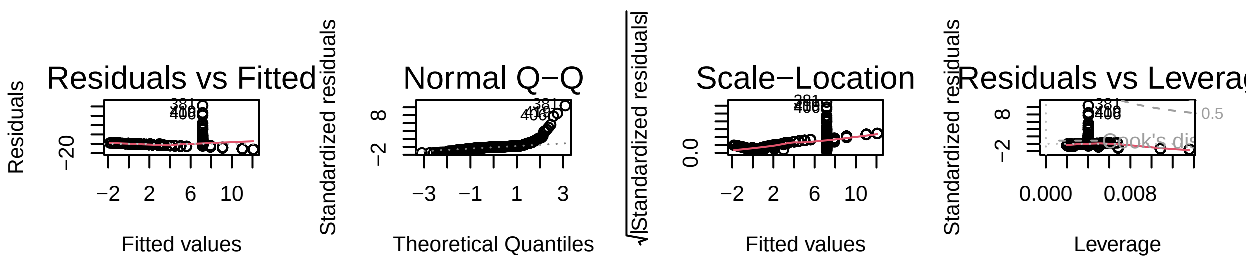
```
##
## Call:
## lm(formula = crim ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    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
```

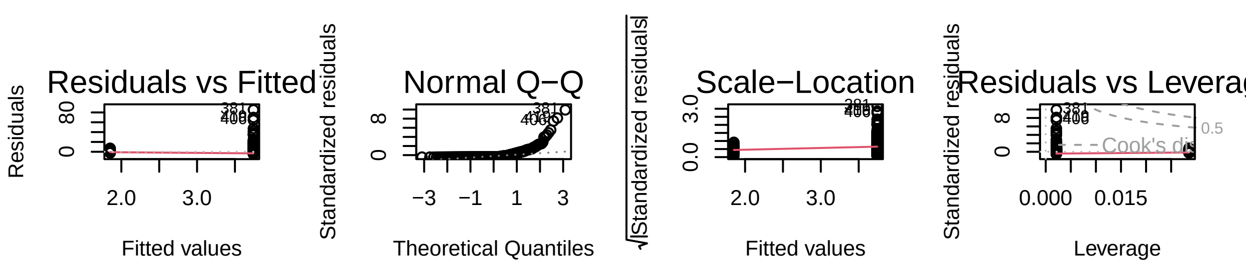
Based on the low p-value, we can say that lstat has a statistically significant association with crim

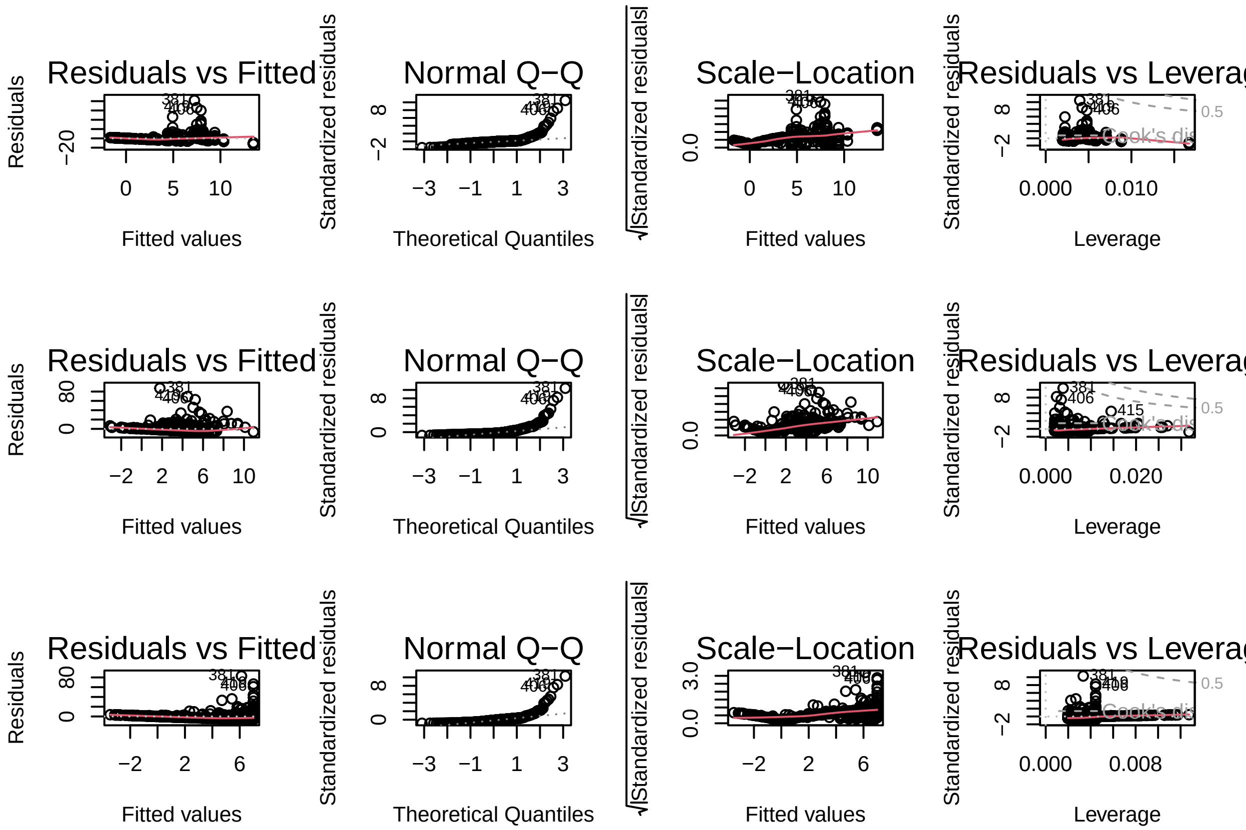
```
##
## Call:
## lm(formula = crim ~ medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839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

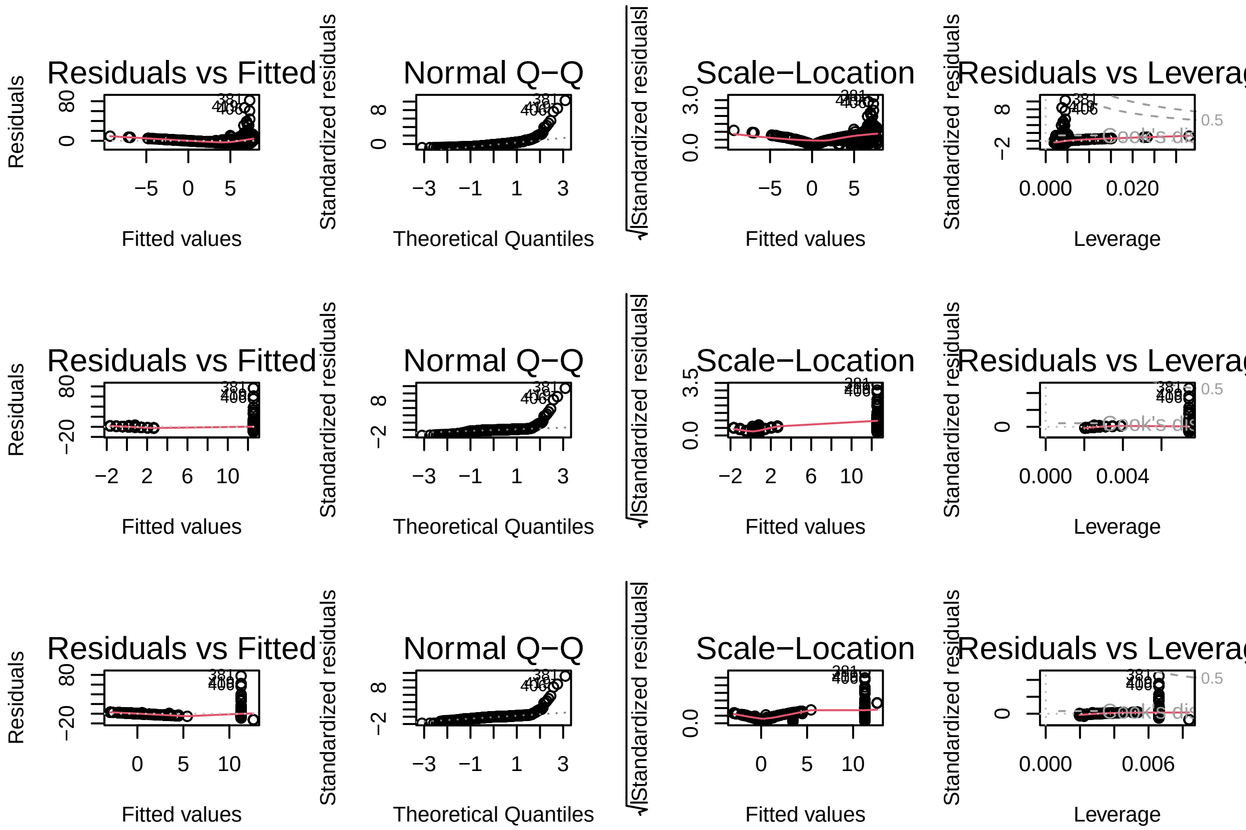
Based on the low p-value, we can say that medv has a statistically significant association with crim

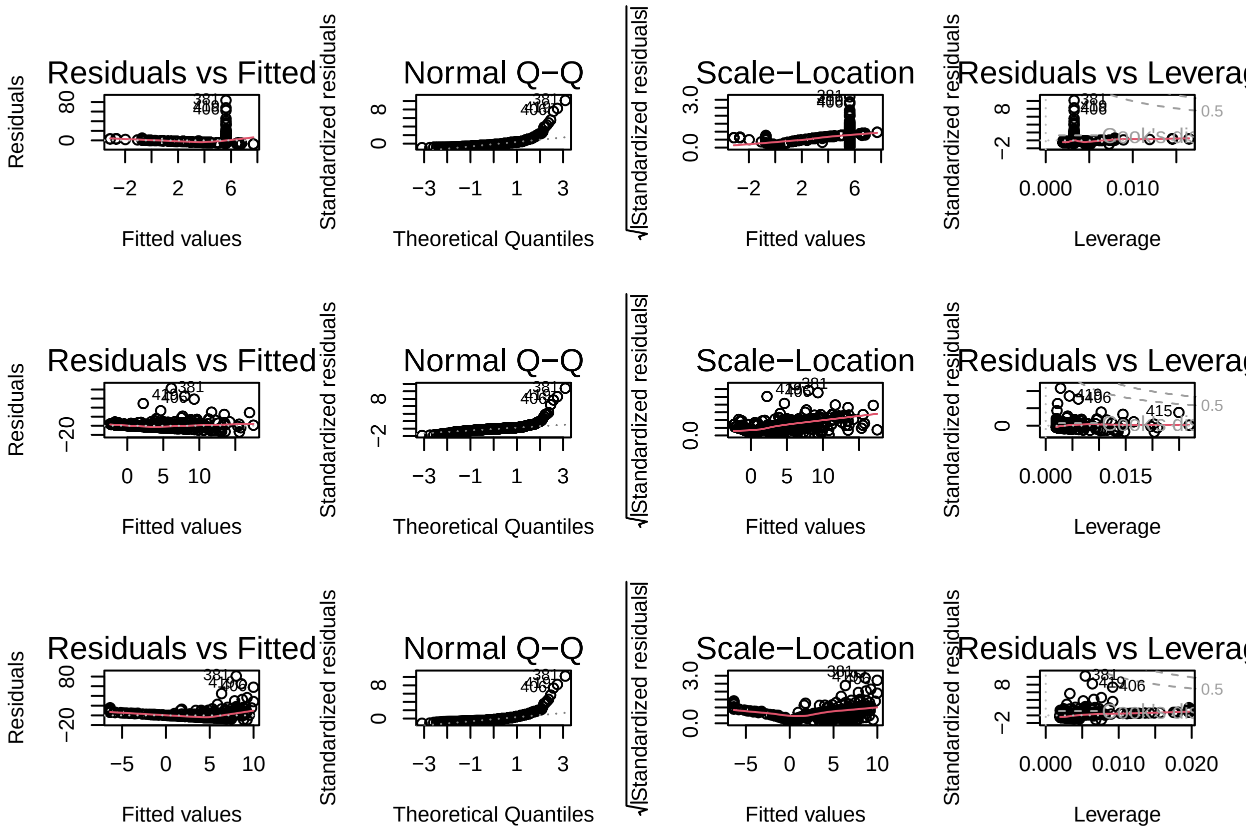







Residuals vs Fitted
Normal Q−Q
Scale−Location
Residuals vs Leverage
Residuals
Standardized residuals
√|Standardized residuals|
Fitted values
Theoretical Quantiles
Leverage
Cook's distance
0.5
381
406
415


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



(b)

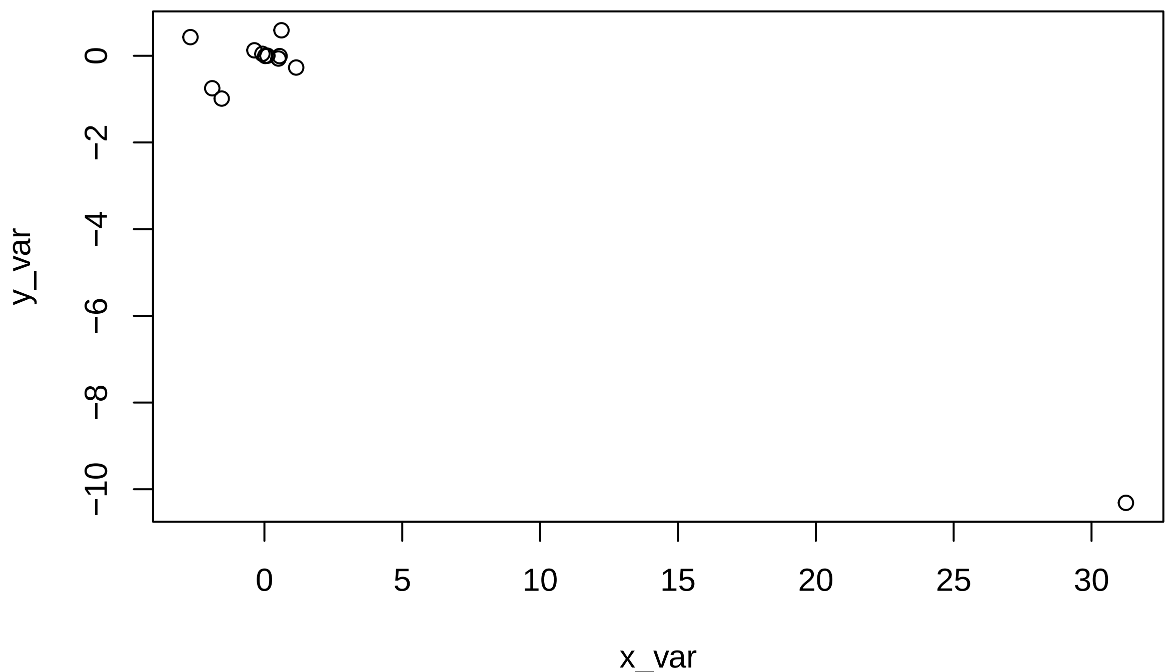
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855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

```
##                     2.5 %        97.5 %
## (Intercept)   2.818109179 31.2483458660
## zn            0.008046562  0.0816638671
## indus        -0.227733150  0.1000235023
## chasY        -3.067882868  1.5696156471
## nox         -20.678894713  0.0518248891
## rm           -0.773956866  1.6342178774
## age          -0.033767600  0.0366708869
## dis          -1.540889544 -0.4334619069
## rad           0.415209611  0.7612075719
## tax          -0.013909700  0.0063496670
## ptratio      -0.637417996  0.0952568794
## black        -0.014754837 -0.0003201725
## lstat        -0.022572584  0.2749953365
## medv         -0.317788478 -0.0799851646
```

Based on the **confidence interval**, we can reject null hypothesis for **zn, dis, rad, black, medv**

(c)



The difference between the **coefficient values** of **Nox** in the **Univariate and Multiple regression model** stands out when compared to the rest of the variables

(d)

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above lower adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a linear model rather than a non-linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

```
##
## 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
```

From the above higher adjusted R-squared value compared with the linear model adjusted R-squared, we can concluded that the above variable has a better fit in a non-linear model rather than a linear model.

We're aren't checking for non-linear relationship between chas and target as chas is binary in nature

Chapter 6 - Q9

(a)

```
## [1] "Training Set Dimensions are the following"
```

```
## [1] 564  18
```

```
## [1] "Test Set Dimensions are the following"
```

```
## [1] 213  18
```

The College data has been split into train and test set in a **75-25 ratio**

(b)

```
## [1] "The Test MSE is the following"
```

```
## [1] 1497009
```

The linear model has been trained on the **training data set** and the above **test MSE** has been computed on the test dataset. Linear Model considers all variables in predicting the value of the target

(c)

```
## [1] "The lambda value that minimizes the test MSE turns out to be:"
```

```
## [1] 43.28761
```

```
## [1] "The test MSE of the Ridge Model is the following"
```

```
## [1] 1723140
```

The Test MSE of the Ridge Model is greater than that of the OLS Model

(d)

```
## [1] "The best lambda is the following"
```

```
## [1] 8.111308
```

```
## [1] "The test MSE of the Lasso Model is the following"
```

```
## [1] 1529757
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -716.52002521
## (Intercept)    .
## PrivateYes  -506.85440904
## Accept         1.29247347
## Enroll         .
## Top10perc     44.47987213
## Top25perc    -12.84792052
## F.Undergrad    0.02808969
## P.Undergrad    0.02876426
## Outstate      -0.05433749
## Room.Board     0.18799436
## Books          0.08155585
## Personal      -0.01249045
## PhD           -9.04176108
## Terminal      -4.37272998
## S.F.Ratio     12.71664883
## perc.alumni   -6.45972486
## Expend         0.10428715
## Grad.Rate      8.25977261
```

The test MSE for the Lasso Model is lower than that of the Ridge Model. As we can see, the Lasso Model doesn't consider the variable 'Enroll' as significant

(e)

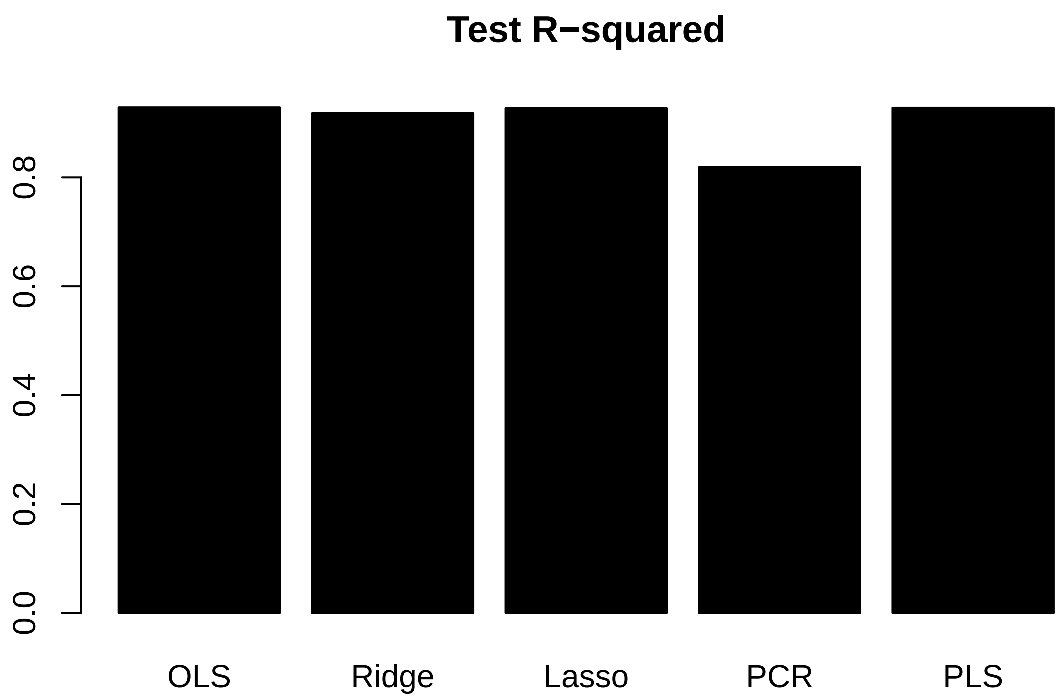
```
## Data:    X dimension: 564 17
##  Y dimension: 564 1
## Fit method: svdpc
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3568     3542     1749     1753     1773     1466     1378
## adjCV         3568     3542     1746     1752     1779     1456     1373
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1355     1306     1242      1258      1253      1252      1262
## adjCV     1354     1289     1239      1256      1250      1249      1259
##        14 comps  15 comps  16 comps  17 comps
## CV         1264      1245      1086      1081
```

```
## adjCV      1260      1244      1082      1076
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X      32.378    57.84    64.86    70.62    75.93    80.74    84.28    87.67
## Apps    4.958    76.89    77.00    77.09    84.73    86.39    86.91    88.29
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       90.72     93.12     95.22     96.97     98.04     98.91     99.40
## Apps    88.51     88.52     88.62     88.78     88.78     88.78     89.38
##       16 comps  17 comps
## X        99.82    100.00
## Apps     91.82     92.08

## [1] 3799601
```

(f)

```
## Data:    X dimension: 564 17
##  Y dimension: 564 1
## Fit method: kernelpls
## 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            3568     1599     1396     1208     1165     1113     1077
## adjCV         3568     1597     1398     1206     1161     1109     1073
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1070     1066     1070      1068      1068      1068      1069
## adjCV     1066     1063     1066      1064      1064      1064      1065
##        14 comps  15 comps  16 comps  17 comps
## CV         1069      1068      1068      1068
## adjCV      1065      1065      1065      1065
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       26.04    48.26    62.84    66.24    69.79    73.89    77.23    80.75
## Apps    80.71    85.42    89.26    90.35    91.39    91.90    92.00    92.02
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       83.90     86.87     89.52     91.74     93.28     95.64     97.02
## Apps    92.04     92.06     92.07     92.08     92.08     92.08     92.08
##       16 comps  17 comps
## X        99.12    100.00
## Apps     92.08     92.08

## [1] 1511867
```

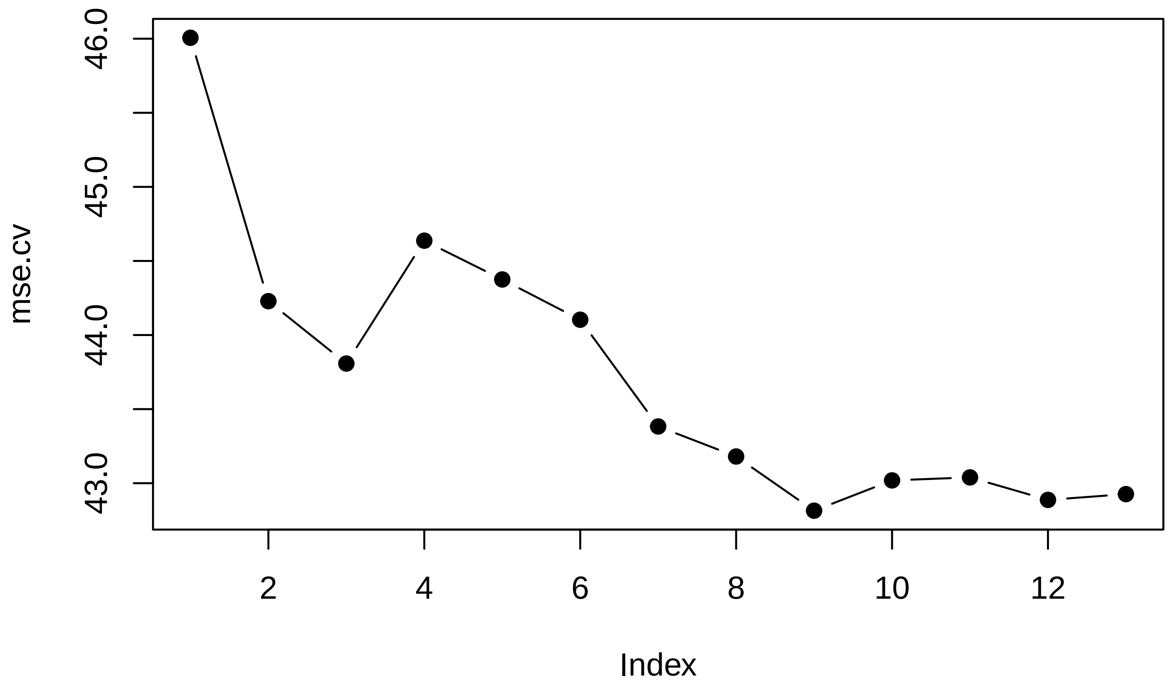
(g)



PCR model seems to **explain** the **least amount of variance** according to the plot seen above. **PLS model** predicts college applications with the **most amount of accuracy**. **Ridge, Lasso and Linear model** are not very far behind the PLS models when it comes to accuracy

Chapter 6 - Question 11

(a)

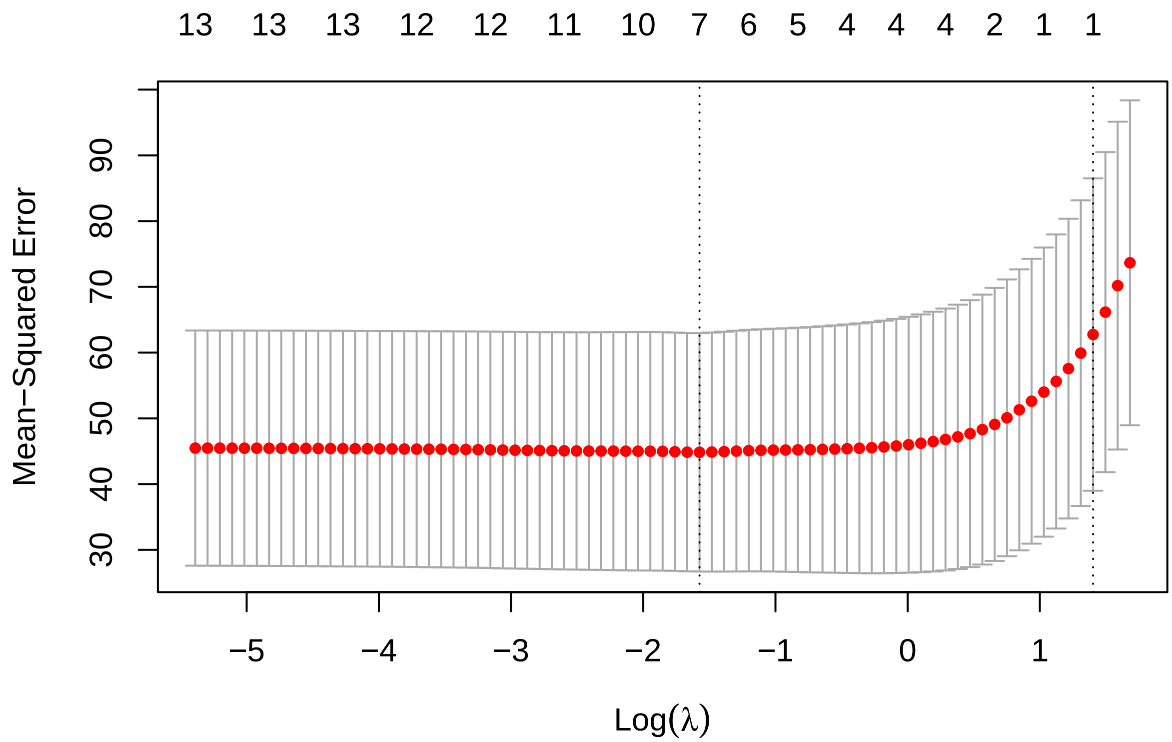



```
## [1] 9
```

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

Cross-Validation selects a **9-variable model**

We have a CV estimate for the **test MSE** equal to **42.815**

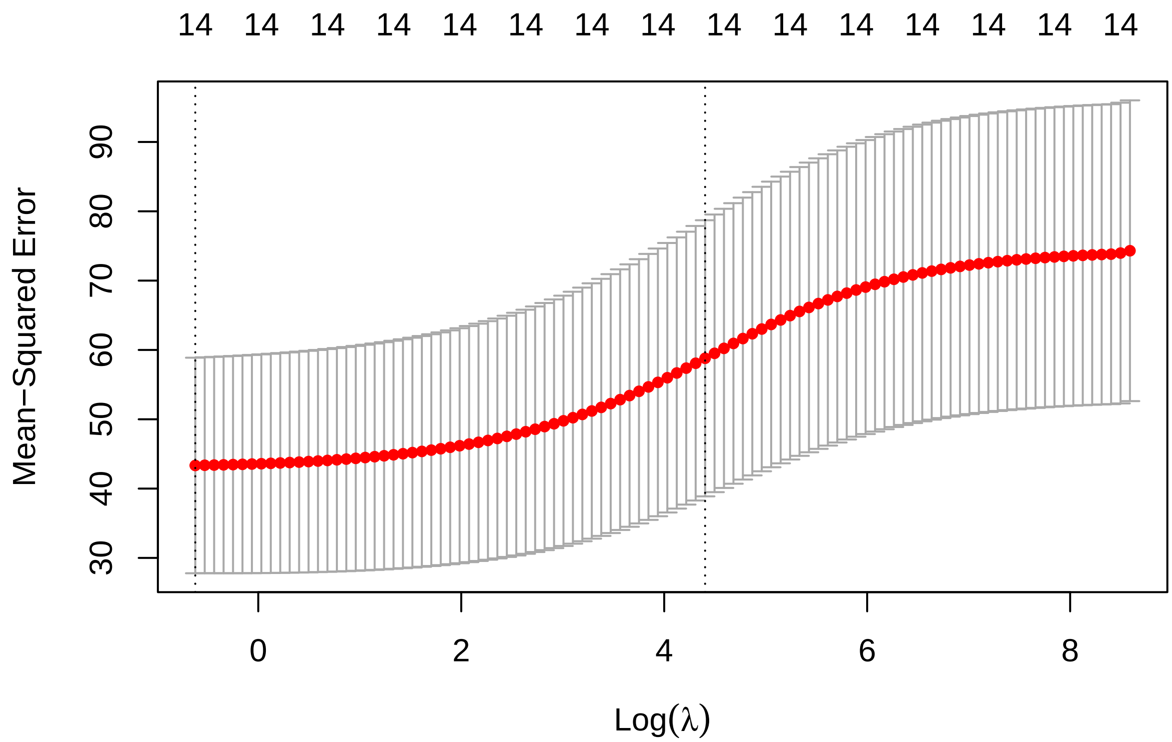



```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                   s1
## (Intercept) 2.176491
## zn          .
## indus       .
## chasN       .
## chasY       .
## nox         .
## rm          .
## age         .
## dis         .
## rad         0.150484
## tax         .
## ptratio     .
## black 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

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

Now, looking at the Lasso model, we will notice that there is only **one variable** being taken into account in the model. The rest are ignored or treated by the model as not significant in the outcome of the dependent variable.

We have a **CV estimate** for the **test MSE** equal to above outputted value



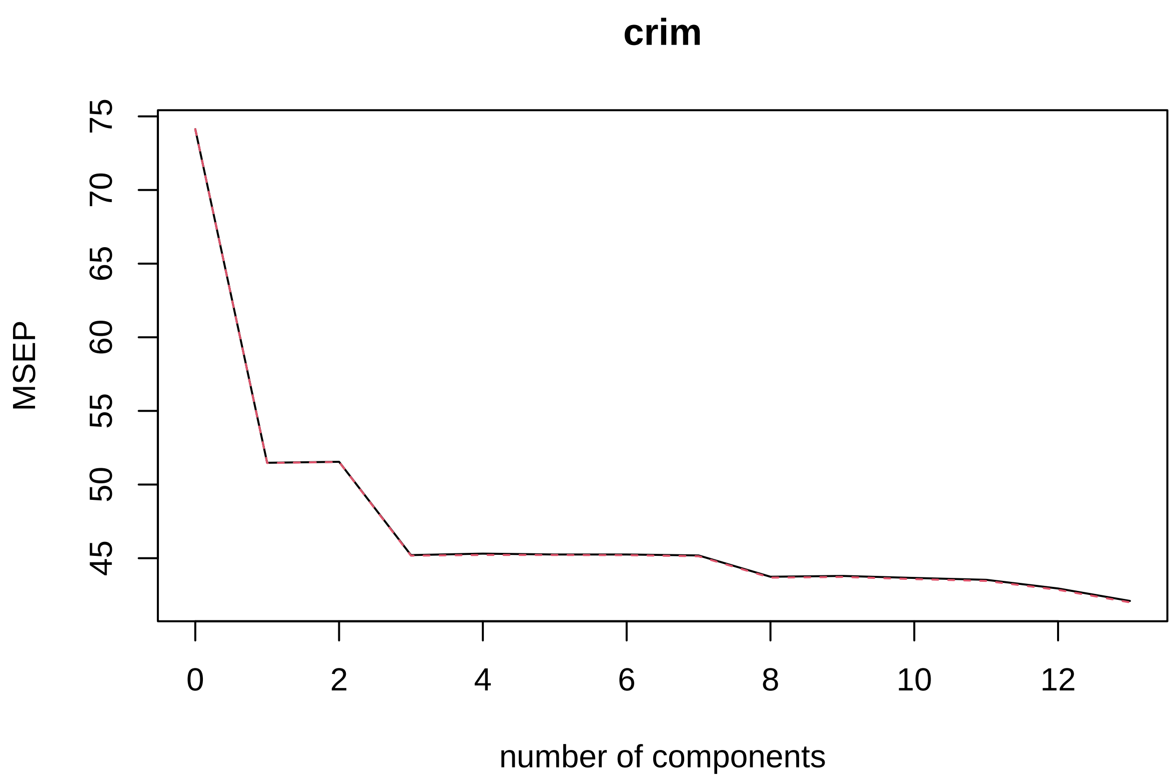

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  1.378868104
## zn          -0.002955708
## indus        0.029308357
## chasN        0.152157898
## chasY       -0.152154852
## nox          1.877361697
## rm          -0.142466331
## age          0.006217963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

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

We have a CV estimate for the **test MSE** equal to the above outputted value

```
## 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
```



Here cross-validation selects M to be equal to **13**

We have a CV estimate for the test MSE equal to the square of rmse value corresponding to M=13 from the above table

(b)

We're choosing the **best subset selection method**. The CV estimate of MSE for best selection method and PCR is comparable and there's a very insignificant difference between them when 13 components are used in the PCR model.

(c)

No, the model chosen by the **best subset selection method** has only **9 features**

Chapter 8 - Problem 8

(a)

```
## [1] "The dimensions of the training set is the following"
```

```
## [1] 286  11
```

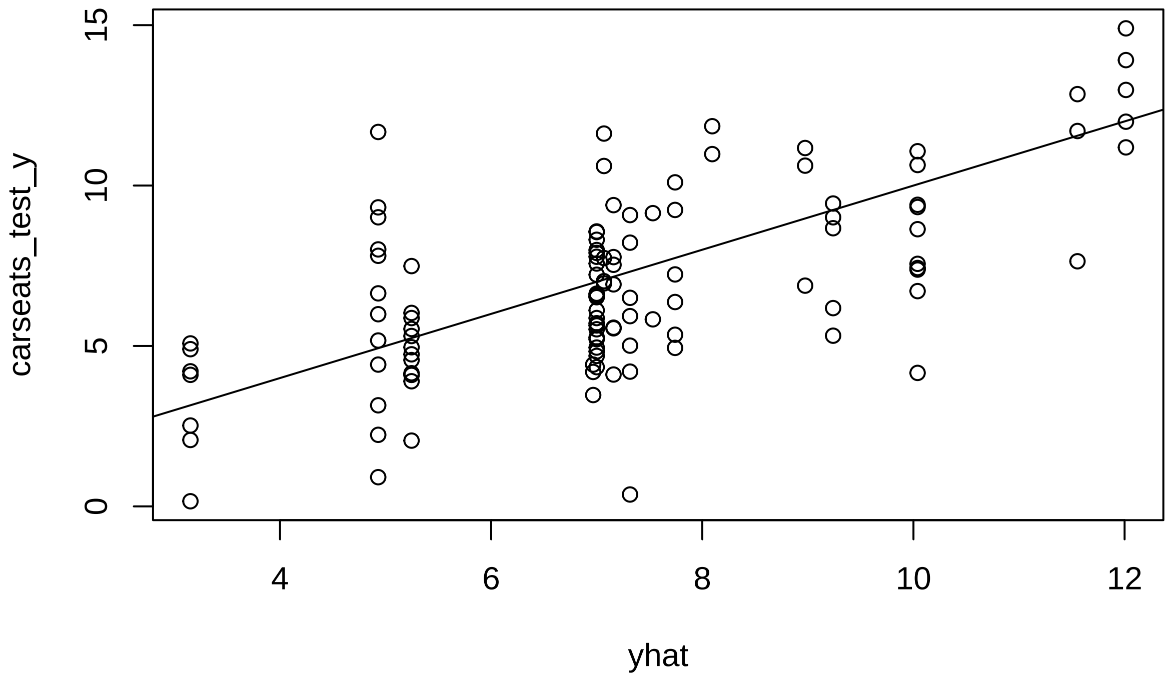
```
## [1] "The dimensions of the test set is the following"
```

```
## [1] 114  11
```

The train and test set is split in a **75-25 ratio**

(b)

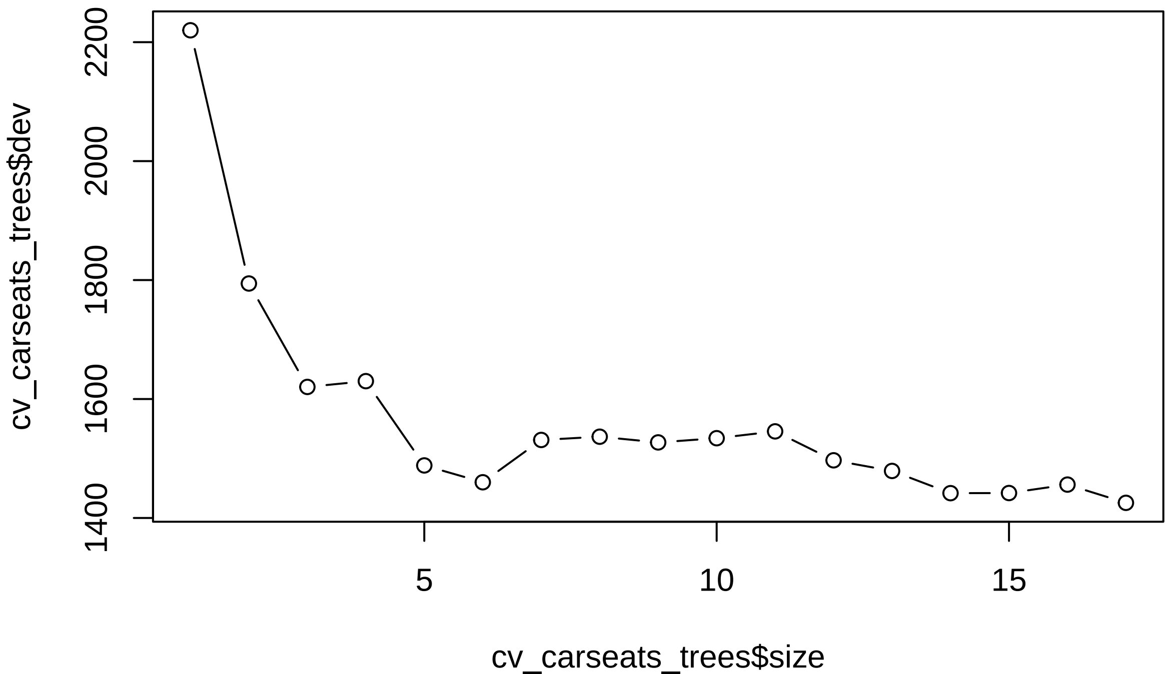
```
##
## Regression tree:
## tree(formula = Sales ~ ., data = carseats_train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Income"      "CompPrice"   "Population"
## [6] "Advertising" "Age"
## Number of terminal nodes:  17
## Residual mean deviance:  2.485 = 668.6 / 269
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.71600 -1.03800  0.03745  0.00000  1.01000  4.25700
```



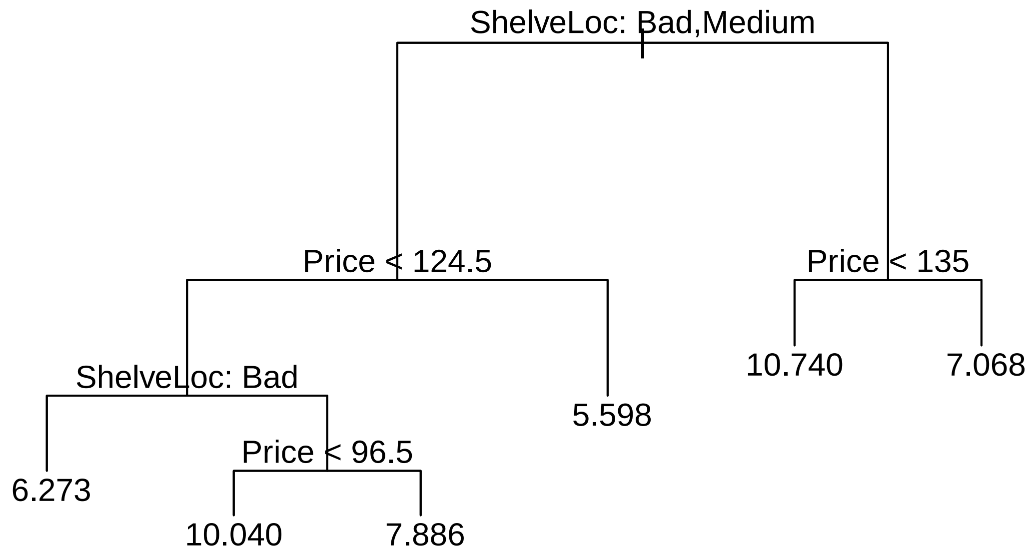

```
## [1] 4.94213
```

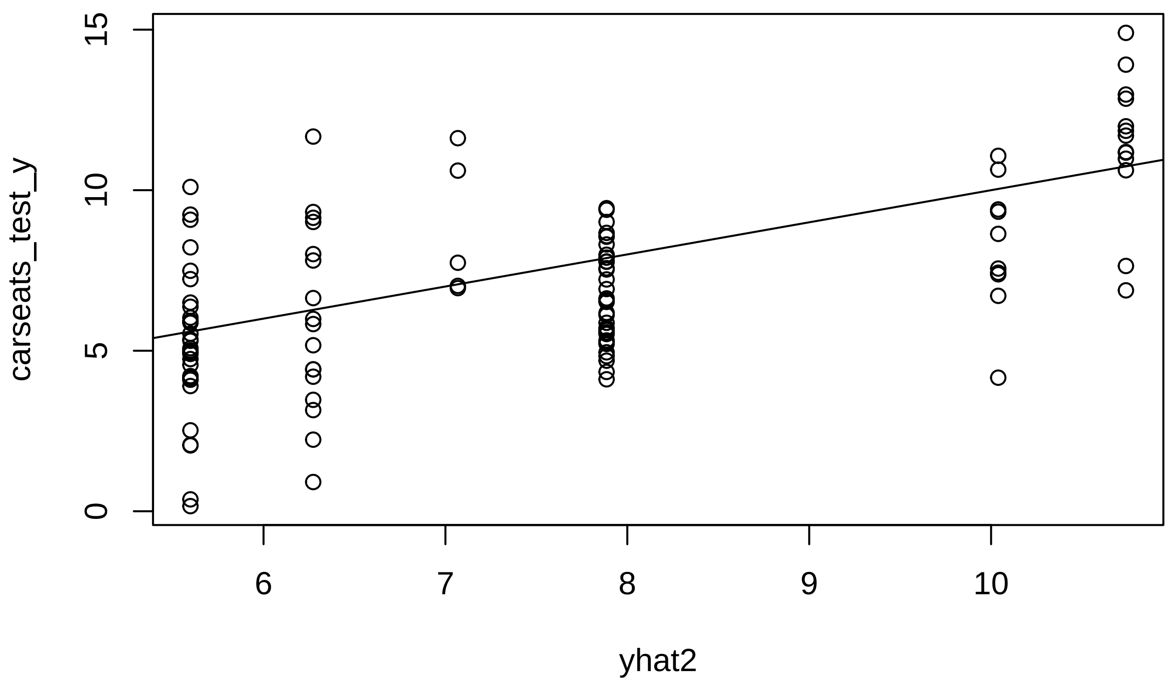
The test error rate obtained is as seen above i.e **4.942**

(c)



We can see from the above plot, the error starts increasing after we hit **6 trees**. Further on, we're going to use this value to prune our tree.


ShelveLoc: Bad,Medium
Price < 124.5
Price < 135
ShelveLoc: Bad
5.598
10.740
7.068
Price < 96.5
6.273
10.040
7.886



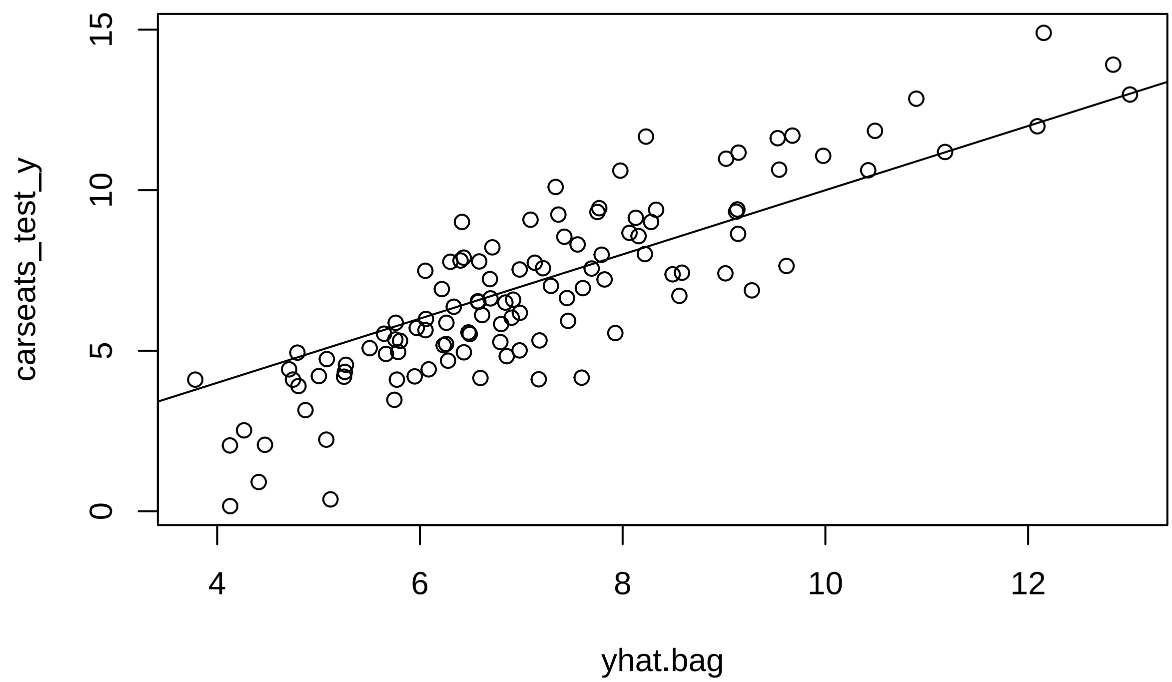

```
## [1] 5.357507
```

Post pruning, we see that the test MSE has increased to **5.358.**

Pruning the trees ended up **increasing** the Test MSE.

(d)

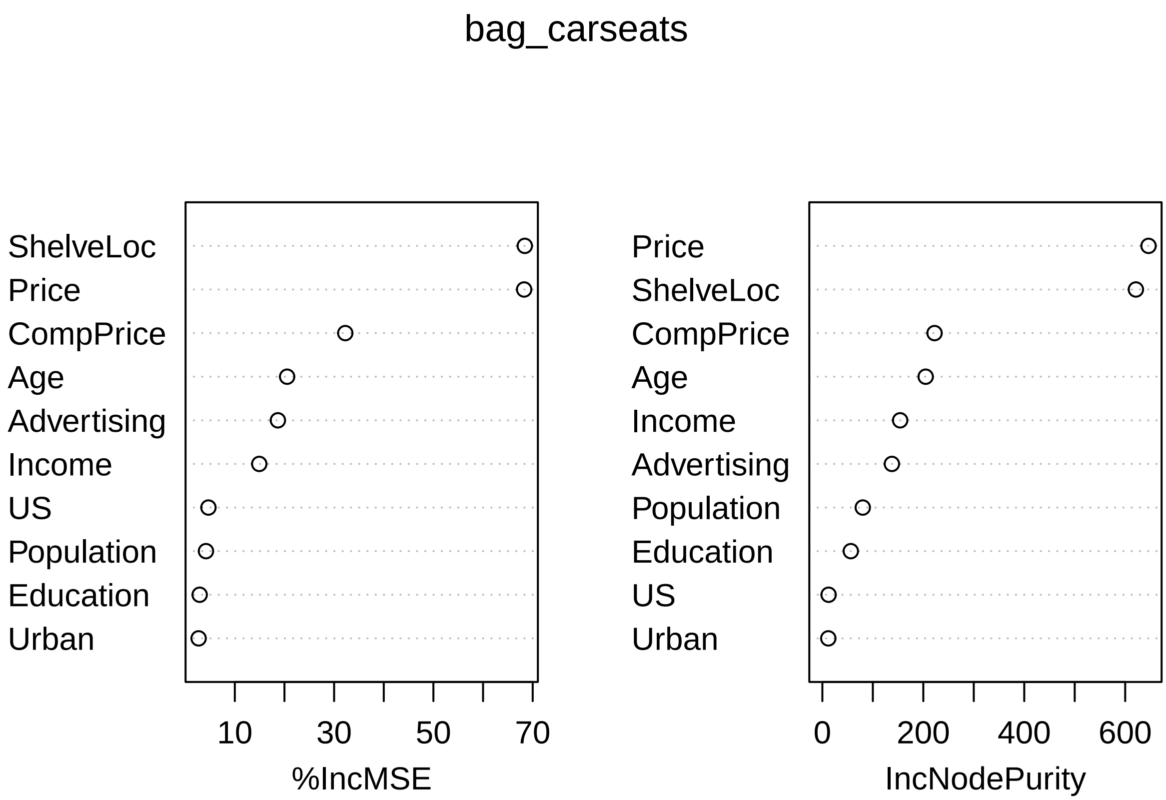
```
##
## Call:
##  randomForest(formula = Sales ~ ., data = carseats_train, mtry = 10,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.448431
##                     % Var explained: 68.29
```




```
## [1] 2.403354
```

Bagging has **improved** the **Test MSE**. The value obtained for **Test MSE** after bagging is **2.403**

```
##               %IncMSE IncNodePurity
## CompPrice   32.251723     222.33057
## Income      14.927098     154.11822
## Advertising 18.679808     137.76261
## Population   4.187291      80.05684
## Price       68.276365     646.31549
## ShelveLoc   68.411227     621.20825
## Age         20.539334     204.70767
## Education    2.921856      56.38866
## Urban        2.709359      11.80821
## US           4.684811      12.39557
```



The most important variables are **ShelveLoc**, **Price** and **CompPrice**

(e)

```
## [1] 2.427966
```

We obtained the lowest test MSE with mtry = **8**

The notice that the test MSE value decreases as we increase the mtry values and after a certain point the change is very small. We hit this point at mtry = 8.

```
##                %IncMSE IncNodePurity
## CompPrice   27.1162522     218.41429
## Income      11.5424941     164.59576
## Advertising 17.9065424     152.41006
## Population   2.0070515      82.60821
## Price       65.6724186     629.03680
## ShelveLoc   68.8937612     608.43570
## Age         20.0008911     214.67219
## Education    2.3349131      59.44528
## Urban        0.1092269      10.93052
## US           3.0271523      13.00809
```

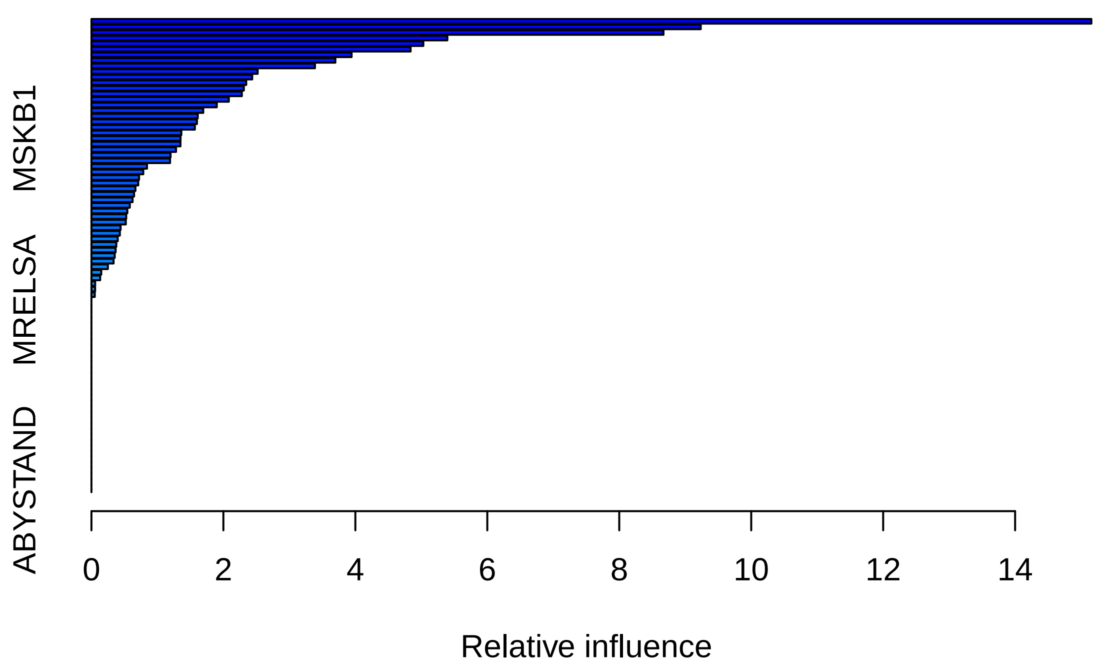
The top 3 features from **Random Forest are ShelveLoc, Price and CompPrice**

Chapter 8 - Problem 11

(a)

```
## [1] "The training set dimensions"

## [1] 1000   86

## [1] "The test set dimensions"

## [1] 4822   86
```

(b)




```
##               var     rel.inf
## PPERSAUT PPERSAUT 15.15534009
## MKOOPKLA MKOOPKLA  9.23499526
## MOPLHOOG MOPLHOOG  8.67017024
## MBERMIDD MBERMIDD  5.39403655
## MGODGE     MGODGE  5.03047673
## PBRAND     PBRAND  4.83740038
## MINK3045 MINK3045  3.94305387
## ABRAND     ABRAND  3.69692919
## MOSTYPE   MOSTYPE  3.38768960
```

```
## PWAPART   PWAPART  2.51970169
## MGODPR     MGODPR  2.43689096
## MSKC         MSKC  2.34594774
## MAUT2       MAUT2  2.30973409
## MFWEKIND MFWEKIND  2.27959503
## MBERARBG MBERARBG  2.08245286
## MSKA         MSKA  1.90020973
## PBYSTAND PBYSTAND  1.69481877
## MGODOV     MGODOV  1.61147668
## MAUT1       MAUT1  1.59879109
## MBERHOOG MBERHOOG  1.56791308
## MINK7512 MINK7512  1.36255296
## MSKB1       MSKB1  1.35071475
## MINKGEM   MINKGEM  1.34913011
## MRELGE     MRELGE  1.28204167
## MAUT0       MAUT0  1.19929798
## MHHUUR     MHHUUR  1.19158719
## MFGEKIND MFGEKIND  0.84203310
## MRELOV     MRELOV  0.78554535
## MZPART     MZPART  0.72191139
## MINK4575 MINK4575  0.70935967
## MSKB2       MSKB2  0.66694112
## APERSAUT APERSAUT  0.64644681
## MGODRK     MGODRK  0.62380797
## MSKD         MSKD  0.58168337
## MINKM30   MINKM30  0.54392696
## PMOTSCO   PMOTSCO  0.52708603
## MOPLMIDD MOPLMIDD  0.52091706
## MGEMOMV   MGEMOMV  0.44231264
## MZFONDS   MZFONDS  0.43037800
## PLEVEN     PLEVEN  0.39901552
## MHKOOP     MHKOOP  0.37672230
## MBERARBO MBERARBO  0.36653424
## MBERBOER MBERBOER  0.35290257
## MINK123M MINK123M  0.33559225
## MGEMLEEF MGEMLEEF  0.24937634
## MFALLEEN MFALLEEN  0.14898856
## MOSHOOFD MOSHOOFD  0.13265308
## MOPLLAAG MOPLLAAG  0.05654615
## MBERZELF MBERZELF  0.05589282
## MAANTHUI MAANTHUI  0.05047841
## 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 variables **PPERSAUT** , **MKOOPKLA** and **MOPLHOOG** are the most important variables

(c)

```
##    boost.pred
##        0    1
##   0 4396  137
##   1  255   34
```

```
## [1] 0.1988304
```

19.88% of the people predicted to make purchase actually end up making one.

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

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

About **14%** of people predicted to make purchase using logistic regression actually end up making one. This is lower than boosting.

Chapter 10 - 7

```
## [1] "Correlation Based Distance"
```

```
##             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
```

```
## [1] "Squared Euclidean Distance"
```

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

From the above output, we can **validated the hypothesis** stated in the question. Both the measures are almost **equivalent**.

For example, let's take a look at Assualt and Murder. Their correlation based distance is **19.813** and their Squared Euclidean distance is **19.416**

Problem 1: Beauty Pays!

(1)

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = df_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
```

From the **regression output above**, we can say that Beauty Score has a **strong positive effect** on the **Course Ratings**. Looking the p-value, we can conclude that this association is **statistically significant**. To ensure that the other determinants are not skewing our analysis, it would make sense to **control for the relevant ones**. In our case all the above stated features are relevant to our analysis

(2)

The key to even beginning address this question running a **natural experiment with blind people**. Armed with the results of this experiment, we would be able to say with certainty **if beauty determines the teaching ability of the teachers or if beautiful teachers are given higher course ratings**. I believe that the output above is implying that this has more to do with the latter.

Problem 2: Mid City!

(1)

```
##
## Call:
## lm(formula = Price ~ brick_dum + N2 + N3 + Offers + SqFt + Bedrooms +
##     Bathrooms, data = Housing_struc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27337.3  -6549.5    -41.7   5803.4  27359.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2159.498   8877.810   0.243  0.80823
## brick_dum   17297.350   1981.616   8.729 1.78e-14 ***
## N2          -1560.579   2396.765  -0.651  0.51621
## N3          20681.037   3148.954   6.568 1.38e-09 ***
## Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
## SqFt           52.994      5.734   9.242 1.10e-15 ***
## Bedrooms     4246.794   1597.911   2.658  0.00894 **
## Bathrooms    7883.278   2117.035   3.724  0.00030 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10020 on 120 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.861
## F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept) -15417.94711 19736.94349
## brick_dum    13373.88702 21220.81203
## N2           -6306.00785  3184.84961
## N3           14446.32799 26915.74671
## Offers      -10415.27089 -6119.70575
## SqFt            41.64034    64.34714
## Bedrooms      1083.04162  7410.54616
## Bathrooms     3691.69572 12074.86126
```

We created **dummy variables** for the **categorical variables i.e Brick and Neighborhood**

To answer the first question, we could refer to the output obtained. Since the **confidence interval** of the coefficient of brick_dum (Brick == 'Yes') **doesn't include zero** and there's a **positive correlation between price and our dummy variable for Brick houses**, we can conclude that it is one of the **important factors** when it comes to predicting price. Hence, **there is a premium for brick houses**

(2)

In the regression output obtained above, we can see the **confidence interval of the coefficients of Neighborhood 3 does not include 0** and there's a **positive correlation** between price and our dummy variable for Neighborhood 3. This is proof enough to conclude that **there is a premium for houses in Neighborhood 3**

(3)

```
##
## Call:
## lm(formula = Price ~ brick_dum + N2 + N3 + Offers + SqFt + Bedrooms +
##     Bathrooms + N3_brick, data = Housing_struc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26939.1  -5428.7   -213.9   4519.3  26211.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3009.993   8706.264   0.346  0.73016
## brick_dum   13826.465   2405.556   5.748 7.11e-08 ***
## N2           -673.028   2376.477  -0.283  0.77751
## N3          17241.413   3391.347   5.084 1.39e-06 ***
## Offers      -8401.088   1064.370  -7.893 1.62e-12 ***
## SqFt           54.065      5.636   9.593  < 2e-16 ***
## Bedrooms     4718.163   1577.613   2.991  0.00338 **
## Bathrooms    6463.365   2154.264   3.000  0.00329 **
## N3_brick    10181.577   4165.274   2.444  0.01598 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9817 on 119 degrees of freedom
## Multiple R-squared:  0.8749, Adjusted R-squared:  0.8665
## F-statistic:   104 on 8 and 119 DF,  p-value: < 2.2e-16
```

```
##                   2.5 %      97.5 %
## (Intercept) -14229.27947 20249.26635
## brick_dum     9063.22323 18589.70668
## N2           -5378.69058  4032.63406
## N3           10526.20666 23956.61921
## Offers      -10508.64698 -6293.52887
## SqFt            42.90493    65.22463
## Bedrooms      1594.33302  7841.99385
## Bathrooms     2197.70794 10729.02197
## N3_brick      1933.91810 18429.23657
```

To check if there is a premium for brick houses in Neighborhood 3, I created an **interaction term** between the Brick houses dummy variables and Neighborhood 3 and **modified my linear model** to include it.

From the regression output above, we can see that there is **positive correlation** between the created interaction term and the price. Additionally the **confidence interval** of the coefficient **does not include 0**. This is proof enough to conclude that **there is a premium for brick houses in Neighborhood 3**

(4)

```
##
## Call:
## lm(formula = Price ~ brick_dum + N2 + N3 + Offers + SqFt + Bedrooms +
##     Bathrooms, data = Housing_struc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27337.3  -6549.5    -41.7   5803.4  27359.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2159.498   8877.810   0.243  0.80823
## brick_dum   17297.350   1981.616   8.729 1.78e-14 ***
## N2          -1560.579   2396.765  -0.651  0.51621
## N3          20681.037   3148.954   6.568 1.38e-09 ***
## Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
## SqFt           52.994      5.734   9.242 1.10e-15 ***
## Bedrooms     4246.794   1597.911   2.658  0.00894 **
## Bathrooms    7883.278   2117.035   3.724  0.00030 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10020 on 120 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.861
## F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept) -15417.94711 19736.94349
## brick_dum    13373.88702 21220.81203
## N2           -6306.00785  3184.84961
## N3           14446.32799 26915.74671
## Offers      -10415.27089 -6119.70575
## SqFt            41.64034    64.34714
## Bedrooms      1083.04162  7410.54616
## Bathrooms     3691.69572 12074.86126
```

From the regression output above, we can see that the **confidence interval** of coefficient corresponding to the Neighborhood 2 **does include 0** and hence according to **null hypothesis Neighborhood** 2 can be considered **unimportant** in predicted price by itself and **can be combined with Neighborhood 1 into a single "older" neighborhood to predict price**

Problem 3: What causes what??

(1)

The main problem here is of the difference between correlation and causuation. What we would be able to tell from the data is that there if there's a correlation between police and crime. But we won't able to determine if more police is causing more crime or more crime is causing more police. Unless there are control variables thrown in, we won't be able to dig deeper into determining the causual relationship.

(2)

The researchers collected data on crime in DC on high alert days i.e days on which there could potentially be a terrorist attack. This was a natural experiment. From the table 2, we can see that there is a statistically significant negetive correlation between the number of police officers and crime on high alert days.

(3)

The reason being that people tend not to step out higher alert days which would translate to lower crime because lesser chances of crime. The table 2 shows that even after controlling for ridership, more police has a negetive impact on crime.This was to ensure that lower number of people wasn't causing a decrease in crime but rather higher number of police officers.

(4)

The model being estimated here dug deeper into the effect of high alert days across districts. They included interaction terms between location and high alert days to support the analysis. It is clear from the values of the coefficients that the effect is predominant is district 1 compared to the other districts. The effect is other districts is negetive but the relative standard error is high and no statistically significant. Hence we can conclude that higher alert in district 1 has a much higher imapct on reducing crimes has compared to the other districts.

Problem 4:

Group no: (2)

Problem Statement:

Predicting the Austin house prices based on a number of features

Approach:

EDA –> Feature Seelection –> Data Preparation –> Modelling –> Model Improvement and Model Selection

My contribution:

My work started off with looking into datasets and determining which would be the most relevant and interesting to proceed with. I used Kaggle mainly to scan for the potential problem statements. We individually proposed ideas for this particular step in the process and ended up narrowing down on the above problem statement. Moving on, I played a role in feature selection. I started off with scanning the dataset for potential features to be included in the model. We had 45 features at start of this process and ended up eliminating 34 of them based on business intuition. This would probably seem harsh at the first look, but proper validation was conducted to ensure that the selected features were indeed the most influential in determining the price of a house in Austin.

Digging deeper, I took up the task of the preparing the data for analysis. This step included a critical piece which had a big influence on the prices i.e adjusting our target variable for inflation. To accomplish the aforementioned task, I relied on the housing activity data found on the website recenter.tamu.edu. I picked up montly Median housing pricing information and used it gain an insight into house appreciation over years and ensure that our targer variable is adjusted for the same. The reason why this step was critical is because our target variable "latestPrice" had prices for the houses as of "latestSaleDate" which ranged from 2018-2021. Hence, using this approaching I adjusted this column to indicate the prices for the current year i.e 2022.

Moving on, the next step in the data prep was to tranform our key features, I encoded the categorical features as 0 & 1. Additionally, created an age column based on the year the house was built as I thought this would be one of the factors determining the price of the house. This hypothesis was validated later on, when we actually did see a trend in prices based on the age feature. We noticed that the houses that were extremely old are very highly prices, following that we had the new houses which were priced higher than the houses which fell in the middle bucket. This made sense to us because the extremely old houses are considered vintage and would sell at high prices as discussed above and new ones are probably ones with a modern built and newer features, whereas, the middle bucket is not new or vintage would sell at lower prices.

Data Preparation also included adding filters to the dataset based on the features being discarded and extreme values of selected features that could potentially skew our analysis. The following filters were added to the following features: 1. HomeType: Single Family
2. City: Austin
3. hasSpa: False
4. hasHeating: True 5. hasCooling: True
6. numOfBathrooms and numOfBedrooms: Non-zero .

My work didn't end there, once the modelling was complete, I collated the results and thought it would be interesting to look at the Geographical significance of all the work put in by our group. Hence I made a side-by-side comparison of the prices predicted by our best performing model plotted against Latitute and Longitute and the map of Austin. This turned out to be very fruitful as the prices predicted by our best performing had a similar trend when compared to what we already knew about Austin.

Lastly, I chimed in documentation of all the work on the deck created by our group. I lay out a skeleton outlining the flow of the deck and ensured that it was coherent and telling a story in a way we wanted it to be told.

In conclusion, I contributed to the following aspects of the project i.e dataset selection, feature selection, data preparation, result validation, deck preparation.