Exam Submission

Mahika Bansal

8/01/2021

1) ISLR Chapter 2 Q10: This exercise involves the Boston housing data set.

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

Ans: The data:

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

## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.5
24 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
```

```
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

The data contains 506 rows and 14 columns which represents crime rate in different Boston suburbs

After checking the data definition for dataset, this data frame contains the following columns:

crim: per capita crime rate by town.

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

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per $10,000.

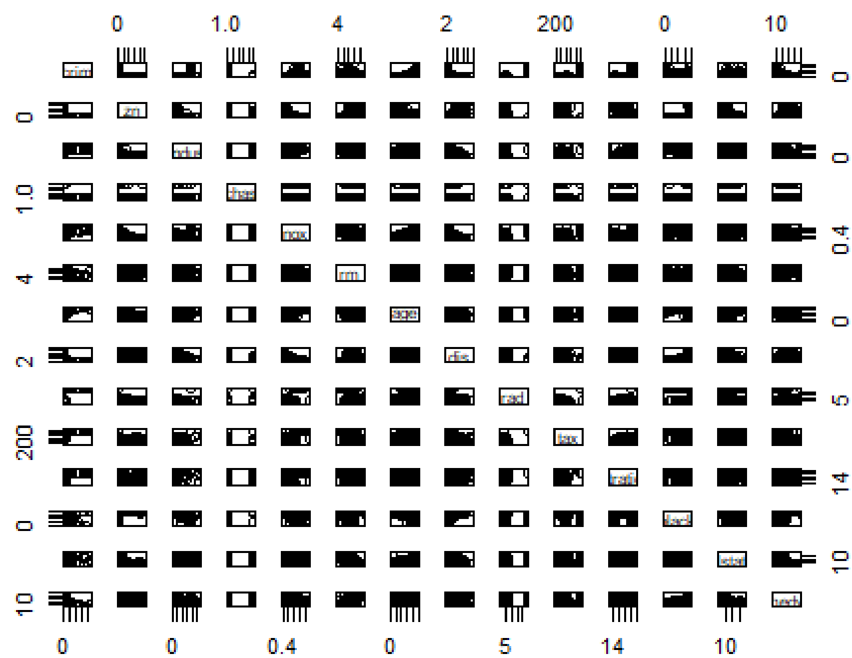
ptratio: pupil-teacher ratio by town.

black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in $1000s.

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




```
##                crim         zn      indus        nox         rm        age
## crim      1.0000000 -0.2004692  0.4065834  0.4209717 -0.2192467  0.3527343
## zn       -0.2004692  1.0000000 -0.5338282 -0.5166037  0.3119906 -0.5695373
## indus     0.4065834 -0.5338282  1.0000000  0.7636514 -0.3916759  0.6447785
## nox       0.4209717 -0.5166037  0.7636514  1.0000000 -0.3021882  0.7314701
## rm       -0.2192467  0.3119906 -0.3916759 -0.3021882  1.0000000 -0.2402649
## age       0.3527343 -0.5695373  0.6447785  0.7314701 -0.2402649  1.0000000
## dis      -0.3796701  0.6644082 -0.7080270 -0.7692301  0.2052462 -0.7478805
## rad       0.6255051 -0.3119478  0.5951293  0.6114406 -0.2098467  0.4560225
## tax       0.5827643 -0.3145633  0.7207602  0.6680232 -0.2920478  0.5064556
## ptratio   0.2899456 -0.3916785  0.3832476  0.1889327 -0.3555015  0.2615150
## black    -0.3850639  0.1755203 -0.3569765 -0.3800506  0.1280686 -0.2735340
## lstat     0.4556215 -0.4129946  0.6037997  0.5908789 -0.6138083  0.6023385
## medv     -0.3883046  0.3604453 -0.4837252 -0.4273208  0.6953599 -0.3769546
##                 dis        rad        tax    ptratio      black      lstat
## crim     -0.3796701  0.6255051  0.5827643  0.2899456 -0.3850639  0.4556215
## zn        0.6644082 -0.3119478 -0.3145633 -0.3916785  0.1755203 -0.4129946
## indus    -0.7080270  0.5951293  0.7207602  0.3832476 -0.3569765  0.6037997
## nox      -0.7692301  0.6114406  0.6680232  0.1889327 -0.3800506  0.5908789
## rm        0.2052462 -0.2098467 -0.2920478 -0.3555015  0.1280686 -0.6138083
## age      -0.7478805  0.4560225  0.5064556  0.2615150 -0.2735340  0.6023385
## dis       1.0000000 -0.4945879 -0.5344316 -0.2324705  0.2915117 -0.4969958
## rad      -0.4945879  1.0000000  0.9102282  0.4647412 -0.4444128  0.4886763
```

```
## tax     -0.5344316  0.9102282  1.0000000  0.4608530 -0.4418080  0.5439934
## ptratio -0.2324705  0.4647412  0.4608530  1.0000000 -0.1773833  0.3740443
## black    0.2915117 -0.4444128 -0.4418080 -0.1773833  1.0000000 -0.3660869
## lstat   -0.4969958  0.4886763  0.5439934  0.3740443 -0.3660869  1.0000000
## medv     0.2499287 -0.3816262 -0.4685359 -0.5077867  0.3334608 -0.7376627
##               medv
## crim    -0.3883046
## zn       0.3604453
## indus   -0.4837252
## nox     -0.4273208
## rm       0.6953599
## age     -0.3769546
## dis      0.2499287
## rad     -0.3816262
## tax     -0.4685359
## ptratio -0.5077867
## black    0.3334608
## lstat   -0.7376627
## medv     1.0000000
```

Crim seems to be highly correlated with rad and tax. Nox, indus; rad, tax are some other coefficients with high correlation

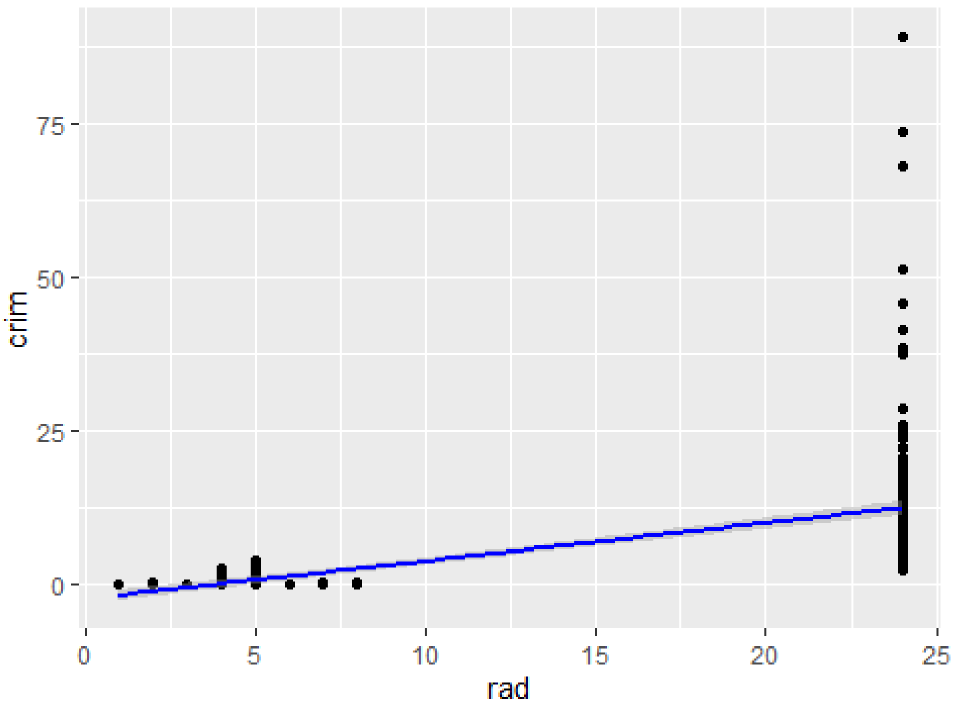
(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```
##         zn      indus        nox         rm        age        dis         r
ad
## -0.2004692  0.4065834  0.4209717 -0.2192467  0.3527343 -0.3796701  0.62550
51
##        tax    ptratio      black      lstat       medv
##  0.5827643  0.2899456 -0.3850639  0.4556215 -0.3883046
```

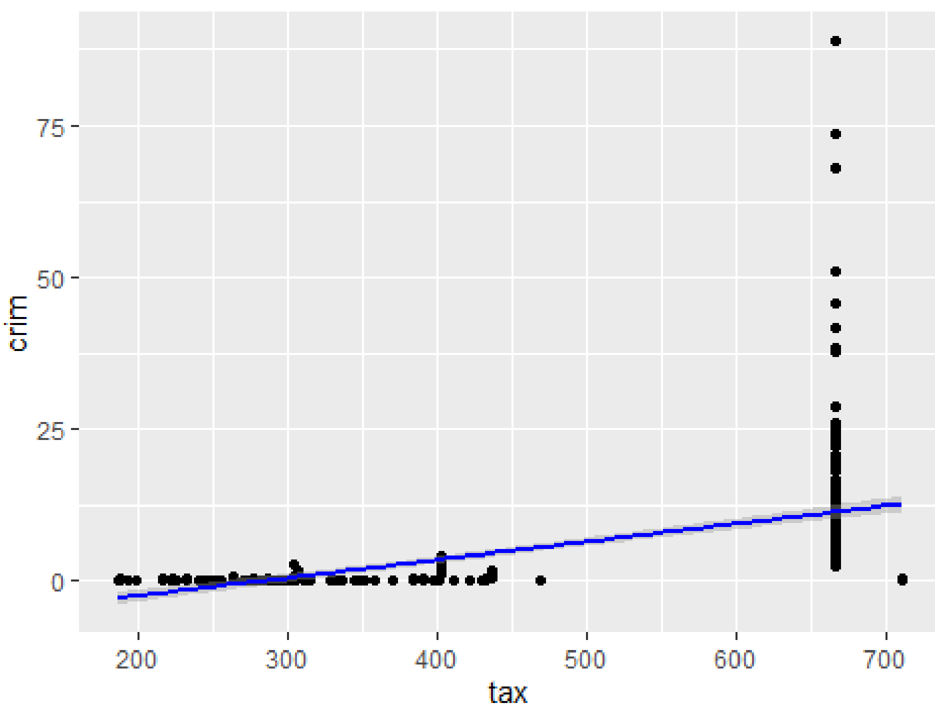
rad and tax seem to be highly correlated to crim.

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

rad vs crim

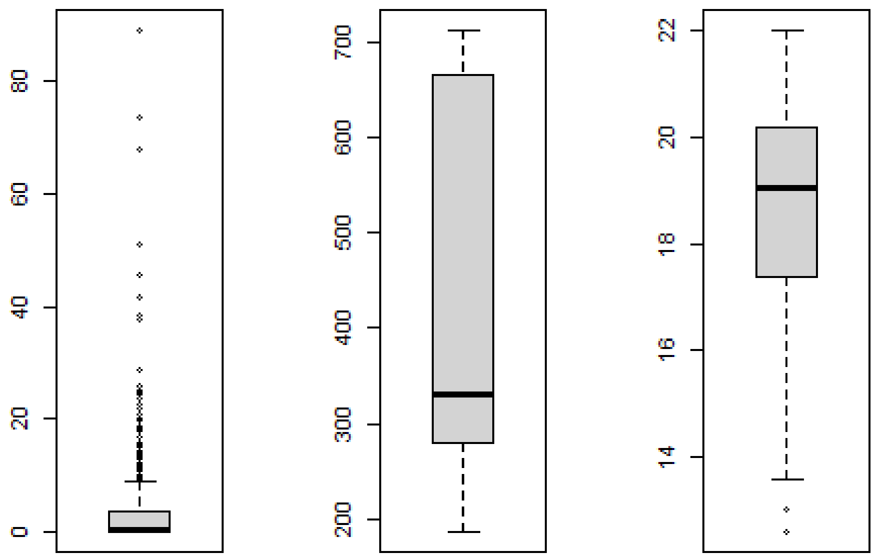


tax vs crim



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

```
##       crim                tax           ptratio
##  Min.   : 0.00632   Min.   :187.0   Min.   :12.60
##  1st Qu.: 0.08205   1st Qu.:279.0   1st Qu.:17.40
##  Median : 0.25651   Median :330.0   Median :19.05
##  Mean   : 3.61352   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 3.67708   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :88.97620   Max.   :711.0   Max.   :22.00
```



For crime rates, some cities tend to have higher crime rates, thus the high amount of outliers. Tax has a large range and thus some suburbs have high tax rate than others. Pupil-teacher ratio still has a smaller range and thus the ratio is less different over different suburbs.

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

```
##
##   0   1
## 471  35
```

35 suburbs are bound by the river.

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

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

Most suburbs have 19.05 as the pupil teacher ratio.

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black ls
tat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30
.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22
.98
##     medv
## 399    5
## 406    5
```

```
##       crim                zn             indus       chas         nox
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.385
0
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.449
0
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.538
0
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.554
7
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.624
0
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.871
0
##        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
```

Two suburbs have lowest median values, with value= 5k$. Compared to overall variables, it has high crim, age and lstat ranges along with indus, nox, rad, tax & ptratio.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

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

64 suburbs have >7 rooms on average per dwelling

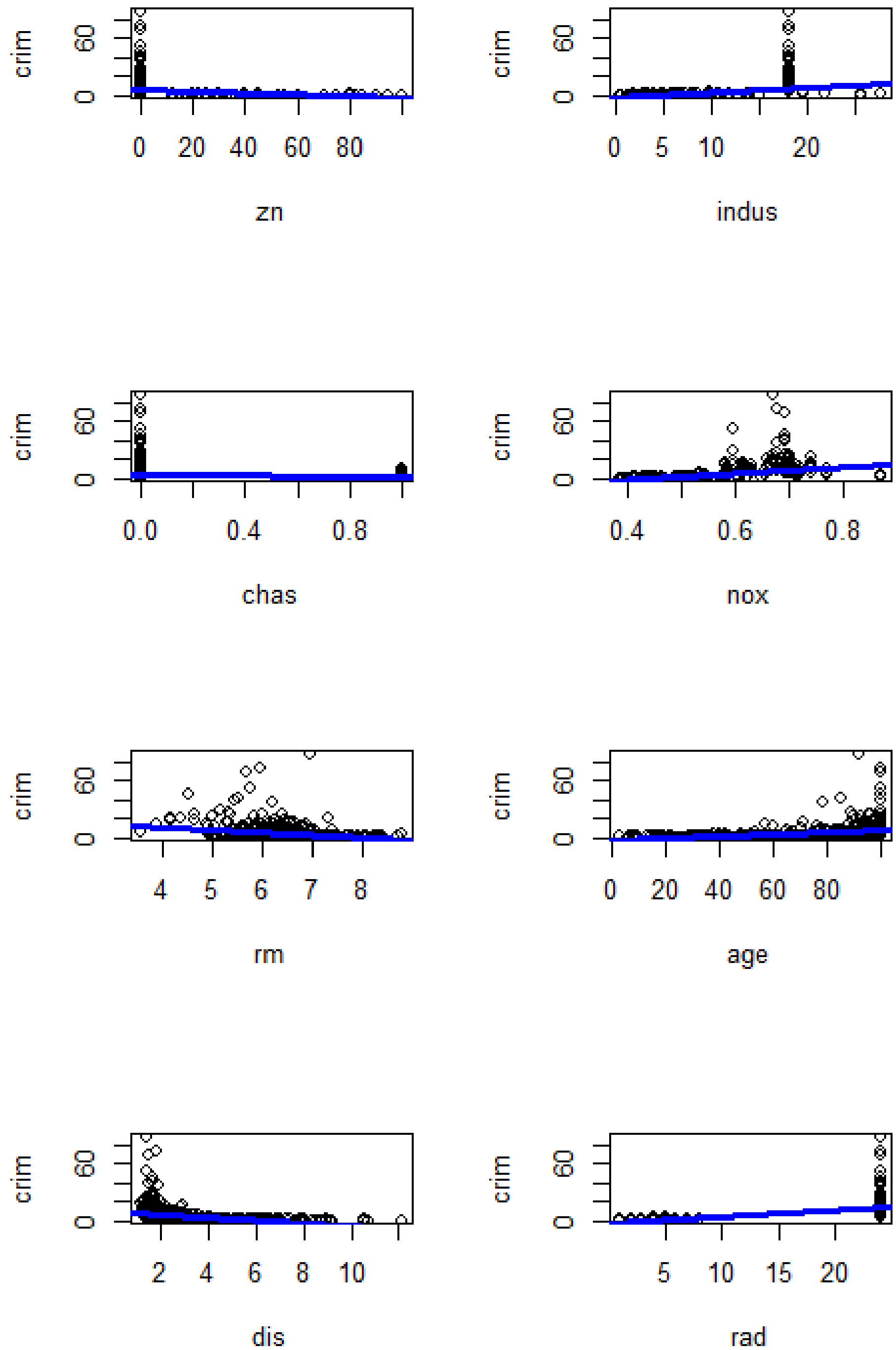
```
## [1] 13
```

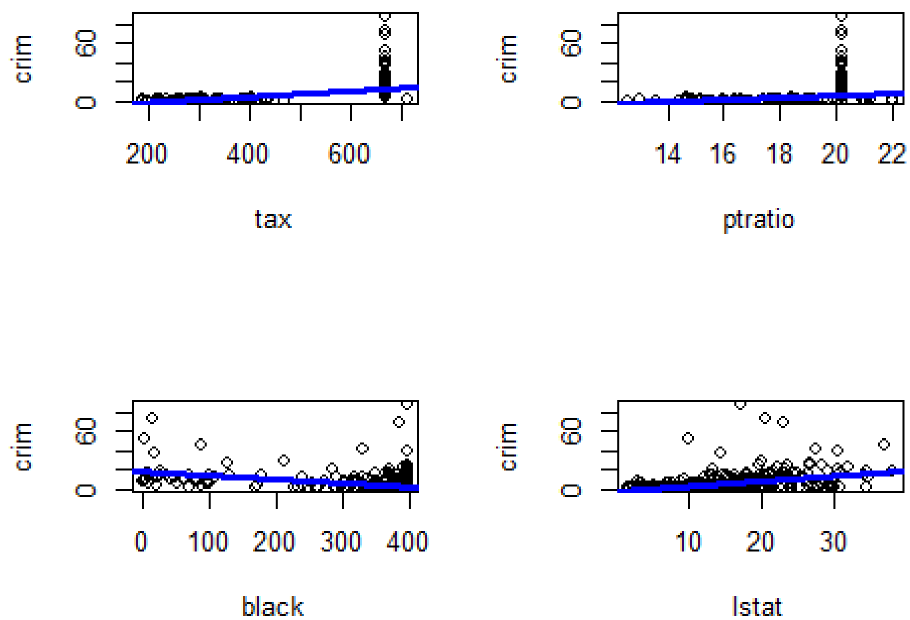
13 suburbs have >8 rooms on average per dwelling

The suburbs with more than 8 rooms have almost an avaerage distribution across variables except lstat and medv, whose values are too low and too high respectively, i.e might be posh suburbs

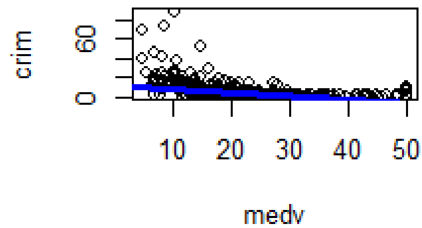
2) ISLR Chapter 3 Q15: This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.


crim
0
60
0 20 40 60 80
zn
crim
0
60
0 5 10 20
indus
crim
0
60
0.0 0.4 0.8
chas
crim
0
60
0.4 0.6 0.8
nox
crim
0
60
4 5 6 7 8
rm
crim
0
60
0 20 40 60 80
age
crim
0
60
2 4 6 8 10
dis
crim
0
60
5 10 15 20
rad




```
##     Variable    R.Square  Intercept       Slope P.Value.Variable
## 1         zn 0.040187908   4.453694 -0.07393498     5.506472e-06
## 2      indus 0.165310070  -2.063743  0.50977633     1.450349e-21
## 3       chas 0.003123869   3.744447 -1.89277655     2.094345e-01
## 4        nox 0.177217182 -13.719882 31.24853120     3.751739e-23
## 5         rm 0.048069117  20.481804 -2.68405122     6.346703e-07
## 6        age 0.124421452  -3.777906  0.10778623     2.854869e-16
## 7        dis 0.144149375   9.499262 -1.55090168     8.519949e-19
## 8        rad 0.391256687  -2.287159  0.61791093     2.693844e-56
## 9        tax 0.339614243  -8.528369  0.02974225     2.357127e-47
## 10   ptratio 0.084068439 -17.646933  1.15198279     2.942922e-11
## 11     black 0.148274239  16.553529 -0.03627964     2.487274e-19
## 12     lstat 0.207590933  -3.330538  0.54880478     2.654277e-27
## 13      medv 0.150780469  11.796536 -0.36315992     1.173987e-19
```



All the variables seem to have a p-value lower than 0.05 except chas (0.2049) and thus significant.

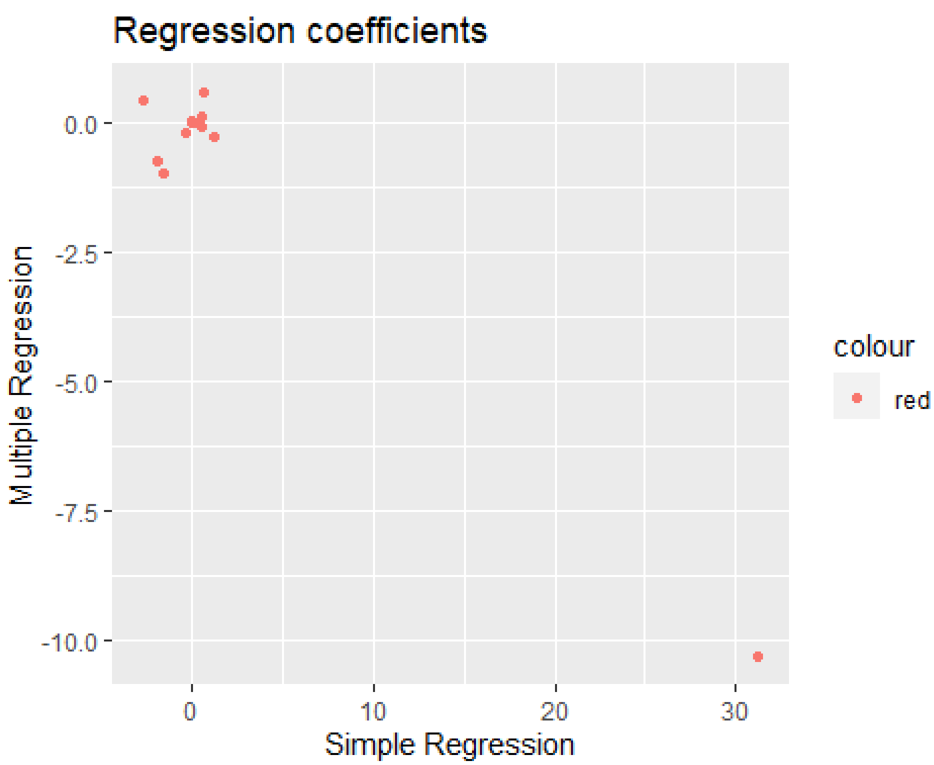
(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

```
##
## 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
## chas          -0.749134   1.180147  -0.635 0.525867
## nox          -10.313535   5.275536  -1.955 0.051152 .
## rm             0.430131   0.612830   0.702 0.483089
## age            0.001452   0.017925   0.081 0.935488
## dis           -0.987176   0.281817  -3.503 0.000502 ***
## rad            0.588209   0.088049   6.680 6.46e-11 ***
```

```
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Null hypothesis can be rejected for zn, dis, rad, black and medv. For other variables, the p-value is above 0.05 and thus not good enough to reject null hypothesis for these values

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.



(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form Y = β0 + β1X + β2X2 + β3X3 + e.

3) ISLR Chapter 6 Q9: In this exercise, we will predict the number of applications received using the other variables in the College data set.

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

```
## [1] 777  18

## [1] 621
```

Out of 777 observations, we've kept 621 in train

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

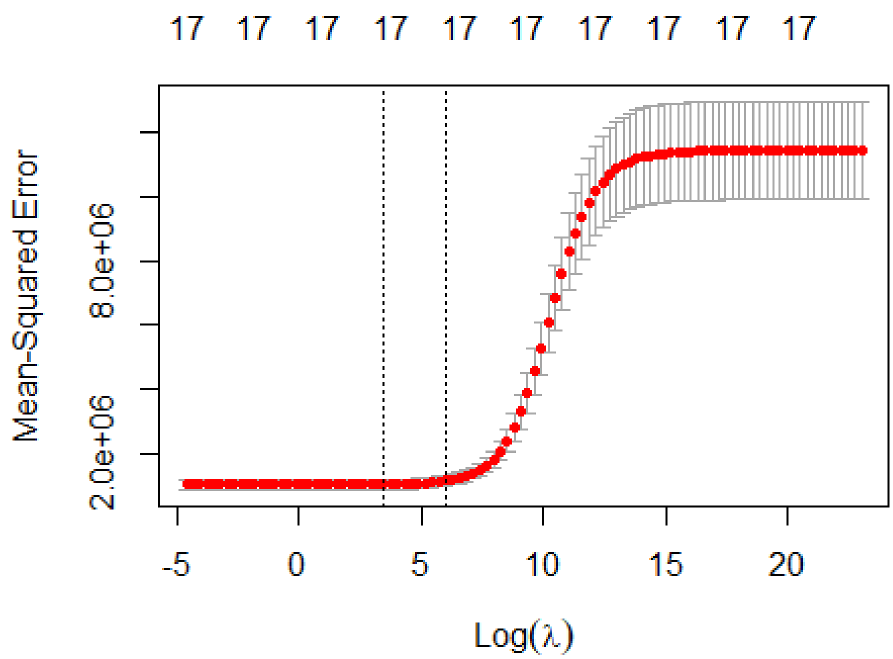
```
##
## Call:
## lm(formula = Apps ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3257.7  -431.1   -57.5   318.8  6581.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.475e+02  4.238e+02  -1.056  0.29141
## PrivateYes  -5.964e+02  1.471e+02  -4.055 5.67e-05 ***
## Accept       1.262e+00  5.474e-02  23.060  < 2e-16 ***
## Enroll      -2.867e-01  1.960e-01  -1.463  0.14402
## Top10perc    4.485e+01  5.787e+00   7.749 3.93e-14 ***
## Top25perc   -1.362e+01  4.713e+00  -2.889  0.00400 **
## F.Undergrad  9.257e-02  3.473e-02   2.665  0.00790 **
## P.Undergrad  4.950e-03  3.319e-02   0.149  0.88150
## Outstate    -5.318e-02  1.962e-02  -2.710  0.00692 **
## Room.Board   1.615e-01  4.929e-02   3.277  0.00111 **
## Books        5.242e-02  2.402e-01   0.218  0.82734
## Personal    -8.572e-03  6.533e-02  -0.131  0.89565
## PhD         -5.727e+00  4.779e+00  -1.199  0.23118
## Terminal    -5.017e+00  5.205e+00  -0.964  0.33546
## S.F.Ratio    3.827e+00  1.342e+01   0.285  0.77560
## perc.alumni -6.235e+00  4.325e+00  -1.442  0.14991
## Expend       7.915e-02  1.270e-02   6.233 8.58e-10 ***
## Grad.Rate    1.064e+01  3.063e+00   3.474  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 971.6 on 603 degrees of freedom
## Multiple R-squared:  0.9192, Adjusted R-squared:  0.9169
## F-statistic: 403.6 on 17 and 603 DF,  p-value: < 2.2e-16

## [1] 1449.199
```

RMSE for linear model using OLS : 1449.199

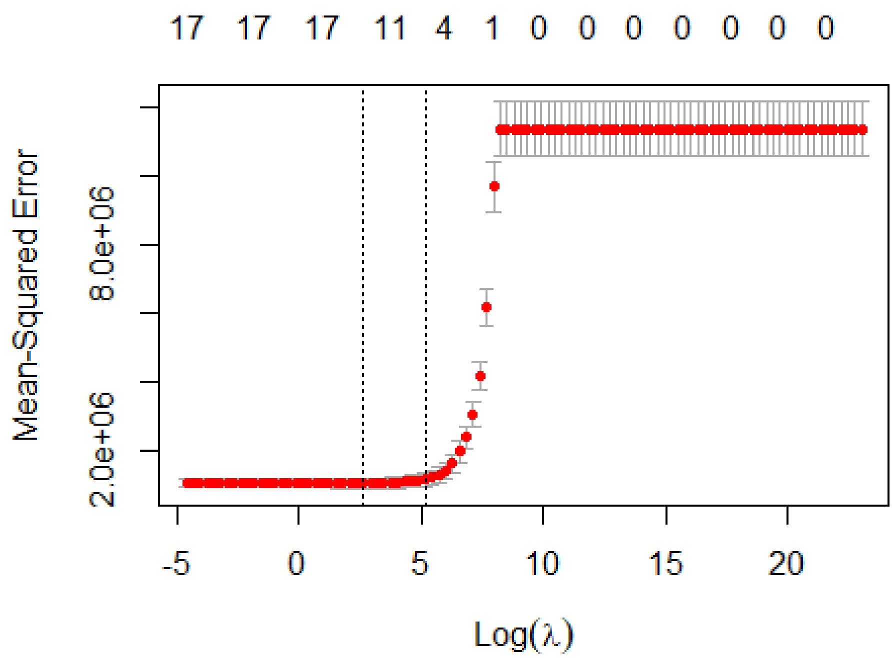
(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

```
## Warning: package 'caret' was built under R version 4.0.5
```




```
## [1] 1554.134
```

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

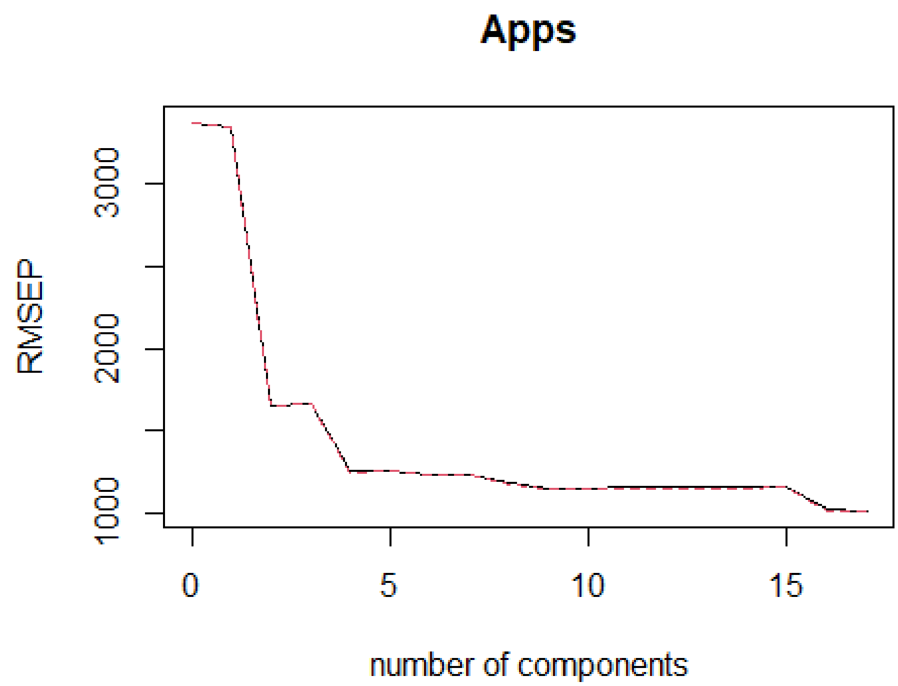



```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                           1
## (Intercept) -589.31720740
## (Intercept)    .
## PrivateYes  -562.10288568
## Accept         1.21313968
## Enroll         .
## Top10perc     35.35449674
## Top25perc     -6.11257774
## F.Undergrad    0.06047836
## P.Undergrad    .
## Outstate      -0.03770090
## Room.Board     0.14305703
## Books          .
## Personal       .
## PhD           -4.24228637
## Terminal      -4.53781138
## S.F.Ratio      .
## perc.alumni   -6.64312557
## Expend         0.07496814
## Grad.Rate      8.41227514
```

```
## [1] 1495.871
```

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

```
## Data:    X dimension: 621 17
##  Y dimension: 621 1
## Fit method: svdpc
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3374     3352     1650     1663     1263     1262     1233
## adjCV         3374     3353     1648     1666     1249     1252     1231
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1232     1178     1152      1152      1153      1154      1155
## adjCV     1231     1173     1150      1150      1151      1152      1153
##        14 comps  15 comps  16 comps  17 comps
## CV         1155      1155      1020      1016
## adjCV      1152      1154      1017      1012
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 com
ps
## X      31.624    57.27    64.26    69.99    75.20    80.18    83.87    87.
34
## Apps    2.251    76.45    76.47    86.58    86.61    87.03    87.13    88.
43
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       90.49     92.95     95.04     96.91     98.02     98.87     99.40
## Apps    89.04     89.07     89.08     89.12     89.16     89.18     89.33
##       16 comps  17 comps
## X        99.81    100.00
## Apps     91.60     91.92
```

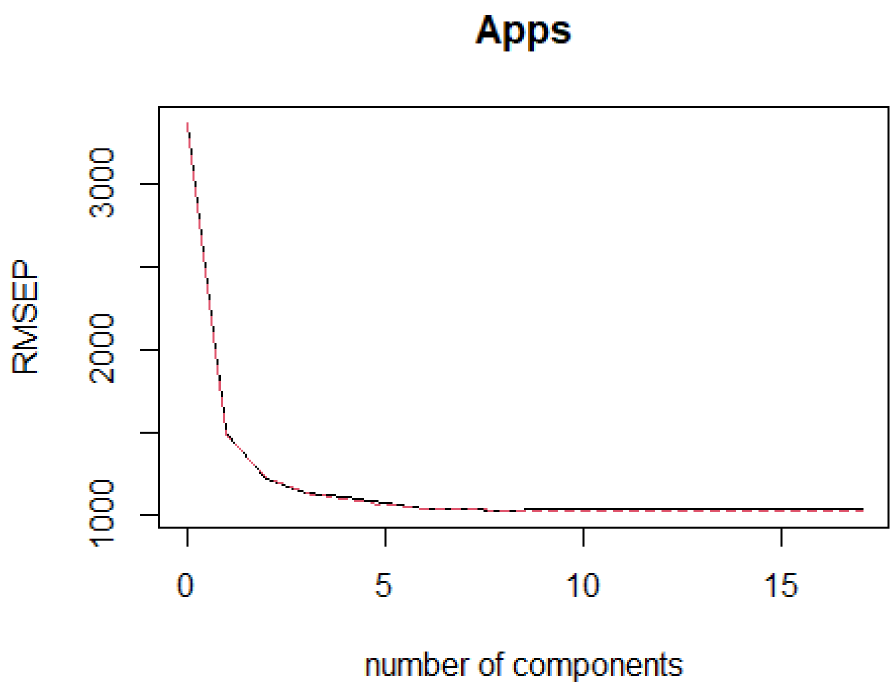



```
## [1] 1449.199
```

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

```
## Data:    X dimension: 621 17
##  Y dimension: 621 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            3374     1493     1221     1134     1104     1067     1037
## adjCV         3374     1490     1224     1132     1100     1053     1033
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1031     1026     1028      1029      1028      1028      1028
## adjCV     1027     1022     1025      1025      1024      1024      1024
##        14 comps  15 comps  16 comps  17 comps
## CV         1028      1028      1028      1028
## adjCV      1024      1024      1024      1024
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 com
ps
```

```
## X       25.76    40.17    62.71    65.98    67.68    72.92    77.11    80.
37
## Apps    81.13    87.45    89.45    90.35    91.60    91.79    91.84    91.
87
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       82.59     84.54     87.05     90.45     92.97     95.09     97.07
## Apps    91.89     91.91     91.92     91.92     91.92     91.92     91.92
##       16 comps  17 comps
## X        98.41    100.00
## Apps     91.92     91.92
```




```
## [1] 1449.211
```

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

PCR, OLS and PLS have similar predictions, but not much difference within the techniques

4) ISLR Chapter 6 Q11: We will now try to predict per capita crime rate in the Boston data set.

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

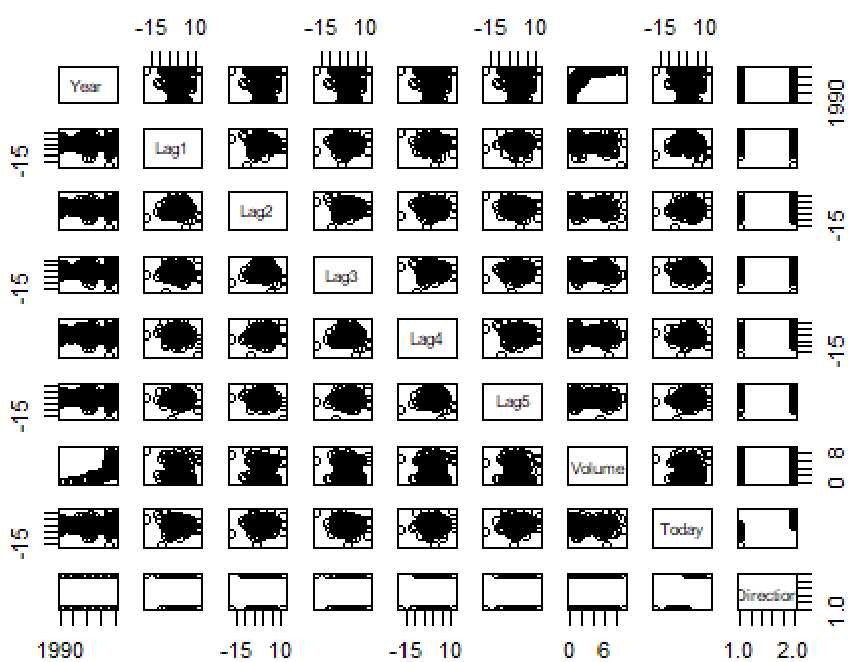
```
## Warning: package 'leaps' was built under R version 4.0.5
```

(b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, crossvalidation, or some other reasonable alternative, as opposed to using training error.

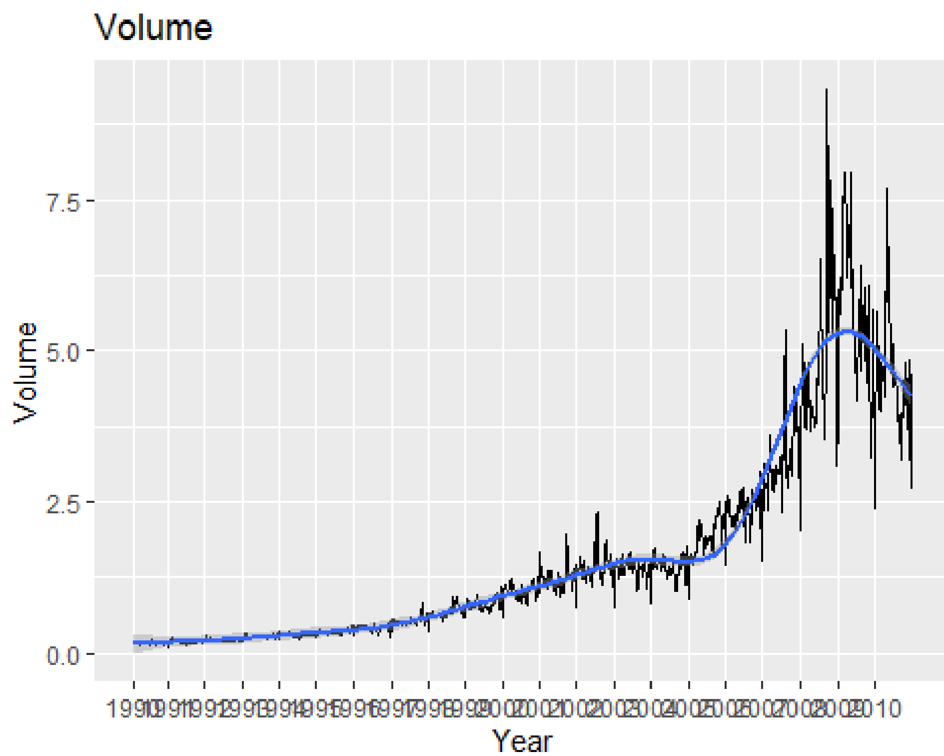
(c) Does your chosen model involve all of the features in the data set? Why or why not?

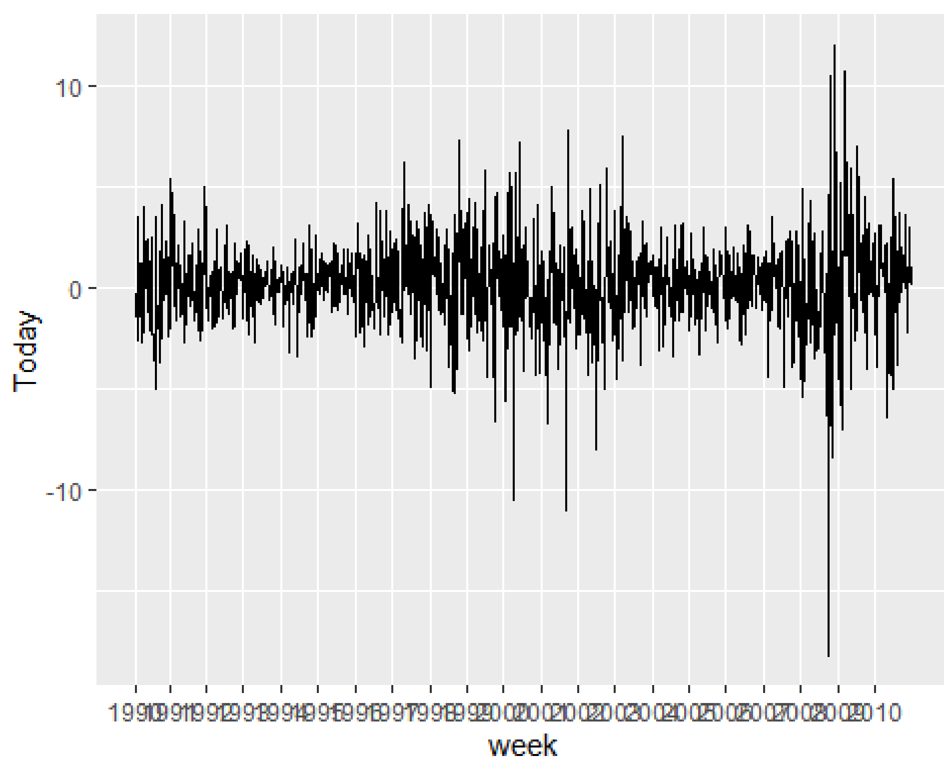
5) ISLR Chapter 4 Q10: This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

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



```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000

##
## Down   Up
##  484  605
```





There seems to be some pattern in the pred intervals.

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

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   54  48
##       Up    430 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.5556
```

```
##     P-Value [Acc > NIR] : 0.369
##
##                   Kappa : 0.035
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.11157
##             Specificity : 0.92066
##          Pos Pred Value : 0.52941
##          Neg Pred Value : 0.56434
##              Prevalence : 0.44444
##          Detection Rate : 0.04959
##    Detection Prevalence : 0.09366
##       Balanced Accuracy : 0.51612
##
##        'Positive' Class : Down
##

## [1] 0.5610652
```

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9  5
##       Up     34 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.2439
##
##                   Kappa : 0.1414
##
##  Mcnemar's Test P-Value : 7.34e-06
##
##             Sensitivity : 0.20930
##             Specificity : 0.91803
##          Pos Pred Value : 0.64286
##          Neg Pred Value : 0.62222
##              Prevalence : 0.41346
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.13462
##       Balanced Accuracy : 0.56367
```

```
##
##        'Positive' Class : Down
##
```

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

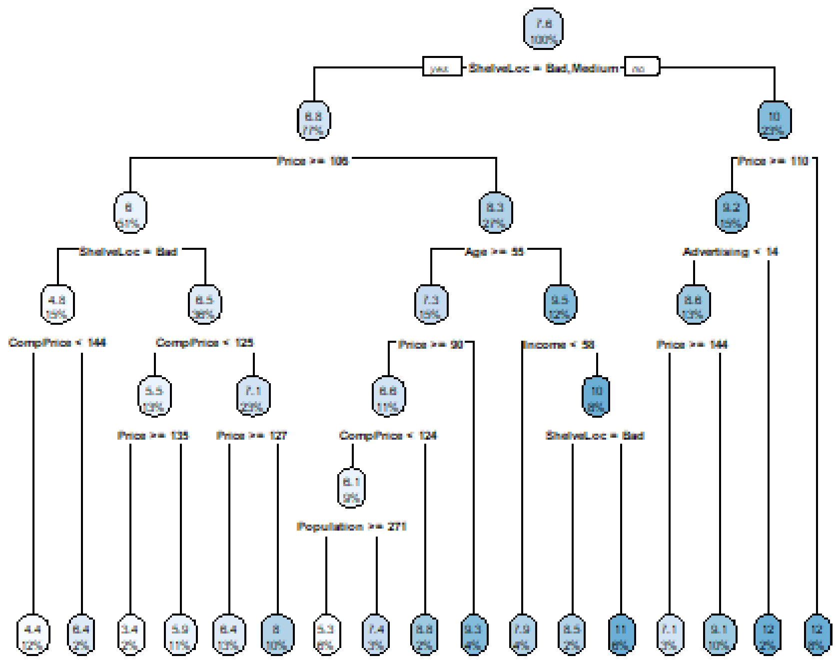
(h) Which of these methods appears to provide the best results on this data?

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

6) ISLR Chapter 8 Q8: In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

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

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?



```
## Call:
## rpart(formula = Sales ~ ., data = train)
##   n= 320
##
##            CP nsplit rel error    xerror       xstd
## 1  0.25230925      0 1.0000000 1.0077092 0.07495569
## 2  0.11111198      1 0.7476907 0.7562111 0.05558115
## 3  0.06776315      2 0.6365788 0.6681456 0.05005373
## 4  0.03974713      3 0.5688156 0.6103059 0.04559716
## 5  0.03958533      4 0.5290685 0.6210473 0.04623989
## 6  0.02629057      5 0.4894832 0.5837891 0.04361248
## 7  0.02617869      6 0.4631926 0.5631657 0.04212497
## 8  0.02470713      7 0.4370139 0.5562131 0.04243727
## 9  0.01756165      8 0.4123068 0.5664121 0.04431057
## 10 0.01628121      9 0.3947451 0.5652330 0.04498939
## 11 0.01615288     10 0.3784639 0.5570172 0.04288226
## 12 0.01425398     11 0.3623110 0.5479911 0.04205973
## 13 0.01397874     12 0.3480571 0.5456586 0.04254927
## 14 0.01157763     13 0.3340783 0.5543549 0.04258965
## 15 0.01080417     14 0.3225007 0.5502428 0.04267747
## 16 0.01023457     15 0.3116965 0.5568318 0.04334267
## 17 0.01000000     16 0.3014620 0.5559352 0.04336971
##
## Variable importance
##   ShelveLoc       Price   CompPrice         Age      Income  Population
##          36          31          13           6           4           4
## Advertising   Education
```

```
##           4           2
##
## Node number 1: 320 observations,    complexity param=0.2523093
##   mean=7.593469, MSE=8.248624
##   left son=2 (247 obs) right son=3 (73 obs)
##   Primary splits:
##       ShelveLoc   splits as  LRL,       improve=0.25230930, (0 missing)
##       Price       < 94.5  to the right, improve=0.15753930, (0 missing)
##       Age         < 61.5  to the right, improve=0.08558690, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.06421279, (0 missing)
##       Income      < 61.5  to the left,  improve=0.04186375, (0 missing)
##   Surrogate splits:
##       Price < 168.5 to the left,  agree=0.775, adj=0.014, (0 split)
##
## Node number 2: 247 observations,    complexity param=0.111112
##   mean=6.80919, MSE=6.032473
##   left son=4 (162 obs) right son=5 (85 obs)
##   Primary splits:
##       Price       < 105.5 to the right, improve=0.19683400, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=0.12077880, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.09439914, (0 missing)
##       Age         < 68.5  to the right, improve=0.08836071, (0 missing)
##       Income      < 61.5  to the left,  improve=0.07943847, (0 missing)
##   Surrogate splits:
##       CompPrice  < 109.5 to the right, agree=0.745, adj=0.259, (0 split)
##       Population < 507.5 to the left,  agree=0.668, adj=0.035, (0 split)
##       Income     < 22.5  to the right, agree=0.664, adj=0.024, (0 split)
##
## Node number 3: 73 observations,    complexity param=0.06776315
##   mean=10.24712, MSE=6.62402
##   left son=6 (49 obs) right son=7 (24 obs)
##   Primary splits:
##       Price       < 109.5 to the right, improve=0.36989670, (0 missing)
##       Age         < 61.5  to the right, improve=0.17051180, (0 missing)
##       Education   < 11.5  to the right, improve=0.11888560, (0 missing)
##       Advertising < 13.5  to the left,  improve=0.11294940, (0 missing)
##       Population  < 345.5 to the left,  improve=0.06314431, (0 missing)
##   Surrogate splits:
##       CompPrice  < 113.5 to the right, agree=0.712, adj=0.125, (0 split)
##       Population < 92.5  to the right, agree=0.712, adj=0.125, (0 split)
##       Age        < 26.5  to the right, agree=0.699, adj=0.083, (0 split)
##       Education  < 11.5  to the right, agree=0.699, adj=0.083, (0 split)
##
## Node number 4: 162 observations,    complexity param=0.03974713
##   mean=6.019877, MSE=4.383369
##   left son=8 (47 obs) right son=9 (115 obs)
##   Primary splits:
##       ShelveLoc   splits as  L-R,       improve=0.14774550, (0 missing)
##       CompPrice   < 124.5 to the left,  improve=0.09958325, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.09712898, (0 missing)
```

```
##       Age         < 65.5  to the right, improve=0.07510875, (0 missing)
##       Price       < 135.5 to the right, improve=0.07474084, (0 missing)
##   Surrogate splits:
##       Population < 15    to the left,  agree=0.722, adj=0.043, (0 split)
##       Age        < 28.5  to the left,  agree=0.716, adj=0.021, (0 split)
##
## Node number 5: 85 observations,    complexity param=0.03958533
##   mean=8.313529, MSE=5.725039
##   left son=10 (47 obs) right son=11 (38 obs)
##   Primary splits:
##       Age       < 54.5  to the right, improve=0.2147179, (0 missing)
##       ShelveLoc splits as  L-R,       improve=0.1736511, (0 missing)
##       CompPrice < 123.5 to the left,  improve=0.1713369, (0 missing)
##       Price     < 88    to the right, improve=0.1347772, (0 missing)
##       Income    < 57.5  to the left,  improve=0.1154715, (0 missing)
##   Surrogate splits:
##       CompPrice   < 124.5 to the left,  agree=0.671, adj=0.263, (0 split)
##       Income      < 72.5  to the right, agree=0.612, adj=0.132, (0 split)
##       Population  < 167   to the right, agree=0.612, adj=0.132, (0 split)
##       Advertising < 13.5  to the left,  agree=0.588, adj=0.079, (0 split)
##       Price       < 75    to the right, agree=0.588, adj=0.079, (0 split)
##
## Node number 6: 49 observations,    complexity param=0.02617869
##   mean=9.151633, MSE=4.818891
##   left son=12 (41 obs) right son=13 (8 obs)
##   Primary splits:
##       Advertising < 13.5  to the left,  improve=0.2926417, (0 missing)
##       Price       < 144   to the right, improve=0.2224431, (0 missing)
##       Age         < 68.5  to the right, improve=0.1852674, (0 missing)
##       US          splits as  LR,        improve=0.1761469, (0 missing)
##       Population  < 345.5 to the left,  improve=0.1538057, (0 missing)
##
## Node number 7: 24 observations
##   mean=12.48375, MSE=2.85679
##
## Node number 8: 47 observations,    complexity param=0.01023457
##   mean=4.761064, MSE=3.514771
##   left son=16 (39 obs) right son=17 (8 obs)
##   Primary splits:
##       CompPrice  < 144   to the left,  improve=0.16353330, (0 missing)
##       Age        < 61.5  to the right, improve=0.15572390, (0 missing)
##       Price      < 143.5 to the right, improve=0.14969230, (0 missing)
##       Population < 283   to the left,  improve=0.12009560, (0 missing)
##       Income     < 101   to the left,  improve=0.09803783, (0 missing)
##   Surrogate splits:
##       Price < 160   to the left,  agree=0.872, adj=0.25, (0 split)
##
## Node number 9: 115 observations,    complexity param=0.02629057
##   mean=6.534348, MSE=3.826058
##   left son=18 (41 obs) right son=19 (74 obs)
```

```
##   Primary splits:
##       CompPrice   < 124.5 to the left,  improve=0.15771830, (0 missing)
##       Income      < 61.5  to the left,  improve=0.13728750, (0 missing)
##       Age         < 49.5  to the right, improve=0.12328360, (0 missing)
##       Advertising < 5.5   to the left,  improve=0.11122190, (0 missing)
##       Price       < 135.5 to the right, improve=0.09066396, (0 missing)
##   Surrogate splits:
##       Price      < 111.5 to the left,  agree=0.739, adj=0.268, (0 split)
##       Population < 499.5 to the right, agree=0.661, adj=0.049, (0 split)
##       Income     < 29.5  to the left,  agree=0.652, adj=0.024, (0 split)
##       Age        < 25.5  to the left,  agree=0.652, adj=0.024, (0 split)
##
## Node number 10: 47 observations,    complexity param=0.02470713
##   mean=7.316596, MSE=4.922095
##   left son=20 (35 obs) right son=21 (12 obs)
##   Primary splits:
##       Price      < 89.5  to the right, improve=0.28190700, (0 missing)
##       ShelveLoc  splits as  L-R,       improve=0.21501930, (0 missing)
##       Income     < 85.5  to the left,  improve=0.21124550, (0 missing)
##       Population < 271   to the right, improve=0.14635480, (0 missing)
##       CompPrice  < 123.5 to the left,  improve=0.08532898, (0 missing)
##   Surrogate splits:
##       Income    < 103.5 to the left,  agree=0.809, adj=0.250, (0 split)
##       CompPrice < 102.5 to the right, agree=0.787, adj=0.167, (0 split)
##
## Node number 11: 38 observations,    complexity param=0.01756165
##   mean=9.546579, MSE=3.968475
##   left son=22 (12 obs) right son=23 (26 obs)
##   Primary splits:
##       Income      < 57.5  to the left,  improve=0.3073898, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=0.2552900, (0 missing)
##       Advertising < 9.5   to the left,  improve=0.2204056, (0 missing)
##       CompPrice   < 124   to the left,  improve=0.1545814, (0 missing)
##       US          splits as  LR,        improve=0.1257007, (0 missing)
##   Surrogate splits:
##       Population < 448.5 to the right, agree=0.711, adj=0.083, (0 split)
##       US         splits as  LR,        agree=0.711, adj=0.083, (0 split)
##
## Node number 12: 41 observations,    complexity param=0.01157763
##   mean=8.627073, MSE=3.712894
##   left son=24 (10 obs) right son=25 (31 obs)
##   Primary splits:
##       Price       < 144   to the right, improve=0.2007496, (0 missing)
##       Age         < 63.5  to the right, improve=0.1975074, (0 missing)
##       Income      < 35.5  to the left,  improve=0.1644589, (0 missing)
##       US          splits as  LR,        improve=0.1272098, (0 missing)
##       Advertising < 0.5   to the left,  improve=0.1073208, (0 missing)
##   Surrogate splits:
##       CompPrice  < 154.5 to the right, agree=0.805, adj=0.2, (0 split)
##       Income     < 102.5 to the right, agree=0.805, adj=0.2, (0 split)
```

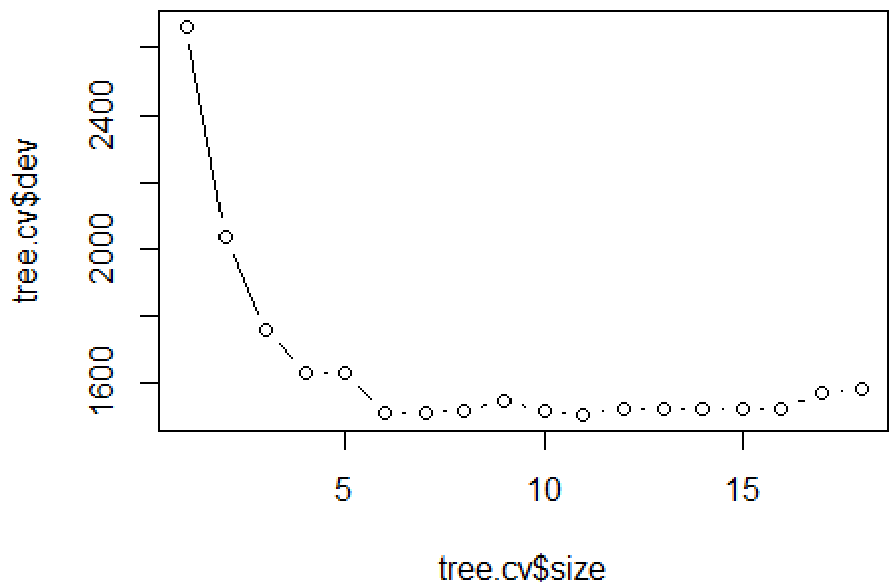
```
##       Population < 144.5 to the left,  agree=0.780, adj=0.1, (0 split)
##       Age        < 33    to the left,  agree=0.780, adj=0.1, (0 split)
##
## Node number 13: 8 observations
##   mean=11.84, MSE=1.8496
##
## Node number 16: 39 observations
##   mean=4.417692, MSE=2.073643
##
## Node number 17: 8 observations
##   mean=6.435, MSE=7.163425
##
## Node number 18: 41 observations,    complexity param=0.01397874
##   mean=5.490732, MSE=3.248402
##   left son=36 (7 obs) right son=37 (34 obs)
##   Primary splits:
##       Price       < 134.5 to the right, improve=0.2770422, (0 missing)
##       Advertising < 5.5   to the left,  improve=0.2754412, (0 missing)
##       US          splits as  LR,        improve=0.1855262, (0 missing)
##       Income      < 83.5  to the left,  improve=0.1625895, (0 missing)
##       Age         < 68    to the right, improve=0.1312075, (0 missing)
##
## Node number 19: 74 observations,    complexity param=0.01628121
##   mean=7.112568, MSE=3.208333
##   left son=38 (42 obs) right son=39 (32 obs)
##   Primary splits:
##       Price       < 127   to the right, improve=0.1810118, (0 missing)
##       Advertising < 13.5  to the left,  improve=0.1688513, (0 missing)
##       Income      < 41    to the left,  improve=0.1293059, (0 missing)
##       Age         < 33.5  to the right, improve=0.1157115, (0 missing)
##       Education   < 16.5  to the right, improve=0.1019203, (0 missing)
##   Surrogate splits:
##       CompPrice   < 133.5 to the right, agree=0.703, adj=0.313, (0 split)
##       Income      < 60.5  to the left,  agree=0.622, adj=0.125, (0 split)
##       Advertising < 4.5   to the right, agree=0.622, adj=0.125, (0 split)
##       Age         < 38    to the right, agree=0.622, adj=0.125, (0 split)
##       Education   < 16.5  to the right, agree=0.608, adj=0.094, (0 split)
##
## Node number 20: 35 observations,    complexity param=0.01615288
##   mean=6.626857, MSE=3.91633
##   left son=40 (28 obs) right son=41 (7 obs)
##   Primary splits:
##       CompPrice  < 123.5 to the left,  improve=0.3110527, (0 missing)
##       ShelveLoc  splits as  L-R,       improve=0.2612636, (0 missing)
##       Population < 271   to the right, improve=0.2400183, (0 missing)
##       Age        < 63.5  to the right, improve=0.2373383, (0 missing)
##       Education  < 11.5  to the left,  improve=0.1598416, (0 missing)
##   Surrogate splits:
##       Price < 103.5 to the left,  agree=0.829, adj=0.143, (0 split)
##
```

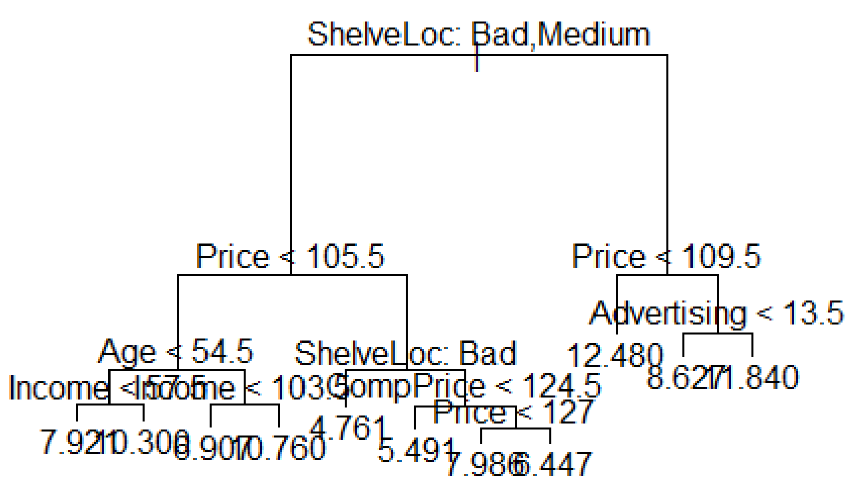
```
## Node number 21: 12 observations
##   mean=9.328333, MSE=2.420914
##
## Node number 22: 12 observations
##   mean=7.920833, MSE=1.545458
##
## Node number 23: 26 observations,    complexity param=0.01425398
##   mean=10.29692, MSE=3.303906
##   left son=46 (8 obs) right son=47 (18 obs)
##   Primary splits:
##       ShelveLoc   splits as  L-R,       improve=0.4379924, (0 missing)
##       Price       < 88    to the right, improve=0.1556226, (0 missing)
##       CompPrice   < 123   to the left,  improve=0.1540437, (0 missing)
##       Advertising < 9.5   to the left,  improve=0.1234244, (0 missing)
##       Age         < 34.5  to the right, improve=0.1061464, (0 missing)
##   Surrogate splits:
##       Education < 10.5  to the left,  agree=0.808, adj=0.375, (0 split)
##
## Node number 24: 10 observations
##   mean=7.107, MSE=3.626381
##
## Node number 25: 31 observations
##   mean=9.117419, MSE=2.755
##
## Node number 36: 7 observations
##   mean=3.4, MSE=2.206314
##
## Node number 37: 34 observations
##   mean=5.921176, MSE=2.377722
##
## Node number 38: 42 observations
##   mean=6.447381, MSE=2.883796
##
## Node number 39: 32 observations
##   mean=7.985625, MSE=2.291312
##
## Node number 40: 28 observations,    complexity param=0.01080417
##   mean=6.075, MSE=3.113168
##   left son=80 (18 obs) right son=81 (10 obs)
##   Primary splits:
##       Population < 271   to the right, improve=0.3271616, (0 missing)
##       ShelveLoc  splits as  L-R,       improve=0.2531328, (0 missing)
##       Education  < 12.5  to the left,  improve=0.2076569, (0 missing)
##       Age        < 68.5  to the right, improve=0.1459641, (0 missing)
##       CompPrice  < 107.5 to the left,  improve=0.1280075, (0 missing)
##   Surrogate splits:
##       Price       < 92    to the right, agree=0.750, adj=0.3, (0 split)
##       Advertising < 14    to the left,  agree=0.679, adj=0.1, (0 split)
##       Age         < 63.5  to the right, agree=0.679, adj=0.1, (0 split)
##       Education   < 15.5  to the left,  agree=0.679, adj=0.1, (0 split)
```

```
##
## Node number 41: 7 observations
##   mean=8.834286, MSE=1.038053
##
## Node number 46: 8 observations
##   mean=8.4925, MSE=2.761894
##
## Node number 47: 18 observations
##   mean=11.09889, MSE=1.454565
##
## Node number 80: 18 observations
##   mean=5.322778, MSE=2.352731
##
## Node number 81: 10 observations
##   mean=7.429, MSE=1.630129

## [1] 2.013859

## [1] 7.593469

## [1] 0.2652094
```

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

```
## Warning: package 'tree' was built under R version 4.0.5
```

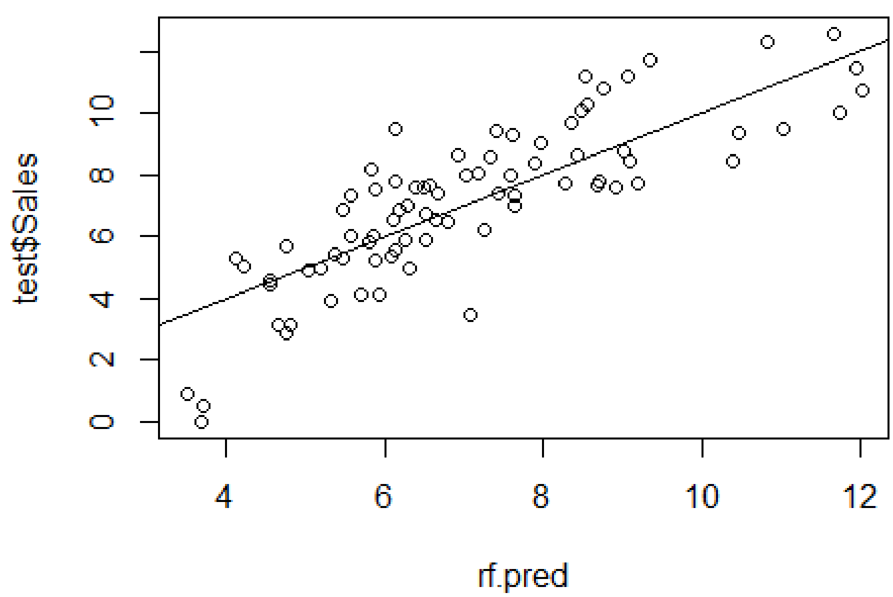

tree.cv$dev
1600
2000
2400
5
10
15
tree.cv$size


ShelveLoc: Bad,Medium
Price < 105.5
Price < 109.5
Age < 54.5
ShelveLoc: Bad
Advertising < 13.5
CompPrice < 124.5
Price < 127
12.480
4.761
5.491

```
## [1] 2.001466
```

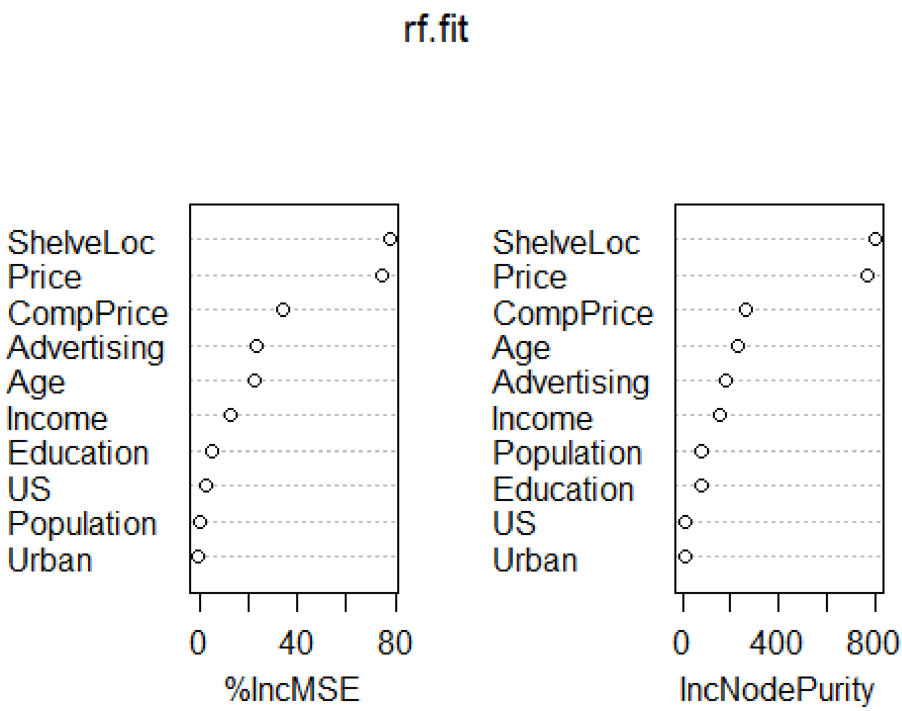
(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = train, mtry = ncol(train) -
1, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.45857
##                     % Var explained: 70.19

## [1] 1.443671
```



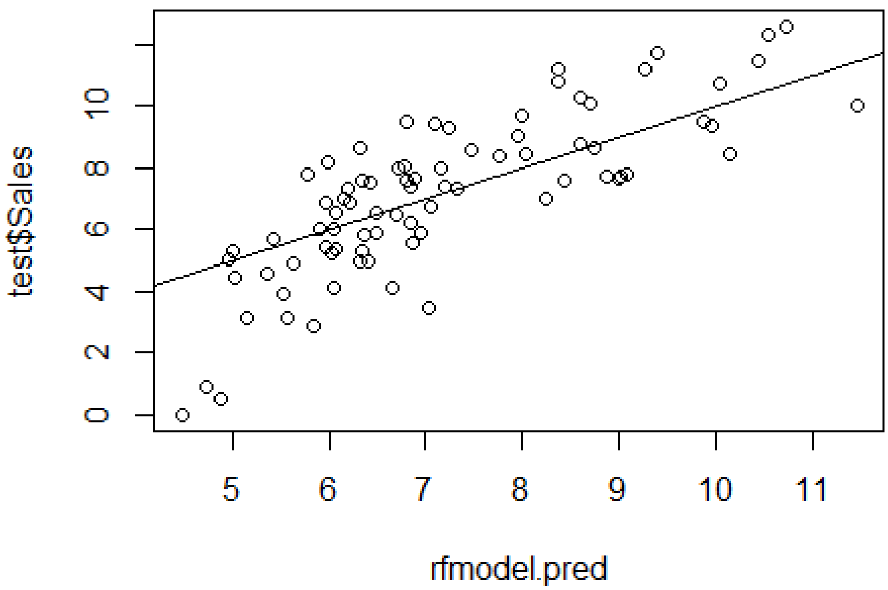

```
##                %IncMSE  IncNodePurity
## CompPrice   34.1596617     267.04557
## Income      12.0724181     153.00015
## Advertising 22.9210685     182.08079
## Population  -0.1591567      79.16700
## Price       74.4337027     766.64681
## ShelveLoc   78.2086163     804.92810
## Age         22.6574307     232.86789
## Education    4.6419444      77.34004
```

```
## Urban       -1.1850577       9.63810
## US           2.1666672      10.15010
```

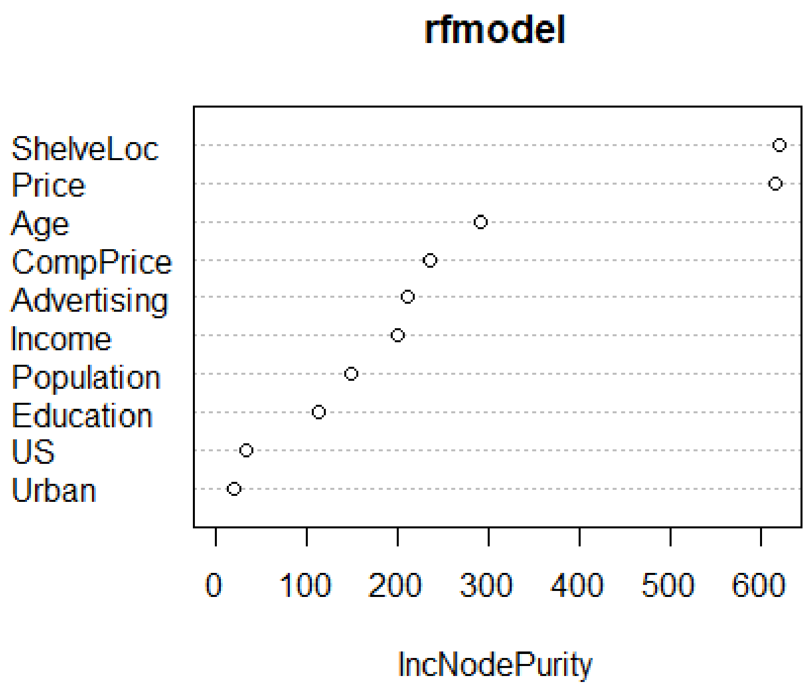


(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables aremost important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = train, control = train_control,
tuneGrid = tunegrid)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 2.813879
##                     % Var explained: 65.89

## [1] 1.616286
```



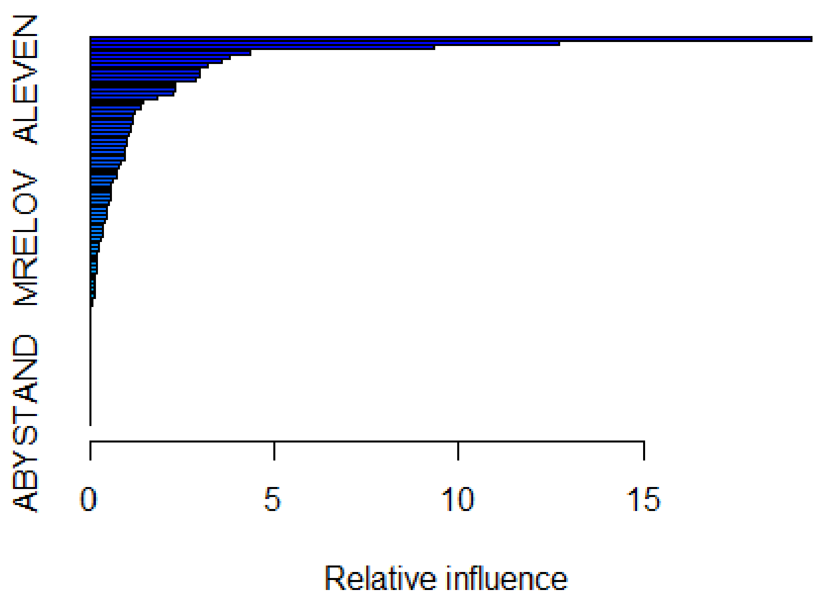

```
##              IncNodePurity
## CompPrice        236.77416
## Income           201.11210
## Advertising      212.40679
## Population       149.75729
## Price            616.57373
## ShelveLoc        620.99515
## Age              291.41276
## Education        114.42523
## Urban             19.83858
## US                32.63020
```


rfmodel
ShelveLoc
Price
Age
CompPrice
Advertising
Income
Population
Education
US
Urban
0
100
200
300
400
500
600
IncNodePurity

7) ISLR Chapter 8 Q11: This question uses the Caravan data set.

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

(b) Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?




```
##                   var     rel.inf
## PPERSAUT PPERSAUT 19.56836803
## PPLEZIER PPLEZIER 12.71687618
## PBRAND     PBRAND  9.33119865
## MOSTYPE   MOSTYPE  4.31396971
## MBERMIDD MBERMIDD  3.77418618
## MOPLLAAG MOPLLAAG  3.57979433
## PWAPART   PWAPART  3.17392245
## MINKGEM   MINKGEM  2.98816114
## MKOOPKLA MKOOPKLA  2.97383667
## ALEVEN     ALEVEN  2.84412447
## MBERARBG MBERARBG  2.32168561
## MBERHOOG MBERHOOG  2.29945785
## MOPLHOOG MOPLHOOG  2.26687545
```

```
## MINKM30   MINKM30  1.83250996
## MOPLMIDD MOPLMIDD  1.45988101
## MINK3045 MINK3045  1.35518473
## PBYSTAND PBYSTAND  1.20286545
## APERSAUT APERSAUT  1.18721095
## MSKA         MSKA  1.16143370
## MSKB1       MSKB1  1.12549065
## MINK7512 MINK7512  1.12085728
## MGODGE     MGODGE  1.05124369
## MGODRK     MGODRK  1.00537920
## MGODPR     MGODPR  0.98736004
## PLEVEN     PLEVEN  0.94827144
## MOSHOOFD MOSHOOFD  0.93590466
## PWAOREG   PWAOREG  0.92135478
## MSKC         MSKC  0.83972599
## MFGEKIND MFGEKIND  0.76077493
## MHHUUR     MHHUUR  0.72849991
## MGEMLEEF MGEMLEEF  0.71637903
## MGODOV     MGODOV  0.61072061
## MAUT1       MAUT1  0.58520614
## MHKOOP     MHKOOP  0.58321736
## PGEZONG   PGEZONG  0.57947265
## MRELGE     MRELGE  0.53173298
## MZPART     MZPART  0.48778859
## MBERZELF MBERZELF  0.47009495
## MFWEKIND MFWEKIND  0.45348781
## PFIETS     PFIETS  0.42456872
## MSKB2       MSKB2  0.37324501
## MINK4575 MINK4575  0.36321987
## MAUT2       MAUT2  0.35992370
## MZFONDS   MZFONDS  0.35099089
## MRELOV     MRELOV  0.26074467
## MBERARBO MBERARBO  0.25657812
## MINK123M MINK123M  0.21541187
## MBERBOER MBERBOER  0.18793450
## PMOTSCO   PMOTSCO  0.18674263
## PTRACTOR PTRACTOR  0.17510146
## AFIETS     AFIETS  0.15692374
## MRELSA     MRELSA  0.15008875
## MAUT0       MAUT0  0.13561988
## ABRAND     ABRAND  0.12219843
## PAANHANG PAANHANG  0.11430633
## MSKD         MSKD  0.11056176
## MFALLEEN MFALLEEN  0.09941437
## MGEMOMV   MGEMOMV  0.08630281
## PWALAND   PWALAND  0.05555946
## MAANTHUI MAANTHUI  0.02005783
## PWABEDR   PWABEDR  0.00000000
## PBESAUT   PBESAUT  0.00000000
## PVRAAUT   PVRAAUT  0.00000000
```

```
## PWERKT     PWERKT  0.00000000
## PBROM       PBROM  0.00000000
## PPERSONG PPERSONG  0.00000000
## PZEILPL   PZEILPL  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
## APERSONG APERSONG  0.00000000
## AGEZONG   AGEZONG  0.00000000
## AWAOREG   AWAOREG  0.00000000
## AZEILPL   AZEILPL  0.00000000
## APLEZIER APLEZIER  0.00000000
## AINBOED   AINBOED  0.00000000
## ABYSTAND ABYSTAND  0.00000000
```

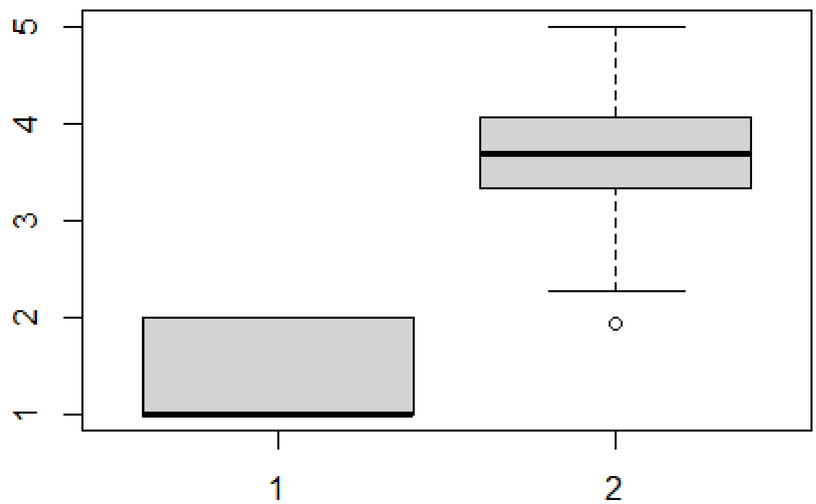
PPERSAUT, PPLEZEIER seem to be some of the most important variables.

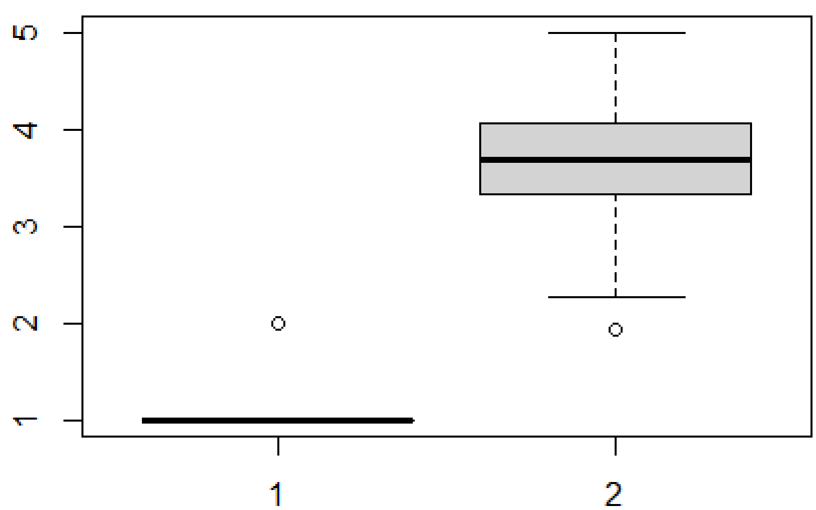
(c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compare with the results obtained from applying KNN or logistic regression to this data set?

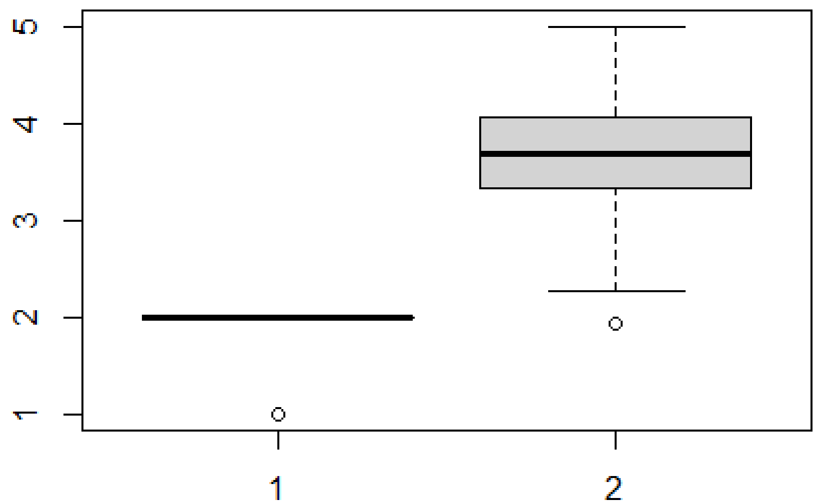
8) Exam Questions - Problem 1: Beauty pays!

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

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


5
4
3
2
1
1
2


5
4
3
2
1
1
2




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

Looking at the results, beauty score has a very high impact on course evaluation with positive correlation.

```
##
## 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 ***
## female1      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower1       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish1  -0.25808    0.08478  -3.044  0.00247 ** 
## tenuretrack1 -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
```

Other factors too seem to have a high impact on course evaluation: butr in a negative way.

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

These results do not confirm that course evaluation solely depends on beauty and factors as gender, english, position or tenure. Other factors like productivity might also be coming in play, thus without analyzing more data around the subset being considered it is highly impossible to make a definite statement on causation of course evaluation.

9) Exam Questions - Problem 2: Housing Price Structure

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

```
## 
## Call:
## lm(formula = Price ~ ., data = mid1[, -c(1:2, 7)])
## 
## 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    
## Nbhd.2      -1560.579   2396.765  -0.651  0.51621    
```

```
## Nbhd.3      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 ***
## BrickYes    17297.350   1981.616   8.729 1.78e-14 ***
## 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
```

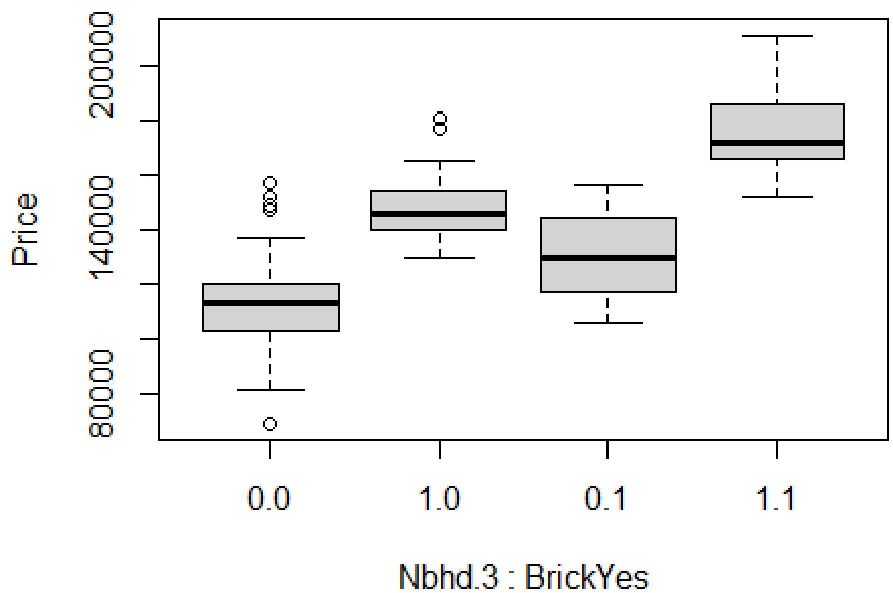
```
##                    2.5 %      97.5 %
## (Intercept) -15417.94711 19736.94349
## Nbhd.2       -6306.00785  3184.84961
## Nbhd.3       14446.32799 26915.74671
## Offers      -10415.27089 -6119.70575
## SqFt            41.64034    64.34714
## BrickYes     13373.88702 21220.81203
## Bedrooms      1083.04162  7410.54616
## Bathrooms     3691.69572 12074.86126
```

Brick is a significant variable with non zero confidence interval. Thus, people might be paying a premium for brick houses

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

Nbhd3 is a significant variable with non zero confidence interval. Thus, people might be paying a premium for living in a better neighborhood as nbhd 3

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



From the plot, it seems price for brick and nbhd 3 houses is high and thus it might be the case that such houses have high premium.

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

As seen previously, Nbhd2 is not such a significant variable, and thus can be clubbed with Nbhd1.

10) Exam Questions - Problem 3: What causes what??

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)

It is not justified to consider no. of cops or no. of crimes to be directly affecting each other. Lesser police can lead to more crimes or more police could have been deployed for higher crimes. Also, the crimes vary in degree of severity, so it might be possible that the task force required for crimes pertaining to certain geography might be different. Also, data from a few different cities itself is a very small data to comment on or base our judgement on.

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

They deployed police for reasons a=other than crime, especifically street crime. They used the terrorist alert system which indicates how much a city is vulnerable to terrorist activity on a day. So on high risk days, there was more police on the roads, the researchers observed this reduced street crime.

From the table, we see that there’s a negative relationship between high alert and number of crimes, the same is also true when controlling for metro ridership as both the models have a negative coefficient on high alert variable.Also the value of Rsquare is very small.

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

They thought it’s possible that on high terrorism risk days, there will be less people traveling in the city and as a result there would be lesser number of people who will be victimized by street crime. So they analyzed the metro ridership data to make sure that the reduced crime was not a result of reduced number of possible victims and not the police.

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

The model being estimated is trying to understand the effect of high alert on diff districts of DC, on total crimes. They introduced 2 variables depicting interaction of high alert with district 1 and one with other districts. We see that effect of high alert on crimes of district 1 is higher than that of other districts, due to the larger magnitude of the negative coefficient on the 1st variable.

11) Exam Questions - Problem 4: Neural Nets

```
## # weights:  76
## initial  value 454.688587
## iter  10 value 2.838153
## iter  20 value 2.303616
## iter  30 value 2.167715
## iter  40 value 2.130205
## iter  50 value 2.094424
## iter  60 value 2.089925
## iter  70 value 2.088112
## iter  80 value 2.077684
```

```
## iter  90 value 2.067918
## iter 100 value 2.066612
## final  value 2.066612
## stopped after 100 iterations

## [1] 0.03810992
```

The RMSE from

12) Exam Questions - Problem 5: Contribution to final group project

I was a part of Group 1- Morning 10-12 batch and we worked on a dataset based on employee attrition (sourced from Kaggle). The data consisted of ~1500 employees and 35 variables related to their job - salary, satisfaction, experience (both past and present), work life balance and personal details like age, distance from home etc. Our methodology was to first indulge in preliminary data preparation and exploration, then modelling, and finally figuring the factors that affected the most to understand business implications of the problems.

We worked on a wide variety of models, knn, trees, regressions; out of which I worked on understanding the data and forecasting via decidion trees and other ensemble methods. After some preliminary analysis and data preparataion, I removed a few and created some variables that intuitively could have impacted attrition like avg tenure of an individual employee based on past experience and working hours per week which proved to highly correlated with attrition.

The approach next was twofold: 1. Model prediction on all the variables 2. Using a subset of variables got from fitting a basic random forest model via feature importance

Also the data was imbalanced so oversampling proved to be helpful.

First I modelled decision trees, which gave a lower bound of accuracy as 75% with both the models. Second, I tried random forest modelling with some tuning and updating class weights to 5:1 to improve recall. This significantly improved the accuracy to 87.7% with the model using all variables, but the model using subset of variables didn't perform so well and gave ~86% accuracy which might be due to loss of information. Last I went for XGBoosting, which gave similar accuracy (~87.5%) post cross validation, but much better recall and thus better prediction.

Finally, I analysed the importance of different factors around attrition. Salary components like monthly income, hourly daily rates etc. , hikes, promotions and stock options , age, avg tenure/ working hours & experience and satisfaction levels for environment/ job/ relationship affected attrition inversely. Factors like work hours and distance from home were some parameters that affected attrition too.