Intro to ML Exam

Sahil Natu

8/1/2021

Book Problems

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

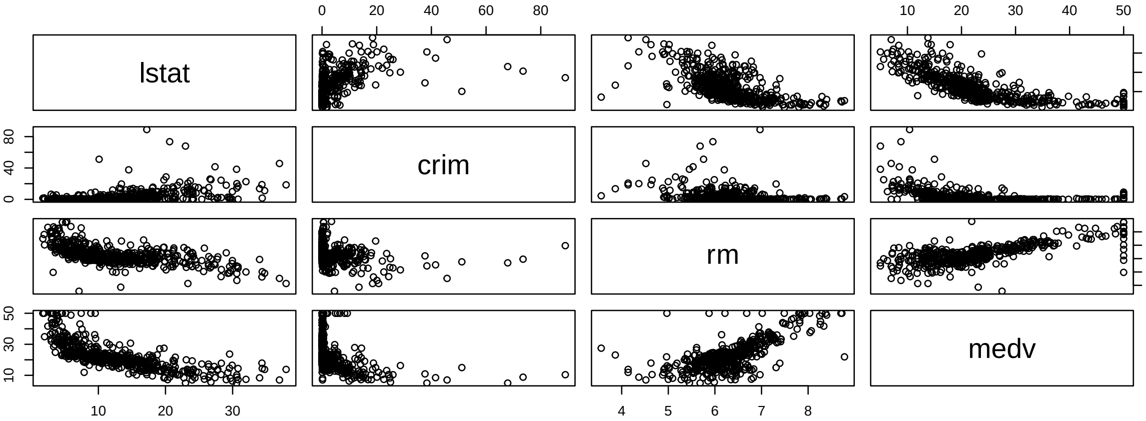
```
## [1] 506
```

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

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

The data set has **506** rows and **14** columns. The rows represent the towns/suburbs in Boston and the columns represent the various attributes measured for each town/suburb.

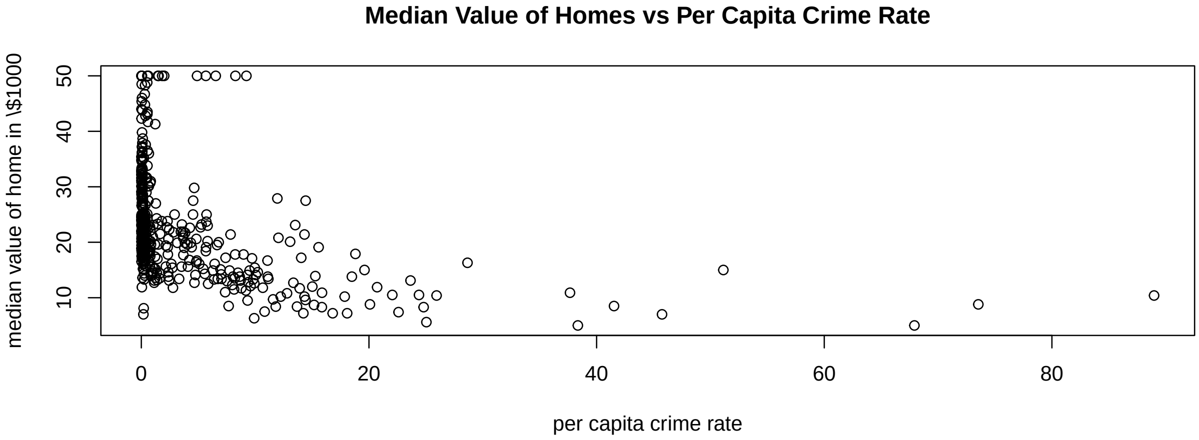
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



Negative correlation can be seen between the following-
- Lower Status of Population & Median House Value
- Lower Status of Population & Rooms per House
- Per Capita Crime Rate & Median House Value

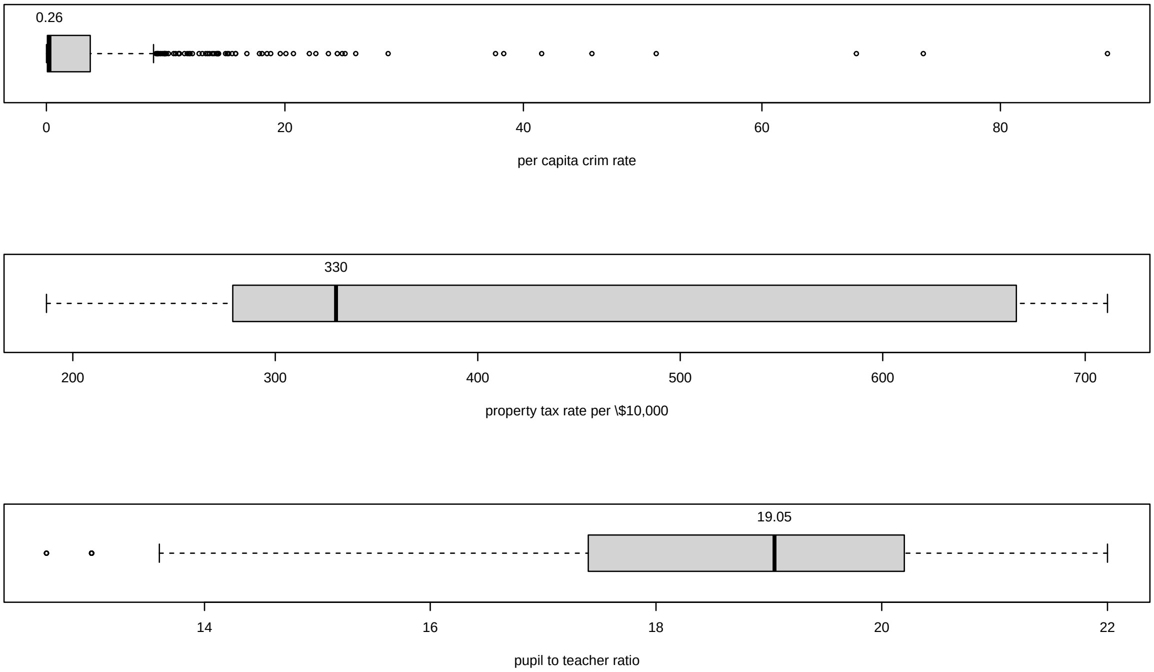
Positive correlation can be seen between the following-
- Rooms per House & Median House Value
- Lower Status of Population & Per Capita Crime Rate

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



The predictor *medv* (Median House Value) is associated with per capita crim rate. It has a negative correlation with per capita crime rate. As the value of houses fall, there is higher per capita crime.

(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.



Some suburbs have particularly high per capita crime rate as evidenced by the boxplot. While the median is 0.26 crimes per capita, some suburbs have crime rate as high as 88.97. There aren't such cases of particularly high tax rates or pupil-teacher ratios as the distribution is more even.

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

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

A total of **35** suburbs in the data set bound the Charles river.

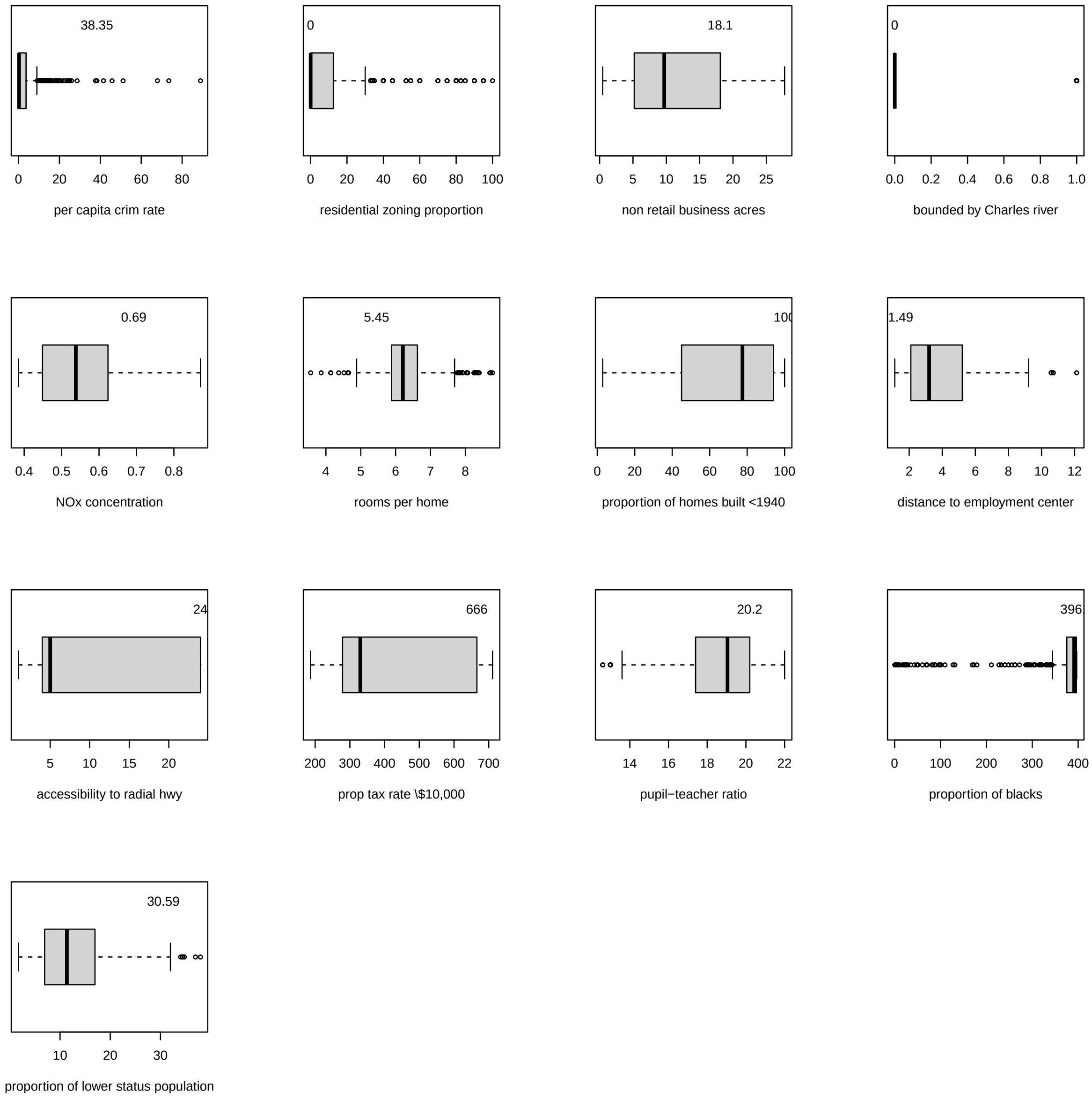
(f) What is the median pupil-teacher ratio among the towns in this data set?

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

The median pupil-teacher ration among the towns in this data set is **19.05**.

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

```
## [1] 399
```



The suburb at index location **399** in the data set has the lowest median house value. Looking at the boxplots, it is apparent that this suburb has higher than average
- Per Capita Crime Rate
- NOx Concentration
- Proportion of Houses built before 1940

And, lower than average
- Rooms per House

Intuitively, the observations suggest that the houses in such a suburb are less desirable and thus are valued lower.

(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
## [1] 13
```

```
##         all_suburbs subset_suburbs
## crim        0.25651        0.52014
## zn          0.00000        0.00000
## indus       9.69000        6.20000
## chas        0.00000        0.00000
## nox         0.53800        0.50700
## rm          6.20850        8.29700
## age        77.50000       78.30000
## dis         3.20745        2.89440
## rad         5.00000        7.00000
## tax       330.00000      307.00000
## ptratio    19.05000       17.40000
## black     391.44000      386.86000
## lstat      11.36000        4.14000
## medv       21.20000       48.30000
```

In this data set, **64** suburbs average more than 7 rooms per dwelling and **13** suburbs average more than 8 rooms per dwelling.
Looking at the table comparing medians for each of the attributes among the full data and data for suburbs averaging over 8 rooms per dwelling, we can conclude the following -
- Suburbs averaging over 8 rooms per dwelling have higher than average per capita crim rate and median house value
- Suburbs averaging over 8 rooms per dwelling have lower than average proportion of non-retail business acres, property tax rate, lower status of population

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 captia 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.

1. **crim~zn**

```
##
## Call:
## lm(formula = crim ~ zn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

```
##        fit       lwr      upr
## 1 4.084019 -12.50623 20.67427
## 2 3.714344 -12.87474 20.30343
## 3 3.344669 -13.24476 19.93410
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between proportion of residential land zoned and per capita crime rate. Rˆ2 value is extremely small, i.e. 0.04, hence *zn* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on zn are very large and thus *zn* isn't a good predictor for *crim*.

2. **crim~indus**

```
##
## Call:
## lm(formula = crim ~ indus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##        fit       lwr      upr
## 1 0.485139 -14.99709 15.96737
## 2 3.034021 -12.43640 18.50445
## 3 5.582902  -9.89195 21.05775
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between proportion of non-retail business acres per town and per capita crime rate. Rˆ2 value is small, i.e. 0.16, hence *indus* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *indus* are very large and thus *indus* isn't a good predictor for *crim*.

3. **crim~chas**

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

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453   <2e-16 ***
## chas         -1.8928     1.5061  -1.257    0.209
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

Since p-value is not sufficiently small, we accept the Null Hypothesis, i.e. there is no relationship between proximity to Charles river and per capita crime rate.

4. **crim~nox**

```
##
## Call:
## lm(formula = crim ~ nox)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## nox           31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##        fit      lwr      upr
## 1 142.5228 112.1580 172.8876
## 2 298.7654 241.0289 356.5019
## 3 455.0081 368.5151 541.5011
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between nitrous oxide concentration and per capita crime rate. R^2 value is small, i.e. 0.17, hence *nox* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *nox* are fairly large and thus *nox* isn’t a good predictor for *crim*.

5. **crim~rm**

```
##
## Call:
```

```
## lm(formula = crim ~ rm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## rm            -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```
##          fit        lwr        upr
## 1   7.061548  -9.513716 23.6368122
## 2  -6.358708 -23.329827 10.6124111
## 3 -19.778964 -38.645074 -0.9128542
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between average number of rooms per house and per capita crime rate. R^2 value is extremely small, i.e. 0.04, hence *rm* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *rm* are very large and thus *rm* isn't a good predictor for *crim*.

6. **crim~age**

```
##
## Call:
## lm(formula = crim ~ age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
##         fit       lwr      upr
## 1 -3.238975 -19.16302 12.68507
## 2 -2.700044 -18.61208 13.21199
## 3 -2.161113 -18.06211 13.73989
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between proportion of owner occupied units built before 1940 and per capita crime rate. Rˆ2 value is fairly small, i.e. 0.12, hence *age* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *age* are very large and thus *age* isn't a good predictor for *crim.*

7. **crim~dis**

```
##
## Call:
## lm(formula = crim ~ dis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.708 -4.134 -1.527  1.516 81.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4993     0.7304  13.006   <2e-16 ***
## dis          -1.5509     0.1683  -9.213   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##          fit       lwr       upr
## 1   1.744753 -13.92519 17.414694
## 2  -6.009755 -21.80847  9.788955
## 3 -13.764264 -29.86147  2.332943
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between weighted mean of distances to 5 Boston employment centers and per capita crime rate. Rˆ2 value is fairly small, i.e. 0.14, hence *dis* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *dis* are very large and thus *dis* isn't a good predictor for *crim.*

8. **crim~rad**

```
##
## Call:
## lm(formula = crim ~ rad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.164  -1.381  -0.141   0.660  76.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## rad          0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:   0.39
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##         fit        lwr      upr
## 1 0.8023952 -12.412451 14.01724
## 2 3.8919498  -9.319368 17.10327
## 3 6.9815045  -6.234893 20.19790
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between accessibility to radial highway and per capita crime rate. R^2 value is about 0.39, hence *rad* does explain to a certain extent the variability in *crim*. However, the prediction intervals with the fit on *rad* are very large.

9. **crim~tax**

```
##
## Call:
## lm(formula = crim ~ tax)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## tax          0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##         fit       lwr      upr
## 1 -8.379658 -22.21754 5.458222
## 2 -8.230947 -22.06692 5.605025
## 3 -8.082235 -21.91632 5.751852
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between full value property tax rates and per capita crime rate. R^2 value is about 0.34, hence *tax* does explain to a certain extent the variability in *crim*. However, the prediction intervals with the fit on *tax* are very large.

10. **crim~ptratio**

```
##
## Call:
## lm(formula = crim ~ ptratio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio       1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```
##           fit       lwr       upr
## 1 -11.8870195 -28.69960  4.925561
## 2  -6.1271056 -22.57495 10.320736
## 3  -0.3671917 -16.61332 15.878939
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between pupil to teacher ratio and per capita crime rate. Rˆ2 value is fairly small, i.e. 0.08, hence *ptratio* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *ptratio* are very large and thus *ptratio* isn’t a good predictor for crim.

11. **crim~black**

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

```
##        fit       lwr      upr
## 1 16.37213 0.5175764 32.22669
## 2 16.19073 0.3425558 32.03891
## 3 16.00933 0.1674464 31.85122
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between proportion of blacks in town and per capita crime rate. Rˆ2 value is fairly small, i.e. 0.14, hence *black* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *black* are very large and thus *black* isn’t a good predictor for *crim*.

12. **crim~lstat**

```
##
## Call:
## lm(formula = crim ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##          fit       lwr      upr
## 1 -0.5865141 -15.67671 14.50368
## 2  2.1575098 -12.91765 17.23267
## 3  4.9015337 -10.17318 19.97624
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between lower status of population and per capita crime rate. R^2 value is fairly small, i.e. 0.2, hence *lstat* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *lstat* are very large and thus *lstat* isn’t a good predictor for *crim*.

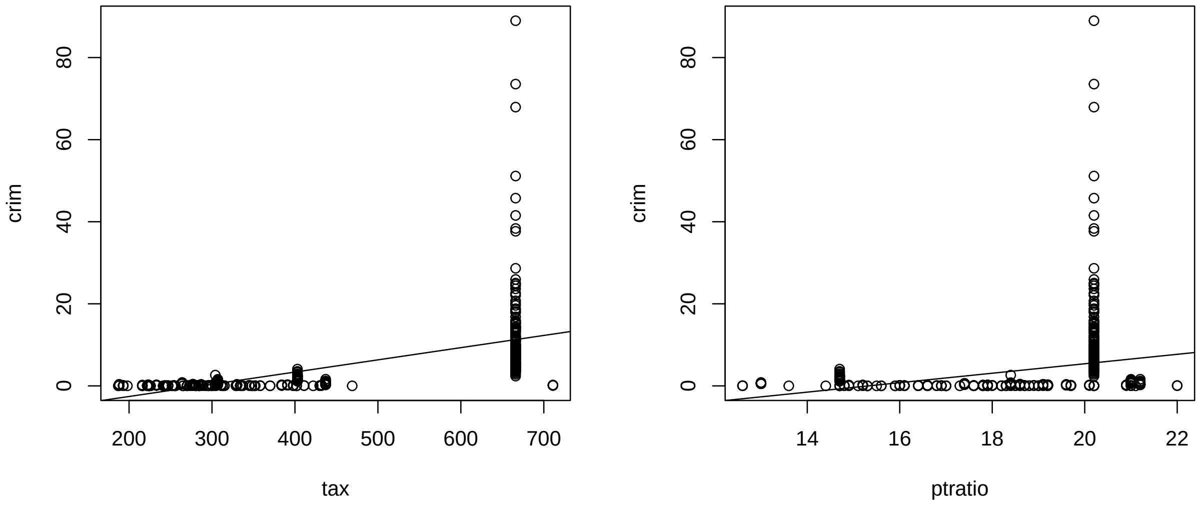
13. **crim~medv**

```
##
## Call:
## lm(formula = crim ~ medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
##        fit       lwr      upr
## 1 9.980736 -5.679268 25.64074
## 2 8.164937 -7.467739 23.79761
## 3 6.349137 -9.265272 21.96355
```

Since p-value is sufficiently small, we reject the Null Hypothesis, i.e. there is relationship between median value of owner occupied homes and per capita crim rate. Rˆ2 value is fairly small, i.e. 0.15, hence *medv* does not explain sufficiently the variability in *crim*, and hence is not a good predictor. The prediction intervals with the fit on *medv* are very large and thus *medv* isn't a good predictor for *crim*.

The predictors 'accessibility to radial highway' and 'full value property tax rates' have a relatively better statistical association with crime rate per capita. Given below are the plots for the two predictors against per capita crime rate.



However, these predictors still fail to sufficiently explain *crim* as evidenced by their prediction intervals and visually by the plots.

(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 following null hypothesis

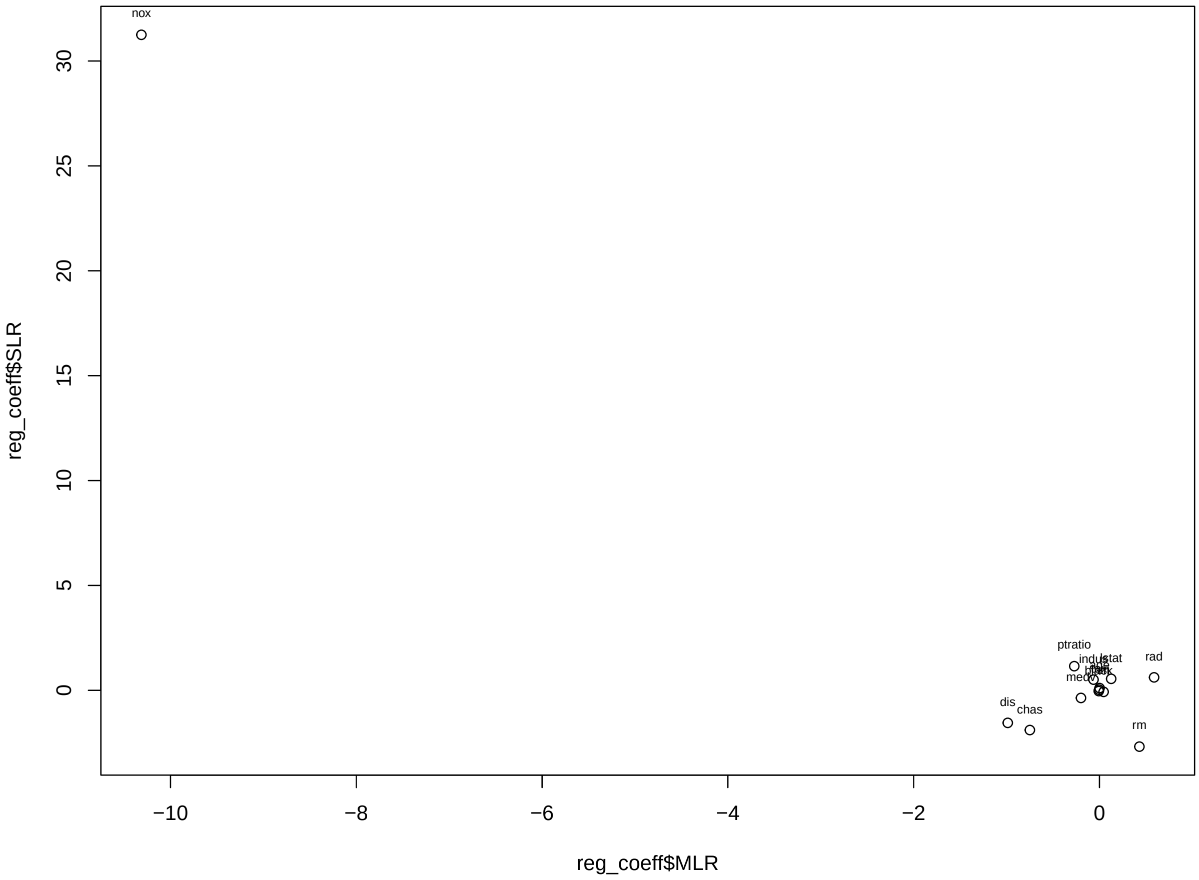
$$H_0 : \beta_j = 0$$

```
##
## 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
```

The model is not particularly good at predicting the variability of crim since the Rˆ2 value is only about 0.45. The F-statistic is sufficiently greater than 1, thus we reject the Null Hypothesis that none of the predictors have a relationship with crim. However, when we look at p-value of individual predictors in this model, it is evident that we can reject the Null Hypothesis for only 2 predictors -
- Distance to 5 Boston Employment Centers
- Accessibility to Radial Hwy

(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.

Multiple Linear Regression model provides a much better fit than any of the individual Simple Linear Regression models as evidenced by the respective Rˆ2 values. While ‘full value property tax rate’ was a statistically important predictor for *crim* in the Simple Linear Regression model, this was not the case in the Multiple Linear Regression Model. Also, ‘distance to 5 Boston employment centers’ has come up as a staistically important predictor in the Multiple Linear Regression model, which was not the case in the Simple Linear Regression model.



(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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

1. **crim~zn**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between proportion of residential land zoned and per capita crime rate.

2. **crim~indus**

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

Since p-value for degree 3 is sufficiently small, there is a non-linear relationship between proportion of non-retail business acres and per capita crime rate.

3. **crim~chas**

```
##
## Call:
## lm(formula = crim ~ chas + I(chas^2) + I(chas^3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients: (2 not defined because of singularities)
##             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
## I(chas^2)         NA         NA      NA       NA
## I(chas^3)         NA         NA      NA       NA
## ---
## 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
```

Since predictor *chas* has only 2 unique values, we cannot fit a model with a 2nd and 3rd degree polynomial of *chas*.

4. **crim~nox**

```
##
## Call:
## lm(formula = crim ~ poly(nox, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.110 -2.068 -0.255  0.739 78.302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3216  11.237  < 2e-16 ***
## poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
## poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
## poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.234 on 502 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
## F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since p-value for degrees 2 and 3 are sufficiently small, there is a non-linear relationship between nitrogen oxide concentration and per capita crime rate.

5. **crim~rm**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between average number of rooms per house and per capita crime rate.

6. **crim~age**

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

Since p-value for degree 2 is sufficiently small, there is a non-linear relationship between proportion of owner occupied units built before 1940 and per capita crime rate.

7. **crim~dis**

```
##
## Call:
## lm(formula = crim ~ poly(dis, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.757  -2.588   0.031   1.267  76.378
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3259  11.087  < 2e-16 ***
## poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
## poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
## poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7.331 on 502 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
## F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since p-value for degrees 2 and 3 are sufficiently small, there is a non-linear relationship between distance to 5 Boston employment centers and per capita crime rate.

8. **crim~rad**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between accessibility to radial highway and per capita crime rate.

9. **crim~tax**

```
##
## Call:
## lm(formula = crim ~ poly(tax, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.273  -1.389   0.046   0.536  76.950
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3047  11.860  < 2e-16 ***
## poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
## poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
## poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.854 on 502 degrees of freedom
## Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
## F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since p-value for degree 2 is sufficiently small, there is a non-linear relationship between full value property tax rate and per capita crime rate.

10. **crim~ptratio**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between pupil to teacher ratio and per capita crime rate.

11. **crim~black**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between proportion of blacks in town and per capita crime rate.

12. **crim~lstat**

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

Since p-value for degrees 2 and 3 are not sufficiently small, there is no non-linear relationship between lower status of population and per capita crime rate.

13. **crim~medv**

```
##
## Call:
## lm(formula = crim ~ poly(medv, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.427  -1.976  -0.437   0.439  73.655
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.614      0.292  12.374  < 2e-16 ***
## poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
## poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
## poly(medv, 3)3  -48.033      6.569  -7.312 1.05e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.569 on 502 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4167
## F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since p-value for degrees 2 and 3 are sufficiently small, there is a non-linear relationship between median value of owner occupied homes and per capita crime rate.

Thus, the predictors *indus*, *nox*, *age*, *tax*, *medv* have a non-linear relationship with *crim*

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] 583
```

```
## [1] 194
```

We have split the data into train and test sets with the train set having 3/4th of all the rows.

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

```
## [1] 941.8909
```

The test error obtained is 941.89

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

```
## [1] 941.8831
```

The test error obtained is 941.88

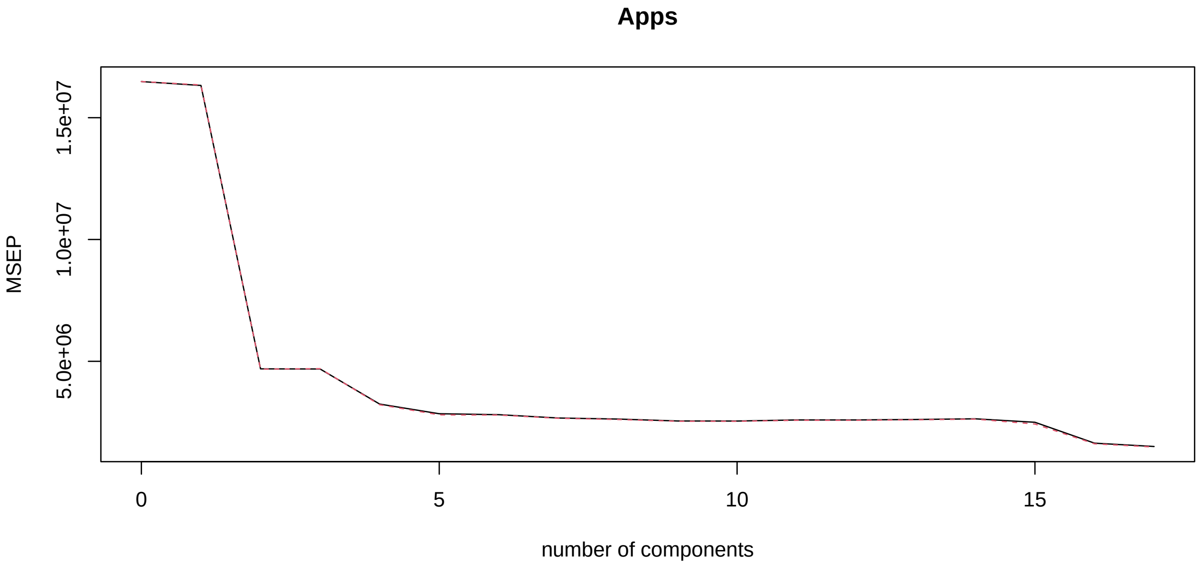
(d) Fit a lasso model on the training set, with lambda chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
## [1] 941.7444
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -471.39372069
## (Intercept)    .
## PrivateYes  -491.04485135
## Accept         1.57033288
## Enroll        -0.75961467
## Top10perc     48.14698891
## Top25perc    -12.84690694
## F.Undergrad    0.04149116
## P.Undergrad    0.04438973
## Outstate      -0.08328388
## Room.Board     0.14943472
## Books          0.01532293
## Personal       0.02909954
## PhD           -8.39597537
## Terminal      -3.26800340
## S.F.Ratio     14.59298267
## perc.alumni   -0.04404771
## Expend         0.07712632
## Grad.Rate      8.28950241
```

The test error obtained is 941.74. No variables have zero coefficient estimates.

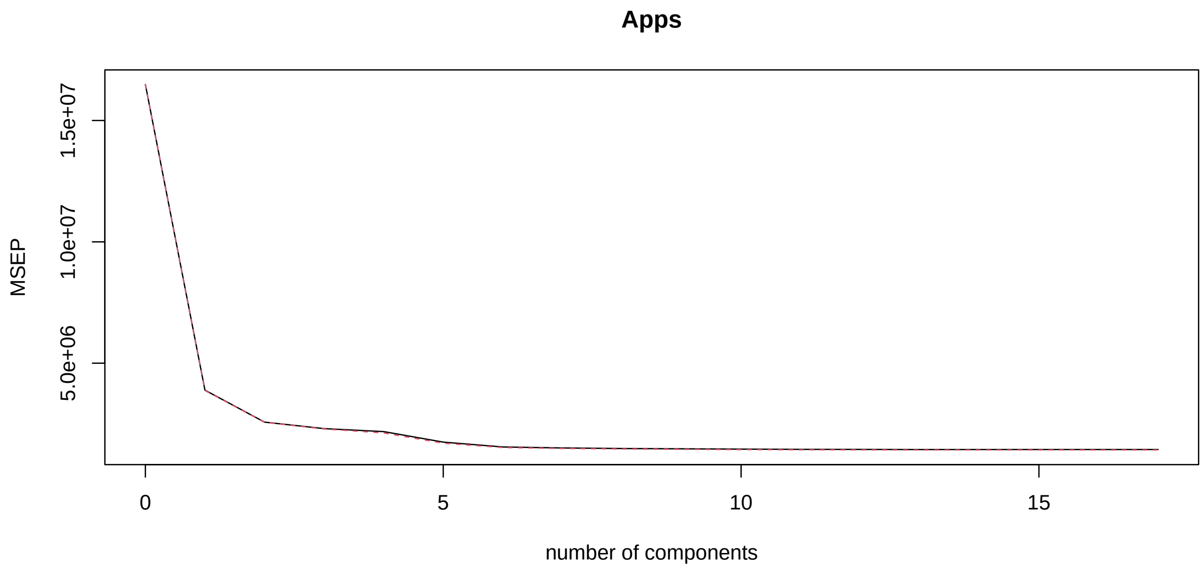
(e) Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.



```
## [1] 1274.351
```

RMSE decreases up to M=5 and is more or less constant thereafter. Hence, we select M=5.
The test error obtained is 1274.35.

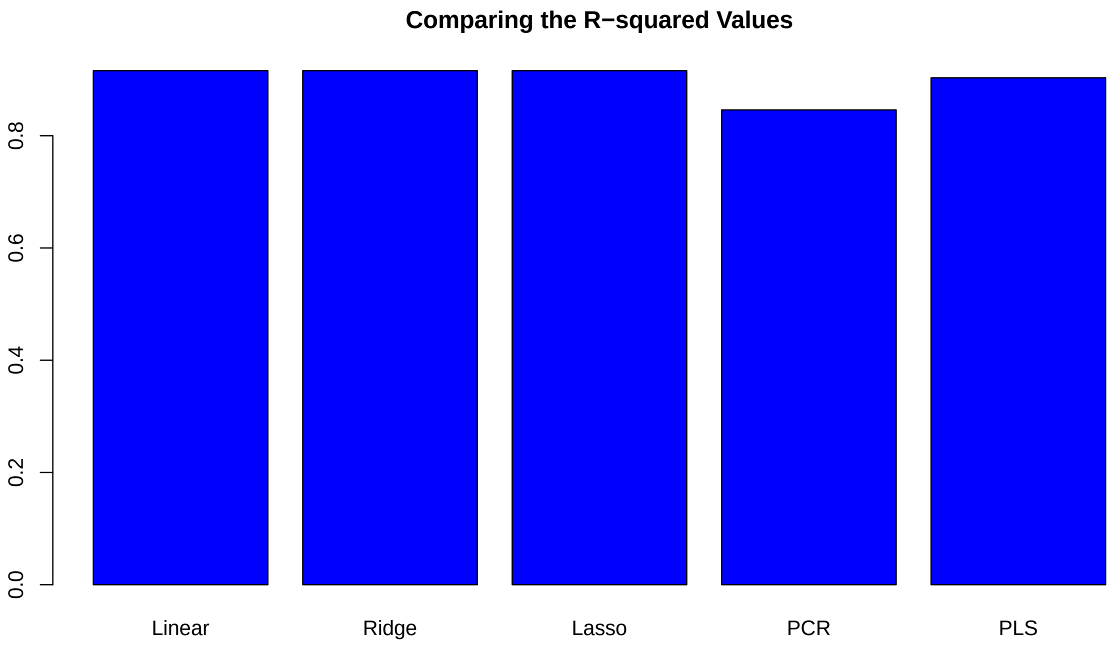
(f) Fit a PLS model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.



```
## [1] 1010.325
```

RMSE decreases up to M=5 and is constant thereafter. Hence, we select M=5.
The test error obtained is 1010.325.

(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?



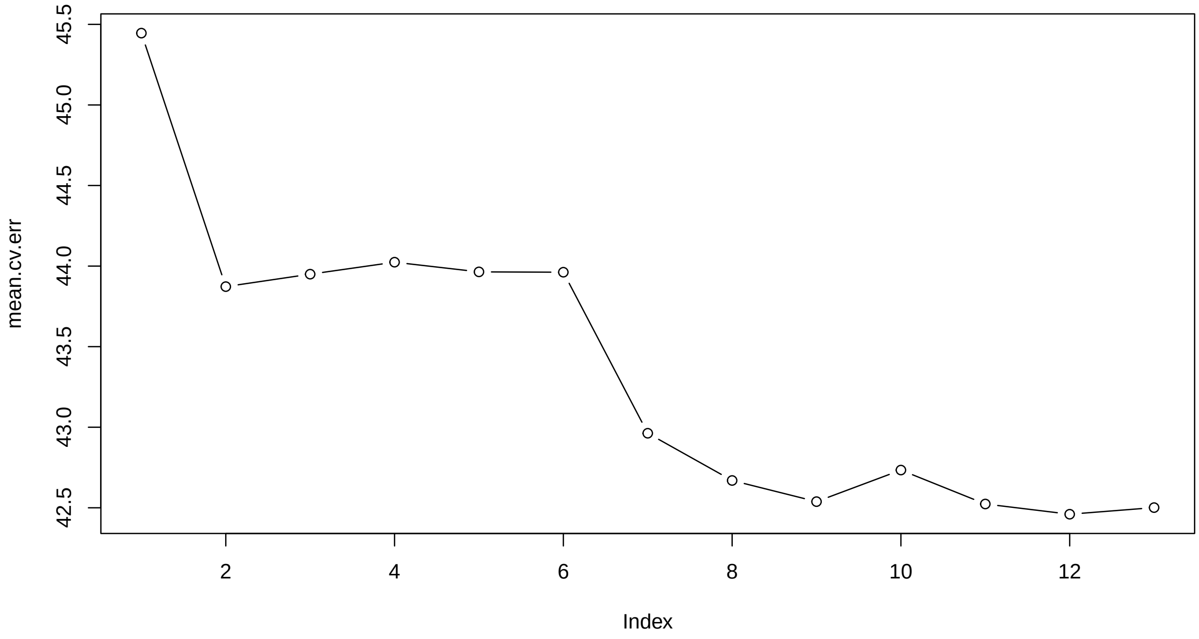
The plot shows that Test R^2 for all models except PCR are around 0.9, while PCR has a smaller Test R^2 of about 0.8. All models except PCR predict college applications with high accuracy.

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.

1. **Subset Selection**



```
## [1] 4.013554
```

Cross Validation has selected model with 12 variables as it has the lowest MSE.
Test RMSE achieved is 4.01.

2. **Lasso Regression**

```
## [1] 3.946214
```

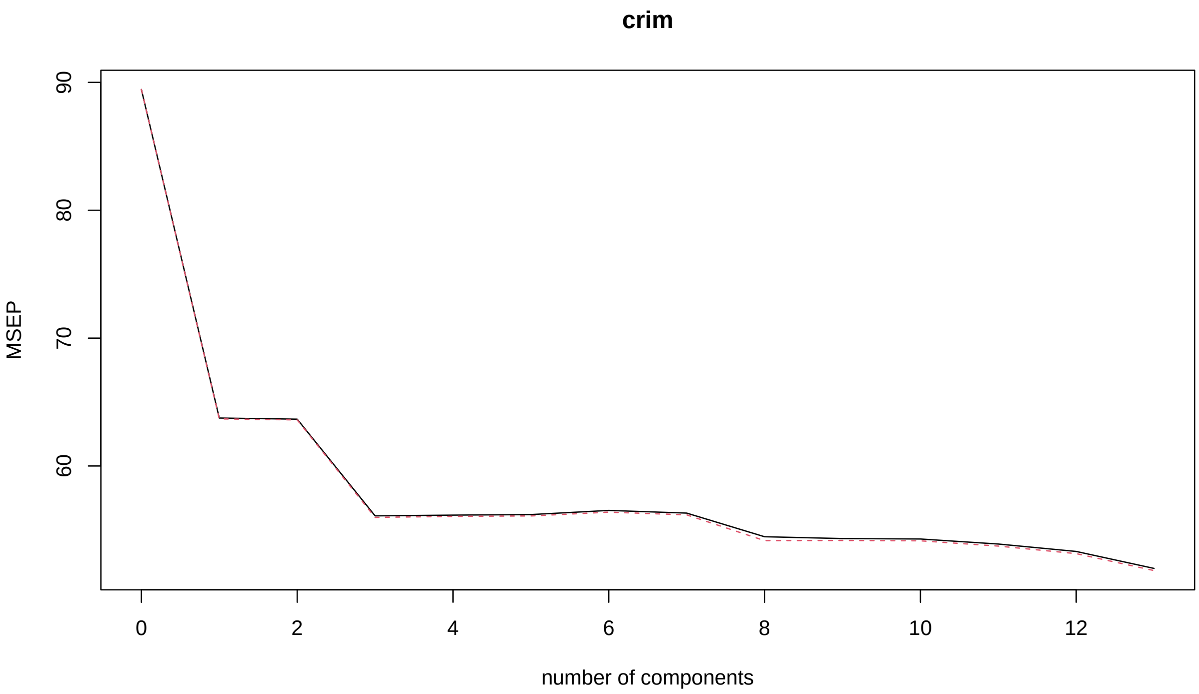
Test error obtained is 8.94.

3. **Ridge Regression**

```
## [1] 3.881068
```

Test error obtained is 3.88.

4. **PCR**



```
## [1] 4.131998
```

RMSE decreases upto M=3 and is more or less constant thereafter. Hence, we select M=3.
Test error obtained is 4.13.

The model that performs the best on the given data set is the one built using Ridge Regression. It has the lowest Test RMSE of all the models at 3.88. This model involves all the features (variables) in the data and it is by design. Ridge Regression shrinks the coefficients of variables to reduce variance but never makes them 0.

Chapter 4 Q10

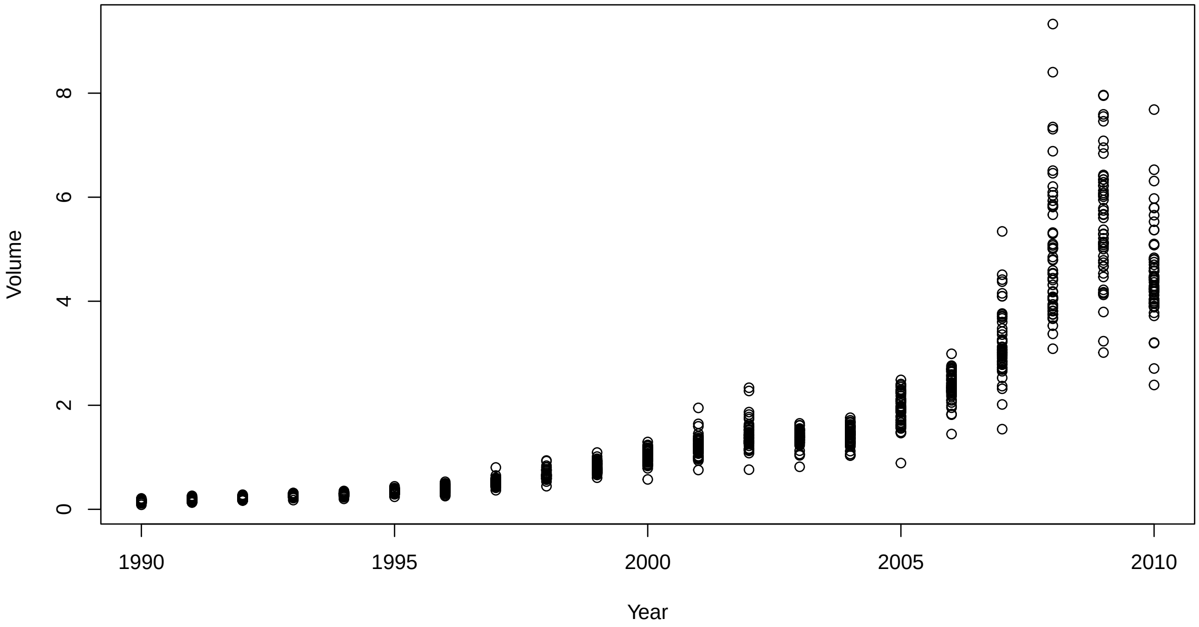
This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

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

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

```
## [1] 1089    9
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume            Today
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```

```
##               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
```



Weekly datasource has 9 columns and 1089 rows. Of the 9 columns, 1 is categorical. From the pairwise correlation matrix, there is no correlation between current week's return and that of any previous weeks. There seems to be a strong correlation between Year and Volume. Upon plotting Year vs Volume, it is evident that Volume of trade has been increasing every year.

(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)
##
## 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
```

Lag2 has the lowest p-value, and has a positive coefficient. Thus direction of market in current week is the same as 2 weeks preceding that week. Other variables have a large p-value and thus are not statistically significant.

(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.

```
##      Up
## Down  0
## Up    1
```

```
##                Direction
## Weekly.glm.pred Down  Up
##            Down   54  48
##            Up    430 557
```

```
## [1] 0.5610652
```

```
## [1] 0.4389348
```

The model correctly predicts UP movement in 557 weeks and DOWN movement in 54 weeks out of a total of 1089 weeks. Fraction of correct predictions is **0.561**. Training error rate is **0.439**, which is too high since training error rate is lower compared to test error rate. The model gives more 'False UPs' than 'True DOWNs' thus being poor at predicting if the market would move down. The model does a decent job of predicting if the market would move up, as it has more than 10x 'True UPs' than 'False DOWNs'.

(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).

```
##
## Weekly.glm.pred.2 Down Up
##              Down    9  5
##              Up     34 56
```

```
## [1] 0.625
```

```
## [1] 0.375
```

The model correctly predicts UP movement in 56 weeks and DOWN movement in 9 weeks out of a total of 104 weeks. Fraction of correct predictions is **0.625**. Training error rate is **0.375**, lower than the previous model. The model gives more 'False UPs' than 'True DOWNs' thus being poor at predicting if the market would move down. The model does a decent job of predicting if the market would move up, as it has more than 10x 'True UPs' than 'False DOWNs'.

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

```
##
## Weekly.knn.pred.3 Down Up
##              Down   22 31
##              Up     21 30
```

```
## [1] 0.5
```

```
## [1] 0.5
```

The KNN model correctly predicts UP movement in 30 weeks and DOWN movement in 22 weeks out of a total of 104 weeks. Fraction of correct predictions is **0.5**. Training error rate is **0.5**, higher than linear regression model. The model gives equal 'False UPs' and 'True DOWNs' as well as equal 'True UPs' and 'False DOWNs', hence it is poor at predicting the response.

(h) Which of these methods appears to provide the best results on this data?
The Logistic Regression Model provides a better result on this data with an error rate of 37.5% as against KNN's error rate of 50%.

(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.
1. **Adding Lag1 as predictor along with Lag2**

```
##
## Weekly.glm.pred.4 Down Up
##              Down    7  8
##              Up     36 53
```

```
## [1] 0.5769231
```

```
## [1] 0.4230769
```

The model correctly predicts UP movement in 53 weeks and DOWN movement in 7 weeks out of a total of 104 weeks. Fraction of correct predictions is **0.577**. Training error rate is **0.423**, higher than the logistic regression model with only Lag2 as predictor. The model gives more 'False UPs' than 'True DOWNs' thus being poor at predicting if the market would move down. The model does a decent job of predicting if the market would move up, as it has more than 10x 'True UPs' than 'False DOWNs'.

2. **Adding Volume as predictor along with Lag2**

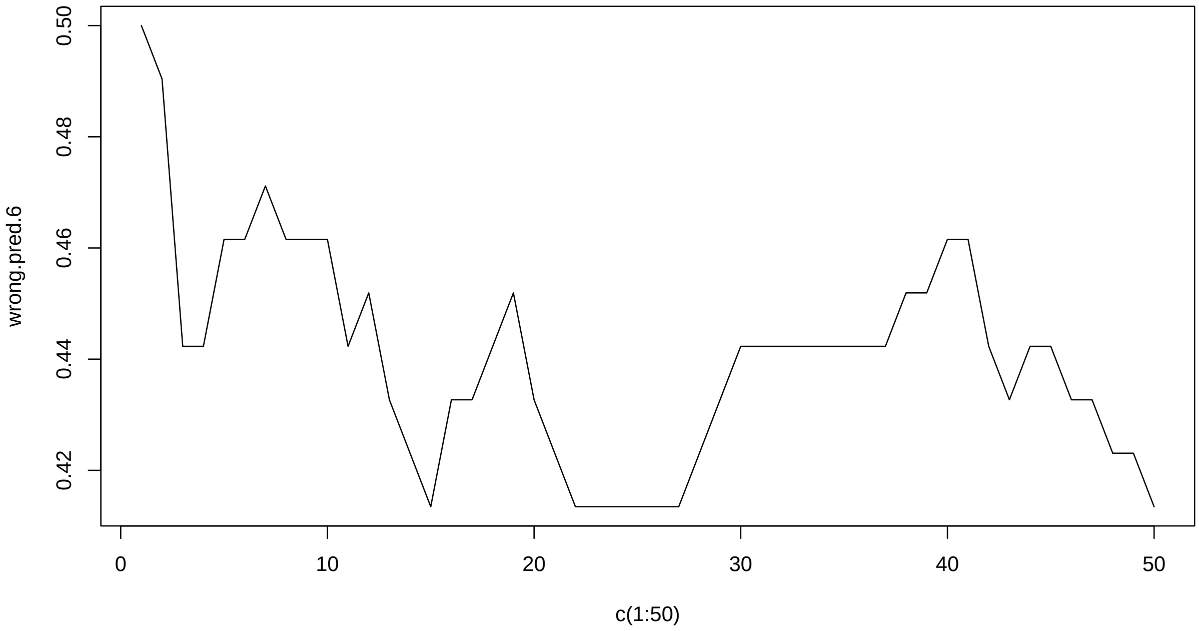
```
##
## Weekly.glm.pred.5 Down Up
##              Down   20 25
##              Up     23 36
```

```
## [1] 0.5384615
```

```
## [1] 0.4615385
```

The model correctly predicts UP movement in 36 weeks and DOWN movement in 20 weeks out of a total of 104 weeks. Fraction of correct predictions is **0.538**. Training error rate is **0.461**, higher than the logistic regression model with only Lag2 as predictor. This model brings down the proportion of 'false UPs' given by the model in comparison to 'true DOWNs' when compared with all previous logistic regression models.

3. **Carrying out KNN with K ranging from 1 to 50**



```
## [1] 15
```

```
## [1] 0.4134615
```

KNN with K=15 has lowest error rate among all attempted KNN models at **0.413**. This is an improvement over the error rate of **0.5** achieved with KNN where K=1.

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.

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

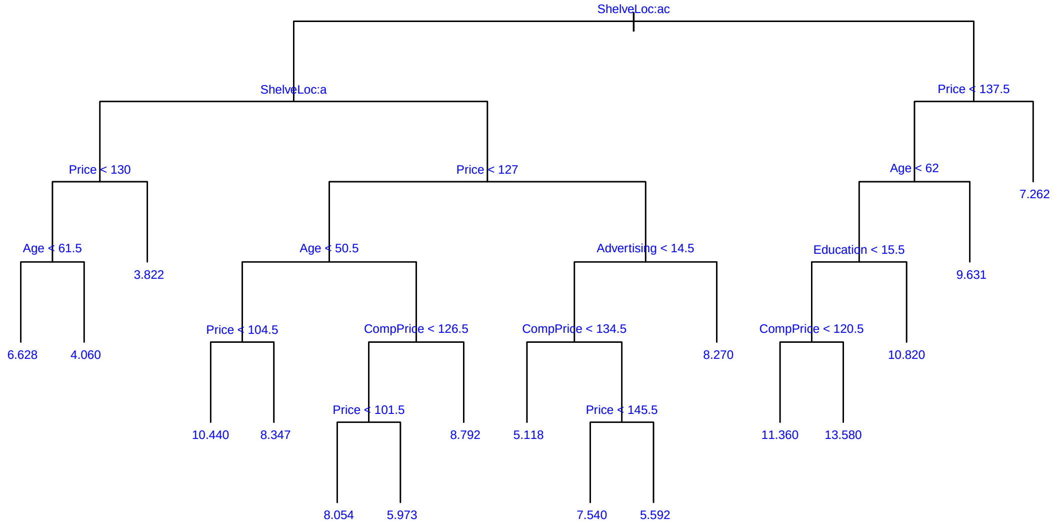
```
## [1] 200
```

The data has been split into training and test sets with 1/2 the rows from original data frame present in the training set.

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

```
## [1] 17
```

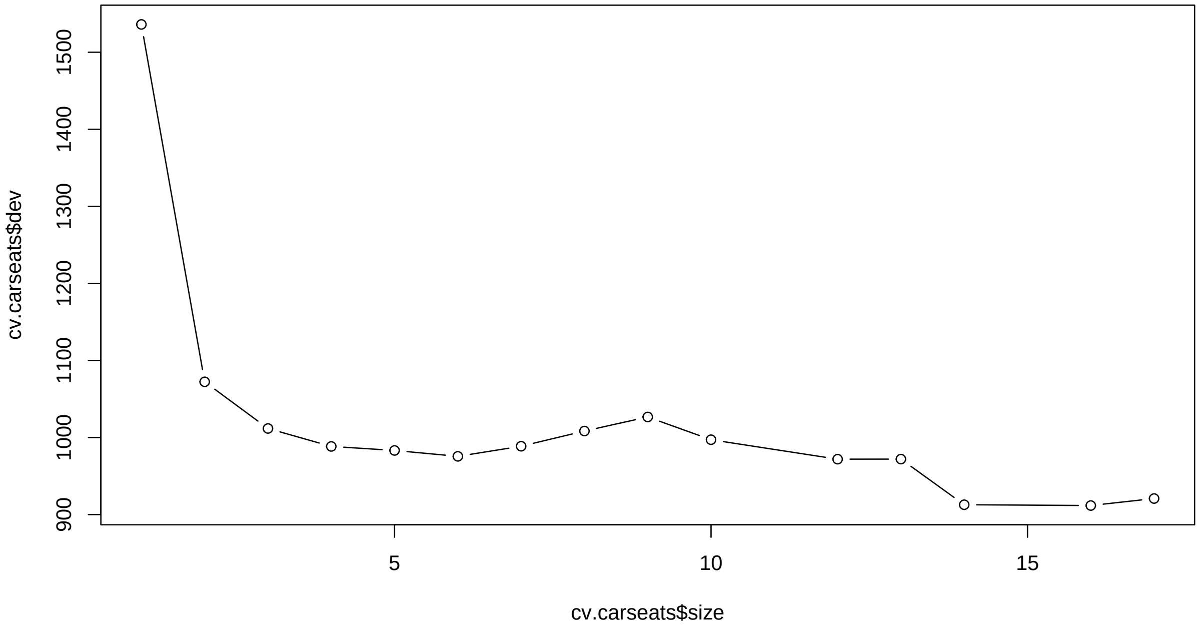
```
##
## Regression tree:
## tree(formula = Sales ~ ., data = df_train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "CompPrice"   "Advertising"
## [6] "Education"
## Number of terminal nodes:  17
## Residual mean deviance:  1.844 = 337.5 / 183
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.81700 -0.89580 -0.01857  0.00000  0.92000  2.91800
```

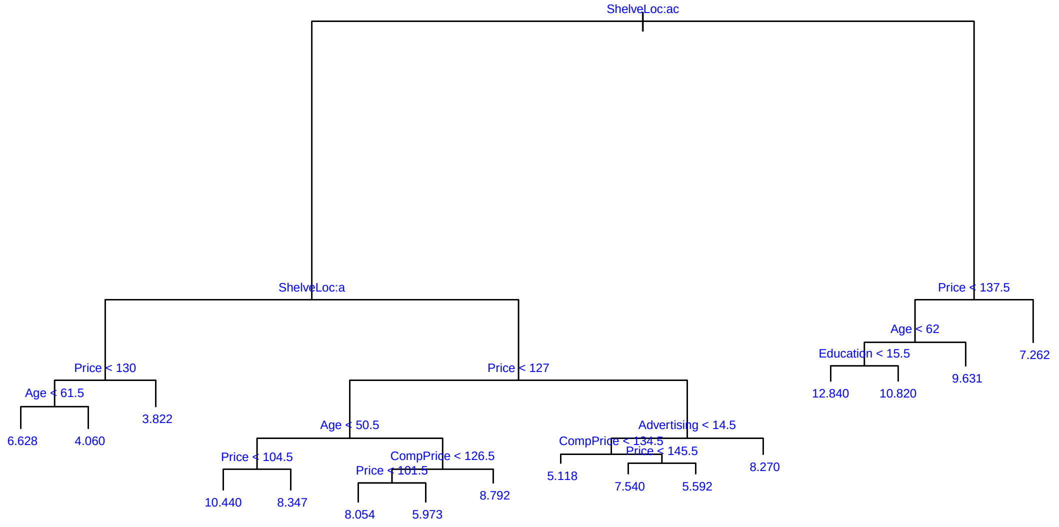


```
## [1] 5.395751
```

7 predictors are used in tree construction - ShelveLoc, Price, CompPrice, Income, Population, Advertising, Age
Number of splits = 16, Number of leaves = 17
Residual Mean Deviance = 2.623
Test MSE = 5.396

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





```
## [1] 5.524571
```

Cross Validation has selected the tree in this case with number of leaves = 15, hence we prune. However, pruning does not improve Test MSE.

(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 = df_train, mtry = ncol(df_train) -      1, importance = TRU
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.508844
##                     % Var explained: 66.28
```

```
##                %IncMSE IncNodePurity
## CompPrice   22.6252650    122.786295
## Income      -0.5752613     54.480928
## Advertising 13.4548844     72.467497
## Population   2.1157044     54.612739
## Price       47.9605574    350.268318
## ShelveLoc   67.7892048    597.768236
## Age         18.8685134    150.445895
## Education    1.4567197     40.065187
## Urban       -0.5901727      5.816261
## US           4.2205104      5.157034
```

10 variables considered at each split, no. of trees constructed = 500
MSE = 2.53
Test MSE has improved to 3.27
The most important variables are - ShelveLoc and Price

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

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = df_train, mtry = m,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 2.732511
##                     % Var explained: 63.27
```

```
## [1] 3.317538
```

```
##                %IncMSE IncNodePurity
## CompPrice   14.2321001    130.050643
## Income      -0.8384576     85.170671
## Advertising  7.7590501     94.454831
## Population  -0.5627058     89.541470
## Price       31.7838871    317.263735
## ShelveLoc   49.3552711    470.795093
## Age         13.1496226    152.088874
## Education    1.4820212     62.677328
## Urban       -0.9108867      8.079763
## US           3.6181113     14.806433
```

4 variables considered at each split, no. of trees constructed = 500
MSE = 2.64
Test MSE has further improved to 3.24
The most important variables are - ShelveLoc and Price
Since we consider only a subset of available predictors for each split in the tree, the trees are more decorrelated than in bagging where all predictors are considered for every split. Combining these uncorrelated trees reduces variance, thereby improving the Test MSE.

Chapter 8 Q11

This question uses the Caravan data set.

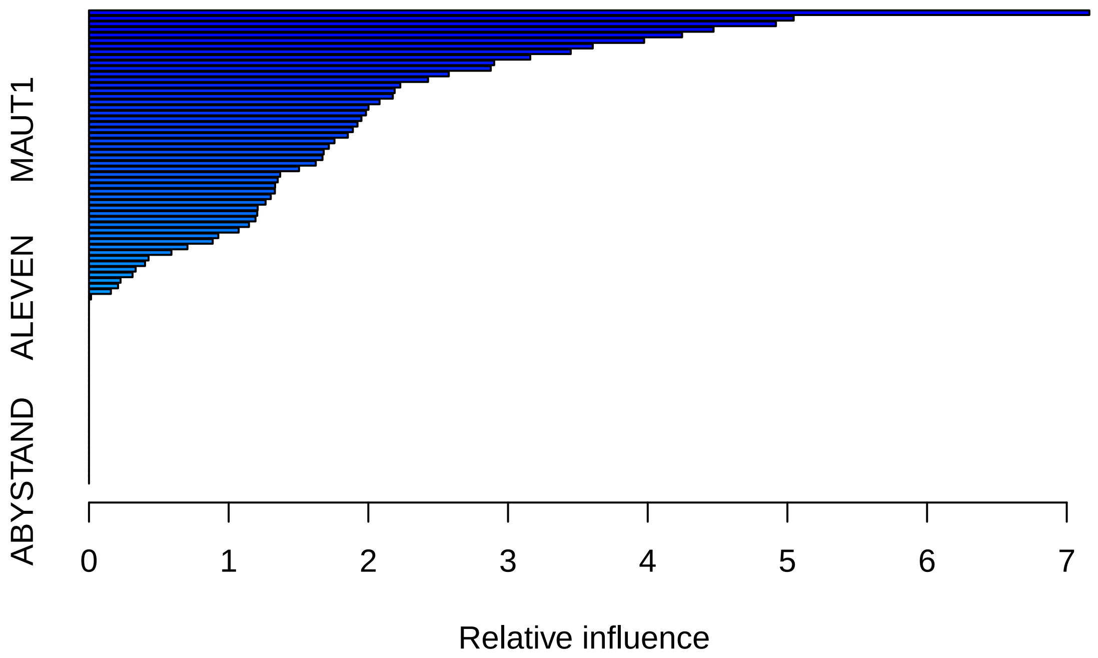
(a) Create a training set consisting of the first 1000 observations, and a test set consisting of the remaining observations.

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

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

The data has been split into training and test sets with first 1000 observations being part of the training set.

(b) Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?



```
##                 var    rel.inf
## PPERSAUT PPERSAUT 7.16175100
## MOSTYPE   MOSTYPE 5.04520919
## MGODGE     MGODGE 4.91815460
## MOPLHOOG MOPLHOOG 4.47167366
## PBRAND     PBRAND 4.24623405
## MKOOPKLA MKOOPKLA 3.97445223
## MBERMIDD MBERMIDD 3.60652310
## MGODPR     MGODPR 3.44917121
## MINK3045 MINK3045 3.15924536
## MSKC         MSKC 2.90076521
## MAUT2       MAUT2 2.87661680
## MBERARBG MBERARBG 2.57600772
## PWAPART   PWAPART 2.42763444
## MOPLMIDD MOPLMIDD 2.22863559
## MFWEKIND MFWEKIND 2.18840265
## MBERARBO MBERARBO 2.17522281
## MSKB1       MSKB1 2.08027636
## MSKA         MSKA 2.00115457
## MINKM30   MINKM30 1.98319453
## MINK7512 MINK7512 1.95059588
## MFALLEEN MFALLEEN 1.92239360
## MAUT1       MAUT1 1.88921867
## MGODOV     MGODOV 1.85329835
## MZFONDS   MZFONDS 1.75762191
## MRELGE     MRELGE 1.71722196
## MRELOV     MRELOV 1.68064894
## MBERHOOG MBERHOOG 1.67179340
## MSKB2       MSKB2 1.62380855
## MGEMLEEF MGEMLEEF 1.50384327
## MAUTO       MAUTO 1.36889564
## MRELSA     MRELSA 1.35142429
## ABRAND     ABRAND 1.33252059
## MINKGEM   MINKGEM 1.33102900
## MHHUUR     MHHUUR 1.30163929
## MZPART     MZPART 1.26442306
## MSKD         MSKD 1.20787514
## MINK4575 MINK4575 1.20474057
## MFGEKIND MFGEKIND 1.19218795
## MGODRK     MGODRK 1.14531889
## MHKOOP     MHKOOP 1.07191305
## MGEMOMV   MGEMOMV 0.92586462
## MOPLLAAG MOPLLAAG 0.88574020
## MBERZELF MBERZELF 0.70495951
## APERSAUT APERSAUT 0.59034465
## PMOTSCO   PMOTSCO 0.42689770
## MOSHOOFD MOSHOOFD 0.40133496
## PLEVEN     PLEVEN 0.33423934
## MBERBOER MBERBOER 0.31198404
## PBYSTAND PBYSTAND 0.22612751
## MINK123M MINK123M 0.20798902
## MAANTHUI MAANTHUI 0.15693996
## ALEVEN     ALEVEN 0.01484142
## 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
## 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
```

Predictor PPERSAUT is the most important one, followed by MOSTYPE,MGODGE,MOPLHOOG,PBRAND.

(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?

```
##
## yhat_boost    0    1
##          0 4337  258
##          1  196   31
```

```
## [1] 0.0941518
```

```
## [1] 0.1365639
```

```
## [1] 0.2571429
```

Purchase Rate = 0.137
KNN Purchase Rate = 0.257
Thus, the model with KNN provides a higher purchase rate.

# Exam Problems

Problem 1: Beauty Pays!

Solution

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

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

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

```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore, data = df)
##
## 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
```

Beauty has a very low p-value, thereby Null Hypothesis is rejected.
Beauty has a positive coefficient meaning that a unit increase in the beauty score leads to a 0.27 increase in course evaluation.
Beauty explains 16.57% of the variance in the Course Evaluations.

Now we fit a multiple linear regression using all variables to predict course evaluation.

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

Beauty has a very low p-value again, thereby Null Hypothesis is rejected.
Beauty has a positive coefficient meaning that course evaluations are indeed positively correlated with beauty.
Keeping everything else constant, a unit increase in beauty score leads to 0.3 increase in course evaluation.
Using all models, 34% of the variance in course evaluations can be explained, which is more than just 16.57% variance explained by beauty alone.

Thus, Dr. Hamermesh is right in pointing out that beauty plays a role in course evaluations, and by extension in labor income. However, beauty alone is not the determinant of course evaluations, there are multiple other factors that affect course evaluations as well.

What Dr. Hamermesh means by the sentence *"Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible"* is that he is puzzled whether beauty leads to more productivity in professors and thus higher evaluation (and income), or whether the professors are being discriminated on the basis of beauty and thus the poor evaluation (and income).

Problem 2: Housing Price Structure

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

```
## [1] 128   8
```

```
## [1] "Home"      "Nbhd"      "Offers"    "SqFt"      "Brick"     "Bedrooms"
## [7] "Bathrooms" "Price"
```

```
##       Home             Nbhd           Offers           SqFt     
##  Min.   :  1.00   Min.   :1.000   Min.   :1.000   Min.   :1450  
##  1st Qu.: 32.75   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1880  
##  Median : 64.50   Median :2.000   Median :3.000   Median :2000  
##  Mean   : 64.50   Mean   :1.961   Mean   :2.578   Mean   :2001  
##  3rd Qu.: 96.25   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2140  
##  Max.   :128.00   Max.   :3.000   Max.   :6.000   Max.   :2590  
##     Brick              Bedrooms       Bathrooms         Price       
##  Length:128         Min.   :2.000   Min.   :2.000   Min.   : 69100  
##  Class :character   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:111325  
##  Mode  :character   Median :3.000   Median :2.000   Median :125950  
##                     Mean   :3.023   Mean   :2.445   Mean   :130427  
##                     3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:148250  
##                     Max.   :5.000   Max.   :4.000   Max.   :211200  
```

```
## 
## Call:
## lm(formula = Price ~ ., data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -24940.6  -8383.0    430.7   7430.4  31371.2 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -9814.663   9858.884  -0.996  0.32149    
## Home            6.187     28.973   0.214  0.83128    
## Nbhd         9832.281   1821.869   5.397 3.47e-07 ***
## Offers      -8351.794   1267.428  -6.590 1.24e-09 ***
## SqFt           49.811      6.769   7.359 2.53e-11 ***
## BrickYes    15601.818   2261.896   6.898 2.66e-10 ***
## Bedrooms     5671.911   1840.979   3.081  0.00256 **
## Bathrooms    8243.545   2449.897   3.365  0.00103 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11540 on 120 degrees of freedom
## Multiple R-squared:  0.8256, Adjusted R-squared:  0.8154
## F-statistic: 81.15 on 7 and 120 DF,  p-value: < 2.2e-16
```

Upon performing multiple linear regression on the data with Price to be predicted as a linear function of all other variables, the p-value for Brick is very small, thereby we reject the Null Hypothesis. Also, the coefficient of Brick (Yes = 1 and No = 0) is positive. In fact, assuming all other predictors are unchanged, brick houses command a premium of USD 15601.82 over non-brick houses.

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

```
##
## Call:
## lm(formula = Price ~ Nbhd.2 + Nbhd.3, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42931 -12310  -1643  11251  51905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   110155       2693  40.904  < 2e-16 ***
## Nbhd.2         15077       3787   3.981 0.000116 ***
## Nbhd.3         49140       3929  12.508  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 17860 on 125 degrees of freedom
## Multiple R-squared:  0.565,  Adjusted R-squared:  0.558
## F-statistic: 81.16 on 2 and 125 DF,  p-value: < 2.2e-16
```

Here, we perform a linear regression on the data with Price to be predicted as a linear function of *Nbhd.2* and *Nbhd.3* variables. *Nbhd* has 3 levels and hence we make 2 dummy variables to run linear regression -
Nbhd.3 == 1 then *Nbhd* is 3
Nbhd.2 == 1 then *Nbhd* is 2
Nbhd.2 == *Nbhd.3* == 0 then *Nbhd* is 1
The p-value for *Nbhd.3* is very small, thus we reject the Null Hypothesis. Also, the coefficient of *Nbhd.3* is positive and greater than that of *Nbhd.2*, thus houses in neighborhood 3 command premium over the other houses.

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

```
##
## Call:
```

```
## lm(formula = Price ~ Brick, data = df.nbhd3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23600  -9415  -2700   8135  36000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   148230       3067  48.325  < 2e-16 ***
## BrickYes       26970       4789   5.632 1.98e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14710 on 37 degrees of freedom
## Multiple R-squared:  0.4616, Adjusted R-squared:  0.447
## F-statistic: 31.72 on 1 and 37 DF,  p-value: 1.981e-06
```

We first take subest of the data for neighborhood 3 and then perform linear regression with Brick as the predictor. The p-value is very small, thus we reject the Null Hypothesis. Also, the coefficient of Brick (Yes = 1 and No = 0) is positive, and a brick house commands a premium of USD 26970 over a non brick house in neighborhood 3.

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

```
##
## Call:
## lm(formula = Price ~ Nbhd.Modified, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48678 -12320  -1786  10368  51905
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          159295       3024   52.67   <2e-16 ***
## Nbhd.Modifiedolder   -41517       3627  -11.45   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 18890 on 126 degrees of freedom
## Multiple R-squared:  0.5098, Adjusted R-squared:  0.5059
## F-statistic:   131 on 1 and 126 DF,  p-value: < 2.2e-16
```

Yes, the neighborhoods 1 and 2 can be combined together as ‘older’ and we still arrive at the same conclusion that the houses in neighborhood 3 command a premium over the houses in older neighborhoods 1 and 2.

Problem 3: What causes what??

Solution
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)

In most of the cases, cities with more crime have more police personnel on the streets. As such, a simple regression to predict crime based on police would yield a positive correlation between the two meaning that more police equals more crime. While there is correlation, there is no causation to prove that more police leads to more crime, in fact it sounds counter intuitive as there would be more police because of more crime. Thus, since we cannot determine any causal relationship between police and crime, we cannot run a simple regression as it would throw up bizarre results such as the lesser police personnel you deploy on the streets, the lower would be the crime rate.

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.
The researchers studied an example from Washington DC where additional police personnel are deployed on the streets irrespective of the crime due to other factors such as the terror alert level. When the terror alert level rises to orange, additional police personnel are deployed on the streets, and in turn crime drops as well. Thus they were able to build a case for a causal relationship between police and crime where rise in police reduces crime. The Table 2 shows negative coefficient for high alert for both regressions, thereby corroborating the finding that crime reduces on high alert days. In the 2nd column of this table, we see that the coefficient for metro ridership is positive, meaning that increased metro ridership (hypothesized as more tourists in the city) leads to more crime.

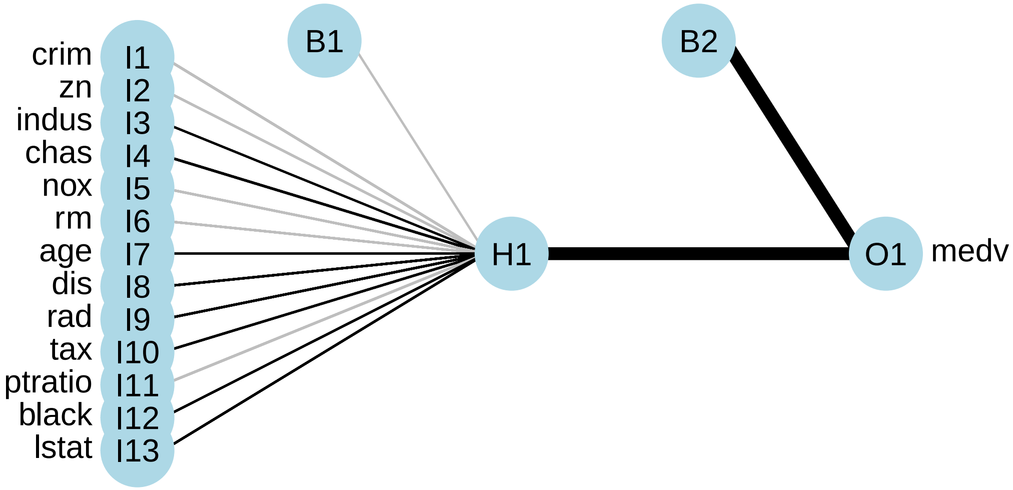
3. Why did they have to control for METRO ridership? What was that trying to capture?
The researchers studied the metro ridership to test the hypothesis that less tourists went out in the city on high alert days and that led to drop in crime rates. As it turns out, the metro ridership numbers remained unaffected on high alert days, thus nullifying this hypothesis. Therefore, the researchers were able to make a strong case for a causal relationship between more police on the streets and less crime.

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 first column in Table 4 provides the Robust Regression Coefficients for multiple variables. Thus, the model used here is Robust Regression. Robust Regression weighs each observation differently so as to reduce the impact of outliers on the final model. It also helps identify influential observations - such as the observation that high alert in District 1 has more substantial effect on reduction of crime than high alert in the other districts. This is inferred based on the magnitude of the coefficients of the first two variables in the table.
To conclude, high alert leads to a reduction in crime rate with greater reduction in crime seen when this alert is issued in District 1. This conclusion seems intuitive as all the high profile targets in Washington DC are within the first police district, and a higher threat level in this district would lead to more deployment of police personnel on the streets, thereby reducing the crime the largest.

Problem 4: Neural Nets

Solution
We fit a neural net to predict *medv* based on all other variables in Boston data set as predictors.

```
## # weights:  29
## initial  value 15819654.958400
## iter  10 value 239820.218471
## iter  20 value 8440.741038
## iter  20 value 8440.741021
## iter  20 value 8440.741021
## final  value 8440.741021
## converged
## a 13-1-1 network with 29 weights
## inputs: crim zn indus chas nox rm age dis rad tax ptratio black lstat
## output(s): medv
## options were - skip-layer connections  linear output units
```



```
## [1] 4.629526
```

A single layer neural net with 1 hidden neuron and decay = 0 gives a Test RMSE of 4.629.

We fit neural nets with different sizes and decay parameters, and choose best neural net based on Test RMSE.

```
## # weights:  29
## initial  value 56791663.868945
## iter  10 value 127063.896989
## final  value 8440.741021
## converged
## # weights:  29
## initial  value 7967730.715098
## iter  10 value 398138.443497
## iter  20 value 8538.146986
## iter  30 value 8515.781592
## iter  40 value 8499.837401
## iter  50 value 8459.698732
## iter  60 value 8191.470651
## iter  70 value 8103.287370
## iter  80 value 7797.122544
## iter  90 value 7725.282511
## iter 100 value 7718.088392
## final  value 7718.088392
## stopped after 100 iterations
## # weights:  29
## initial  value 41586323.116373
## iter  10 value 265268.065931
## iter  20 value 10452.342618
## iter  30 value 8501.387105
## iter  40 value 8452.725034
## iter  50 value 8448.113843
## iter  60 value 8431.395416
## iter  70 value 8427.355664
## iter  80 value 8424.373258
## iter  90 value 8419.727462
## iter 100 value 8417.202740
## final  value 8417.202740
## stopped after 100 iterations
## # weights:  29
## initial  value 57104106.542887
## iter  10 value 251690.831545
## iter  20 value 8529.579155
## iter  30 value 8448.462146
## iter  40 value 8446.883599
## iter  50 value 8440.489638
## iter  60 value 8440.066466
## iter  70 value 8421.915341
## iter  80 value 8405.207605
## iter  90 value 8365.616486
## iter 100 value 8331.823285
## final  value 8331.823285
## stopped after 100 iterations
## # weights:  29
## initial  value 30504355.607305
## iter  10 value 85497.373852
## iter  20 value 8440.851855
## final  value 8440.851461
## converged
## # weights:  29
## initial  value 492977.436019
## iter  10 value 94876.695883
## iter  20 value 8440.760287
## iter  20 value 8440.760233
## iter  20 value 8440.760233
## final  value 8440.760233
## converged
## # weights:  44
## initial  value 6307789.390203
## iter  10 value 1761476.691631
## iter  20 value 101495.743311
## iter  30 value 8470.482253
## iter  40 value 8408.389056
## iter  50 value 8358.265405
## iter  60 value 8357.311658
## final  value 8357.297198
## converged
## # weights:  44
## initial  value 5367676.002897
## iter  10 value 1884223.269026
## iter  20 value 8519.981313
## iter  30 value 8498.307840
## iter  40 value 8427.074952
## iter  50 value 8404.369060
## iter  60 value 7847.185616
## iter  70 value 7571.470924
## iter  80 value 7542.177212
## iter  90 value 7529.077905
## iter 100 value 7521.090114
## final  value 7521.090114
## stopped after 100 iterations
## # weights:  44
## initial  value 2692844.781941
## iter  10 value 1391275.476346
## iter  20 value 201308.145347
## iter  30 value 62851.644526
## iter  40 value 18169.135979
## iter  50 value 8452.234020
## iter  60 value 8435.149625
## iter  70 value 8426.652655
## iter  80 value 8409.887092
## iter  90 value 8353.826595
## iter 100 value 8235.047753
## final  value 8235.047753
## stopped after 100 iterations
## # weights:  44
## initial  value 4865286.873521
## iter  10 value 2131518.479954
## iter  20 value 8441.847167
## iter  20 value 8441.847135
## iter  20 value 8441.847135
## final  value 8441.847135
## converged
## # weights:  44
## initial  value 47488446.405746
## iter  10 value 328362.160800
## iter  20 value 8459.018909
## iter  30 value 8440.868536
## final  value 8440.843122
## converged
## # weights:  44
## initial  value 15804319.570806
## iter  10 value 476100.241215
## iter  20 value 39655.534473
## iter  30 value 8414.494202
## iter  40 value 8409.741990
## iter  50 value 8408.616818
## iter  60 value 8408.588995
## iter  70 value 8406.114858
## iter  80 value 8405.919285
## iter  90 value 8405.780641
## iter 100 value 8325.929318
## final  value 8325.929318
## stopped after 100 iterations
## # weights:  89
## initial  value 40910462.481848
## iter  10 value 162846.995881
## iter  20 value 65494.970377
## iter  30 value 46583.266123
## iter  40 value 9968.649647
## iter  50 value 8390.310151
## iter  60 value 8056.236674
## iter  70 value 8038.726982
## iter  80 value 8017.527060
## iter  90 value 7892.392834
## iter 100 value 7859.920706
## final  value 7859.920706
## stopped after 100 iterations
## # weights:  89
## initial  value 2581782.001109
## iter  10 value 745722.626011
## iter  20 value 33564.325155
## iter  30 value 8793.337636
## iter  40 value 8735.412869
## iter  50 value 8724.187205
## iter  60 value 8521.076607
## iter  70 value 8360.349829
## iter  80 value 8084.916252
## iter  90 value 7797.617563
## iter 100 value 7429.130056
## final  value 7429.130056
## stopped after 100 iterations
## # weights:  89
## initial  value 6152560.870130
## iter  10 value 1356643.181588
## iter  20 value 492457.438917
## iter  30 value 12895.667469
## iter  40 value 11534.130292
## iter  50 value 11202.239790
## iter  60 value 11088.320616
## iter  70 value 10749.291126
## iter  80 value 10447.196666
## iter  90 value 10396.490075
## iter 100 value 10191.852714
## final  value 10191.852714
## stopped after 100 iterations
## # weights:  89
## initial  value 28131376.988229
## iter  10 value 45602.132069
## iter  20 value 11674.786087
## iter  30 value 8371.162269
## iter  40 value 8365.871461
## iter  50 value 8364.669726
## iter  60 value 8339.418748
## iter  70 value 8327.326889
## iter  80 value 8302.127635
## iter  90 value 8163.533304
## iter 100 value 7862.090295
## final  value 7862.090295
## stopped after 100 iterations
## # weights:  89
## initial  value 12208937.164215
## iter  10 value 102213.758485
## iter  20 value 22842.693244
## iter  30 value 8472.574056
## iter  40 value 8409.516949
## iter  50 value 8409.226155
## iter  60 value 8409.135638
## iter  70 value 8393.818914
## iter  80 value 8393.128283
## iter  90 value 8392.578284
## iter 100 value 8365.508937
## final  value 8365.508937
## stopped after 100 iterations
## # weights:  89
## initial  value 9103113.927309
## iter  10 value 642045.138762
## iter  20 value 550574.833621
## iter  30 value 8756.988240
## iter  40 value 8402.163497
## iter  50 value 8402.037508
## iter  60 value 8397.401266
## iter  70 value 8396.794606
## iter  80 value 8396.756306
## final  value 8396.734039
## converged
## # weights:  164
## initial  value 5764869.938282
## iter  10 value 233770.149992
## iter  20 value 56441.007696
## iter  30 value 34297.020101
## iter  40 value 17814.149765
## iter  50 value 10111.701701
## iter  60 value 8284.467873
## iter  70 value 8166.605712
## iter  80 value 8074.507789
## iter  90 value 7976.954552
## iter 100 value 7900.058036
## final  value 7900.058036
## stopped after 100 iterations
## # weights:  164
## initial  value 47542094.668262
## iter  10 value 277997.159385
## iter  20 value 67479.629013
## iter  30 value 15300.176142
## iter  40 value 10073.208870
## iter  50 value 8884.078596
## iter  60 value 8558.997422
## iter  70 value 8550.539258
## iter  80 value 8523.269370
## iter  90 value 8457.818883
## iter 100 value 8440.134813
## final  value 8440.134813
## stopped after 100 iterations
## # weights:  164
## initial  value 5600309.773220
## iter  10 value 86032.935247
## iter  20 value 53171.281168
## iter  30 value 12943.702178
## iter  40 value 9322.191749
## iter  50 value 8447.049189
## iter  60 value 8432.025013
## iter  70 value 8275.967694
## iter  80 value 8240.919294
## iter  90 value 8155.861187
## iter 100 value 8144.090479
## final  value 8144.090479
## stopped after 100 iterations
## # weights:  164
## initial  value 2591493.624197
## iter  10 value 559006.597238
## iter  20 value 108876.371664
## iter  30 value 89167.209130
## iter  40 value 8942.148978
## iter  50 value 7262.797710
## iter  60 value 7201.487246
## iter  70 value 7198.575711
## iter  70 value 7198.575645
## final  value 7198.575645
## converged
## # weights:  164
## initial  value 552648.603756
## iter  10 value 127064.965743
## iter  20 value 92937.334935
## iter  30 value 40613.040705
## iter  40 value 10213.821209
## iter  50 value 8257.766143
## iter  60 value 8037.466299
## iter  70 value 7891.050193
## iter  80 value 7889.153704
## iter  90 value 7886.592744
## iter 100 value 7885.804869
## final  value 7885.804869
## stopped after 100 iterations
## # weights:  164
## initial  value 963634.173835
## iter  10 value 62558.850444
## iter  20 value 53445.709678
## iter  30 value 16929.274037
## iter  40 value 8480.045128
## iter  50 value 8386.998625
## iter  60 value 8339.909336
## iter  70 value 8322.316509
## iter  80 value 8311.183111
## iter  90 value 8296.482099
## iter 100 value 8292.878507
## final  value 8292.878507
## stopped after 100 iterations
## # weights:  314
## initial  value 6462391.268699
## iter  10 value 51539.965017
## iter  20 value 22741.328714
## iter  30 value 13983.740620
## iter  40 value 10280.397644
## iter  50 value 9609.891582
## iter  60 value 8662.937930
## iter  70 value 7939.628914
## iter  80 value 7589.655504
## iter  90 value 7514.819785
## iter 100 value 7513.744432
## final  value 7513.744432
## stopped after 100 iterations
## # weights:  314
## initial  value 14112889.565262
## iter  10 value 512667.695354
## iter  20 value 377568.245287
## iter  30 value 261304.997680
## iter  40 value 214094.228655
## iter  50 value 133924.127740
## iter  60 value 106035.135090
## iter  70 value 90017.699010
## iter  80 value 20723.692495
## iter  90 value 9411.447257
## iter 100 value 8486.824155
## final  value 8486.824155
## stopped after 100 iterations
## # weights:  314
## initial  value 2426623.257434
## iter  10 value 128619.162161
## iter  20 value 25815.980931
## iter  30 value 16413.877027
## iter  40 value 9930.198623
## iter  50 value 8545.566393
## iter  60 value 8372.003746
## iter  70 value 8329.982989
## iter  80 value 8211.506898
## iter  90 value 8156.175612
## iter 100 value 8071.120986
## final  value 8071.120986
## stopped after 100 iterations
## # weights:  314
## initial  value 66199327.398091
## iter  10 value 158135.623628
## iter  20 value 143784.496418
## iter  30 value 34243.159536
## iter  40 value 26270.724705
## iter  50 value 16669.710954
## iter  60 value 13145.992720
## iter  70 value 11132.190331
## iter  80 value 8326.143840
## iter  90 value 7165.308189
## iter 100 value 6759.981641
## final  value 6759.981641
## stopped after 100 iterations
## # weights:  314
## initial  value 2764355.761086
## iter  10 value 281307.420926
## iter  20 value 45104.822341
## iter  30 value 28353.047434
## iter  40 value 15924.393006
## iter  50 value 12068.891427
## iter  60 value 9593.943458
## iter  70 value 8399.806294
## iter  80 value 8001.365547
## iter  90 value 7753.085063
## iter 100 value 7564.006964
## final  value 7564.006964
## stopped after 100 iterations
## # weights:  314
## initial  value 4981076.271006
## iter  10 value 917686.125252
## iter  20 value 804502.608203
## iter  30 value 231419.436978
## iter  40 value 173787.589456
## iter  50 value 139014.055912
## iter  60 value 17913.939754
## iter  70 value 8304.664378
## iter  80 value 8255.790186
## iter  90 value 8244.062142
## iter 100 value 8223.904719
## final  value 8223.904719
## stopped after 100 iterations
## # weights:  464
## initial  value 8893668.257786
## iter  10 value 341611.850347
## iter  20 value 245587.302926
## iter  30 value 22893.331469
## iter  40 value 13602.223637
## iter  50 value 10224.868958
## iter  60 value 8401.813581
## iter  70 value 8178.397649
## iter  80 value 8093.647458
## iter  90 value 8087.128975
## iter 100 value 8080.597394
## final  value 8080.597394
## stopped after 100 iterations
## # weights:  464
## initial  value 23121310.099318
## iter  10 value 572518.092493
## iter  20 value 34796.078807
## iter  30 value 22092.049128
## iter  40 value 11283.272555
## iter  50 value 9932.985168
## iter  60 value 8517.396484
## iter  70 value 7912.847552
## iter  80 value 7768.528007
## iter  90 value 7735.998656
## iter 100 value 7704.566503
## final  value 7704.566503
## stopped after 100 iterations
## # weights:  464
## initial  value 22751243.072930
## iter  10 value 1363729.030515
## iter  20 value 1226295.397192
## iter  30 value 790643.948857
## iter  40 value 233209.032843
## iter  50 value 32379.036887
## iter  60 value 11011.130062
## iter  70 value 9564.503695
## iter  80 value 9340.152383
## iter  90 value 9230.558269
## iter 100 value 9220.711627
## final  value 9220.711627
## stopped after 100 iterations
## # weights:  464
## initial  value 44915873.729132
## iter  10 value 601756.876650
## iter  20 value 164026.773023
## iter  30 value 15243.414305
## iter  40 value 11060.895378
## iter  50 value 9164.275288
## iter  60 value 8896.276711
## iter  70 value 8259.110309
## iter  80 value 8197.635469
## iter  90 value 7993.085302
## iter 100 value 7296.127549
## final  value 7296.127549
## stopped after 100 iterations
## # weights:  464
## initial  value 3352313.606429
## iter  10 value 310836.471150
## iter  20 value 232669.818427
## iter  30 value 44945.504653
## iter  40 value 21387.196879
## iter  50 value 13332.420815
## iter  60 value 10190.260959
## iter  70 value 8754.703176
## iter  80 value 8023.609872
## iter  90 value 7915.132286
## iter 100 value 7879.031495
## final  value 7879.031495
## stopped after 100 iterations
## # weights:  464
## initial  value 68975087.174099
## iter  10 value 60142.159532
## iter  20 value 22185.280128
## iter  30 value 12370.161455
## iter  40 value 10129.857649
## iter  50 value 9339.573810
## iter  60 value 8707.229423
## iter  70 value 8127.230174
## iter  80 value 8005.584933
## iter  90 value 7909.352335
## iter 100 value 7790.138938
## final  value 7790.138938
## stopped after 100 iterations
```

```
## [1] 0.1
```

```
## [1] 5
```

```
## [1] 4.407205
```

Best Test RMSE of 4.407 is obtained for a neural net with size = 5 and decay = 1

Problem 5: Final Project

Solution

My group project was finding out the variables that determine 3-year returns of small cap mutual funds. I, along with Meha Mehta, Charan Musunuru, Jacob Pammer and Chandler Wann, together worked on this project.

Upon deciding on our problem statement, each of us individually started hunting for relevant data sets. I and Chandler jointly discovered the small cap mutual funds data set on the Fidelity Investments website. The data was part of 7 separate workbooks each with 7 worksheets, for a total of 49 worksheets. I was tasked with unifying the data into a single data frame. Upon creating a unified data frame, I further proceeded with data cleaning. This involved eliminating duplicate columns and dropping rows where the predictors had null values. After we had the relevant columns, each of the columns had to be further transformed to make the data ready for feeding into our statistical models. For this, prefixes such as the '$' sign and suffixes such as the '%' sign had to be removed and the strings had to be parsed as numbers. Further, new calculated columns were created based on existing columns in the data, such as creation of 'Life of Fund' column based on 'Fund Start Date'.

Apart from data sourcing and data cleaning, I worked on fitting a Random Forest regression model on the data. The model was run with m=4 (total predictors were 12) and was able to achieve a test root mean squared error of 2.32%. Further, important predictors were identified, and results were corroborated with other team members who had worked on different regression models.

Lastly, I jointly worked on the presentation deck that was used for presenting our problem, work and the outcomes with the class. This involved selecting appropriate formatting and theme for the deck so as to make our presentation crisp and comprehensible.