Introduction

This submission consists of 12 questions and solutions of the Exam. The R code can be found in a separate file attached along with this pdf.

Chapter 2 | Question 10

a). Load in the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

Solution:

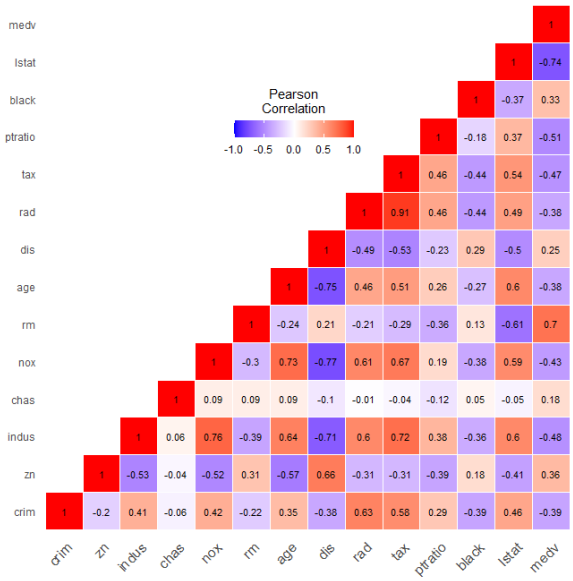
- The Boston data set consists of 14 columns and 506 rows
- The columns represent the different variables that are present in the Boston data set. The rows represent the values for the different variables in the Boston data set.

```
> dim(Boston)
[1] 506  14
```

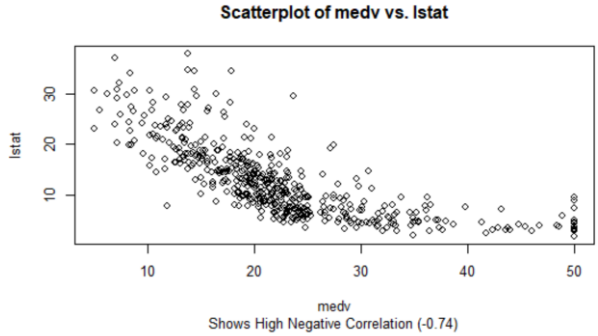
b). Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

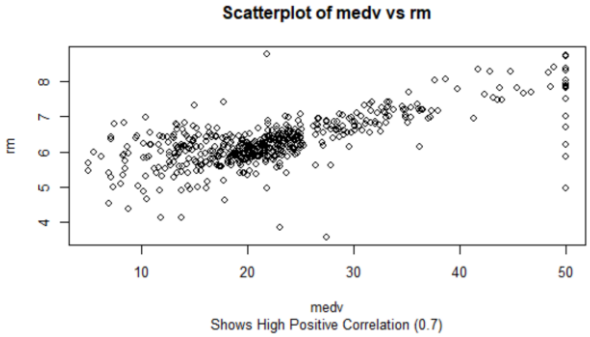
Solution:

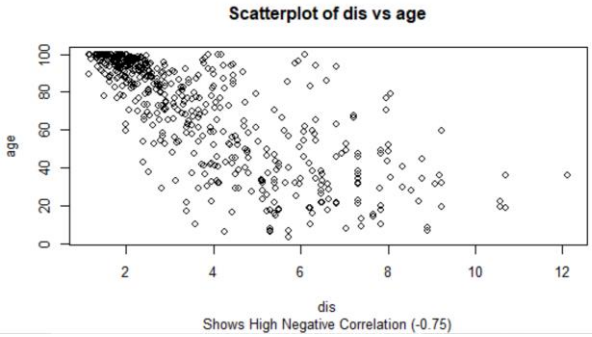
- This data consists of 14 different predictors. In order to choose some predictors for the scatter plots, first a correlation matrix is plotted.

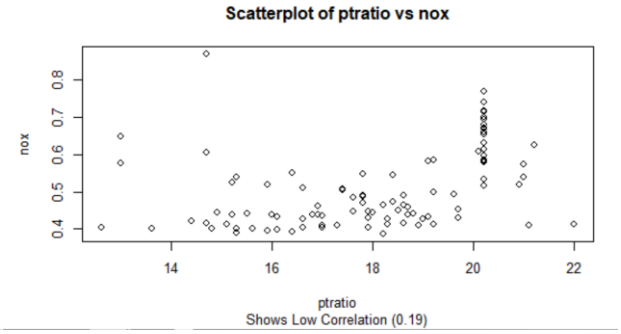


- The correlation matrix shows that
    - Medv vs. Lstat shows a High Negative Correlation of -0.74
    - Medv vs. rm shows a High Positive Correlation of 0.7
    - Dis vs. age shows a High Negative Correlation of -0.75
    - Ptratio vs. nox shows a low correlation of 0.19







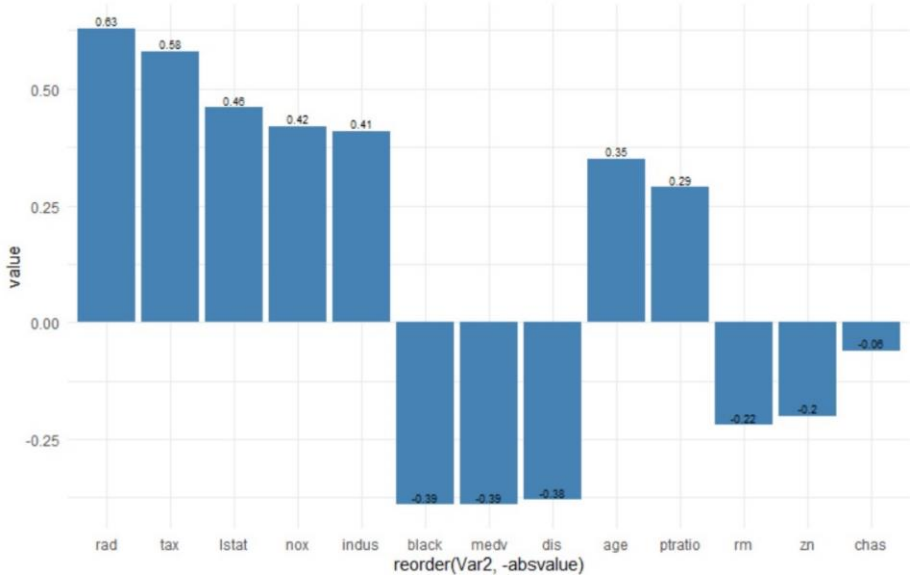


c). Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Solution:

- The below predictors are associated with per capita crime rate
    - rad: 0.63
    - tax: 0.58
    - lstat: 0.45
    - nox: 0.42
    - indus: 0.41
    - black: -0.35
    - medv: -0.35

- dis: -0.35

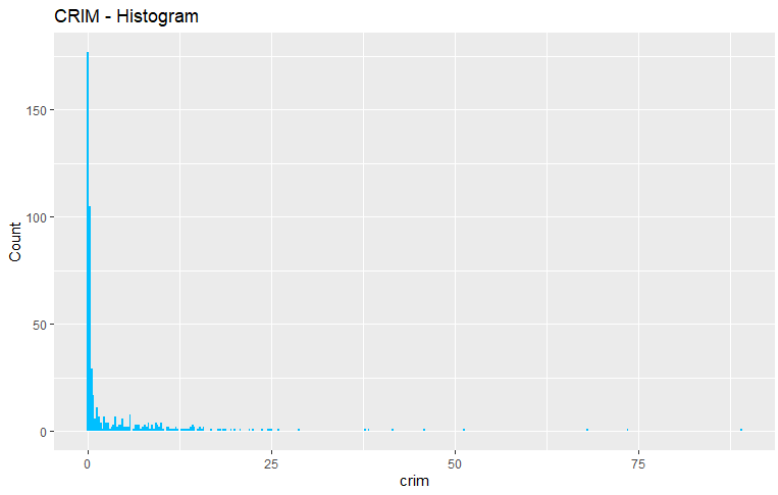


d). Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Solution:

- The 3rd Quartile Value is 3.677 and Max Value is 88.97. As inferred from the Histogram there are suburbs with extremely high crime rates.

```
> summary(crim)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```



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

Solution:

- There are 35 suburbs bound by the Charles River.

```
> table(Boston$chas)

  0   1
471  35
```

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

Solution:

- The median Pupil- Teacher ratio is 19.05.

```
> median(ptratio)
[1] 19.05
```

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

Solution:

- Suburbs with ID 399 and 406 have the lowest median value of owner-occupied homes.

```
> Boston[medv == min(medv),]
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
```

- The crime rates are obviously on the very upper end of the city, although not the suburbs with the highest crime.
- The proportion of residential land zoned for lots over 25,000 sq.ft. is 0, which is the minimum of the city and indicating that there is little investment in these suburbs.
- The proportion of non-retail business acres per town, lies within the 3rd quartiles for both suburbs. this indicates there are viable businesses and potentially employment in the area.
- The Charles River dummy variable only indicates the suburbs do not lie on the river side.
- The nitrogen oxides concentration (parts per 10 million) are in the upper quartile of the city, perhaps since the suburbs are so close to the highways.

- The average number of rooms per dwelling is in the lower quartile indicating smaller apartments or houses, however it is not at a minimimum. Perhaps the areas closer to the city have smaller apartments instead of houses.
- The age proportion of owner-occupied units built prior to 1940, is at the maximum indicating old housing units and no new builds
- The weighted mean of distances to five Boston employment centres is well into the lower quartile and close to the minimum. this indicates an area of high unemployment.
- The index of accessibility to radial highways is at the maximum indicating that the areas lit on or very near a highway.
- The full-value property-tax rate per $10,000 is quite high and in the upper quartile. Perhaps this is due to the small apartments/houses, in a city where the tax and unit area is not linearly correlated - so larger units actually pay less per square feet.
- The pupil-teacher ratio in these areas is in the upper quartile suggesting some relative under investment in schooling. H
- The variable black is around the median of all suburbs, indicating not a predominantly black population compared to the rest of the city.
- The lower status of the population (percent) of population is close the maximum and high in the upper quartile.

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.

Solution:

- 64 Suburbs average more than 7 rooms per dwellling.

```
> nrow(Boston[rm>7,])
[1] 64
```

- 13 Suburbs average more than 8 rooms per dwelling.

```
> nrow(Boston[rm>8,])
[1] 13
```

Chapter 3 | Question 15

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

Solutions:

- There are a total of 13 simple linear regression models to be run. The following table summarizes the output for each.

Dependent variable	t value	p value	|t|>2
zn	-4.59	5.51E-06	Yes
indus	9.991	2.00E-16	Yes
chas	-1.25	0.209	No
nox	10.4	2.00E-16	Yes
rm	-5.04	6.34E-07	Yes
age	8.463	2.85E-16	Yes
dis	-9.2	2.00E-16	Yes
rad	17.99	2.00E-16	Yes
tax	16.1	2.00E-16	Yes
ptratio	6.8	2.90E-11	Yes
black	-9.36	2.00E-16	Yes
lstat	11.49	2.00E-16	Yes
median	-9.46	2.00E-16	Yes

- Therefore, we can conclude that most variable have a statistically significant association with the predictor (crim), except for the 'chas' variable. This has a t-value equal to -1.25, which may result in coefficient value 0 for several datapoints.

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?

Solution:

Following Table is the summary of the output from the Multiple Regression Model.

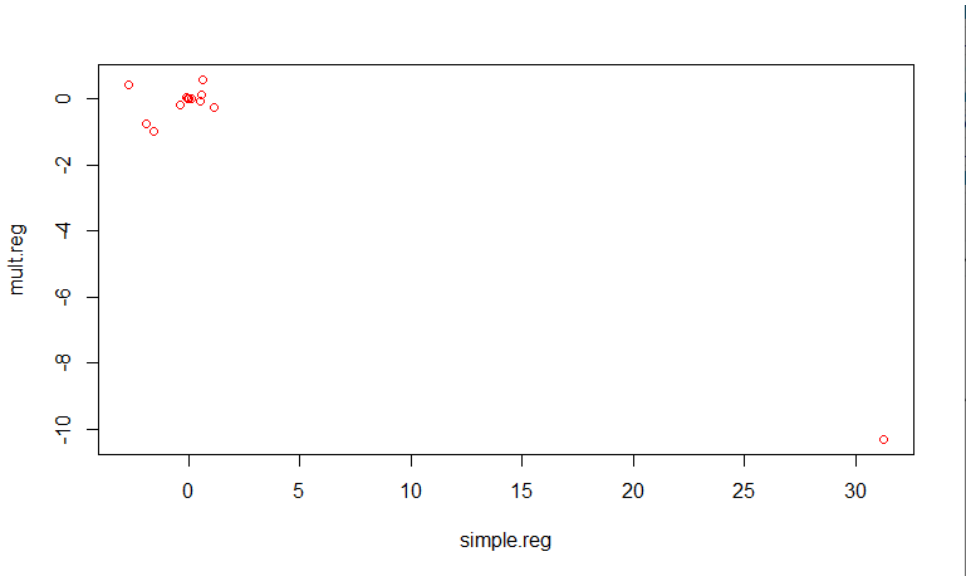
```
> fit.all <- lm(crim ~ ., data = Boston)
> summary(fit.all)

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
```

- We may reject the null hypothesis for all variables that have a **|t| value >2.** Therefore, we may reject the null hypothesis for the variables
    - zn
    - dis
    - rad
    - black
    - medv

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 on 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.

Solution:

- There is a difference between the simple and multiple regression coefficients.
    - In simple regression case, the slope term represents the average effect of an increase in the predictor, ignoring other predictors.
    - In contrast, in the multiple regression case, the slope term represents the average effect of an increase in the predictor, while holding other predictors fixed.



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 $Y=\beta 0+\beta 1X+\beta 2X2+\beta 3X3+\varepsilon$.

Solution:

- The below model is run for every Predictor variable against crim.

```
> fit.medv2 <- lm(crim ~ poly(medv, 3))
> summary(fit.medv2)

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 ***
---
```

- The following table summarizes the results from each model being fit.

	Linear Coeff	Quadratic Coeff	Cubic Coeff
Dependent variable	t-value	t-value	t-value
zn	-4.7	2.85	-1.2
indus	10.58	-3.28	-7.29
chas	NA	NA	NA
nox	11.249	-3.98	-8.34
rm	-5.088	3.19	-0.662
age	8.697	4.78	2.794
dis	-10.01	7.69	-5.814
rad	18.093	2.62	0.703
tax	16.436	4.68	-1.16
ptratio	6.9	3.05	-2.74
black	-9.35	0.74	0.6
lstat	11.543	2.08	-1.15
median	-11.42	13.41	-7.312

- The table below shows whether there is/isn't evidence of a non-linear relationship between the per capita crime rate and the predictors. All variables with a | t | values less than 2 are marked in red.
- Please Note: We do not run the non-linear model for chas as it is a binary variable, and receive er ror **'degree' must be less than number of unique points**

Chapter 6 | Question 9

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

Solution:

- The College data consists of 777 observations and 18 variables. 75% of the data (543 observations) were split into the train data set and 25% of the data (234 observations) were split into the test data set.

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

Solution:

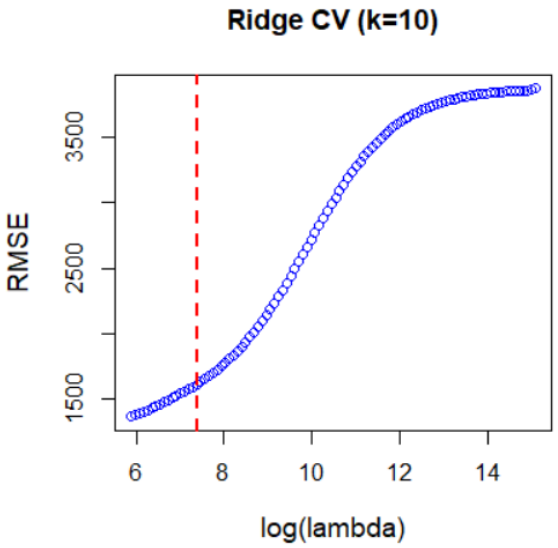
- The resultant test error obtained after fitting the linear model using least squares on the training set is: **1548033**

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

Solution:

- The value of λ chosen by cross validation for ridge regression is: **910**
- The test error obtained is: **3934873**

```
> ridge_coef
18 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -2.461715e+03
PrivateYes  -5.122564e+02
Accept       6.503878e-01
Enroll       7.447330e-01
Top10perc    1.802371e+01
Top25perc    6.812102e+00
F.Undergrad  1.172724e-01
P.Undergrad  4.126669e-02
Outstate     2.368708e-03
Room.Board   2.152576e-01
Books        2.948032e-01
Personal     1.033914e-02
PhD          1.956731e+00
Terminal    -5.328236e-01
S.F.Ratio    1.544653e+01
perc.alumni -1.028975e+01
Expend       6.283698e-02
Grad.Rate    1.121700e+01
```

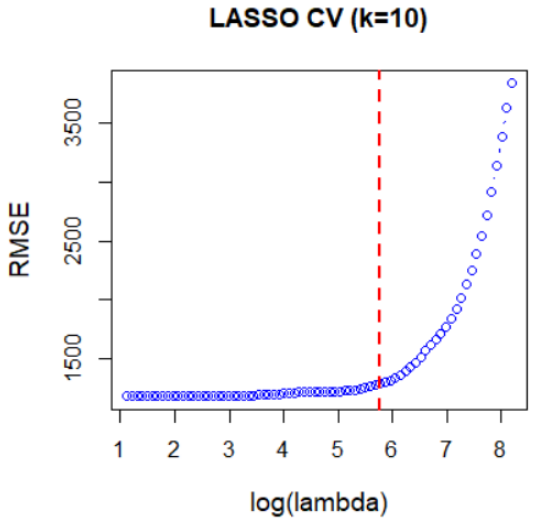


d). Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates

Solution:

- The value of λ chosen by cross validation for lasso model is: 183
- The test error obtained is: 1631729
- The number of non-zero coefficients are: 5

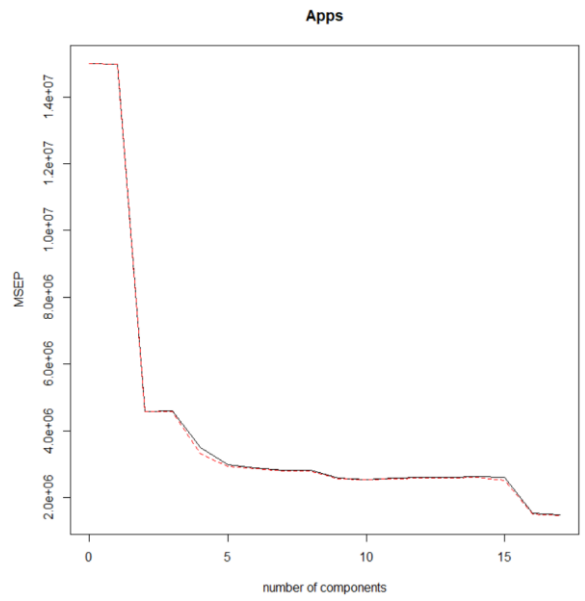
```
> lasso_coef
18 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) -426.25453561
PrivateYes     .
Accept         1.34875434
Enroll         .
Top10perc     20.17578100
Top25perc      .
F.Undergrad    .
P.Undergrad    .
Outstate       .
Room.Board     .
Books          .
Personal       .
PhD            .
Terminal       .
S.F.Ratio      .
perc.alumni    .
Expend         0.01397043
Grad.Rate      .
```



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

Solution:

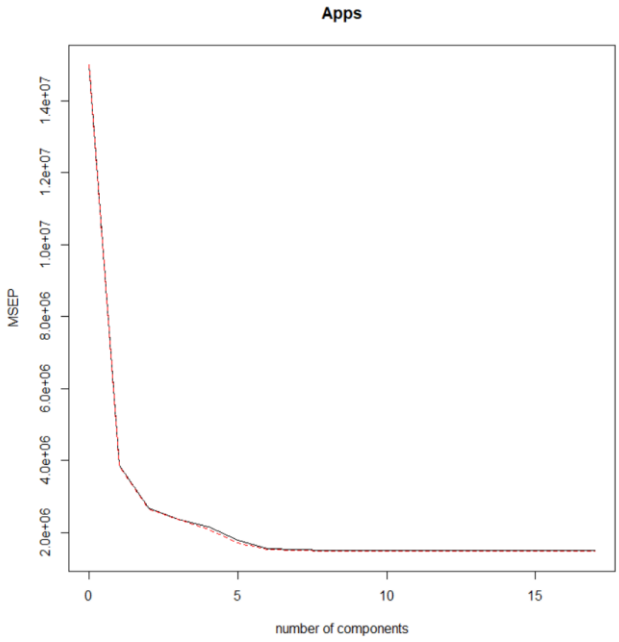
- The value of M chosen by cross validation for PCR is: 17
- The test error obtained is: 3908533



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

Solution:

- The value of M chosen by cross validation for PLS is: 6
- The test error obtained is: 1655849



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?

Solution:

- The model performances are as shown below. The linear model showed the least testing error **(1540833),** whereas Ridge showed the highest testing error **(3934873).**

Model	RMSE	R squared
Linear	1548033	0.9012
Lasso	1631729	0.901
PLS	1655849	
PCR	3908533	
Ridge	3934873	0.90

- The R squared value for all the models is equivalent to 0.9 and is quite similar. Which indicates that all of them can predict the number of college applications received accurately.

Chapter 6 | Question 11

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.

Solution:

- The model performances are as shown below. The Best subset selection showed the least testing error **(68.933),** whereas PCR showed the highest testing error **(715.372).**

Model	Cross-Validation RMSE
Best Subset selection	68.933
Linear model	69.187
Lasso	92.549
Ridge	96.985
PCR	715.372

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

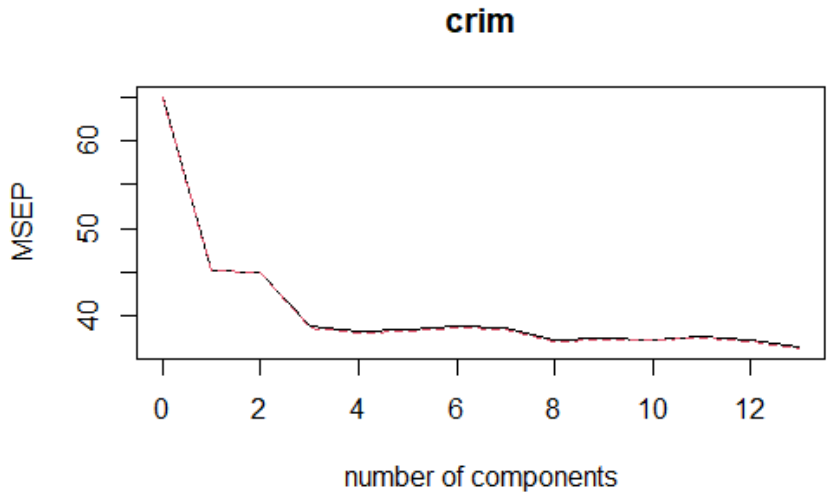
Solution:

- As indicated above, the best subset selection showed the least cross validation testing error of 68.933 and would be the best model for this data set. Alternatively, the Linear model can also be used.

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

Solution:

- The chosen model is best subset selection, and it does not include all the features. As shown in the plot below, choosing 13 variables gives us the least cross validation test error rate.



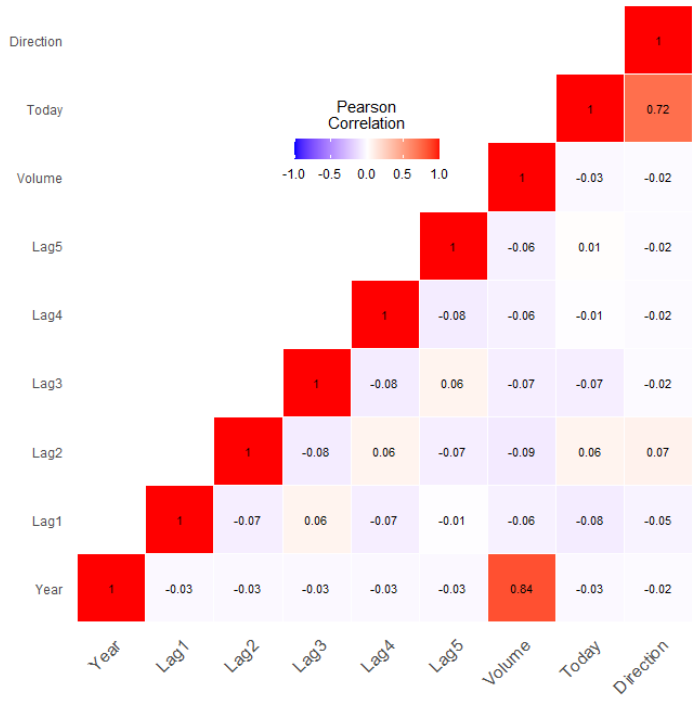
Chapter 4 | Question 10

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

Solution:

- The correlation matrix provides some insights about the Weekly Data.
    - Volume vs. Year has a high positive correlation (0.84)
    - Direction vs. Today has a high positive correlation (0.72)

  The correlation matrix does not provide any other insights that the other variables may be linearly related.



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?

Solution:

- Using the summary function, the only predictor that is statistically significant is Lag 2. The values of Lag2 with a t-value of 2.175 and a p-value of 0.02986

```
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
```

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

Solution:

- The confusion matrix for the logistic regression model is as shown below

```
> table(Weekly$Direction, pred.testw1)
   pred.testw1
      0   1
  0  54 430
  1  48 557
```

- The percentage of correct predictions is **56%**

- The confusion matrix for the logistic regression model shows that it made a correct prediction **56%** of the times. However, if we separate the 0 (down) and 1(up) prediction, the model correctly predicts the up predictions **92%** of the times and correctly predicts the down predictions only **11%** of the times.

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).

Solution:

- The confusion matrix for the logistic regression model is as shown below

```
> table(pred.testw2, Direction[!train])

pred.testw2 Down Up
       Down    9  5
       Up     34 56
```

The percentage of correct predictions for the test data is **62.5%**

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

Solution:

- The confusion matrix for the knn model is as shown below

```
> table(Weekknn.pred,Direction[!train])

Weekknn.pred Down Up
       Down   21 30
       Up     22 31
```

The percentage of correct predictions for the test data is **50%**

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

Solution:

- The logistic regression model provides the best results with 62.5% correct predictions.

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

Solution:

- Interaction of Lag2 squared, gave the best accuracy rate of 62.5%
- Logistic regression had the best accuracy rate of 60%.
- For the knn models, K = 10 had the best accuracy rate of 55%

Chapter 8 | Question 8

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

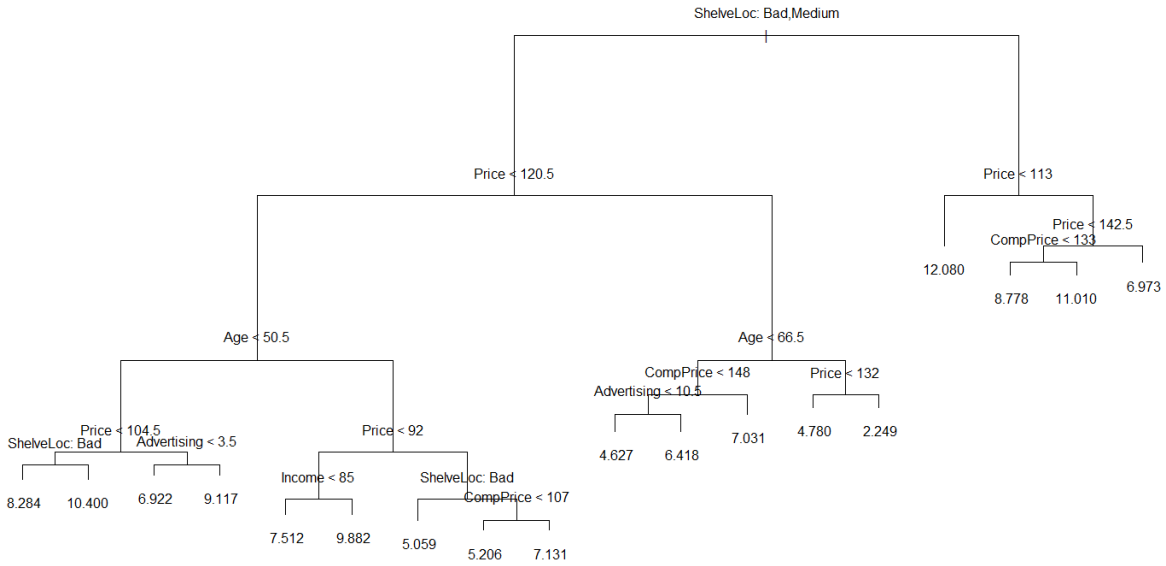
Solution:

- Data has been split in the ratio 50 : 50, Train: Test Ratio.

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

Solution:

The following is the representation of the tree



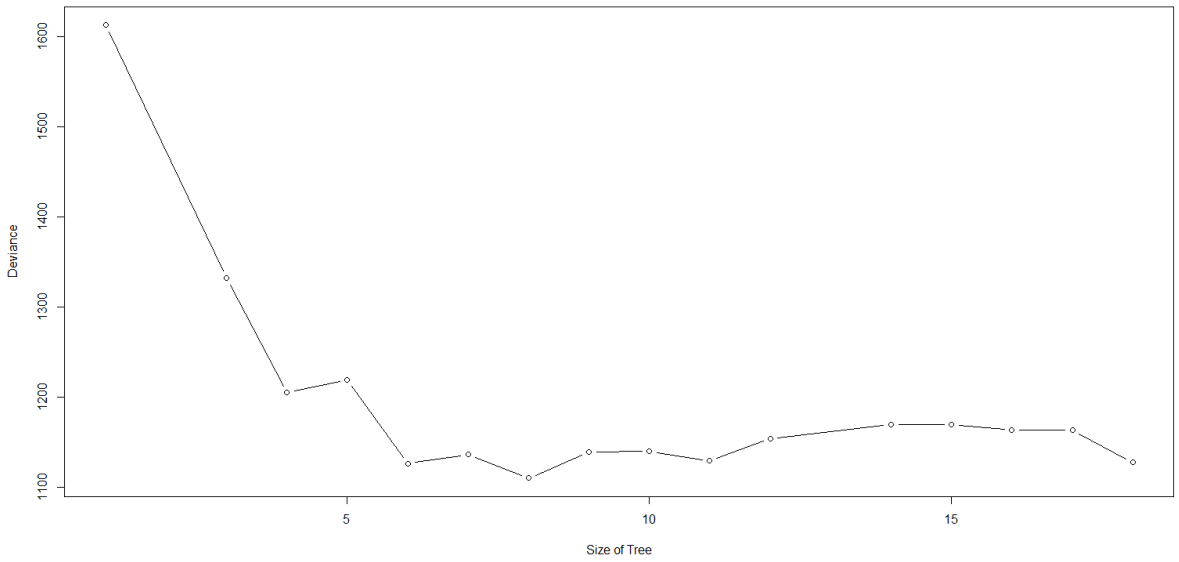
- We could infer from this that ShelveLoc and Price are the two most important factors in predicting car seat sales, since they appear at the top of the tree (because they provided the best split of the data).
- The tree has a total of 18 terminal nodes.
    - The OOS RMSE is : 4.148897

```
> summary(car.model.train)

Regression tree:
tree(formula = Sales ~ ., data = car.train)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Age"         "Advertising" "Income"
[6] "CompPrice"
Number of terminal nodes:  18
Residual mean deviance:  2.36 = 429.5 / 182
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.2570 -1.0360  0.1024  0.0000  0.9301  3.9130
> |
```

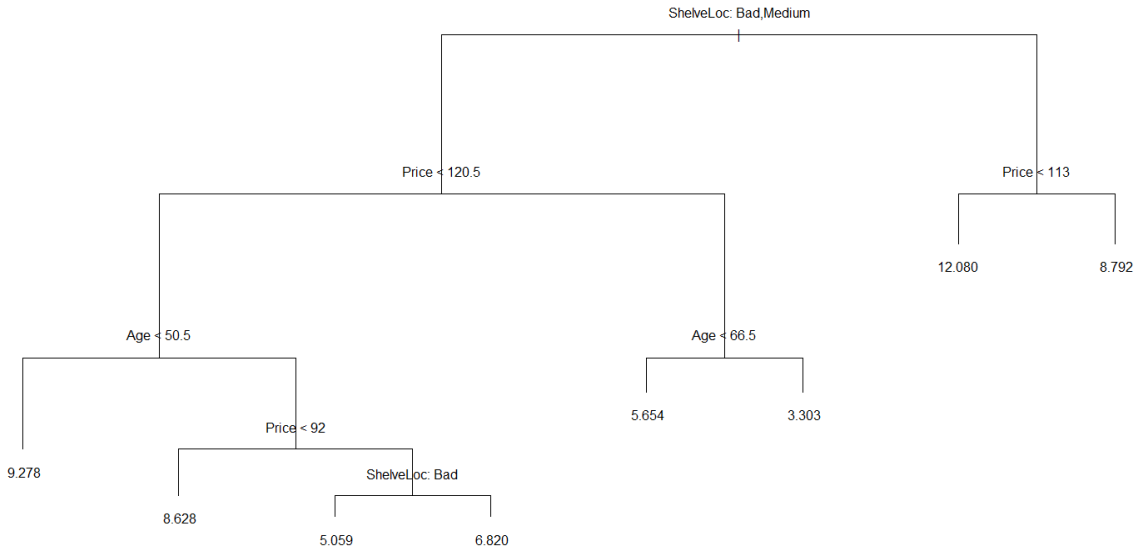
c). Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?

Solution:

- The following data frame below indicates the mean of cross-validation results (10fold) for n leaves. The selected model with the lowest cross-validation error appears to be the model with 17 terminal nodes:



- Best Deviance is found when Size of Tree = 8
- Pruning Tree : For size = 8.



```
> mean((prune.predict-car.test$Sales)^2)
[1] 5.09085
```

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

RMSE for out of Sample prediction using Bagging is : 2.368674

```
> mean((test_pred - test$Sales)^2)
[1] 2.368674
```

Result of Importance Function:

	varname	%IncMSE	IncNodePurity
1	Price	55.76266629	493.804969
2	ShelveLoc	53.87451311	446.816951
3	CompPrice	25.47984338	173.982449
4	Age	12.07366106	117.502364
5	Advertising	13.97464644	96.929928
6	Income	1.72616791	71.465227
7	Population	1.01449985	68.297498
8	Education	0.08382003	37.513944
9	Urban	-3.06299457	6.909530
10	US	0.14346468	5.985091

- The first measure, %IncMSE, is calculated by recording the prediction error of every tree on the OOB portion of the data (for regression, the MSE). A predictor is then scrambled (permuted) and the OOB MSE is then re-assessed. The difference between these MSE's is averaged across all trees (and normalized by the standard deviation of the differences).
- The second measure, IncNodePurity, is the total decrease in node impurity from splitting on that variable (averaged across all the trees). Node impurity is measured by Gini and RSS for classification and regression respectively.
- Whichever measure is used, we can see that for bagged trees, Price and ShelveLoc are the most important variables. This was also the conclusion from inspecting the top splits of the single regression tree produced in part b).

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 mm, the number of variables considered at each split, on the error rate obtained.

To run Random Forest, we run the model with various models of mtry from 1 to 10. That is all features included.

```
> data.frame(mtry = 1:10, test_MSE = test_MSE)
   mtry test_MSE
1     1 4.879483
2     2 3.440958
3     3 2.966130
4     4 2.759282
5     5 2.628801
6     6 2.536844
7     7 2.446945
8     8 2.438825
9     9 2.395672
10   10 2.368674
```

We see that the best MSE is when we include all ten variables. Therefore the Best MSE is 2.368674.

```
+   arrange(desc(IncNodePurity))
        varname     %IncMSE IncNodePurity
1         Price 55.76266629    493.804969
2     ShelveLoc 53.87451311    446.816951
3     CompPrice 25.47984338    173.982449
4           Age 12.07366106    117.502364
5   Advertising 13.97464644     96.929928
6        Income  1.72616791     71.465227
7    Population  1.01449985     68.297498
8     Education  0.08382003     37.513944
9         Urban -3.06299457      6.909530
10           US  0.14346468      5.985091
> |
```

We see the same set of importance variables, as the data set is run for the same features as bagging

Chapter 8 | Question 11

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

Solution:

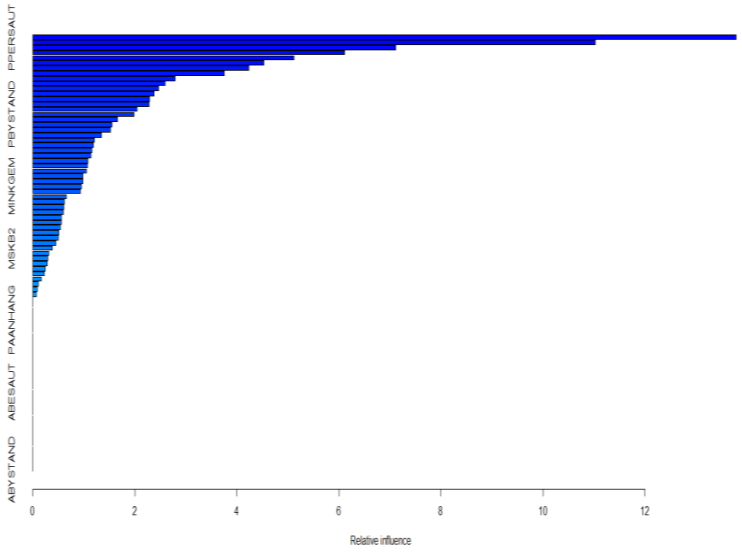
- There are 5822 records. We take the first 1000 records as sample.

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?

Solution:

- Most Important variables as per the Boosting Model. Table indicates top 10 Values.

	var	rel.inf
PPERSAUT	PPERSAUT	13.78940042
MKOOPKLA	MKOOPKLA	11.02790575
MOPLHOOG	MOPLHOOG	7.10896347
MBERMIDD	MBERMIDD	6.11219556
PBRAND	PBRAND	5.11377144
MGODGE	MGODGE	4.52422513
ABRAND	ABRAND	4.23046849
MINK3045	MINK3045	3.74847373
PWAPART	PWAPART	2.79375331
MOSTYPE	MOSTYPE	2.58918249



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 dataset?

Solution:

- Confusion Matrix for Boosting:

```
> table(Caravan.test$Purchase, pred.test1)
   pred.test1
      0    1
0 4494   39
1  276   13
```

Note : We have decided to compare values for logistic regression instead of KNN, as logistic regression is better suited towards classification in the presence of categorical independent variables.

- Confusion Matrix for Logistic Regression:

```
> table(Caravan.test$Purchase, pred.test2)
   pred.test2
      0    1
0 4183  350
1  231   58
```

Conclusion : Boosting is a better model, as it gives us a prediction accuracy of 93% as opposed to the logistic regression model that gives an accuracy of 87%.

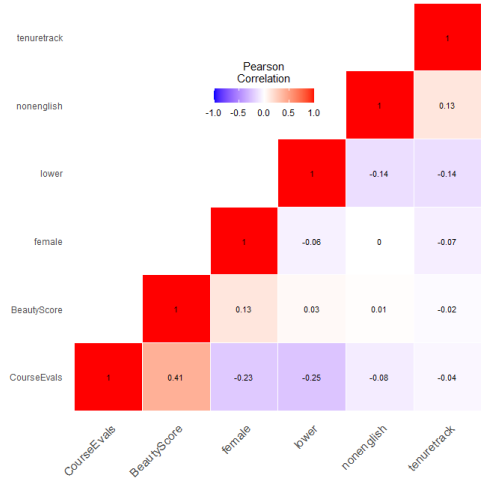
Problem 1

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

Solution:

- A three-step analysis is performed to estimate the effect of beauty into course ratings.
  - Linear Correlation
  - Multiple Linear Regression
  - Stepwise feature selection (Forward and Backward)

- **Linear Correlation:** If we look at the correlation between each of the independent variables, without taking the interactions of these variables among each other into consideration, we will get a correlation matrix as shown below.



- This matrix shows that Beauty score is positively co-related to CourseEvals, all the other variables being constant.

- **Multiple Linear Regression:** We need to perform a multiple linear regression to confirm the causation between the variables. The summary from MLR as shown below, proves that CourseEvals has some dependency on BeautyScore, Female, Lower and NonEnglish

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.293890   0.006493 199.267  < 2e-16 ***
BeautyScore  0.067964   0.006680  10.174  < 2e-16 ***
female      -0.047583   0.006643  -7.163 4.83e-12 ***
lower       -0.038794   0.006715  -5.777 1.70e-08 ***
nonenglish  -0.021448   0.006644  -3.228  0.00137 **
tenuretrack -0.007230   0.006708  -1.078  0.28189
```

- Stepwise Feature Selection: To further confirm the hypothesis, and to establish which are the most important variables, we perform a forward and backward selection of the variables.
    - Forward Selection : The forward selection shows the least AIC for a combination of BeautyScore + Female + Lower + NonEnglish

      ```
      Step:  AIC=-1451.38
      Y ~ BeautyScore + female + lower + nonenglish
      ```

    - Backward Selection : The backward selection shows the least AIC for a combination of BeautyScore + female + lower + lower.nonenglish

      ```
      Step:  AIC=-1450.36
      Y ~ BeautyScore + female + lower + lower.nonenglish
      ```

b). 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?

Solution:

- The problem to be addressed here is, are beautiful people actually more productive (better teachers in this case), or are they being considered better teachers because of their looks? The data provided here can’t really answer this question.
- Had the survey been conducted after blindfolding the students or making sure the physical appearances of the professors were not disclosed to the students before getting their inputs, then the outcome of the results would give us a clear picture of whether this represents productivity or discrimination.
- But currently, it has hard to separate the two and come up with a conclusive answer just based on the results of our analysis for this dataset.

Problem 2

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

Solution:

- Brick houses is very important as it has a t value of 8.729. Indicating that coefficient would be zero only for points occurring 8.7 Standard Errors away from the Estimate 17.297, which is less than a probability of 1.78*e^-14. **This means that there is a premium for brick houses everything else being equal.**

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.159      8.878   0.243  0.80823
dn2                 -1.561      2.397  -0.651  0.51621
dn3                 20.681      3.149   6.568 1.38e-09 ***
BR                  17.297      1.982   8.729 1.78e-14 ***
SqFt                52.994      5.734   9.242 1.10e-15 ***
MidCity.Bedrooms     4.247      1.598   2.658  0.00894 **
MidCity.Bathrooms    7.883      2.117   3.724  0.00030 ***
MidCity.Offers      -8.267      1.085  -7.621 6.47e-12 ***
```

b). Is there a premium for houses in neighborhood 3?

Solution:

- Neighborhood 3 is very important as it has a t value of 6.568. Indicating that coefficient would be zero only for points occurring 6.568 Standard Errors away from the Estimate 20.681, which is less than a probability of 1.38*e^-09. **This means that there is a premium for houses in Neighborhood 3 everything else being equal.**

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.159      8.878   0.243  0.80823
dn2                 -1.561      2.397  -0.651  0.51621
dn3                 20.681      3.149   6.568 1.38e-09 ***
BR                  17.297      1.982   8.729 1.78e-14 ***
SqFt                52.994      5.734   9.242 1.10e-15 ***
MidCity.Bedrooms     4.247      1.598   2.658  0.00894 **
MidCity.Bathrooms    7.883      2.117   3.724  0.00030 ***
MidCity.Offers      -8.267      1.085  -7.621 6.47e-12 ***
```

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

Solution:

- The interaction between houses in neighborhood 3 and brick houses is fairly important as it has a t value of 2.444. Indicating that coefficient would be zero only for points occurring 2.444 Standard Errors away from the Estimate 10.182, which is less than a probability of 0.01598. **This**

means that there could be an extra premium for brick houses in Neighborhood 3 everything else being equal.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.010      8.706   0.346  0.73016
dn2                 -0.673      2.377  -0.283  0.77751
dn3                 17.241      3.391   5.084 1.39e-06 ***
BR                  13.826      2.406   5.748 7.11e-08 ***
SqFt                54.065      5.636   9.593  < 2e-16 ***
MidCity$Bedrooms     4.718      1.578   2.991  0.00338 **
MidCity$Bathrooms    6.463      2.154   3.000  0.00329 **
MidCity$Offers      -8.401      1.064  -7.893 1.62e-12 ***
dn3:BR              10.182      4.165   2.444  0.01598 *
```

d). For the purpose of prediction could you combine the neighborhoods 1 and 2 into a single older neighborhood?

Solution:

- The summaries provided below indicate that houses in Neighborhood 2 and Neighborhood 1 don't have much individual importance, so they can be combined into a single neighborhood.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.159      8.878   0.243  0.80823
dn2                 -1.561      2.397  -0.651  0.51621
dn3                 20.681      3.149   6.568 1.38e-09 ***
BR                  17.297      1.982   8.729 1.78e-14 ***
SqFt                52.994      5.734   9.242 1.10e-15 ***
MidCity.Bedrooms     4.247      1.598   2.658  0.00894 **
MidCity.Bathrooms    7.883      2.117   3.724  0.00030 ***
MidCity.Offers      -8.267      1.085  -7.621 6.47e-12 ***

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5989     9.5522   0.063  0.95011
dn1                 1.5606     2.3968   0.651  0.51621
dn3                22.2416     2.5318   8.785 1.32e-14 ***
BR                 17.2973     1.9816   8.729 1.78e-14 ***
SqFt               52.9937     5.7342   9.242 1.10e-15 ***
MidCity.Bedrooms    4.2468     1.5979   2.658  0.00894 **
MidCity.Bathrooms   7.8833     2.1170   3.724  0.00030 ***
MidCity.Offers     -8.2675     1.0848  -7.621 6.47e-12 ***
---
```

Problem 3

a). 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)

Solution:

- By running a regression of “Crime” on “Police”, we cannot understand how more cops in the streets affect crime, since this would only explain correlation between the two and not the causation. To understand the causation, we would also need to look at other contributing factors.

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

Solution:

- To isolate the effect, the researchers looked at a scenario where the number of police in a city is independent of the current crime rate of the city. They took days of high terrorism alert as an independent event from crime in the city.
- To ensure that there is no relationship between high terrorism and reduced crime rate, they also considered that high terrorism alert causes reduced number of tourists which further lowers crime.
- The result of their analysis reveals that at the 5% level, there is a statistically significant negative relationship between high alert and crime rate.
- In the second model, the researchers validated that given a fixed level of metro ridership, the relationship between crime rate and high alert continues to be inverse and significant. Moreover, there is a positive relationship between midday ridership and crime.
- This makes it important to evaluate the relationship between high alert and mid-day ridership, as done by the researchers

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

Solution:

- Data on police and crime cannot tell the difference between more police officers leading to crime or more crime leading to increase in police officers.
- What would be useful is to randomly place cops on the street of a city on different days and see what happens to crime.
- The researchers at UPENN found a different method to approach this. They were able to collect data on crime in DC and relate that to days in which there was a higher alert for potential terrorist attacks. The decision has nothing to do with crime so it works essentially as an experiment.
- From table 1 we see that controlling for ridership in the METRO, days with a high alert (this was a dummy variable) have lower crime as the coefficient is negative. Why do we need to control for ridership in the subway? Well, if people were not out and about during the high alert days there would be fewer opportunities for crime and hence less crime (police not having any effect on the same). The result from the table tells us that holding metro ridership fixed, police and crime are

negatively correlated. However, we still, we can't for sure prove that more cops lead to less crime.

d). 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?

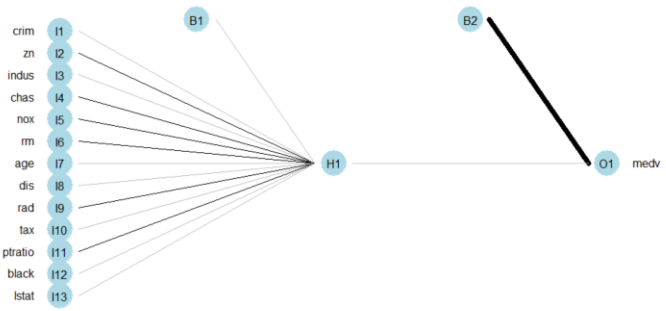
Solution:

- In table 4 they just refined the analysis a little further to check whether or not the effect of high alert days on crime was the same in all areas of town. Using interactions between location and high alert days they found that the effect is only clear in district 1. Again, this makes a lot of sense as most of the potential terrorists targets in DC are in District 1 and that's where more cops are most likely deployed to. The effect in the

Problem 4

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

Solution:



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

Problem 5

Describe your contribution to the final group project

Our final project was based on a classification problem. We tried to the determine the probability of a person (given indicators of age, employment, salary etc.) opening a term deposit with a well-known Portuguese bank.

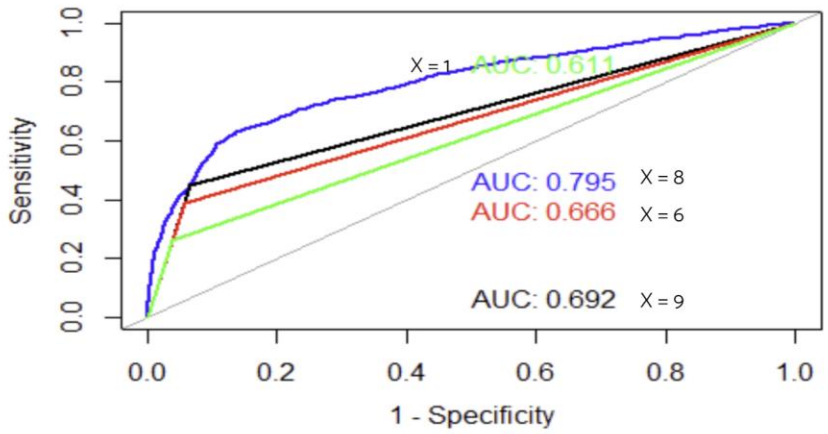
The three models we used to analyze this classification problem were: KNN, Random forests and Logistic regression

My contribution to the project was to run the Logistic Regression model for this classification problem. This involved a few steps of Data prep. I had to perform one hot encoding for our predictor variables, and a label encoding for the dependent variable. Since our dependent variable was extremely skewed (The ratio of Yes to No was 1:10), I also performed a stratified oversampling of the dependent variable to bring this ratio down to 1:4.

We used ROC curves, and the area under ROC curves (AUC) to determine which model would suit our problem statement the most.

Once all the data was in place, I ran the logistic regression analysis on our data set with all variables first. The AUC value for all variables was approximately ~0.602. In order to increase the efficiency, I tried a step-wise forward feature selection to determine the AUC for multiple combinations of the variables.

This analysis provided me with the below results:



This helped me gain an understanding that my maximum AUC was generated when I used 8 of the of the independent 19 variables that were provided to us. When I went further up to 9,10 and 11 variables, the AUC decreased, indicating that 8 variables would give me the maximum test AUC. The 8 variables were age + Euribor + employed + job + campaign + education + day_of_week + month.

For the combination of these 8 variables I calculated the Accuracy (87%), Precision (46%), Recall (47%) and F-score (47%) for the logistic regression model