Predictive Models Take Home Exam

Evan David

Carvalho STA S380

Book Problems

Chapter 2: #6

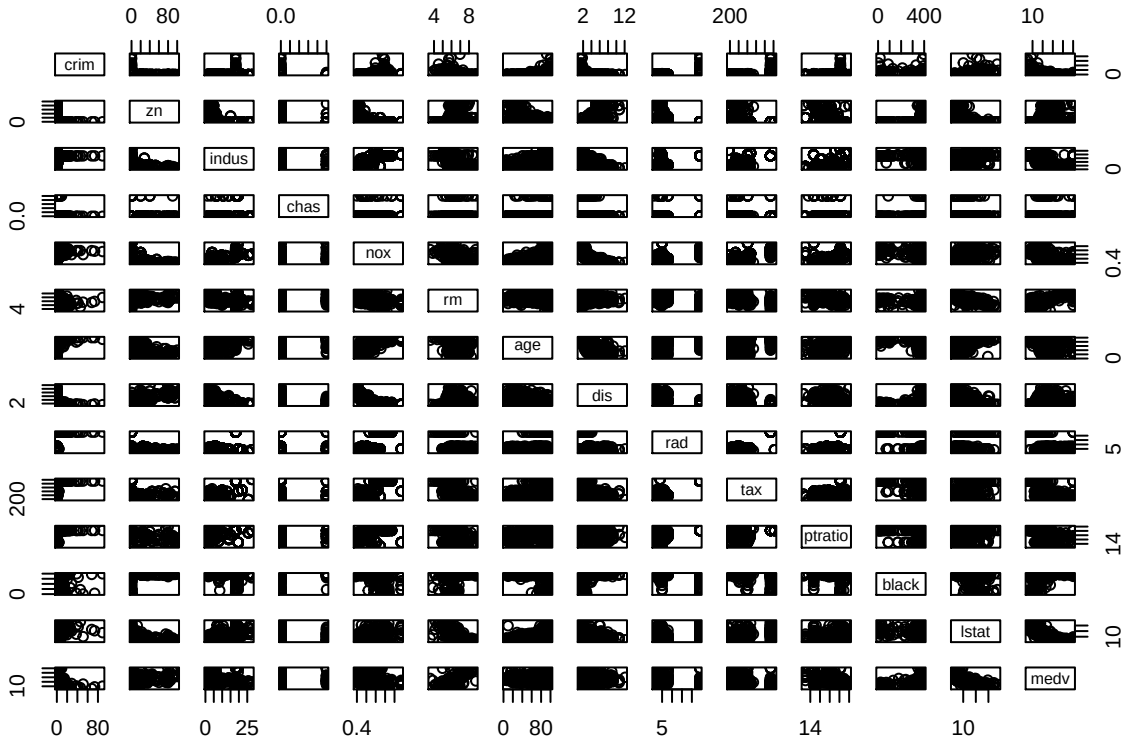
a)

Looking at the data

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

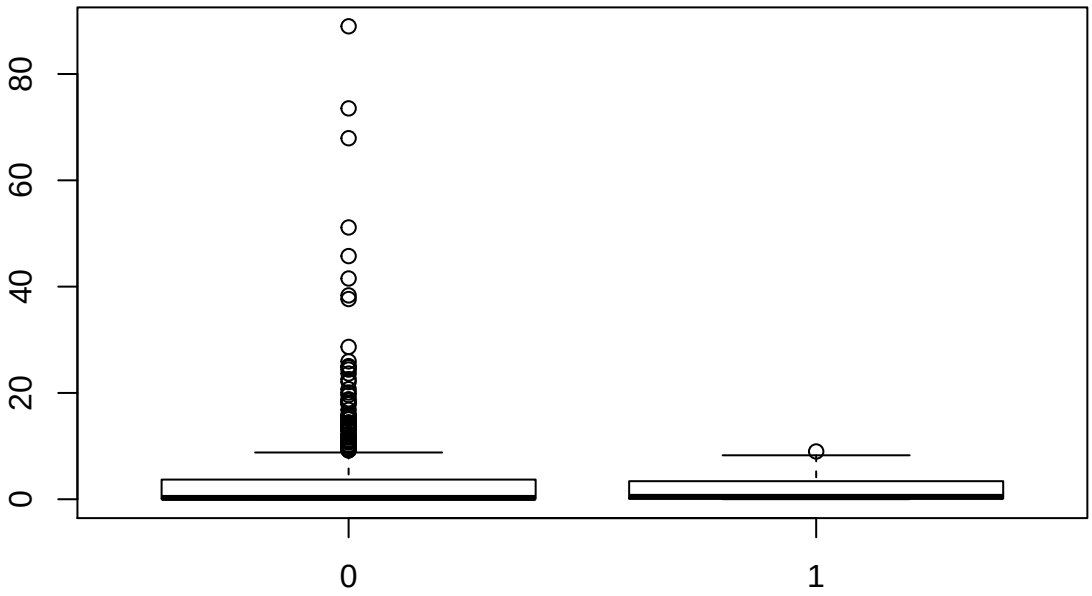
The data set has 506 rows and 14 columns. The rows represent each of the observations in this Boston housing data set, and the columns represent each of the predictor variables.

b)

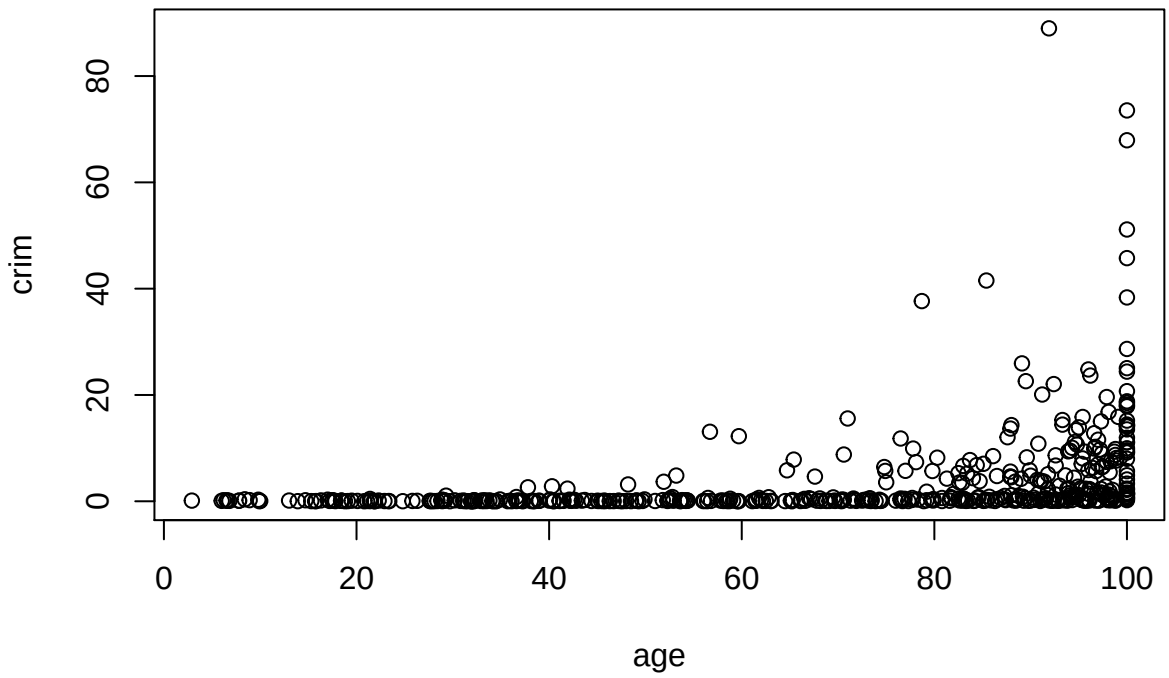


Based on the pairwise plots, there seems to be some relationship between some of the variables, such as between dis and nox, as well as between lstat and medv. It's tougher to tell from these plots what the relationships may be like for the other variables.

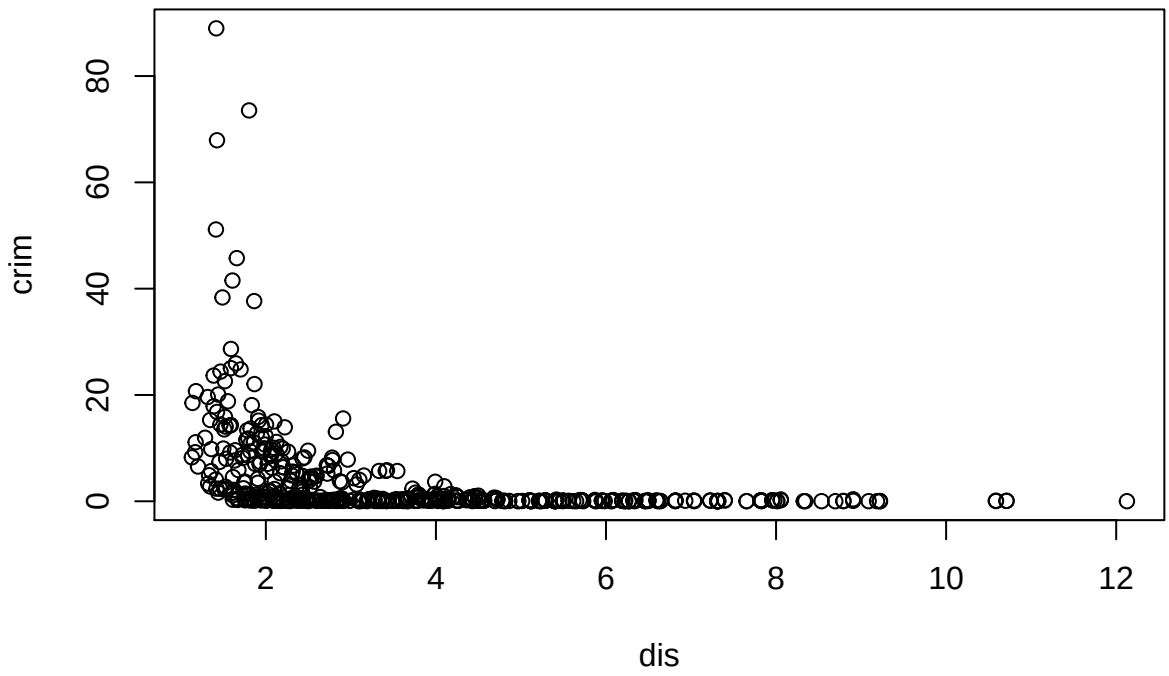
c)



There is a relationship between crim and chas - so it seems that there is more crime further away from the river.



There also seems to be a relationship between crim and age, so there is more crime in older neighborhoods.



Additionally, the relationship between crim and dis signifies that there is more crime closer to the five Boston employment centers.

d)

Some of the suburbs appear to have alarmingly high crime rates, compared to the average.

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
```

```
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

From the summary data we gather that there is a mean of about 4% crime rate, however the maximum is 89% so this range is very large, and in fact most of the suburbs have a low crime rate and just a few have very high crime rate. This is also true of the tax rate, though not as large of a separation as the average tax rate is 408 and the maximum is 711. It looks to be largely the same suburbs where the crime rate is very high as is the tax rate. The pupil-teacher ratios are much more evenly distributed and there doesn't seem to be a difference depending on the suburb, and the range of the ratios are small.

e)

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

Here we find that 35 suburbs bound the Charles River, as the chas variable is a dummy variable with 1 meaning that suburb bounds the river, and 0 meaning it doesn't. So, by summing up the observations we get the total number of suburbs bounding the Charles River.

f)

The median pupil-teacher ratio among the towns is 19.05, meaning there are about 19 pupils for every teacher. This can be gathered from the summary data of the ptratio variable.

g)

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

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9
##     lstat medv
## 399 30.59    5
```

The suburb with the lowest median value of owner-occupied homes is suburb #399, with a median value of $5,000. This area has a much higher crime rate than the average, 38% compared to the mean of 4%. There is no land zoned for lots over 25,000 sq. ft. There is a pretty high ratio of non-retail business in this area. This suburb does not bound the Charles River, and there is a pretty high concentration of nitrogen oxides here. There is about 1 fewer room in this suburb than the average, and these are also the oldest houses in the data set. The suburb is very close to Boston employment centers and has the highest index of accessibility to radial highways among the suburbs. The tax rate and pupil-teacher ratio are both very high here, and the proportion of blacks is not much different than the median. This suburb has a high percentage of people in the lower status of the population, more than double the average.

h)

There are 64 suburbs with an average of more than 7 rooms per dwelling, and 13 suburbs with an average of more than 8 rooms per dwelling. These suburbs in general have low crime rate, low tax rate, and very high

median value of houses. The proportion of non-retail business is low and the distance from the highways is high, indicating that these are residential areas. As would be expected these suburbs also have a low percentage of people in the lower status of the population.

Chapter 3: #15

a)

After fitting a linear regression model for each of the predictors with per capita crime rate as the response variable, it was found that all variables were statistically significant except for the chas variable.

```
##
## Call:
## lm(formula = crim ~ as.factor(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 ***
## as.factor(chas)1  -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
```

This is to say that for each predictor aside from chas, there was a p-value of less than 5%. There seems to be no significant relationship between per capita crime rate and whether a suburb bounds the Charles river.

b)

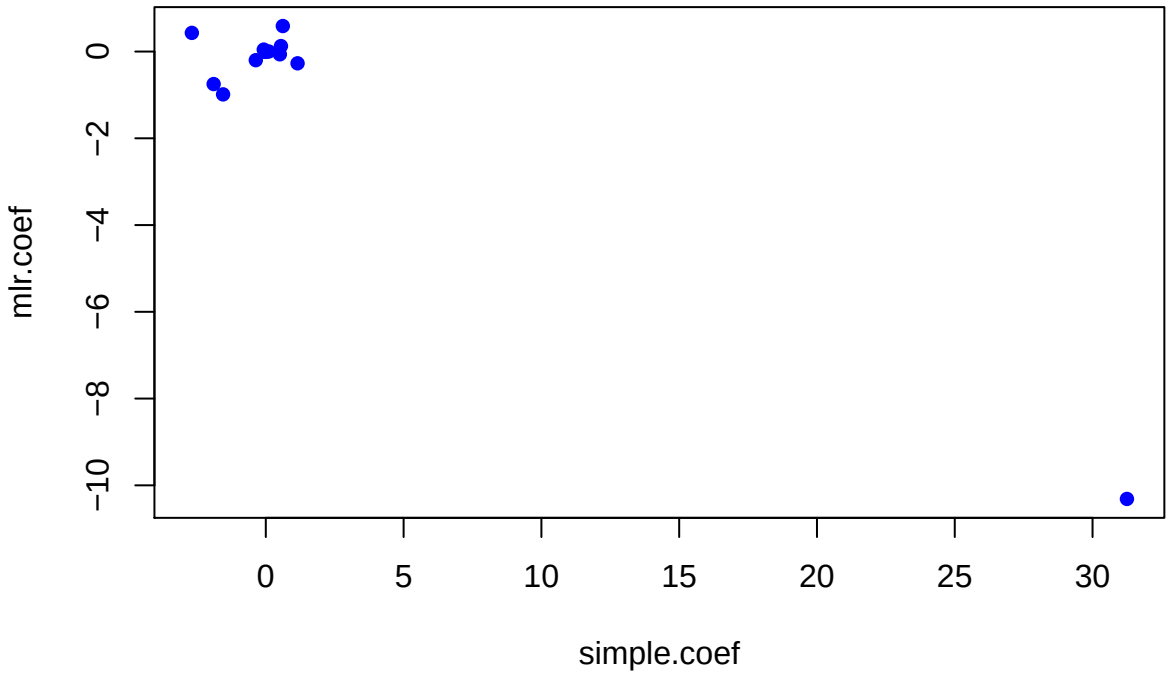
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855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

After fitting a multiple linear regression model using all predictors, the significant variables appear to be dis, rad, medv, black, and zn. This means we can reject the null hypothesis for these predictors because their p-value is less than 0.05 so they are statistically significant.

c)

These models show different results, and this can be accounted for by the fact that the multiple regression model shows the effect of a predictor holding all other variables fixed, and the simple linear model shows the relationship for only one variable on the response. Some variables may be correlated with each other, thus seeming to have a relationship in the simple linear model when in fact the effect of that relationship was due to another variable. An example of this is the effect of age, which is significant in the simple linear model but is really due to the effect of dis and that shows up in our multiple regression model. Here is a plot of the coefficients:



d)

After fitting a polynomial model for each of the predictors on the response, there is evidence of a non-linear association in all the variables except for the "black" variable. In the models for the variables indus, nox, age, dis, ptratio, and medv the cubic fit seems appropriate when looking at the p-values. For the variables zn, rm, rad, tax, and lstat, the coefficients for the cubic fit are not significant, but the quadratic fit seems appropriate. And again, for the "black" variable there is no statistical significance in either the quadratic or the cubic coefficients suggesting it doesn't have a non-linear relationship.

Chapter 6: #9

a)

Splitting the data into a 50/50 training and test set.

b)

Fitting the linear model and then testing the predictions, the test error or RMSE obtained is about 1215 in this case.

c)

Fitting a ridge regression model, the best lambda obtained by cross-validation is about 357 in this case, and the observed RMSE is about 1573, even higher than for least squares.

d)

Fitting a lasso model, the best lambda obtained is about 4, and the observed test error is about 1228, in between the least squares and the ridge test error. The number of non-zero coefficients in this case is 16 - all of the coefficients were kept, so none of them were zero.

e)

Fitting a PCR model, the RMSE is about 1735 - the highest test error of any of the models so far, and the value of M chosen by cross-validation is M=17.

f)

Fitting a PLS model, the RMSE is about 1220, and the value of M chosen by cross-validation is M=9.

g)

Most of the models will pretty accurately predict the number of applications received. PCR and ridge performed slightly worse than the other models based on their higher RMSE and lower values of R^2.

Chapter 6: #11

a)

I began by setting the seed in order for my results to be reproducible. I then tried different methods such as best subset selection, ridge regression, lasso, and PCR, and found that they all performed about the same, with ridge regression performing the best.

b)

The ridge regression performs the best of the models that were tried, as it gives an RMSE of 6.001 after cross-validation, the lowest test error of any model.

c)

The selected model does involve all the features in the data set, because ridge regression uses all the variables.

Chapter 4: #10

a)

Looking at the summary data:

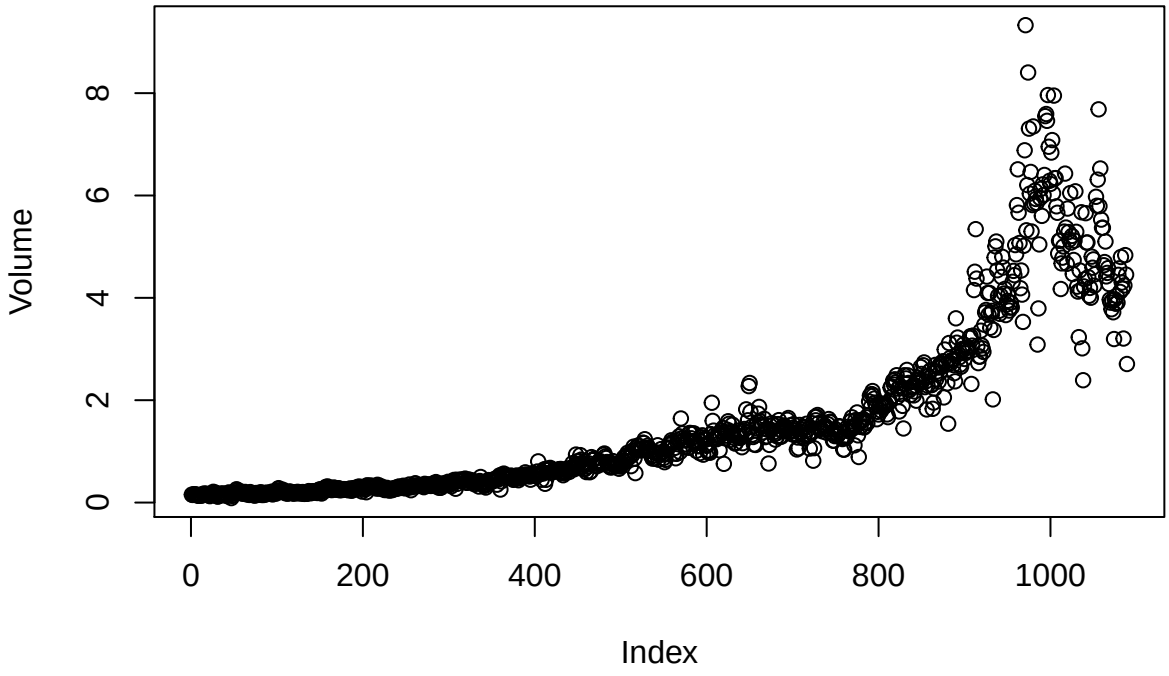
```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  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       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260
```

And some correlations:

```
##               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
```

Looking at volume over time:



Based on summary data and the correlations between the variables, it seems the only relationship is between Year and Volume, with the other lag variables having correlations close to zero. Looking into this relationship, it appears that Volume is increasing over time, perhaps in a non-linear fashion.

b)

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
```

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

Only the Lag2 predictor is statistically significant, with a p-value of 0.0296.

c)

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

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

```
## [1] 0.9206612
```

```
## [1] 0.1115702
```

From the confusion matrix, the prediction accuracy is about 56%, so a misclassification error rate of about 44%. It can also be seen that the model performs well on weeks when the market goes up, correctly predicting 92% of the time, but when the market goes down, the model only predicts correctly around 11% of the time.

d)

```
##          Direction.20092010
## pred.glm2 Down Up
##      Down    9  5
##      Up     34 56
```

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

```
## [1] 0.9180328
```

```
## [1] 0.2093023
```

From the confusion matrix, the prediction accuracy is 62.5% this time, so a misclassification error rate of 37.5%. Again, looking deeper into the confusion matrix, on weeks when the market goes up the model predicts correctly about 92% of the time, and on weeks when the market is down the model predicts correctly only about 21% of the time.

g)

```
##         Direction.20092010
## pred.knn Down Up
##     Down   21 30
##     Up     22 31
```

For the KNN model, we get a prediction accuracy of 50%.

h)

The best method is logistic regression using only Lag2 as a predictor variable. This method yielded the highest prediction accuracy.

i)

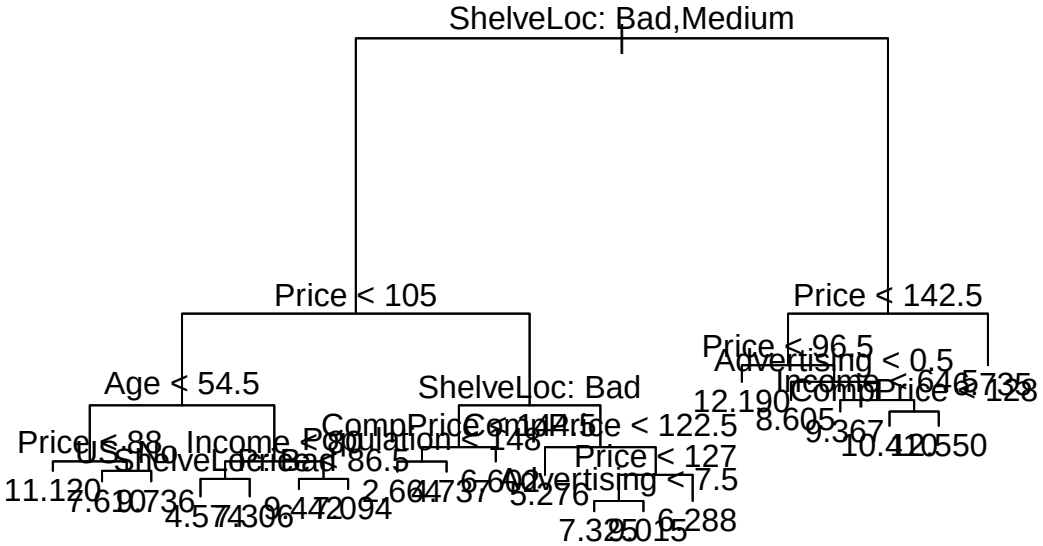
Trying different combinations for logistic regression, and different KNN models (k=10, k=100), the best model is still the original logistic regression from the previous answer as it yielded the best prediction accuracy by far.

Chapter 8: #8

a)

Splitting the data into a 50/50 train and test set.

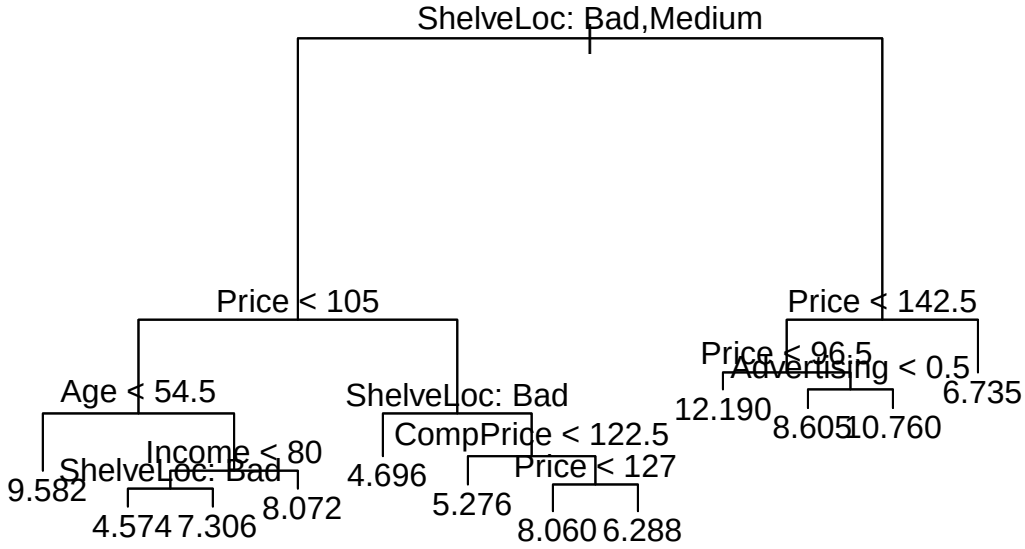
b)




```
## [1] 4.247095
```

The resulting regression tree is one with 17 terminal nodes, with 7 of the variables used in the tree. The test MSE from this tree is computed to be about 4.25.

c)




```
## [1] 4.173681
```

Cross-validation gives the optimal tree with 11 terminal nodes, and pruning the tree gives a test MSE of 4.17, so pruning the tree did improve the test MSE.

d)

```
##                %IncMSE IncNodePurity
## CompPrice   25.100299    133.415726
## Income       6.281301     79.807791
## Advertising 16.808168     96.016520
## Population  -1.384437     56.124138
## Price       49.076084    371.282172
## ShelveLoc   62.375209    514.722651
## Age         14.287146    126.222261
## Education    1.053640     37.892871
## Urban       -3.899254      5.346573
## US           6.188616     16.139124
```

```
## [1] 2.819432
```

Using the bagging approach, the test MSE improved to 2.8. The importance function shows ShelveLoc and Price to be the most important variables.

e)

```
##                %IncMSE IncNodePurity
## CompPrice   10.2296094     116.96860
## Income       1.2727783     107.94633
## Advertising 12.9818325     148.37184
## Population  -0.7297195      95.20080
## Price       29.4591017     300.70264
## ShelveLoc   41.7697381     377.38971
## Age         11.7890546     145.20989
## Education    0.7730010      63.06423
## Urban       -0.4370588      14.42387
## US           6.4437073      24.77031
```

```
## [1] 3.457287
```

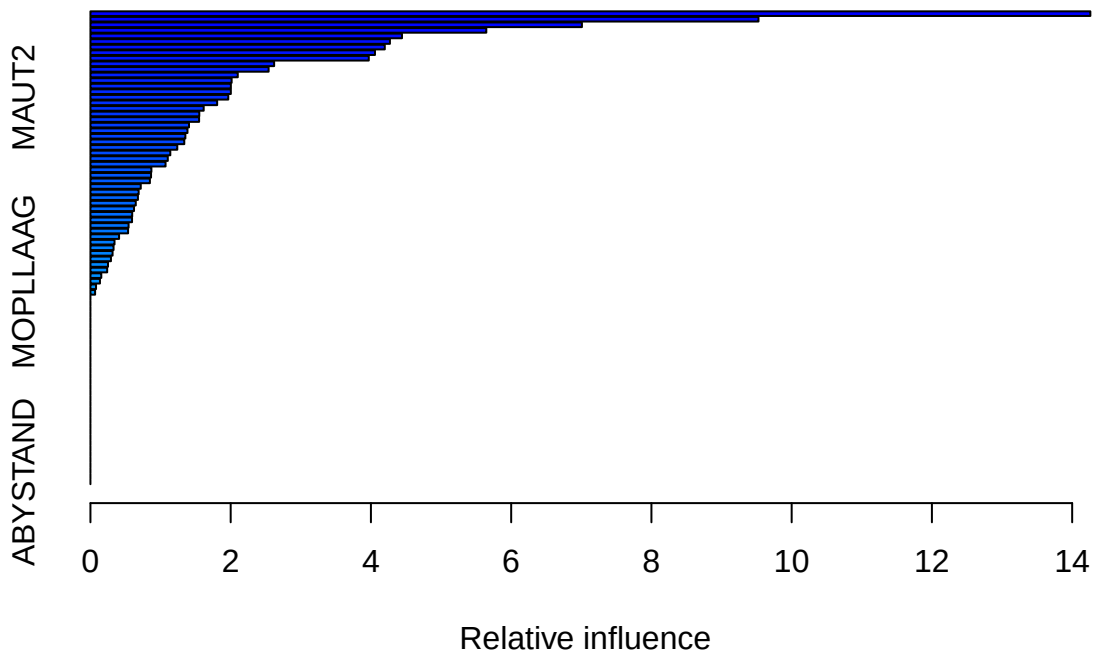
Using random forests to analyze the data, the test MSE is 3.45. The most important variables from this model are again the ShelveLoc and Price variables. The random forests model takes an m value of the square root of p, the number of predictors. So in the random forest, m=3, while in the bagging method m=10, and because the test MSE was better for bagging, it would appear that the higher values of m provide a lower test MSE in this case.

Chapter 8: #11

a)

Setting up the training set of the first 1000 observations and the test set of the remaining observations.

b)




```
##               var   rel.inf
## PPERSAUT PPERSAUT 14.259505
## MKOOPKLA MKOOPKLA  9.526890
## MOPLHOOG MOPLHOOG  7.008897
## MBERMIDD MBERMIDD  5.645143
## ABRAND     ABRAND  4.442858
```

After fitting a boosting model to the training set using 1000 trees and a shrinkage of 0.01, the most important variables appear to be PPERSAUT, MKOOPKLA, and MOPLHOOG.

c)

```
##    pred.test
##        0    1
##   0 4499   34
##   1  280    9
```

```
## [1] 0.2093023
```

For the boosting model, the fraction of people predicted to make a purchase that actually do make one is 0.209 or about 21%.

```
##    pred.test2
##        0    1
##   0 4499   34
```

```
##   1  280    9
```

```
## [1] 0.2093023
```

For logistic regression, the result is the same - we get that 21% of people predicted to make a purchase actually do.

Other Problems

Problem 1: Beauty Pays!

1.

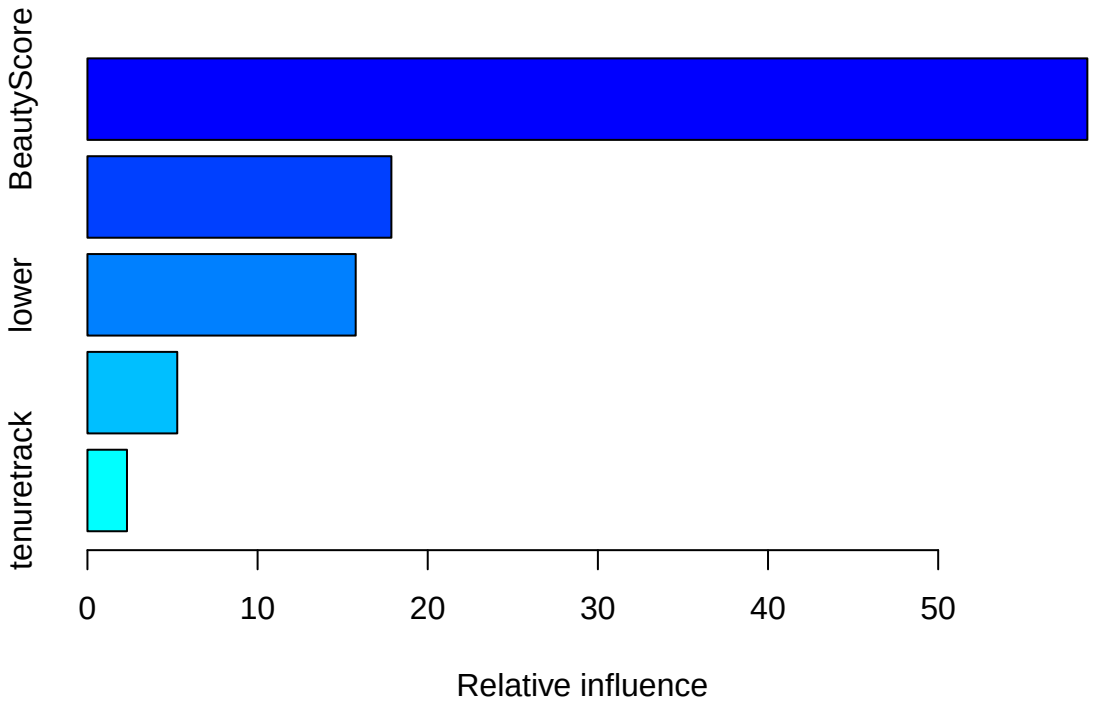
After looking at the data and applying different regression models to estimate the effect of "beauty" on course ratings, I conclude that beauty does play a significant role in determining course ratings. Doing some initial exploratory analysis, I looked at correlations between the variables, as well as looking into the relationship of the beauty score and the course evaluation score, and I found that the beauty score seemed to be fairly correlated with course evaluations. I wanted to look further into this relationship so I created some regression models. For all the models, I split the data using 60% for a training set and 40% for a test set. I first ran a couple of linear regression models, first using just BeautyScore as a predictor and then using all the variables in a multiple regression model. Additionally I tried out ridge regression, the lasso, classification trees, random forests, bagging, and boosting. Here are the results I obtained for out-of-sample RMSE from each of the models:

```
##                     RMSE
## Simple Linear 3.8476819
## Multiple      3.8610497
## Ridge         0.4342260
## Lasso         0.4342643
## Trees         0.4799599
## Bagging       0.4894225
## Random Forest 0.4590718
## Boosting      0.4368712
```

From the table it can be seen that the simple linear regression model and the multiple linear regression model performed poorly compared to the other models, so I ignored those models in my analysis, although it's worth noting that the multiple regression model signified CourseEvals, female, and lower to be statistically significant variables. The ridge regression model had the best RMSE, and looking into this model all of the variables are pretty equally significant aside from tenuretrack:

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept)  4.02206019
## (Intercept)  .
## BeautyScore  0.30417940
## female      -0.30360085
## lower       -0.31731715
## nonenglish  -0.35611246
## tenuretrack -0.06858246
```

Additionally, the lasso, which had an RMSE nearly identical to ridge, kept all variables and did not shrink any coefficients to zero. All of the random forest models, including bagging and boosting, showed the beauty score to be much more significant than the other variables, as shown by this result from a boosting model:




```
##                     var   rel.inf
## BeautyScore BeautyScore 58.767124
## female           female 17.864210
## lower             lower 15.766177
## nonenglish   nonenglish  5.278442
## tenuretrack tenuretrack  2.324046
```

I conclude that "beauty" has a very significant impact on the course evaluations that professors get.

2.

When thinking about the results of this analysis, it's important to consider the potential pitfalls. You can't necessarily say that beauty is a direct cause of higher course evaluations, because there are many other factors at play. Even though my analysis considers other determinants, it does not consider ALL of the other determinants. There are many characteristics that professors may have that could lead to better evaluations such as how well liked they are, how well they actually teach the material, etc. Additionally, the evaluation made by the student can vary greatly in what method that student may choose. A student may evaluate a professor on how well they taught, how well it was received by the student, whether they liked the professor, or even beauty may play some role in a particular student's evaluation. Whether that student did well in the course could also impact their evaluation - so there can be many factors involved in the course evaluations. When thinking about beauty as a predictor, it may be correlated with other factors and could be a proxy for another variable that is a true predictor of better evaluations. So this correlation does not imply causation. In that respect it cannot be directly concluded that there is any discrimination of professors based on beauty, with so many other factors at play.

Problem 2: Housing Price Structure

1.

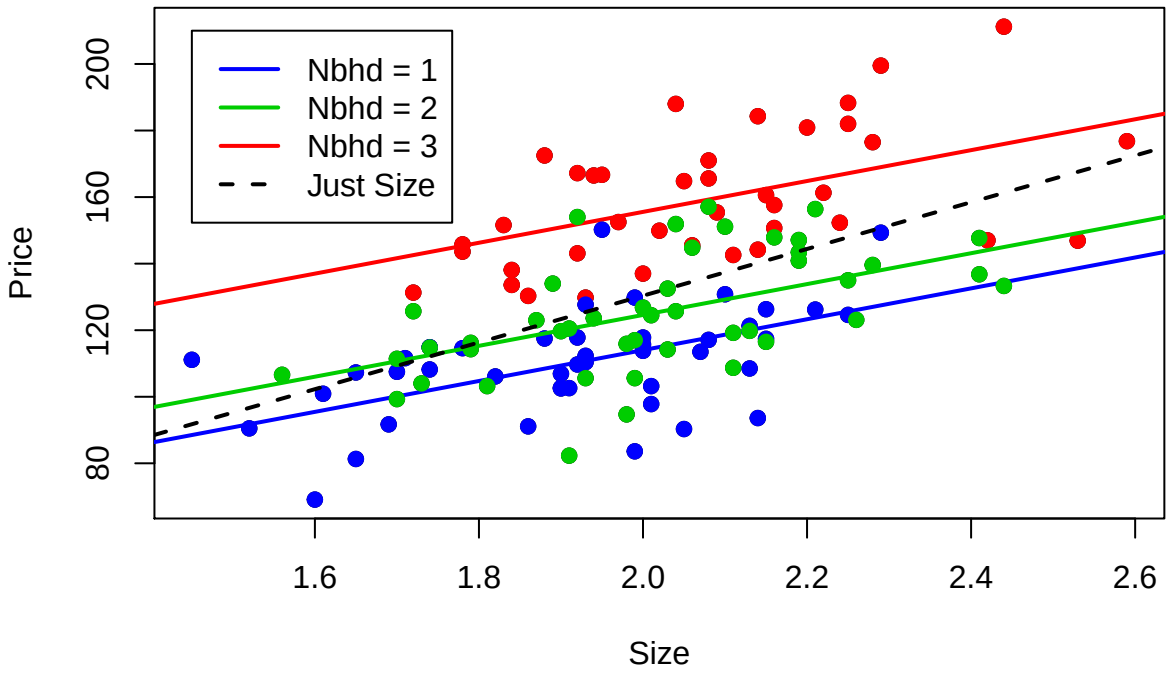
All else being equal, there is a premium for brick houses among this population. Based on a multiple regression model considering all the variables, a house that is brick will cause the price of the house to increase versus a house that is not brick, if you ignore all of the other variables. Here is a simple linear regression showing this point:

```
##
## Call:
## lm(formula = Price ~ ., data = midcity)
##
## Coefficients:
## (Intercept)         Home         Nbhd       Offers         SqFt
##   -9814.663        6.187     9832.281    -8351.794       49.811
##    BrickYes     Bedrooms    Bathrooms
##   15601.818     5671.911     8243.545
```

Holding all other variables equal, having a brick house or not would seem to have the most impact on the price of the house due to it having the largest magnitude of coefficient.

2.

There is a premium for houses in neighborhood 3, as evidenced by the regression models. Here is an example I have recreated from the class lecture demonstrating the impact of houses in different neighborhoods:



It's clearly seen from this that a house being in neighborhood 3 has a much higher effect on the price of the house than a house in neighborhood 1 or 2.

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt + Brick)
##
## Coefficients:
## (Intercept)        Nbhd2        Nbhd3         SqFt     BrickYes
##      18.725        5.556       36.770       46.109       19.152
```

This is true across all different variations of models.

3.

There is not an extra premium for brick houses in neighborhood 3.

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt + Nbhd:Brick)
##
## Coefficients:
##    (Intercept)           Nbhd2           Nbhd3            SqFt
##         20.736           5.821          33.023          45.562
## Nbhd1:BrickYes  Nbhd2:BrickYes  Nbhd3:BrickYes
##         13.107          16.374          26.160
```

It can be seen from the analysis that a brick house in neighborhood 3 may have a higher positive impact on price than for brick houses in the other neighborhoods, but it has a lesser impact when looking at just houses in neighborhood 3 in general. So from this we can say that because the houses area already in neighborhood 3, the affect on their price is probably mostly going to come from the neighborhood they are in rather than if the house is brick or not.

4.

You could combine neighborhoods 1 and 2 into a single "older" neighborhood because it is clear from the data that houses being in neighborhood 3, the newer and more moder neighborhood, have a much bigger impact on their price than if houses are in neighborhoods 1 or 2. Additionally there isn't much difference in the effect of being in neighborhood 1 versus neighborhood 2, so you could combine these into a single neighborhood to better see the effect of the newer versus older neighborhoods.

Problem 3: What causes what?

1.

You can't just get the data from a few different cities to figure out the effect of the number of cops in the street affecting crime because you are unable to hold other factors equal when you do this. For example, a small town in east Texas may have a very low crime rate, and probably very few cops, but a city like Chicago inherently has a high crime rate and thus is going to have more cops. So it's understandable that just adding or reducing the number of cops for a city isn't going to directly affect the crime rate in that city.

2.

The researchers from UPENN were able to isolate this effect by setting up a controlled experiment. They did this by collecting data for a situation in which the number of police would be unrelated to the amount of crime going on. By doing this, the researchers essentially hold other variables fixed while changing the amount of cops, and seeing the effect on the crime rate. This works because the number of cops will be changed based on what day it is, whether it was a high alert day for terrorism or not, and not based on the level of crime. The results show that on high alert days, the level of crime actually did decrease.

3.

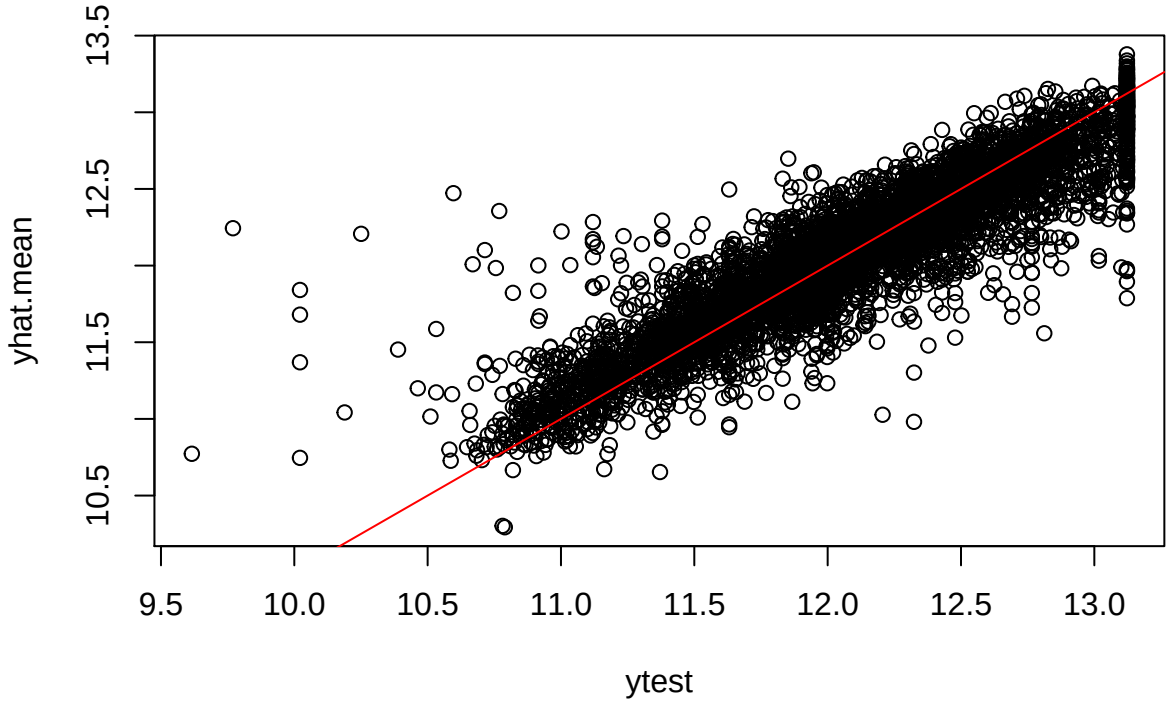
The researchers had to account for the level of METRO ridership because it could have been that on high alert days, fewer passenger(i.e. potential victims of crime) would want to go out and ride the METRO. They wanted to make sure that the ratio of potential victims stayed constant because that could mess up the experiment. It turns out that the METRO ridership level was unchanged on these high alert days from normal days, so the researchers were able to control this variable.

4.

They are estimating a simple linear regression model, with effect of the high alert variable on the crime rate being examined. The conclusion is that on days of high alert, the crime rate decreased, as evidenced by the negative coefficient for the "High Alert" variable. The R^2 value is less important to consider in this table, and it is more sensible to examine the standard error given. With the standard error considered, the coefficient would be negative even for two standard deviations, or 95% confidence. So the researchers would be at least 95% confident that crime rate went down on the high alert days, when more police were present on the streets.

Problem 4: BART

After running a BART model on the California Housing dataset, it can be concluded that BART does not perform as well as random forests or boosting. From class, it was seen that boosting had a minimum out-of-sample loss, or RMSE, of 0.231, and random forests had a minimum out-of-sample RMSE of 0.233. Here is the plot of predictions vs validation data for BART:



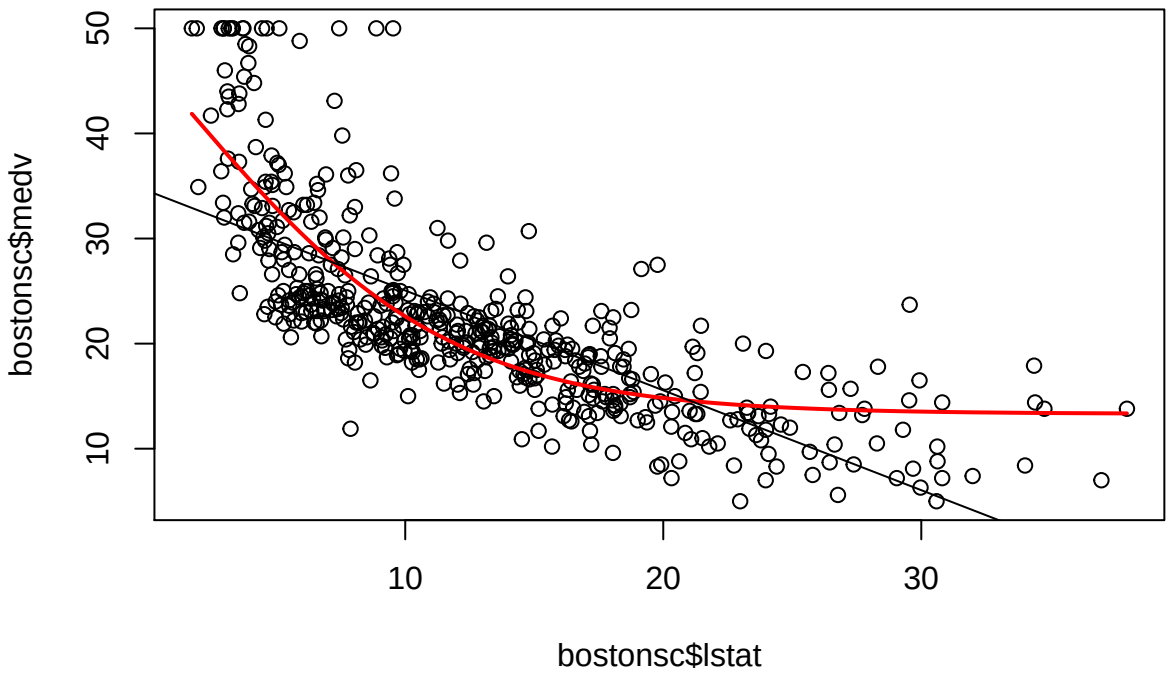
And the RMSE:

```
## [1] 0.7783646
```

The BART model gives an RMSE of 0.778, so it does not outperform either the random forests or boosting methods.

Problem 5: Neural Nets

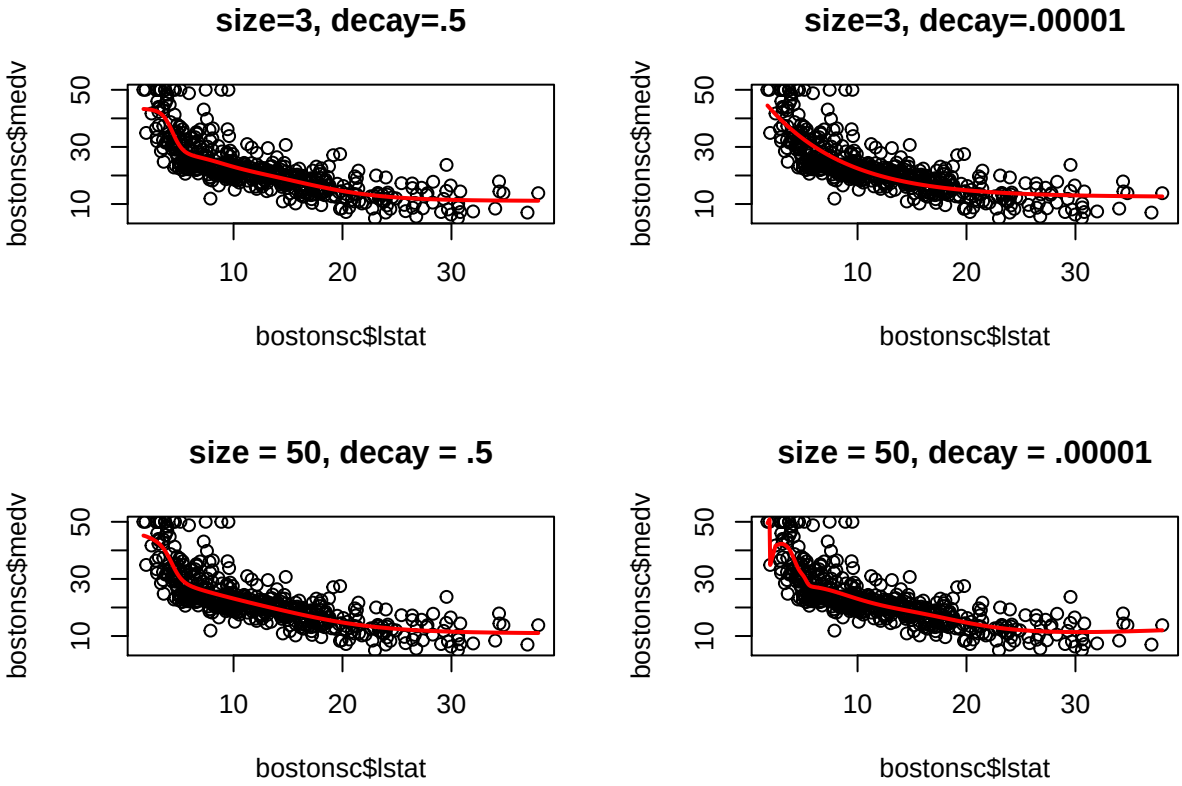
After running the Boston housing data using a single layer neural net, there are a few important insights and takeaways. The neural net does pretty well overall as a fit for non-linear relationships between variables. Here is an example of a single layer neural net fit on the "lstat" vs "medv" relationship in the data:



It's easy to see that the fit does much better than a simple linear regression fit, and here's the associated RMSE for this fit:

```
## [1] 5.398339
```

In this case the fit was done using a size of 3 and a decay parameter of 0.1. I then examined different neural net fits for several variations of size and decay parameters, and the results can be seen below:



Out of these different fits, the last one (bottom right), using a size of 50 and a decay parameter of 0.00001, proved to have the best RMSE at 5.07, but it's worth noting that all the fits were very similar in terms of RMSE. The last fit may be overfitting the data which is why it had the best RMSE, and I would say the fit that had a size of 50 and decay of 0.5 looks like a better fit and the RMSE is still very good. Single layer neural nets are pretty good for fitting a model with only one predictor variable, but they are harder to visualize when you bring in more variables, so I have only included analysis of one predictor variable on the response here. I did try a neural net with all predictor variables used to see the results and get the RMSE, and it did perform worse than the other fits that only used the lstat variable.

Problem 6: Final Project

For the final project, I was in group 12 looking at the World Cup dataset. In our group's initial meeting, I was involved in the idea of looking for datasets involving sports and in particular the World Cup because I am a huge sports fan and the World Cup was fresh in my mind, so I was excited for the group to be up for doing our project on the World Cup. I was also leading the group on cleaning the data as far as which variables to keep or throw out. Due to my knowledge of soccer, I knew which variables would be more interesting for our analysis and that we could throw out some of the categorical variables such as the date, the site of the match, and the round in which the match took place. I was also in charge of some exploratory analysis of the data - I examined some of the relationships between the predictor variables and our response variable, using box plots as they were more suitable for our classification problem. We were able to use some of that analysis to make early predictions about what variables would prove to be significant in the models and what to look for. I then was tasked with running a KNN model on our dataset as we had learned in class that KNN would be a good choice for a classification problem. I constructed the model using significant variables we had found in our feature selection, and also tried a model using all predictors but found that one to have very poor performance. Our group ended up using my KNN model in our presentation, to ultimately show that it in

fact performed worse than logistic regression and random forests. I was also heavily involved in the overall structure of our project and presentation, wanting to include a little bit of background about the World Cup as well as doing the reflections or takeaways that we had. I thought it would be important to consider and comment on the fact that we assumed a level playing field for the purposes of our analysis, and to talk about the things we could improve on for future analysis, which our group incorporated into our presentation.