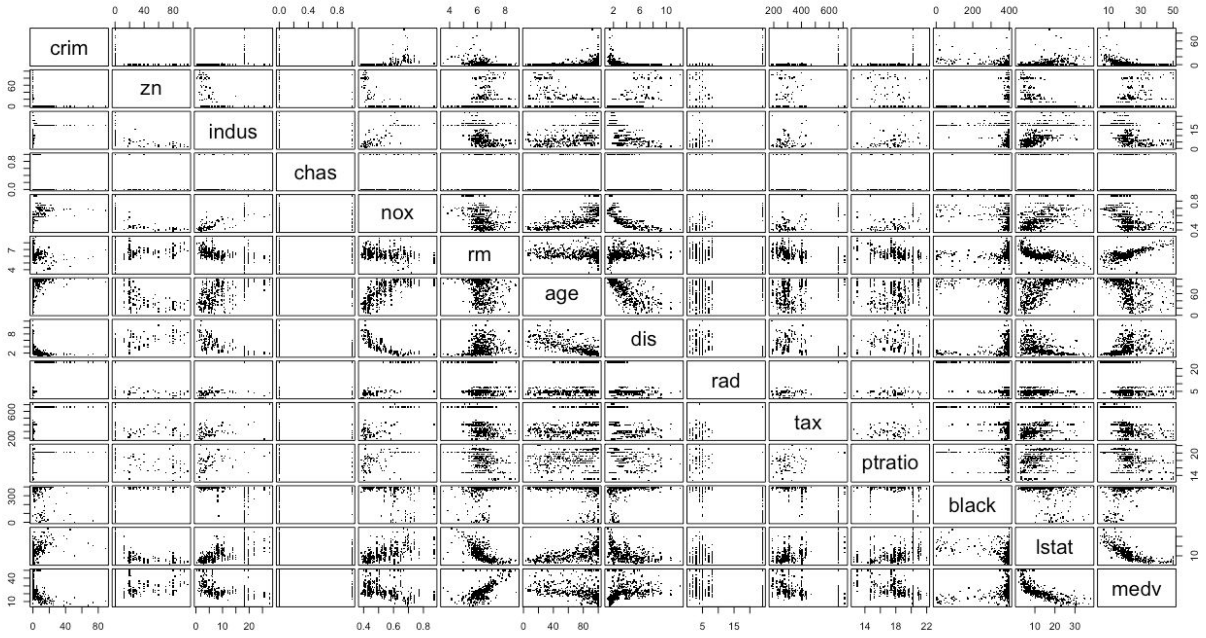
2.10

a. There are 506 rows and 14 columns. The rows represent a suburb of Boston. The columns represent recorded geographic and socioeconomic characteristics like black proportion, tax rate, and distance from highways

b.



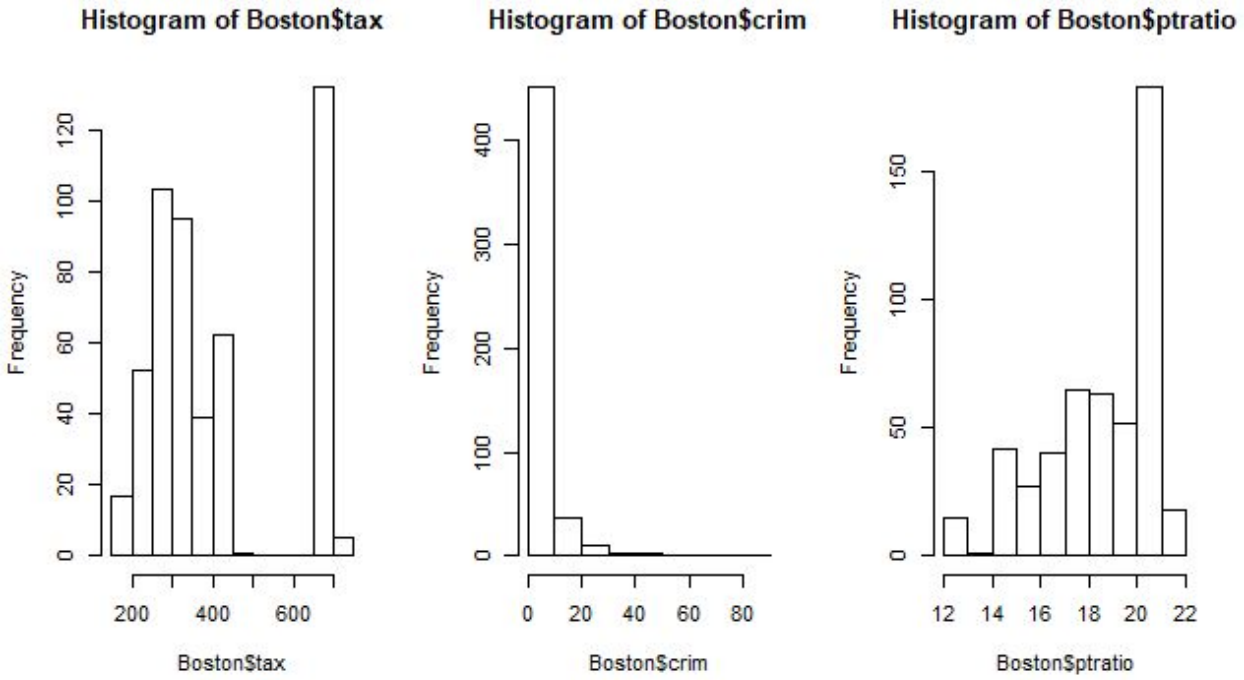
In general, there appears to be pairwise correlation among many predictors such as age and distance, lstat and medv, crime and medv/lstat/dis. Such correlations appear linear as well as nonlinear. There are categorical AND continuous variables present

c. Crime appears to have negative patterns with rm, dis, and medv. On the other hand, crime has positive correlations with age and lstat.

d. On the next page are histograms of tax, crim and pt to view the distribution of each column across all suburbs. Over 80% of the suburbs recorded have per capita crime rates below 20 so we will count all suburbs over this range, of which there are only 18 suburbs. For tax rates, there is a sudden spike in frequency past the tax = 500 mark where 137 suburbs reside (roughly 25%). Finally for ptratio, the distribution spikes for ptratio > 20 which contains 201 suburbs (39.7%)

```
> range(Boston$crim)
[1]  0.00632 88.97620
> range(Boston$tax)
[1] 187 711
> range(Boston$ptratio)
[1] 12.6 22.0
```

Regarding ranges of each predictor, the max value of crim appears to be a drastic outlier relative to the rest of crim data discussed above. For tax, only 5 suburbs have tax rates greater than 700 suggesting many of the suburbs in the tax spike pay less than 700. Ptratio's histogram shows many suburbs fall between the range



e. There are 35 suburbs located along the Charles River

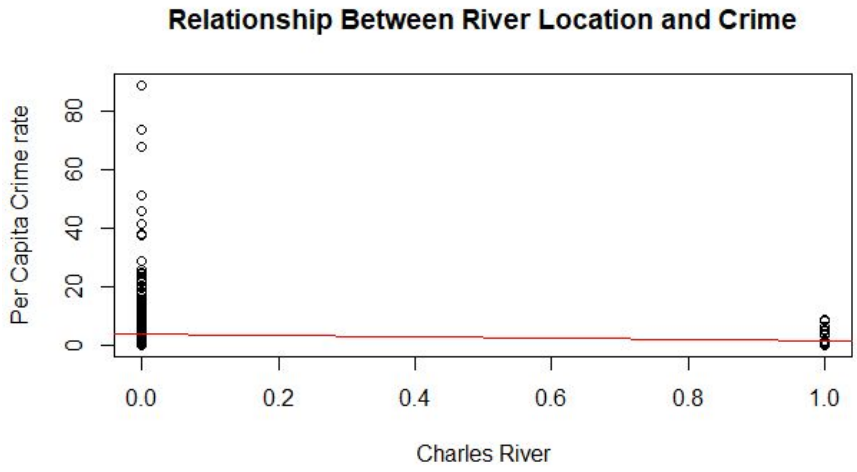
f. The median pupil-teacher ratio is 19.05

g. This suburb contains a crim value > 20 which is uncommon, the minimum zn value, an average indus value, average nox, average rm, maximum age, average dis, maximum rad, common tax rate and ptratio, maximum black proportion, and is in an lstat bin with only 11 other suburbs. In conclusion, this suburb's medval makes sense given the percentage of low status residents and high crime rate. Furthermore, all buildings of this suburb were built prior to 1940 and the suburb contains no zoned off large plots of land (>25,000 sq. ft)

h. There are 13 suburbs who average over 8 rooms per dwelling. Such suburbs see lower crime rates, industrial business, lstat percent, ptratio, distance to highways, and tax rates. These same suburbs have higher median home value, and black proportion. This makes sense since we would expect suburbs with higher priced homes to have low crime rates, distance to highways, and low status population. These suburbs are also more demanded due to their preferable property tax rates and pupil-teacher ratio.

3.15

a. There were 12 out of 13 significant predictors based on t-value. The only insignificant predictor is chas makes sense when plotting chas against crim alongside its low t-stat

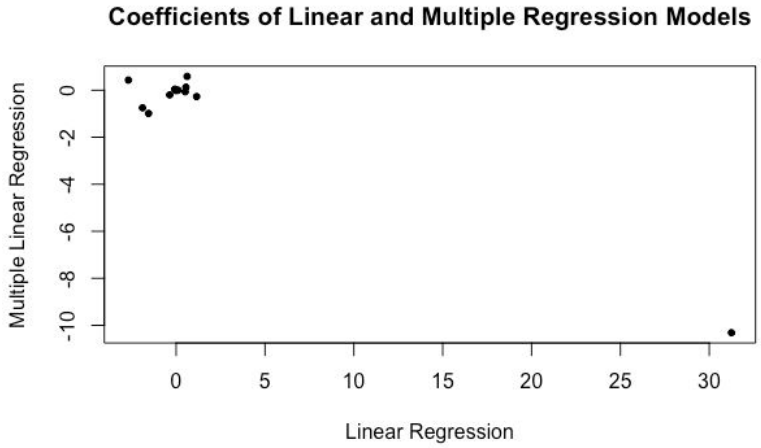


b. The only significant predictors remaining are zn, dis, rad, black and medv although nox is extremely close to the threshold with t-value 1.955.

```
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 **
```



c.

We expect multivariate model to have more insignificant predictors than univariate models. Since there is correlation among the predictors as we stated in 2.10 then weaker predictors were seen as significant because they were correlated to a strong predictor. For example, crime had a positive correlation with rm although it is likely that the effect of rm was acting as a proxy to the effect of medv since we expect higher medv homes to have more average rooms as well.

d.

```
Call:
lm(formula = crim ~ poly(zn, 3))

Residuals:
   Min     1Q Median     3Q    Max
-4.821 -4.614 -1.294  0.473 84.130

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared:  0.05824,   Adjusted R-squared:  0.05261
F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

We repeat this cubic fit for every variable and conclude the following based on t-statistic:
Zn, rm, rad, tax, and lstat are significant up until their cubic term
Indus, nox, age, dis, ptratio, and medv are significant through the cubic term
Black is significant only on the linear term
Chas is excluded because it’s binary variable so this fit’s output would be unaffected

6.9

a.
```
train = sample(1:nrow(College), nrow(College)/2)
test <- -train
College.train <- College[train,]
College.test <- College[test,]
```

b. OLS MSE = 1108531
Note: MSE is referencing test MSE not training

c. .01149757 is our lambda chosen by CV. RR MSE = 1108512
Note: The lambda search grid ranges from 10^{10} to 10^{-2} for RR and the lasso.

d. 28.48036 is our lambda chosen via CV. Lasso MSE = 1028718. Non-zero coefficients:

```
(Intercept) -4.248125e+02
(Intercept)  .
PrivateYes  -4.955003e+02
Accept       1.540306e+00
Enroll      -3.900157e-01
Top10perc    4.779689e+01
Top25perc   -7.926581e+00
F.Undergrad -9.846932e-03
P.Undergrad  .
Outstate    -5.231286e-02
Room.Board   1.880308e-01
Books        1.265938e-03
Personal     .
PhD         -4.137294e+00
Terminal    -3.184316e+00
S.F.Ratio    .
perc.alumni -2.181304e+00
Expend       3.193679e-02
Grad.Rate    2.877667e+00
```

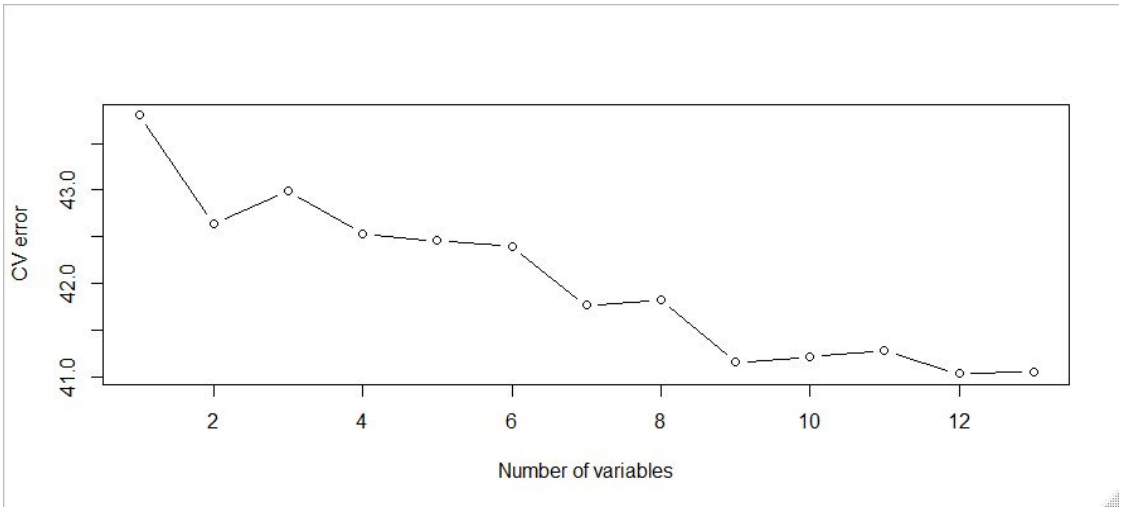
e. PCR MSE = 1505718. M = 16 based on CV error

f. PLS MSE = 1134531. M = 14 based on CV error

g. To compare prediction accuracy we calculate the test R^2 of each model. All models except PCR (R^2 = .86) had roughly R^2 = .9 although LASSO was best performer by a marginal amount.

6.11

a. Let's try best subset, the lasso, and ridge regression with this dataset to predict crim. I go straight into optimizing the tuning parameter of each model using 10-fold CV
b. ***Best Subset***: Using the approach provided in chapter 6 lab. We obtain test MSE = 41.03457 on a 12 variable model. The excluded variable is 'age'



LASSO: Used the approach provided in the lab using cv.glmnet() and search 10^{10} to 10^{-2}. Test MSE is 38.3096. The excluded variables are 'age' and 'tax' when fitting onto the entire dataset. Our optimal lambda is .09979553

Note: The MSE in this plot is training MSE not test MSE since we train THEN optimize

Ridge Regression: Similar as above except alpha=0. Test MSE is 38.36587. Our optimal lambda is .7908625

c. No. Our strongest model is the lasso which zeroed out 'age' and 'tax' coefficients

4.10

a.

```
> cor(Weekly[,-9])
             Year         Lag1        Lag2        Lag3         Lag4         Lag5      Volume        Today
Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923 -0.030519101  0.84194162 -0.032459894
Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876 -0.008183096 -0.06495131 -0.075031842
Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535 -0.072499482 -0.08551314  0.059166717
Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865  0.060657175 -0.06928771 -0.071243639
Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000 -0.075675027 -0.06107462 -0.007825873
Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027  1.000000000 -0.05851741  0.011012698
Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617 -0.058517414  1.00000000 -0.033077783
Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873  0.011012698 -0.03307778  1.000000000
> summary(Weekly)
      Year          Lag1               Lag2               Lag3               Lag4               Lag5         
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580   1st Qu.: -1.1580   1st Qu.: -1.1660  
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410   Median :  0.2380   Median :  0.2340  
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472   Mean   :  0.1458   Mean   :  0.1399  
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4050  
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
     Volume            Today          Direction  
 Min.   :0.08747   Min.   :-18.1950   Down:484  
 1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605  
 Median :1.00268   Median :  0.2410             
 Mean   :1.57462   Mean   :  0.1499             
 3rd Qu.:2.05373   3rd Qu.:  1.4050             
 Max.   :9.32821   Max.   : 12.0260             
```

The only strong pattern present is between Year and Volume

b.
```
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
```

Lag2 is the only significant predictor by z-stat.

c.
```
          Direction
logit.pred Down  Up
      Down   54  48
      Up    430 557
```
We have 56% prediction accuracy on the entire dataset

d.
```
          Direction.test
logit.preds Down Up
       Down    9  5
       Up     34 56
```
This logistic model has 62.5% test set prediction accuracy

g.
```
        Direction.test
knn.pred Down Up
    Down   21 30
    Up     22 31
```
KNN with k=1 has 50% test set prediction accuracy

h. Our logistic regression outperforms KNN by 12.5% test set accuracy

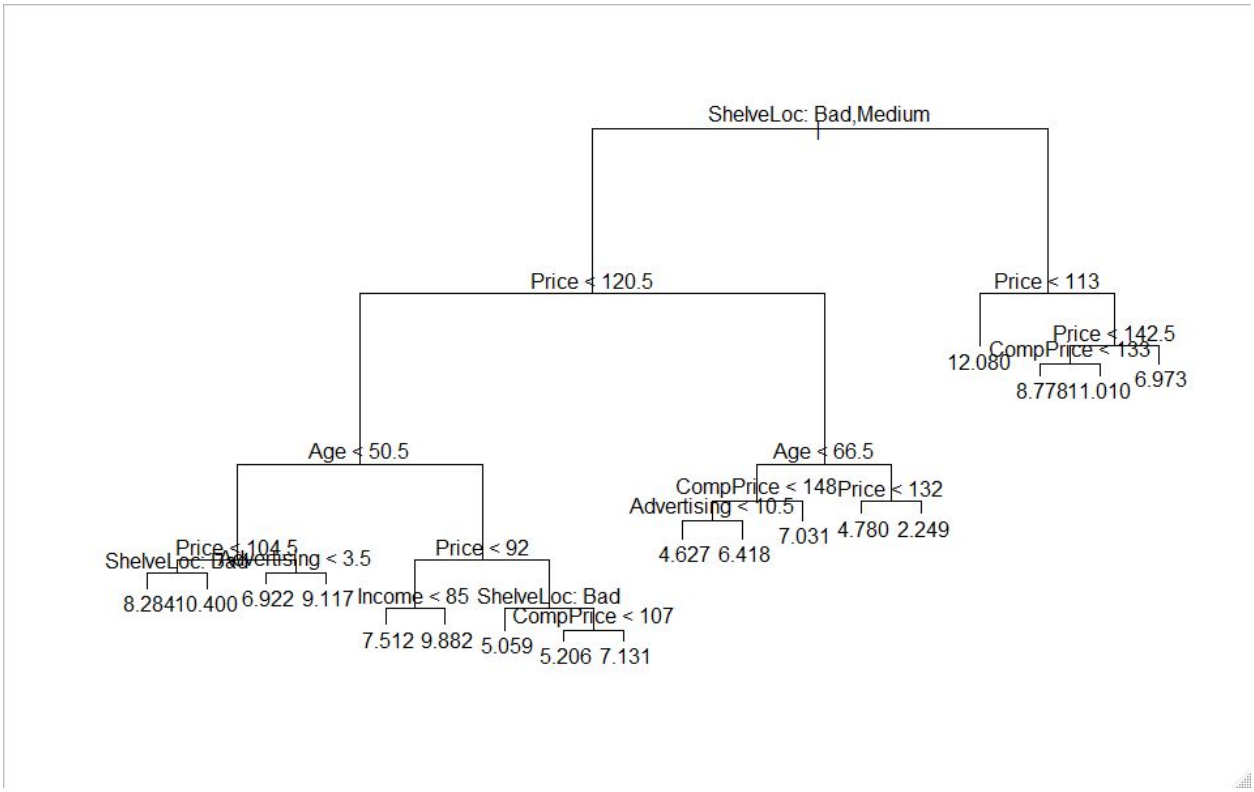
i. For our logistic regressions we considered a Lag1+Lag2 model and Lag1+Lag2+Lag1*Lag2. Neither models were able to outperform our 1st logistic fit on only Lag2.
When testing KNN, we search through K=1 to K=100 with a for loop while calculating prediction accuracy at each iteration. Our accuracy maxes out at 61.5% at K=47.

8.8

a.

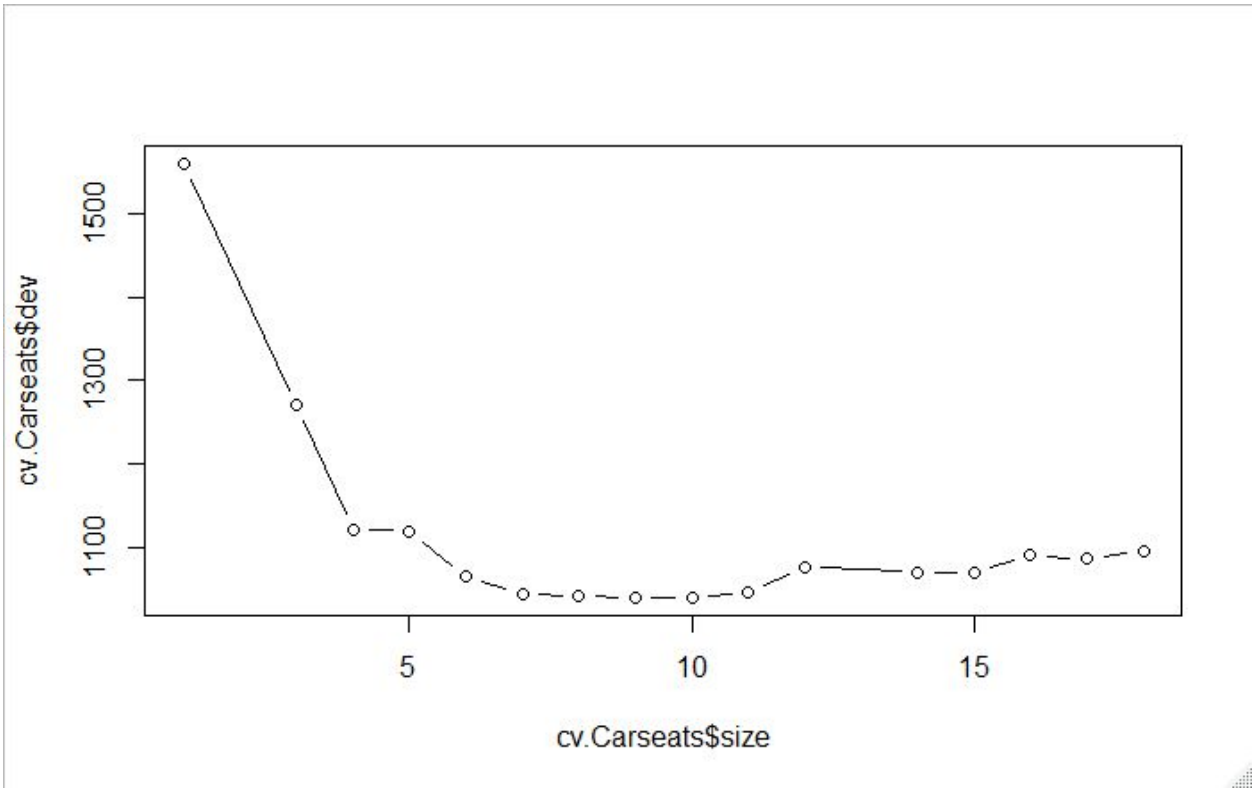
```
train=sample(1:nrow(Carseats), nrow(Carseats)/2)
test = -train
Carseats.train = Carseats[train,]
Carseats.test = Carseats[test,]
```

b.



The tree has 18 leaves and only uses 6 of the 10 possible predictors. Urban, US, Education, and Population were all excluded from this tree. Big Tree MSE = 4.148897

c. CV plot indicates the tree can be pruned for better results with optimal tree size 8. Pruned Tree Test MSE = 5.09085



d. Bagging Test MSE = 2.633915. Price and ShelveLoc are by far the most important

	%IncMSE	IncNodePurity
CompPrice	16.9874366	126.852848
Income	3.8985402	78.314126
Advertising	16.5698586	123.702901
Population	0.6487058	62.328851
Price	55.3976775	514.654890
ShelveLoc	42.7849818	319.133777
Age	20.5135255	185.582077
Education	3.4615211	42.253410
Urban	-2.5125087	8.700009
US	7.3586645	18.180651

e. RF (m=round(sqrt[p])) Test MSE = 3.321154. Price and ShelveLoc are still most important

	%IncMSE	IncNodePurity
CompPrice	7.443405	130.87552
Income	3.227858	127.18662
Advertising	13.388259	139.53499
Population	-1.031306	102.32154
Price	36.616911	369.59534
ShelveLoc	31.284175	233.49549
Age	17.622273	206.09959
Education	1.454555	70.41374
Urban	-1.864781	15.13225
US	6.193082	35.74746

I looped through every value of m (1 through 10) and find that MSE ranges between 2.52063 (m=10) to 5.078495 (m=1). Since m=10 is bagging, the results of this loop and the MSE in part d confirm that bagging outperforms single tree and random forest models.

8.11

a.
```
train = 1:1000
Caravan$Purchase = ifelse(Caravan$Purchase == 'Yes', 1, 0)
Caravan.train = Caravan[train,]
Caravan.test = Caravan[-train,]
```

b. I make no limit to interaction depth and obtain MOSTYPE, APERSAUT, PPERSAUT as my 3 most important predictors. These results are the same without any interaction depth. I also specify the distribution as bernoulli since Purchase has only levels of 'Yes' and 'No'

c.
```
   boost.preds2
      0    1
  0 4410  123
  1  256   33
```
Only ~21% of those predicted to purchase actually purchase

```
  Caravan.logit.preds
     0    1
0 4183  350
1  231   58
```
Only ~14% of those predicted to purchase actually purchase

Thus boosting outperforms logistic regression in this case

Problem 1: Beauty Pays!

a. I evaluated 3 models: the full model, just BeautyScore, and full model + full interactions using GLMs. The model with full interactions shows the lowest residual deviance although the full model has only a marginal increase in deviance suggesting interactions are not significant. This is further evidenced by t-stats of interactions which all fall below 1.96. Furthermore, the AIC of the full model is lower than the interaction model. Therefore, by virtue of interpretation we prefer the full model most. Thus, we interpret BeautyScore's coefficient as follows: for every marginal

increase in BeautyScore, we expect CourseEvals to go up by .31.

b. It is all too likely that we are not controlling for other important predictors of CourseEvals. For example, race and location/region could be important predictors for 2 different reasons: 1) it is possible that race/location can affect CourseEvals significantly but also 2) that race/location may have an effect on one's BeautyScore rating. In summary, it is likely that our model's "noise" contains the "signal" of important predictors we have not controlled/captured

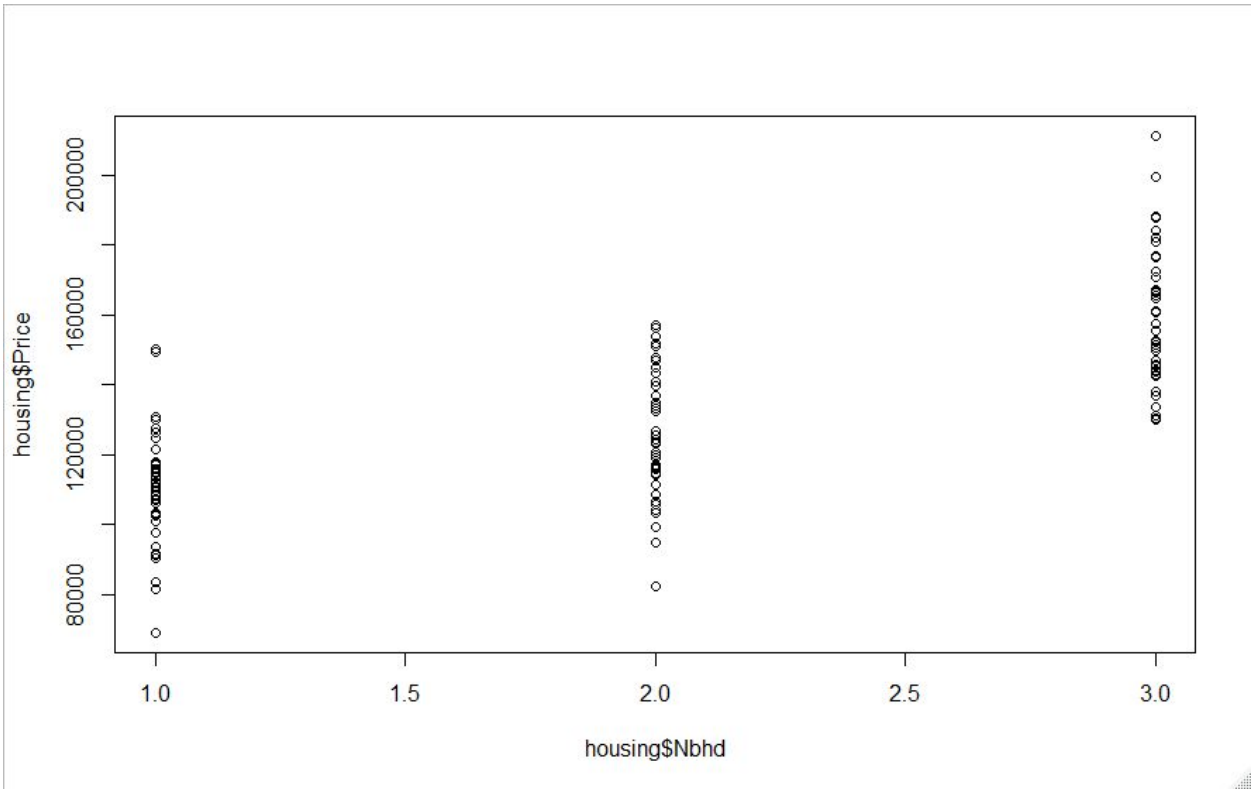
Problem 2: Housing Price Structure

a. Yes. There is a $15,603.19 premium for brick homes versus non-brick homes when ceteris paribus

b. Yes. We can interpret the nbhd coefficient as follows: as nbhd marginally increases, we see housing price go up by $9,790.38 per level of neighborhood. So for any homes in neighborhood 3, we expect there to be a $29,371.14 price increase relative to homes not in neighborhoods 1 through 3.

c. To check this effect we create an interaction term of nbhd on brick. We can now control for different premiums of brick homes in different neighborhoods. The coefficient of this interaction is $7,701.08. Therefore, for neighborhood 3, we see there is a $23,103.24 premium for brick homes in that neighborhood.

d. I plot nbhd vs. price (next page) to check if there is reasoning behind the decision and conclude there is no reason. While there is overlap between the house prices, neighborhood 2 contains houses with prices not typical of those in neighborhood 1 suggesting differences between the houses of these neighborhoods.



Problem 3: What causes what?

a. The only things controlled for in our comparison between cities is crime rate and police size. This would not account for socioeconomic differences or systemic differences between the cities. Furthermore, police size and crime rate is correlated because cities with already high crime rates have incentive to hire more police. As a result, naive interpretations would say that police size has a causal effect that increases crime rates.

b. Researchers would use the terrorist alert system to find situations in which police size was high for reasons aside from crime. Based on table 2, during such high alert periods they concluded the total number of daily crimes would lower by 7.316 (significant at the 95% confidence level).

c. One could argue that crime rates were low since tourists were not as active during high alert periods. To control for this factor, the researchers used metro ridership to check if there would be differences in tourist activity level on high alert days versus non-high alert days.

d. The model uses 4 variables: midday ridership (numeric), high alert (binary indicator), district 1 (binary indicator), and other districts (binary indicator). This suggests that the crime during high alert days varies depending on the district. Furthermore, the decrease in crime in other districts is not as pronounced as in district 1 although it is

still an overall decrease in daily crime. Thus, it is more likely that larger police size does in fact reduce daily crime rates.

Problem 4: BART

a. First I setup the CAhousing data and create my X and Y matrices

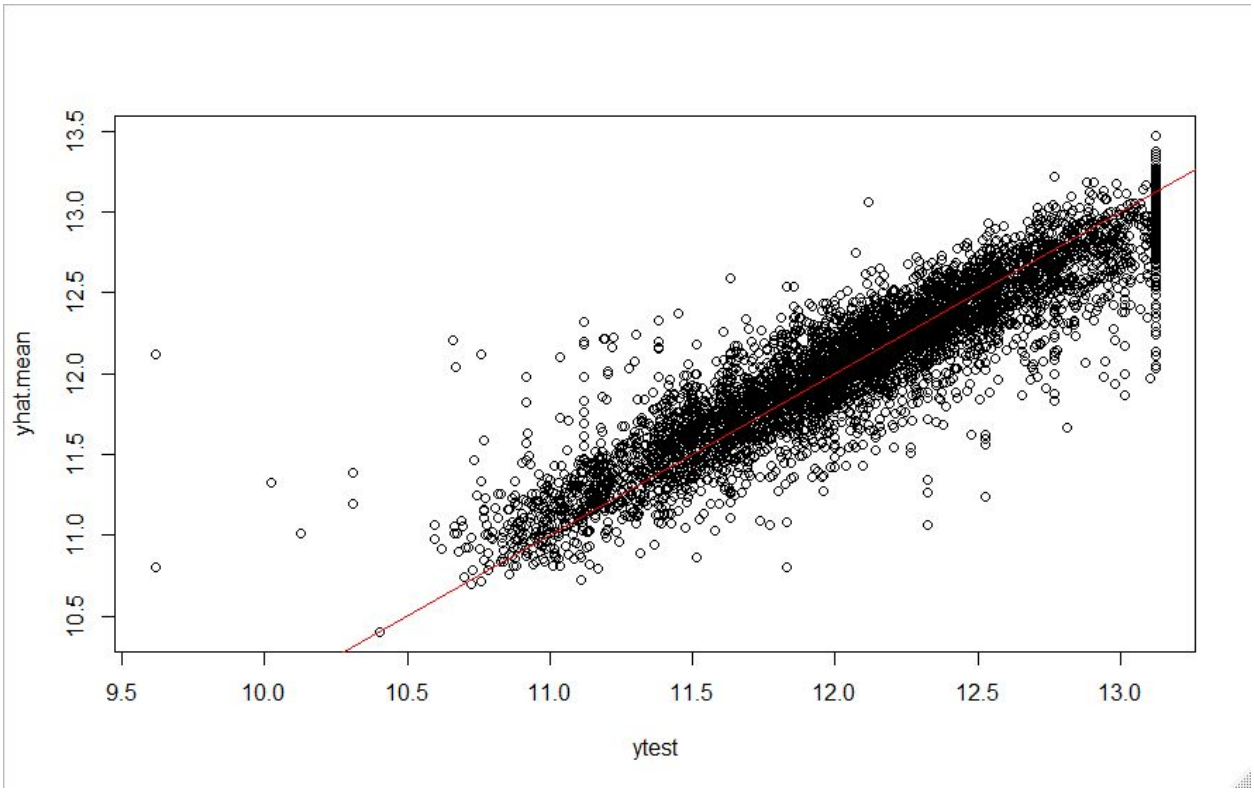
```
ca = read.csv('CAhousing.csv', header=TRUE)
ca$AveBedrms <- ca$totalBedrooms/ca$households
ca$AveRooms <- ca$totalRooms/ca$households
ca$AveOccupancy <- ca$population/ca$households
logMedVal <- log(ca$medianHouseValue)
ca <- ca[,-c(4,5,9)] # lose lmedval and the room totals
ca$logMedVal = logMedVal
x = ca[,1:9]
y = ca$logMedVal # median value
head(cbind(x,y))
```

Then we create our training and test splits

```
ii = sample(1:n,floor(.75*n)) # indices for train data, 75% of data
xtrain=x[ii,]; ytrain=y[ii] # training data
xtest=x[-ii,]; ytest=y[-ii] # test data
```

Finally we fit our data to BART and compute RMSE of ~.238.

```
set.seed(1)
bf_train2 = wbart(xtrain, ytrain, xtest)
RMSE3 = sqrt(mean((bf_train2$yhat.test.mean - ytest)^2))
```



Thus a default BART fit is slightly outperformed by the best RF and boosting models of this dataset (RMSE = .23 and .231 respectively)

Problem 5: Neural Nets

a. First I standardized all x's before splitting the scaled data into training and test set.

```
#standardize the x's
minv = rep(0,13)
maxv = rep(0,13)
Bostonsc = Boston
for(i in 1:13) {
  minv[i] = min(Boston[[i]])
  maxv[i] = max(Boston[[i]])
  Bostonsc[[i]] = (Boston[[i]]-minv[i])/(maxv[i]-minv[i])
}
```

```
train = sample(1:nrow(Bostonsc), nrow(Bostonsc)/2)
Bostonsc.train = Bostonsc[train,]
Bostonsc.test = Bostonsc[-train,]
```

I tested models with decay=.5 vs decay=.00001 as well as size=3 vs size=50 for a total of 4 models.

```
znn1 = nnet(medv~.,Bostonsc.train,size=3,decay=.5,linout=T)
znn2 = nnet(medv~.,Bostonsc.train,size=3,decay=.00001,linout=T)
znn3 = nnet(medv~.,Bostonsc.train,size=50,decay=.5,linout=T)
znn4 = nnet(medv~.,Bostonsc.train,size=50,decay=.00001,linout=T)
```

To compare the performance of each model I use the correlation between the model's predictions on the test set and the actual values of the test set.

```
> znnf1 = predict(znn1,Bostonsc.test)
> znnf2 = predict(znn2,Bostonsc.test)
> znnf3 = predict(znn3,Bostonsc.test)
> znnf4 = predict(znn4,Bostonsc.test)
> temp=data.frame(y=Bostonsc.test$medv, znn1=znnf1, znn2=znnf2, znn3=znnf3, znn4=znnf4)
> print(cor(temp$y,temp$znn1))
[1] 0.9165034
> print(cor(temp$y,temp$znn2))
[1] 0.9272405
> print(cor(temp$y,temp$znn3))
[1] 0.934798
> print(cor(temp$y,temp$znn4))
[1] 0.9058997
```

Thus we can conclude that our best model of the 4 is a neural net using size=50 and decay=.5

Edward Eustachon
epe245

Problem 6: Final Project

I contributed a variety of ideas and actions during the course of our project. Initially, when we had chosen our Lending Club dataset I advised the group to use a different dataset pertaining to Lending Club. This is because our initial dataset had left out FICO scores which we all agreed would be an important predictor of interest rate for loans on the site. During this process, the group looked towards me for assistance in filling any gaps of knowledge regarding the course material and statistical theory.

Afterwards during our model selection process, I was tasked with fitting random forests and a decision trees to our dataset. I also provided the process of how we would go about selecting the best model of all the model families we had tried. I outlined the need to cross-validate each model to optimize our tuning parameters before calculating test set RMSE. Additionally, during this process, my teammates often came to me for assistance in debugging their script as well as providing a final passthrough and interpretation of their code and results.

Finally, during our creation of the presentation and report, I was given the freedom to revise the interpretation of our results as I deemed fit. Overall, the team tasked me with making sure that we had a firm understanding in how to explain and interpret our results. For instance, when a model did not output as expected, I was able to communicate to the team some reasons why that may be.