DS-6030 Homework Module 1

Matt Scheffel

DS 6030 | Spring 2022 | University of Virginia

1. Flexible vs Inflexible Methods

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

For this example, we would expect the performance of a flexible statistical learning method to be better than an inflexible method. This is because when a large dataset is present, a flexible method will fit the data better and come closer to its true distribution.

(b) The number of predictors p is extremely large, and the number of observations n is small.

For this example, we would expect the performance of a flexible statistical learning method to be worse than an inflexible method. This is due to the issue of overfitting with the smaller dataset.

(c) The relationship between the predictors and response is highly non-linear.

For this example, we would expect the performance of a flexible statistical learning method to be better than an inflexible method. This is because when there are more degrees of freedom, a flexible method fits the dataset better.

(d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

For this example, we would expect the performance of a flexible statistical learning method to be worse than an inflexible method. This is due to the issue of overfitting with the “noise” of the error terms having a large impact on the fit.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is a regression problem where we are most interested in inference.

N = 500 and P = 3

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each prod- uct we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

This is a classification problem where we are most interested in prediction.

N = 20 and P = 14

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

This is a regression problem where we are most interested in prediction.

N = 52 and P = 4

6. Describe the differences between a parametric and a non-parametric statistical learning approach.

What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

A parametric statistical learning approach assumes a linear function for the model when estimating fit. A non-parametric model makes no assumption, but thus requires a larger sample size. This demonstrates an advantage of the parametric model (in comparison to a non-parametric model): it requires less data/ a smaller sample size. However, a disadvantage is that it may assume the wrong form of the model and result in overfitting that leads to an inaccurate estimate.

8. This exercise relates to the College data set, which can be found in the file College.csv on the book website.

It contains a number of variables for 777 different universities and colleges in the US. The variables are

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10 % of high school class
- `Top25perc` : New students from top 25 % of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree • S.F.Ratio : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

(a) Use the `read.csv()` function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
setwd("~/Desktop/MSDS/DS 6030/ALL CSV FILES - 2nd Edition")
college <- read.csv("College.csv")
head(college)
```

```
#>                              X Private Apps Accept Enroll Top10perc Top25perc
#> 1 Abilene Christian University     Yes 1660   1232    721        23        52
#> 2           Adelphi University     Yes 2186   1924    512        16        29
#> 3               Adrian College     Yes 1428   1097    336        22        50
#> 4          Agnes Scott College     Yes  417    349    137        60        89
#> 5    Alaska Pacific University     Yes  193    146     55        16        44
#> 6            Albertson College     Yes  587    479    158        38        62
#>   F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
#> 1        2885         537     7440       3300   450     2200  70       78
#> 2        2683        1227    12280       6450   750     1500  29       30
#> 3        1036          99    11250       3750   400     1165  53       66
#> 4         510          63    12960       5450   450      875  92       97
#> 5         249         869     7560       4120   800     1500  76       72
#> 6         678          41    13500       3335   500      675  67       73
#>   S.F.Ratio perc.alumni Expend Grad.Rate
#> 1      18.1          12   7041        60
#> 2      12.2          16  10527        56
#> 3      12.9          30   8735        54
#> 4       7.7          37  19016        59
#> 5      11.9           2  10922        15
#> 6       9.4          11   9727        55
```

(b) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```r
rownames(college) <- college[, 1]
View(college)
```

```
#> Error in check_for_XQuartz(): X11 library is missing: install XQuartz from xquartz.macosforge.org
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```r
college <- college[, -1]
View(college)
```

```
#> Error in check_for_XQuartz(): X11 library is missing: install XQuartz from xquartz.macosforge.org
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

(c)

i. Use the `summary()` function to produce a numerical summary of the variables in the data set.

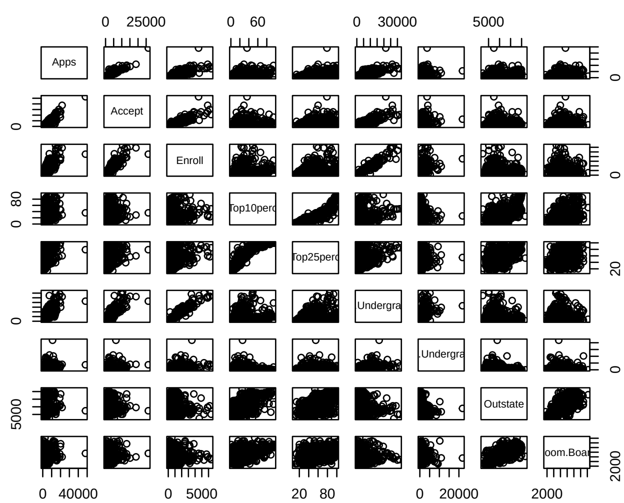
```r
summary(college)
```

```
#>    Private               Apps           Accept          Enroll    
#>  Length:777         Min.   :   81   Min.   :   72   Min.   :  35  
#>  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242  
#>  Mode  :character   Median : 1558   Median : 1110   Median : 434  
#>                     Mean   : 3002   Mean   : 2019   Mean   : 780  
#>                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902  
#>                     Max.   :48094   Max.   :26330   Max.   :6392  
#>    Top10perc       Top25perc      F.Undergrad     P.Undergrad    
```

```
#>  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0  
#>  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0  
#>  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0  
#>  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3  
#>  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0  
#>  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0  
#>     Outstate       Room.Board       Books          Personal   
#>  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250  
#>  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850  
#>  Median : 9990   Median :4200   Median : 500.0   Median :1200  
#>  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341  
#>  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700  
#>  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800  
#>       PhD            Terminal       S.F.Ratio      perc.alumni   
#>  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00  
#>  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00  
#>  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00  
#>  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74  
#>  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00  
#>  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00  
#>      Expend        Grad.Rate     
#>  Min.   : 3186   Min.   : 10.00  
#>  1st Qu.: 6751   1st Qu.: 53.00  
#>  Median : 8377   Median : 65.00  
#>  Mean   : 9660   Mean   : 65.46  
#>  3rd Qu.:10830   3rd Qu.: 78.00  
#>  Max.   :56233   Max.   :118.00
```

ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

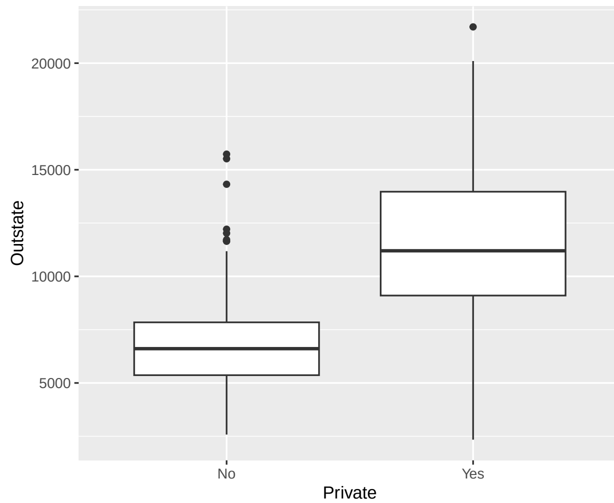
```
pairs(college[,2:10])
```



iii. Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

```
library(ggplot2)

ggplot(college, aes(x = Private, y = Outstate))+
  geom_boxplot()
```



iv. Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

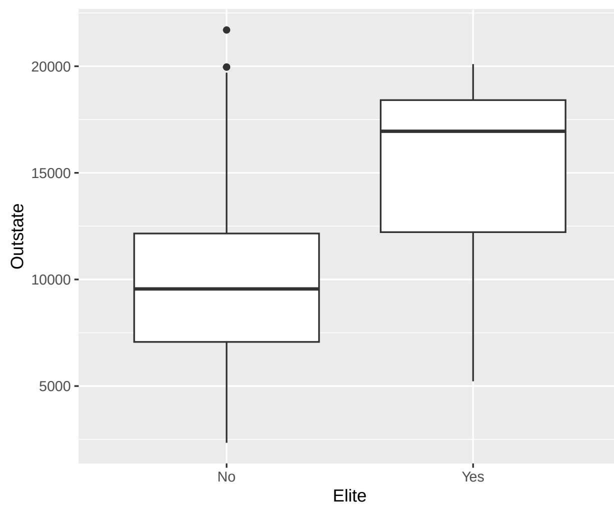
```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
```

Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

```
summary(college$Elite)
```

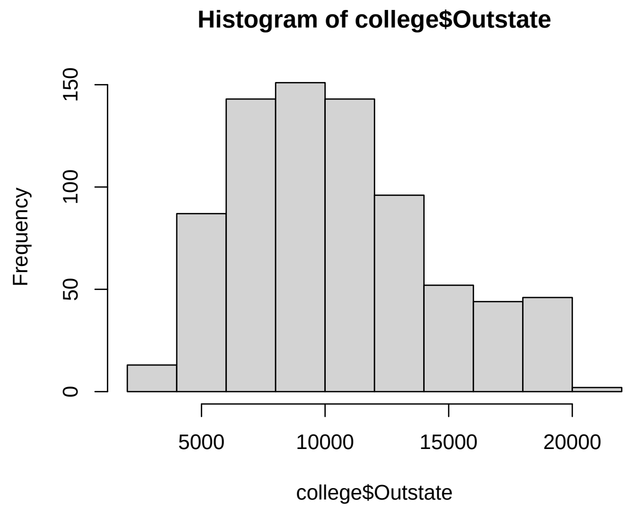
```
#>  No Yes
#> 699  78
```

```
ggplot(college, aes(x = Elite, y = Outstate))+
  geom_boxplot()
```

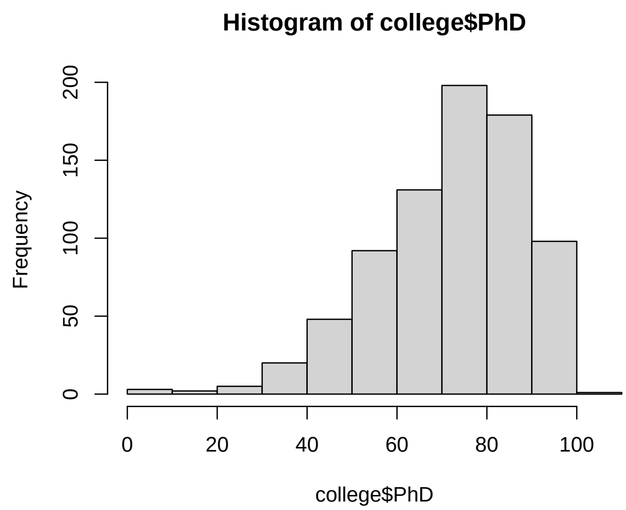


v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow = c(2, 2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

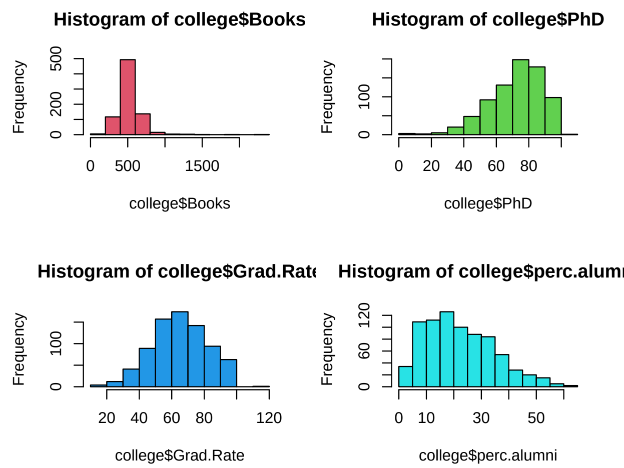
```r
hist(college$Outstate)
```




```r
hist(college$PhD)
```




```r
par(mfrow = c(2,2))
hist(college$Books, col = 2)
hist(college$PhD, col = 3)
hist(college$Grad.Rate, col = 4)
hist(college$perc.alumni, col = 5)
```



vi. Continue exploring the data, and provide a brief summary of what you discover.

I discovered a number of things from this dataset. Public schools tend to have higher raw numbers than private schools. Schools labeled as "Elite" unsurprisingly perform better in many categories.

10. This exercise involves the Boston housing data set.

(a) To begin, load in the `Boston` data set. The Boston data set is part of the ISLR2 library.

```
install.packages("ISLR2")
```

```
#> Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(ISLR2)
```

Now the data set is contained in the object Boston.

```
Boston
```

Read about the data set:

```
?Boston
```

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

```
head(Boston)
```

```
#>      crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
#> 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
#> 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
#> 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
#> 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
#> 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
#> 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

506 rows and 14 columns (with 13 variables).

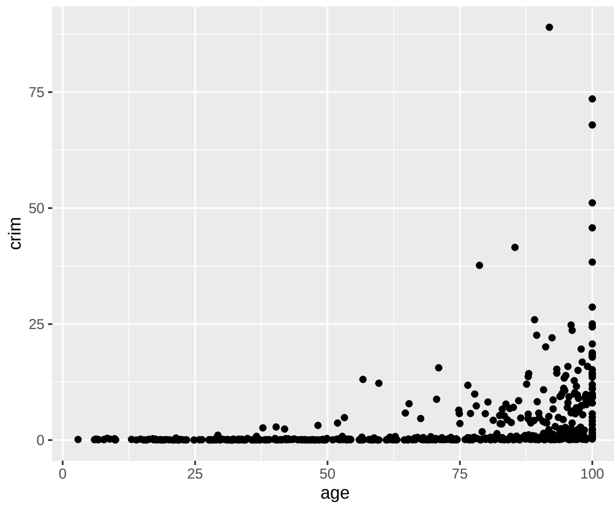
Rows represent the 506 Boston suburbs.

Columns:

crim - per capita crime rate by town. zn - proportion of residential land zoned for lots over 25,000 sq.ft. indus - proportion of non-retail business acres per town. chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). nox - nitrogen oxides concentration (parts per 10 million). rm - average number of rooms per dwelling. age - proportion of owner-occupied units built prior to 1940. dis - weighted mean of distances to five Boston employment centres. rad - index of accessibility to radial highways. tax - full-value property-tax rate per $10,000. ptratio - pupil-teacher ratio by town. lstat - lower status of the population (percent). medv - median value of owner-occupied homes in $1000s.

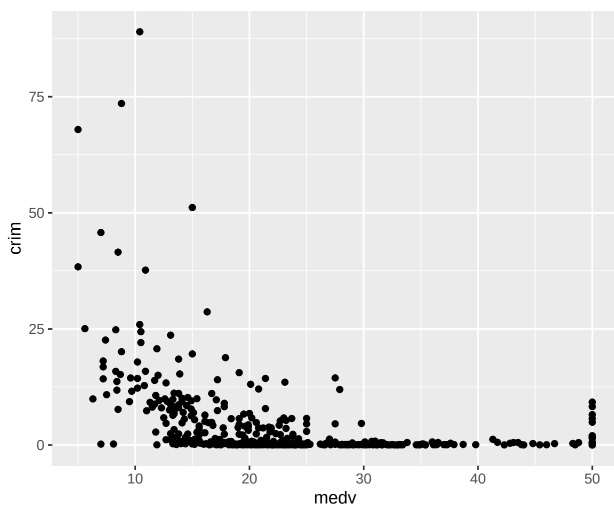
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
ggplot(Boston, aes(x= age, y = crim))+
  geom_point()
```



Crime tends to increase in areas with older houses.

```
ggplot(Boston, aes(x= medv, y = crim))+
  geom_point()
```



Crime tends to decrease as median home value goes up.

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

Yes, housing predictors tend to be associated with the per capita crime rate. We see crime increase as house ages increase and we see crime decrease as median house values increase.

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

```r
summary(Boston$crim)
```

```
#>     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#>  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```r
which.max(Boston$crim)
```

```
#> [1] 381
```

```r
range(Boston$crim)
```

```
#> [1]  0.00632 88.97620
```

```r
summary(Boston$tax)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   187.0   279.0   330.0   408.2   666.0   711.0
```

```r
which.max(Boston$tax)
```

```
#> [1] 489
```

```r
range(Boston$tax)
```

```
#> [1] 187 711
```

```r
summary(Boston$ptratio)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   12.60   17.40   19.05   18.46   20.20   22.00
```

```r
which.max(Boston$ptratio)
```

```
#> [1] 355
```

```r
range(Boston$ptratio)
```

```
#> [1] 12.6 22.0
```

Suburb 381 has the highest crime rate. Range extends far beyond the median value.

Suburb 489 has the highest tax rate. Range tends to stretch pretty far beyond the median, more than double.

Suburb 355 has the highest pupil-teacher ratio. Max/min do not extend very far beyond themedian.

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

```r
sum(Boston$chas == 1)
```

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

35 census tracts in this data set bound the Charles River.

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

```r
median(Boston$ptratio)
```

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

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

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

```r
which.min(Boston$medv)
```

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

Census tract 399 has lowest median value of owner-occupied homes.

```r
Boston[which.min(Boston$medv),]
```

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

In addition to the lowest median home value, census tract 399 has a high crime rate and older homes on average

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

```r
sum(Boston$rm > 7)
```

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

64 census tracts average more than seven rooms per dwelling.

```r
sum(Boston$rm > 8)
```

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

13 census tracts that average more than eight rooms per dwelling.

```r
summary(Boston[Boston$rm > 8,])
```

```
#>       crim               zn            indus             chas       
#>  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
#>  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
#>  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
#>  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
#>  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
#>  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
#>       nox               rm             age             dis       
#>  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
#>  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
#>  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
#>  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
#>  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
#>  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
#>       rad              tax           ptratio          lstat           medv     
#>  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :2.47   Min.   :21.9  
#>  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:3.32   1st Qu.:41.7  
#>  Median : 7.000   Median :307.0   Median :17.40   Median :4.14   Median :48.3  
#>  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :4.31   Mean   :44.2  
#>  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:5.12   3rd Qu.:50.0  
#>  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :7.44   Max.   :50.0  
```

It looks like dwellings with more than 8 rooms tend to be older and have higher crime rates.