# STATS 101C - Statistical Models and Data Mining - Homework 1

*Darren Tsang, Discussion 4B*

Produced on Saturday, Oct. 17 2020 @ 01:39:21 AM

## Question 1 (Exercise 5 from Section 2.4)

**What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?**

An advantage to using a flexible approach is that you can find many different possible forms for your model. For example, an inflexible approach might only be able to estimate linear relationships, but a flexible approach can estimate exponential relationships. A disadvantage to using a very flexible approach is that there is a high chance of overfitting since a flexible model requires estimating a lot of parameters. This is a big flaw because your model will not be generalizable to new data.

An advantage to using a less flexible approach is that it is much less likely that overfitting will happen. This allows your model to be able to be generalizable to new, unseen data. You also will be able to interpret the relationships more clearly with a less flexible approach. Furthermore, a less flexible approach does not need as much data as a more flexible approach does. A disadvantage to using a less flexible approach is that you can not really explore complex relationships.

The choice of using a flexible vs a less flexible approach really depends on your situation. For example, if you only care about the predictions and not the interpretability of the model, you might consider using a flexible approach. On the other hand, if you want to be able to interpret your model and what it's doing, a less flexible approach might be for you. Also, if you have a large dataset, it would be good to consider a flexible approach because overfitting is less likely to happen.

## Question 2 (Exercise 6 from Section 2.4)

**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 nonparametric approach)? What are its disadvantages?**

In a paramteric approach, you first start off by making an assumption for the true $f$. An example is to assume that the true $f$ is a liner model. Then you will need to find the coefficients for this linear model. This boils the problem down to only having to estimate parameters because you already assumed the form of the true $f$ to be a linear model. This is a big disadvantage because your assumption might be very, very different from the true $f$, which means your model will perform poorly to new data.

On the other hand, with a nonparametric approach, you do not have to make any assumptions about the true $f$, which completely eliminates the problem seen in a parametric approach. As a result, a nonparametric approach can fit a wider range of potential shapes for $f$. A downside to a nonparametric approach is that you need way more data than a parametric approach because you are not making any assumptions.

## Question 3 (Exercise 10 from Section 2.4)

### Part A

```r
library(MASS)
data(Boston)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```r
dim(Boston)
```

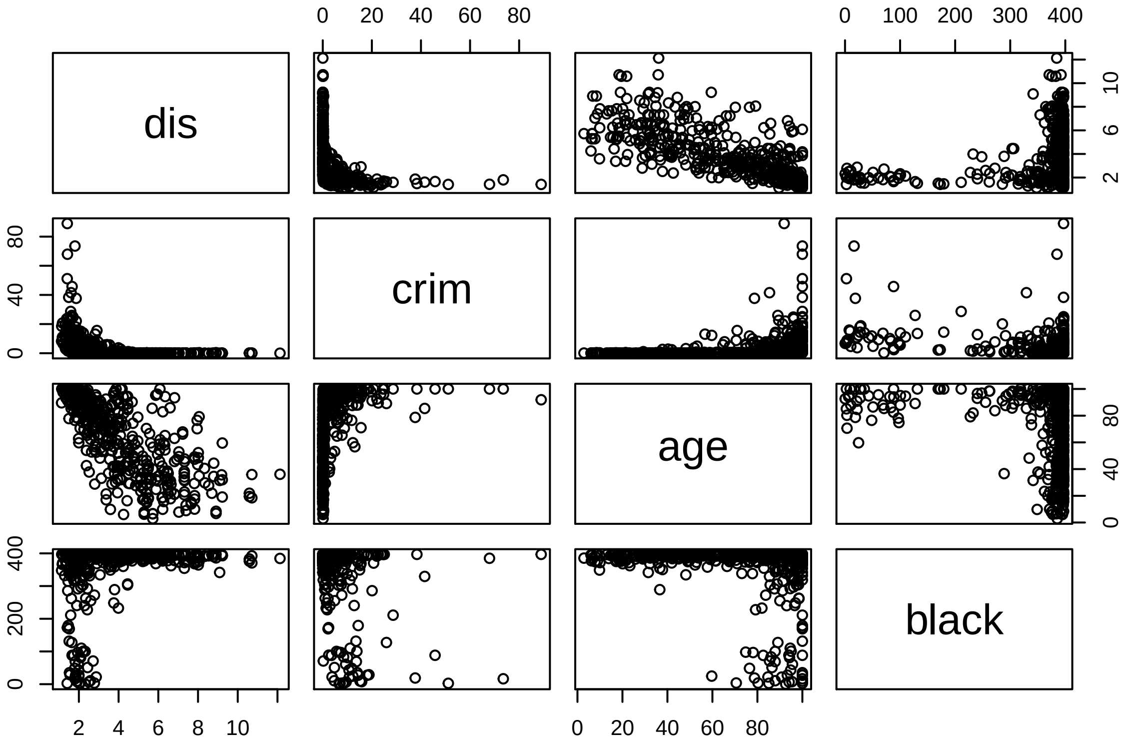
```
## [1] 506  14
```

There are 506 rows and 14 columns in the data set.

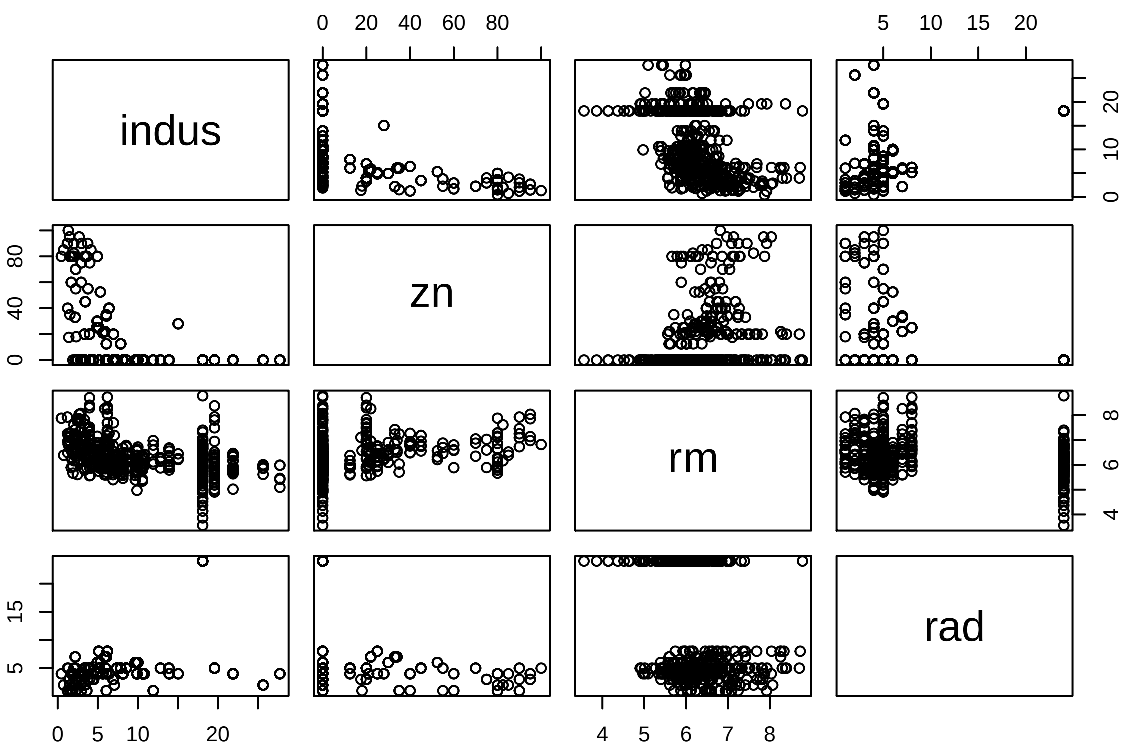
The rows represent a town in Boston. The columns are *crim*, which represents the per capita crime rate, *zn*, which represents the proportion of residential land zoned for lots over 25,000 square feet, *indus*, which represents the proportion of non-retail business acres, *chas*, which represents if the town bounds the Charles River, *nox*, which represents the nitrogen oxide concentration, *rm*, which represents the average rooms per dwelling, *age*, which represents the proportion of owner-occupied units built before 1940, *dis*, which represnets the weighted mean of distances to the five Boston employment centers, *rad*, which represents the index of accessibility to radial highways, *tax*, which represents the full-values property tax per $10,000, *ptratio*, which represents the pupil-teacher ratio, *black*, which is found by the formula $(1000(Bk - 0.63)^2)$ where Bk is the proportion of blacks, *lstat*, which represents the lower status of the population as a percent, and *medv*, which represents the median value of owner-occupied homes in $1000s.

### Part B

```r
plot(Boston[, c("dis", "crim", "age", "black")])
```




```r
plot(Boston[, c("indus", "zn", "rm", "rad")])
```




Above are a couple of pairwise scatterplots for some of the predictors.

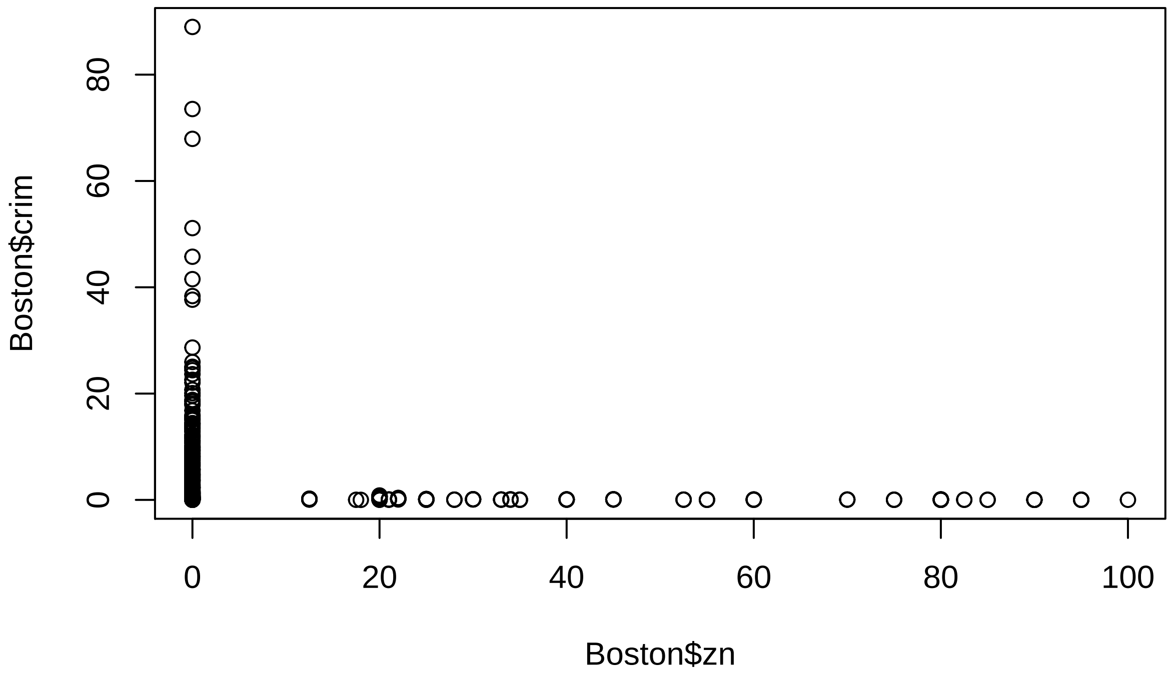
I notice that as *age* increases, *crim* increases slightly and *dis* decreases linearly. Furthermore, *black* is pretty high for every age level, but it is important to note that there is a small cluster of low *black* when *age* is past

80. Also, *crim* is bunched together at low values of *dis*. Similarly, *black* is bunched together at low values of *dis*.

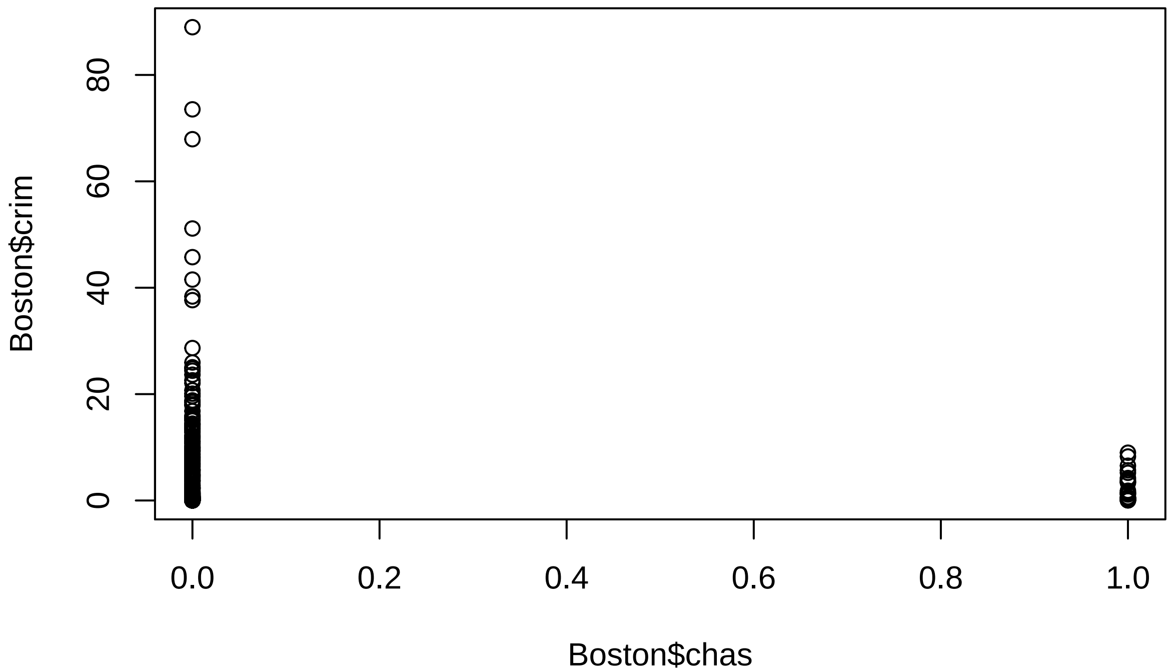
Furthermore, *indus* and *rm* seems to be have a wide range when *zn* is 1, and for other values of *zn*, *indus* and *rm* are relatively low. The values for *rad* seem to be small no matter the value of *indus* and *zn*.

## Part C

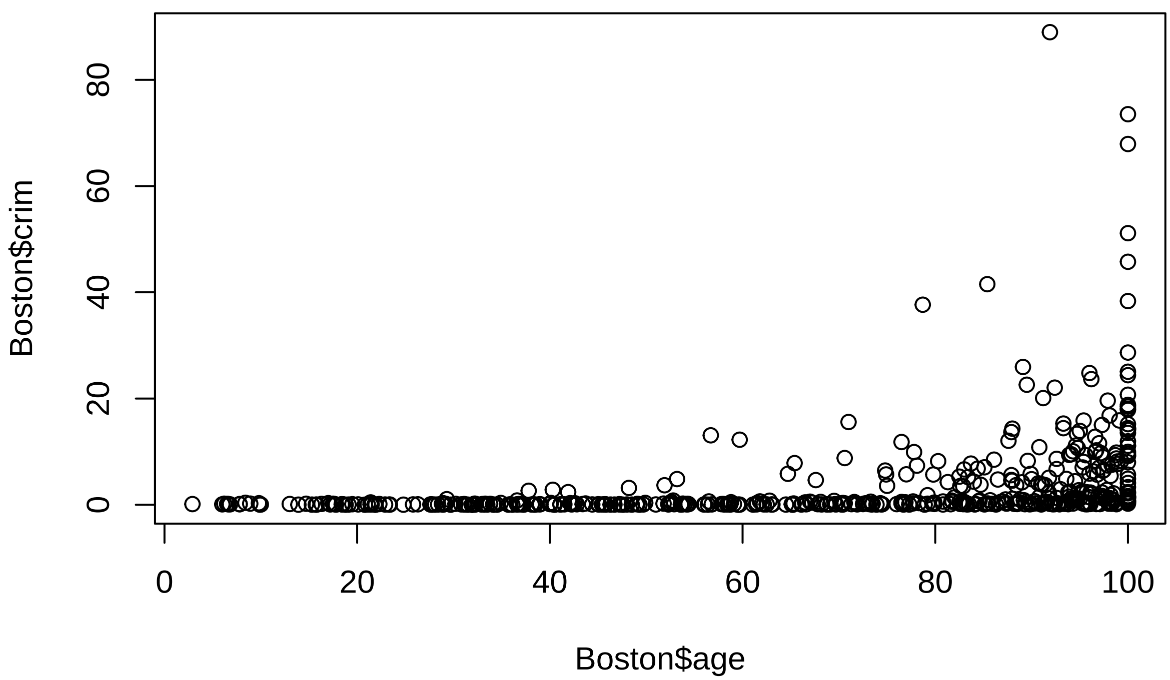
```r
plot(Boston$zn, Boston$crim)
```



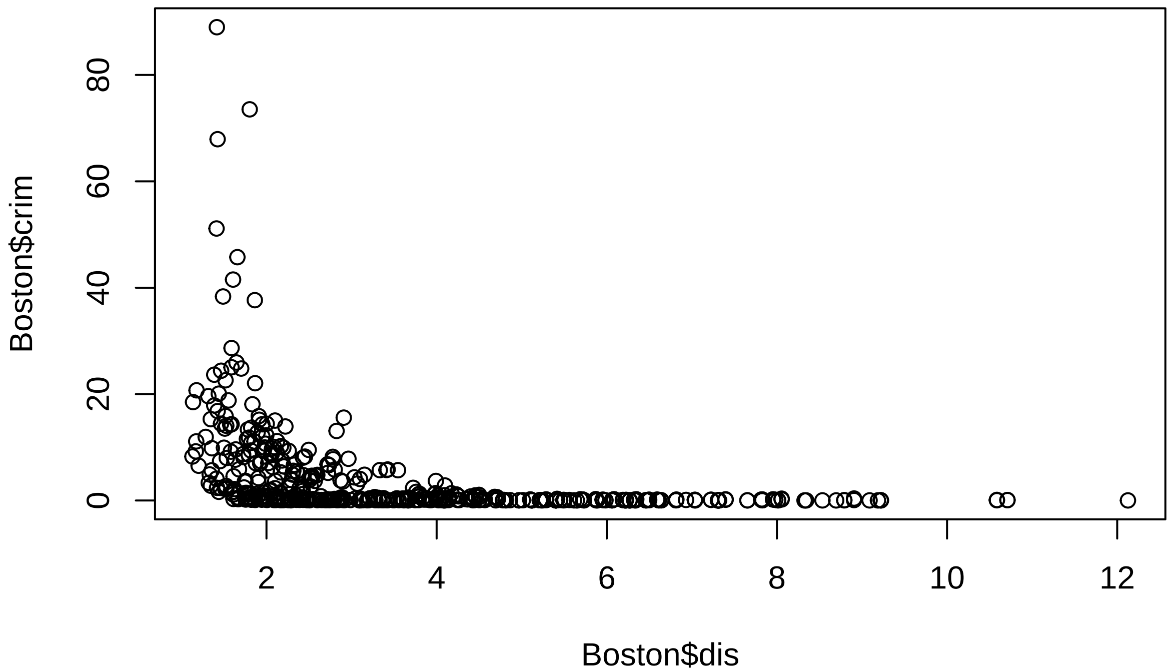

```r
plot(Boston$chas, Boston$crim)
```




```r
plot(Boston$age, Boston$crim)
```



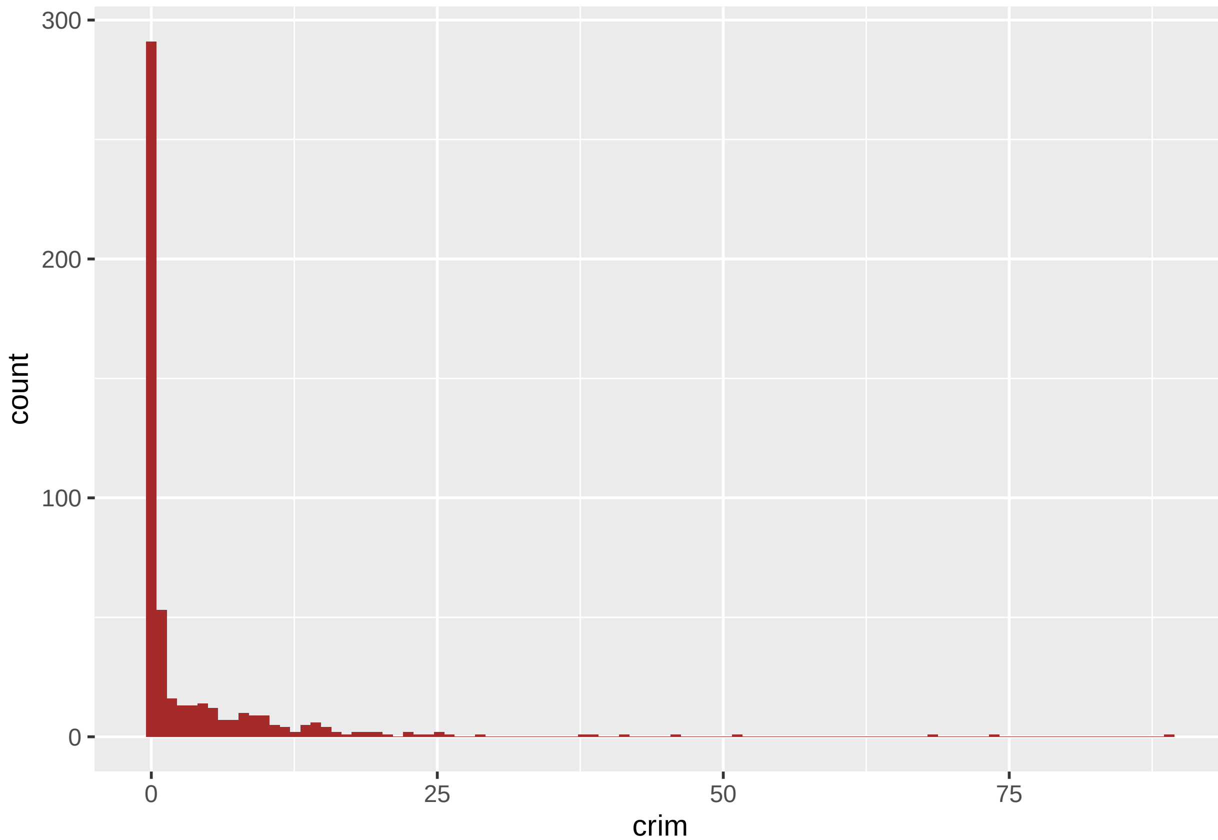

```r
plot(Boston$dis, Boston$crim)
```



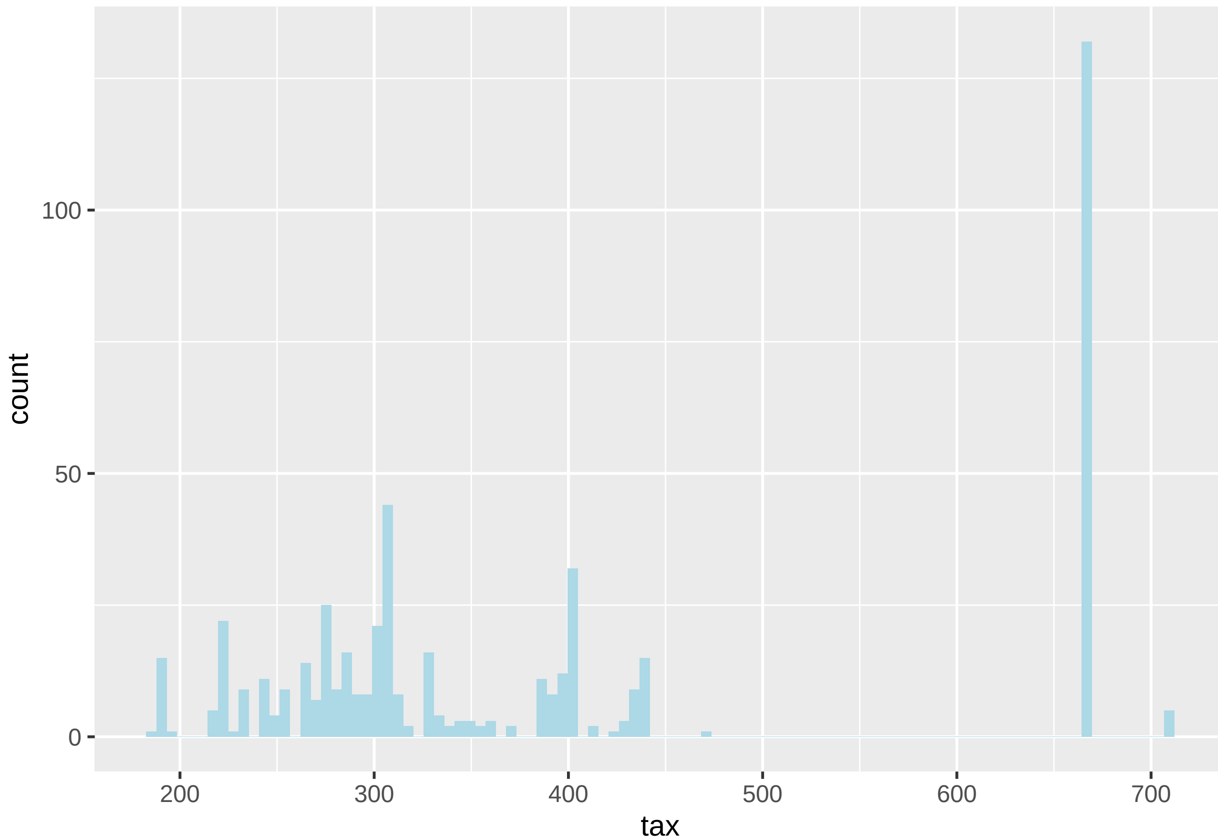

Yes, there are predictors associated with per capita crime rate, as seen in the above plots. When *zn* is low, there is a higher chance of a higher *crim*. When *chas* is 1, *crim* is pretty low, but when *chas* is 0, there is a wider range of *crim*. There seems to be a larger range of values for *crim* past *age* values of 80. When *dis* is low, there is a larger range of values for *crim*.

## Part D

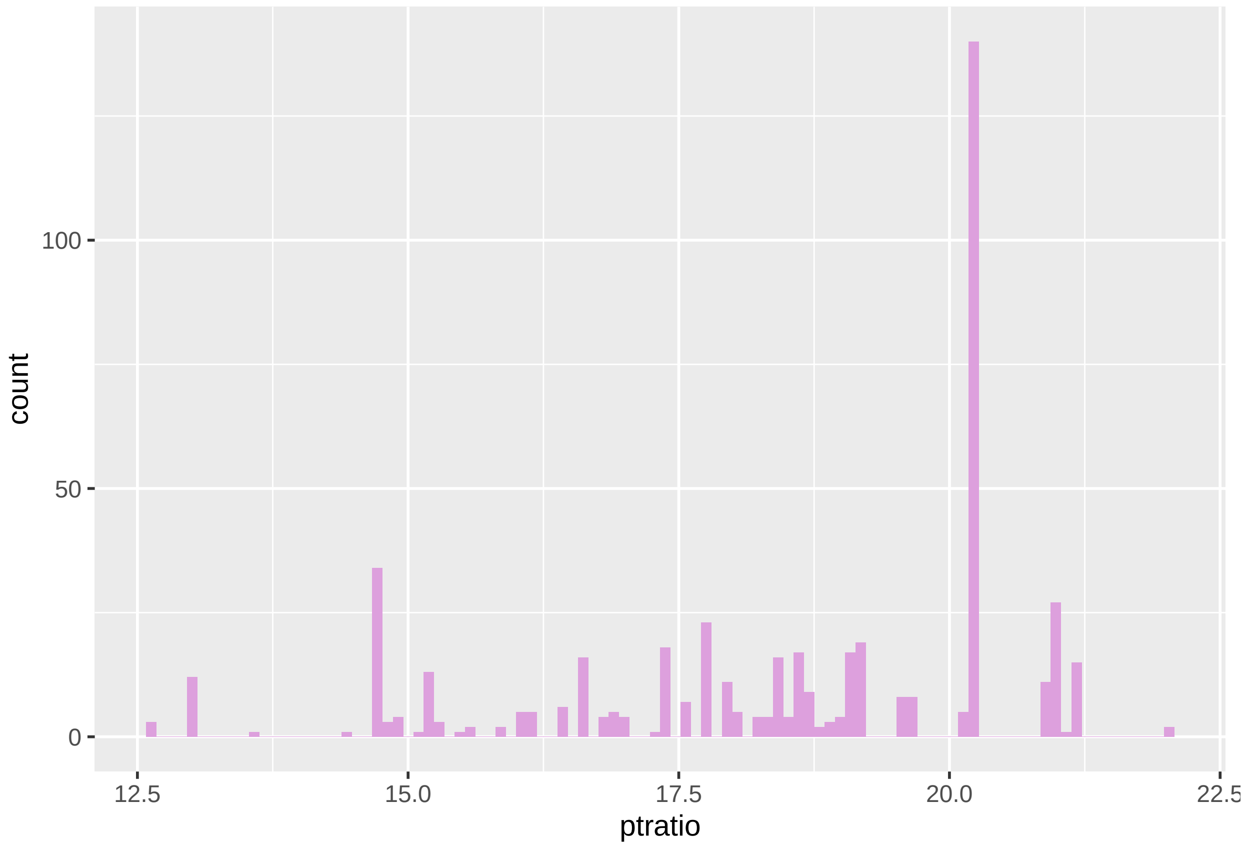
```
library(ggplot2)
ggplot(Boston) +
  aes(x = crim) +
  geom_histogram(bins = 100, fill = "brown")
```




```
ggplot(Boston) +
  aes(x = tax) +
  geom_histogram(bins = 100, fill = "lightblue")
```




```r
ggplot(Boston) +
  aes(x = ptratio) +
  geom_histogram(bins = 100, fill = "plum")
```




```r
data.frame(min = sapply(Boston, min),
           max = sapply(Boston, max))
```

```
##                min      max
## crim      0.00632  88.9762
## zn        0.00000 100.0000
## indus     0.46000  27.7400
## chas      0.00000   1.0000
## nox       0.38500   0.8710
## rm        3.56100   8.7800
## age       2.90000 100.0000
## dis       1.12960  12.1265
## rad       1.00000  24.0000
## tax     187.00000 711.0000
## ptratio  12.60000  22.0000
## black     0.32000 396.9000
## lstat     1.73000  37.9700
## medv      5.00000  50.0000
```

There are not many suburbs with an abnormally high *crim*, but there are some. For *tax*, there is a huge spike around 675, but for the most part, *tax* ranges from around 200 to 450. For *ptratio*, besides the huge spike around 20.5, the values seem to be pretty evenly spread out from 13.5 to 21.

The range of all the predictors is shown in the table above.

## Part E

```r
dim(Boston[Boston$chas == 1,])
```

```
## [1] 35 14
```

35 of the suburbs in this data set bound the Charles River.

## Part F

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

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

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

## Part G

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

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

```r
colMeans(Boston)
```

```
##         crim           zn        indus         chas          nox           rm
##   3.61352356  11.36363636  11.13677866   0.06916996   0.55469506   6.28463439
##          age          dis          rad          tax      ptratio        black
##  68.57490119   3.79504269   9.54940711 408.23715415  18.45553360 356.67403162
##        lstat         medv
##  12.65306324  22.53280632
```

The 399th suburb has the lowest median value of owner occupied homes. As we can see above, the suburb of Boston that has lowest median value of owner occupied homes has: a much higher *crim* than average, much higher *indus* than average, a higher *nox* than average, a lower *rm* than average, much higher *age* than average, a much lower *dis* than average, a much higher *rad* than average, a much higher *tax* than average, about average *ptratio*, a slightly less *black* than average, a much higher *lstat* than average, and a much lower *medv* than average.

## Part H

```r
over_seven <- which(Boston$rm > 7)
length(over_seven)
```

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

```r
over_eight <- which(Boston$rm > 8)
length(over_eight)
```

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

```r
colMeans(Boston[over_eight, ])
```

```
##        crim          zn       indus        chas         nox          rm 
##   0.7187954  13.6153846   7.0784615   0.1538462   0.5392385   8.3485385 
##         age         dis         rad         tax     ptratio       black 
##  71.5384615   3.4301923   7.4615385 325.0769231  16.3615385 385.2107692 
##       lstat        medv 
##   4.3100000  44.2000000
```

```r
colMeans(Boston)
```

```
##         crim           zn        indus         chas          nox           rm 
##   3.61352356  11.36363636  11.13677866   0.06916996   0.55469506   6.28463439 
##          age          dis          rad          tax      ptratio        black 
##  68.57490119   3.79504269   9.54940711 408.23715415  18.45553360 356.67403162 
##        lstat         medv 
##  12.65306324  22.53280632
```

There are 64 suburbs that average more than seven rooms per dwelling, and there are 13 suburbs that average more than eight rooms per dwelling.

For the suburbs averaging more than eight rooms, their average *crim* is lower, *indus* is lower, *rm* is higher, *tax* is lower, *black* is higher, *lstat* is lower, *medv* is higher, when they are compared to the overall average. For all the other categories, their average is about the same as the overall average.