Assignment_10_MAT3373

John Bute

2024-10-02

```
library(MASS)
str(Boston)

## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

?Boston
```

Taking a look at the Help file, we can see that they have the following columns:

- Crim: per capita crime rate per town - zn: proportion of resedential land zoned for lots over 25,000 sq.ft
- indus: proportion of non-retail businesses 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 centers
- rad: index of accessibility to radial highways
- tax: full-value property-tax rate per 10,000$ - ptratio: pupil-teacher ratio by town
- black: 1000(Bk - 0.63)^2 where proportion of blacks by town
- lstat: lower status of the population
- medv: median value of owner occupied homes in $1000s.

In total there are 14 columns and 506 rows. The rows represent a Boston suburb or town.

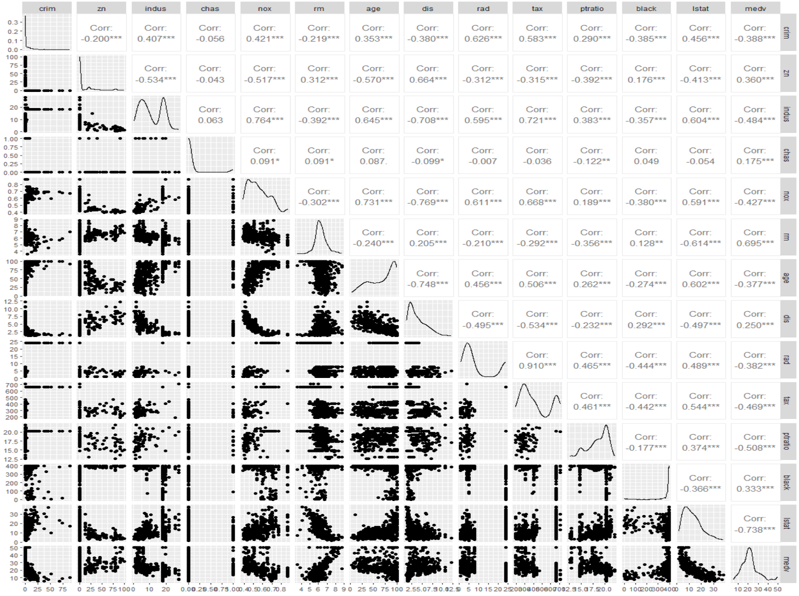
The ggpairs function provides a 14 x 14 scatter plot matrix, along with the correlation for each scatter plot.

```
library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggpairs(Boston) +
  theme(axis.text.y = element_text(size = 8),
        axis.text.x = element_text(size = 8))
```



Crime Rate (CRM):

Crime rate is positively correlated with RAD, and tax, therefore showing that a higher crime rate leads to a bigger access to radial highways, but higher taxes. Meanwhile, crime rate is negatively correlated with larger lot sizes (ZN) and areas with more employment centers (dis).

Proportion of residential land zoned for lots over 25,000 sq.ft. (ZN):

Whenever ZN increases, the proportion of owner-occupied units built prior to 1940 decrease, which intuitively makes sense as bigger plots generally mean newer buildings. Furthermore, the proportion of non-retail business acres per town decreases as lots of 25,000 sq.ft. or more are generally reserved for big retail-like stores. Meanwhile, the weighted mean of distances to five boston employment centers increase as bigger zoned lots increase (which is interesting). Finally, the median value of owner-occupied homes (in $1000s) increase as well, since access to bigger stores generally increase home value.

Proportion of non-retail business acres per town (Indus):

When it comes to proportion of non-retail business acres per town, we notice that it is positively correlated with more nitrogen oxides concentration and full-value property tax rate. Although the nitrogen oxides concentration is suprising, the increase in full-value property tax-rate is not, as the more stores that are available to suburbs, the higher value, and therefore higher tax rate is.

On the other hand, whenever the proportion of non-retail business acres per town is high, median value of houses is low, as there are less amenities available. Furthermore, there seems to be a larger distance from Employment centers when there are less non-retail business acres per town, as employment centers tend to be in areas where there are more employment opportunities available.

Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). (CHAS)

Every correlation seems very close to 0, meaning that aspects of a town do not vary or change depending on if it bounds the Charles river or not.

Nitrogen oxides concentration (parts per 10 million). (NOX)

Whenever nitrogen oxides concentration increases, the accessibility to radial highways and the poportion of units built prior to 1940 increase. This is in part due to cars being a polluting factor, therefore it would make sense to have more highways near areas of higher nitrogen oxides. Furthermore, units built prior to 1940 might not run efficiently or be considered green, which can cause even more pollution.

Furthermore, whenever Nitrogen oxides concentration increases, the weighted mean of distances to five Boston employment centers decreases, which is very interesting.

Average number of rooms per dwelling. (RM)

Whenever the number of rooms per dwelling increases, so does the median value of owner-occupied homes, since they are positively correlated. On the contrary, the lower status population percentage increases whenever number of rooms decrease.

Proportion of owner-occupied units built prior to 1940. (Age)

When it comes to the owner-occupied units built prior to 1940, we found that it is positively correlated with the percentage of lower status population, but negatively correlated to the mean of distancesto five Boston employment centers, which makes sense as employment centers are found within newer areas of towns with more opportunities (which come with newer buildings and installations)

Index of accessibility to radial highways (RAD)

There is a very high positive correlation between property tax rates and accessibility to radial highways, which indicates that areas with better highway access tend to have higher property taxes.

pupil-teacher ratio by town (ptratio)

Ptratio and MEDV are negatively correlated. Thus, towns who has less students per teacher tend to have homes with higher values.

lower status of the population (percent). (LSTAT)

Lower status population percentage tend to be extremely low whenever the median value of houses within a town are higher, which makes sense as lower status population refer to people in the lower class, who do not have access to money to pay for expensive houses.

Q3: Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Positive correlations with CRIM:

NOX (0.421): Higher air pollution is linked to higher crime rates, often in industrial areas.

RAD (0.626): Better highway access correlates with higher crime, likely due to urbanization.

TAX (0.583): Higher property taxes are associated with more crime, possibly urban areas.

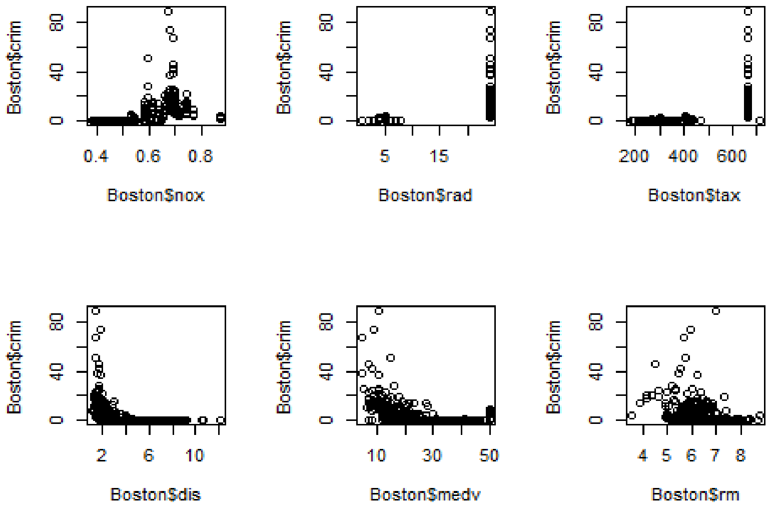
Negative correlations with CRIM:

DIS (-0.380): Areas farther from employment centers tend to have lower crime rates.

MEDV (-0.388): Wealthier areas with higher home values have less crime.

RM (-0.219): Larger homes are linked to lower crime rates

```r
par(mfrow=c(2,3))
plot(Boston$nox, Boston$crim)
plot(Boston$rad, Boston$crim)
plot(Boston$tax, Boston$crim)
plot(Boston$dis, Boston$crim)
plot(Boston$medv, Boston$crim)
plot(Boston$rm, Boston$crim)
```



Q4: Do any of the suburbs of Boston appear to have particularly high crime rates, tax rates, or pupil-teacher ratios? Comment on the range of each predictor.

```
library(ggplot2)
columns <- c("crim", "tax", "ptratio")
ranges <- sapply(Boston[columns], function(x){
  median_value = median(x)
  names(median_value) <- "Median"
  limits <- range(x)
  names(limits) <- c("Lower Limit", "Higher Limit")
  return(c(median_value, limits))
})
ranges

##                  crim tax ptratio
## Median        0.25651 330   19.05
## Lower Limit   0.00632 187   12.60
## Higher Limit 88.97620 711   22.00
```

Some towns have extremely high crime rates, as the highest range is 88, but the median is 0.25, meaning that certain towns and suburbs exceed by almost 350% increase in crime rate. Meanwhile, tax rate seem to have a higher range, but lower median, meaning that there are a smaller group of towns (compared to those with less property tax) who's property taxes are higher, which can be attributed to neighborhoods with nicer houses (which are not seen as often). Finally, the pupil-teacher ratio seems to be reasonable, except for the lower limit being at 12.60, indicating that it is less likely that there is a town in Boston who has the resources necessary to furnish an education environment where students receive more attention from teachers.

Q5: How Many Suburbs Bound the Charles River?

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

## [1] 35
```

Q6: What is the median pupil-teacher ratio among the towns in this dataset? It is 19.05 (as seen in the dataset above)

Q7: Which suburb of Boston has the lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors?

```
library(ggplot2)
min_medv <- Boston[Boston$medv == min(Boston$medv), ]
min_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    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
```

CRIM is extremely high at 38.35 and 67.92, far above the median of 0.26, indicating a high-crime area.
ZN is 0, meaning no large residential lots, which is typical as the median is also 0.
INDUS is 18.10, suggesting a high proportion of non-retail business land, likely lowering

home values.
NOX is 0.693, indicating poor air quality due to industrial activity or highways.
RM values of 5.453 and 5.683 show smaller homes compared to the median of 6.21.
AGE is 100%, meaning all homes are older, built before 1940.
DIS values of 1.49 and 1.43 indicate close proximity to employment centers, though industrial zones may reduce attractiveness.
RAD is the maximum value of 24, showing very high highway access, which may increase crime and pollution.
TAX is 666, much higher than the median of 330, making it a high-tax area.
PTRATIO is 20.2, slightly above the median, indicating larger class sizes.
LSTAT is 30.59, significantly higher than the median, correlating with lower home values and higher crime rates.

Q8

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

## [1] 64

print(sum(Boston$rm > 8))

## [1] 13

eight_rm <- Boston[Boston$rm > 8, ]

ranges_rm <- sapply(eight_rm, function(x){
  median_value = median(x)
  names(median_value) <- "Median"
  limits <- range(x)
  names(limits) <- c("Lower Limit", "Higher Limit")
  return(c(median_value, limits))
})
ranges_rm

##                  crim zn indus chas    nox    rm  age    dis rad tax ptratio
## Median       0.52014  0  6.20    0 0.5070 8.297 78.3 2.8944   7 307    17.4
## Lower Limit  0.02009  0  2.68    0 0.4161 8.034  8.4 1.8010   2 224    13.0
## Higher Limit 3.47428 95 19.58    1 0.7180 8.780 93.9 8.9067  24 666    20.2
##                black lstat medv
## Median       386.86  4.14 48.3
## Lower Limit  354.55  2.47 21.9
## Higher Limit 396.90  7.44 50.0
```

If we were to compare houses with eight rooms versus every house, we find that the crime rate is extremely low, with less proportion of lower status population, and a higher median value of homes within the suburb/town.