Assignment on predictive analysis

Chandrani Sengupta

2026-01-20

PROBLEM SET 1

Download “Boston” housing data from MASS library in R. Complete the task given below and submit the report using R markdown. You need to copy each question as well.

```
library(MASS)

## Warning: package 'MASS' was built under R version 4.5.2

x=Boston
head(x)

##      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
```

1. Report the “class” of the data set. How many rows and columns are in this data set? What do the rows and columns represent?

```
class(Boston)

## [1] "data.frame"
```

The Boston dataset is of dataype data.frame

```
dim(Boston)

## [1] 506  14
```

The Boston dataset has 506 rows and 14 columns

```
str(x)

## '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 ...
```

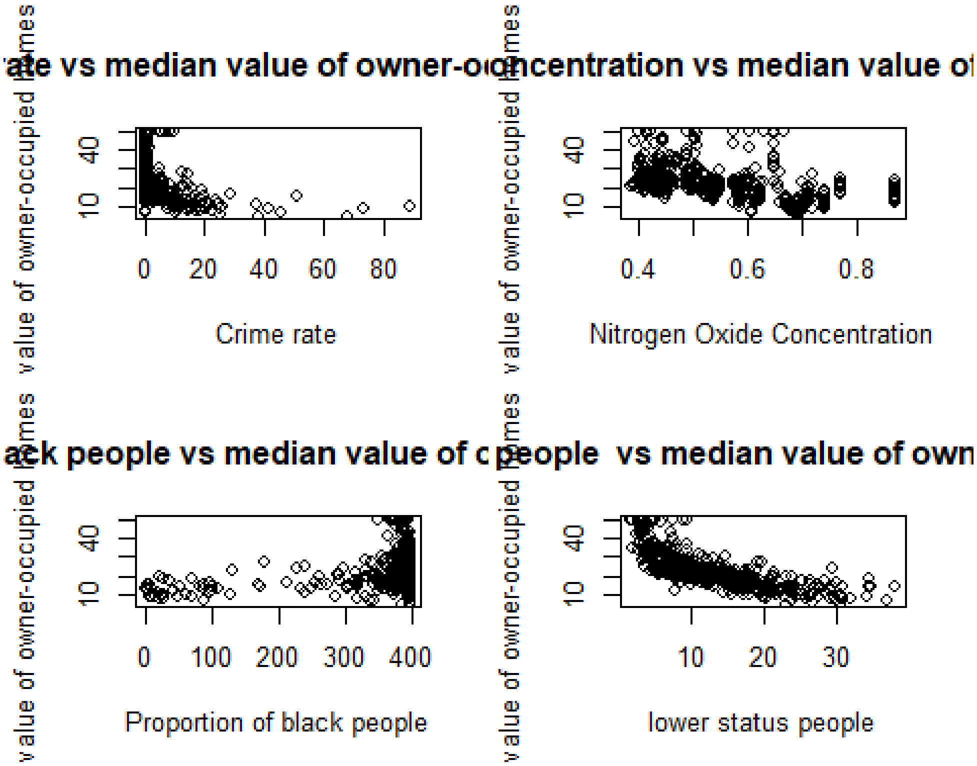
Each row represents a suburb of Boston . The columns represent the variables

2. Create a smaller data set with the variables median value of owner-occupied homes, per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population. Choosing median value of owner occupied homes as the response and the rest as the pre- dictors, make scatter plots of the response versus each predictor. Present the scatter plots in different panels of the same graph. Comment on your findings.

```
x=data.frame(Boston$medv,Boston$crim,Boston$nox,Boston$black,Boston$lstat)
head(x)

##   Boston.medv Boston.crim Boston.nox Boston.black Boston.lstat
## 1        24.0     0.00632      0.538       396.90         4.98
## 2        21.6     0.02731      0.469       396.90         9.14
## 3        34.7     0.02729      0.469       392.83         4.03
## 4        33.4     0.03237      0.458       394.63         2.94
## 5        36.2     0.06905      0.458       396.90         5.33
## 6        28.7     0.02985      0.458       394.12         5.21

par(mfrow=c(2,2))
plot(x$Boston.crim,x$Boston.medv,xlab="Crime rate",ylab="median value of
owner-occupied homes in $1000s",main="Scatterplot of Crime rate vs median
value of owner-occupied homes in $1000s")
plot(x$Boston.nox,x$Boston.medv,xlab="Nitrogen Oxide
Concentration",ylab="median value of owner-occupied homes in
$1000s",main="Scatterplot of nitrogen oxides concentration vs median value of
owner-occupied homes in $1000s ")
plot(x$Boston.black,x$Boston.medv,xlab="Proportion of black
people",ylab="median value of owner-occupied homes in
$1000s",main="Scatterplot of proportion of black people vs median value of
owner-occupied homes in $1000s ")
plot(x$Boston.lstat,x$Boston.medv,xlab="lower status people",ylab="median
value of owner-occupied homes in $1000s",main="Scatterplot of lower status
people  vs median value of owner-occupied homes in $1000s ")
```



From Scatterplot of crime rate vs median values we thus see that there is a negative strong linear relationship between them indicating higher crime rates are associated with lower house prices

From Scatterplot of Nitrogen Oxide Concentration vs median values we thus see that there is a negative strong linear relationship between them indicating higher pollution is associated with lower house prices

From Scatterplot of proportion of black people vs median values we thus see that there is a positive linear relationship between them indicating higher the proportion of black people,higher are the house prices

From Scatterplot of lower status of population vs median values we thus see that there is a negative strong linear relationship between them indicating more number of lower status people in the popluation are associated with lower house prices

3. Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors mentioned in (2), for that suburb. How do these values compare to the overall ranges for those pre- dictors? Comment on your findings. Hint: Mention which percentile these values belong to.

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

## [1] 399

low_medv_suburb=Boston[min, c("medv", "crim", "nox", "black", "lstat")]
low_medv_suburb

##     medv    crim   nox black lstat
## 399    5 38.3518 0.693 396.9 30.59
```

The 399 th suburb has the lowest median . The values of the suburb for the predictors in 2 are given above

```
percentile = function(x, value) {
  mean(x <= value) * 100
}


percentiles = sapply(
  c("crim", "nox", "black", "lstat"),
  function(v) percentile(Boston[[v]], low_medv_suburb[[v]])
)

percentiles

##      crim       nox     black     lstat
##  98.81423  85.77075 100.00000  97.82609
```

The lowest median value (medv = 5) corresponds to one (or more) suburb(s).

For this suburb:

Crime rate (crim) lies in a very high percentile, indicating unusually high crime.

Nitric oxide concentration (nox) is also in the upper percentile, showing high pollution.

Lower status population (lstat) lies in the highest percentile, indicating socio-economic disadvantage.

Black proportion (black) is relatively low compared to its maximum, but interpretation should be cautious.

The suburb with the lowest house prices also ranks very poorly in crime, pollution, and socio-economic indicators, explaining the depressed housing value.