Predictive Analysis

Problem Set 1: An Introduction

ATRIJO ROY

2026-01-19

Problem Statement

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.

Loading the data

```
rm(list=ls())
library(MASS)

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

data=Boston
attach(data)
head(data)

##      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
```

Problem 1

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

```r
class(data) #To find the "class" of the dataset
```

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

```r
dim(data) #To find the number of rows and the columns present in the data
```

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

We can see there are 506 rows and 14 columns in the dataset.

Each of the 506 rows represent each of the suburbs of the town Boston and the 14 columns represent fourteen variables describing each suburb.

Problem 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 predictors, make scatter plots of the response versus each predictor. Present the scatter plots in different panels of the same graph. Comment on your findings.

```r
df=data.frame(medv,crim,black,nox,lstat) #partitioning the data into smaller data
head(df)
```

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

```r
#scatter plots
par(mfrow=c(2,2))
plot(crim,medv,main="Scatterplot of medv against crim")
plot(black,medv,main="Scatterplot of medv against black")
plot(nox,medv,main="Scatterplot of medv against nox")
plot(lstat,medv,main="Scatterplot of medv against lstat")
```

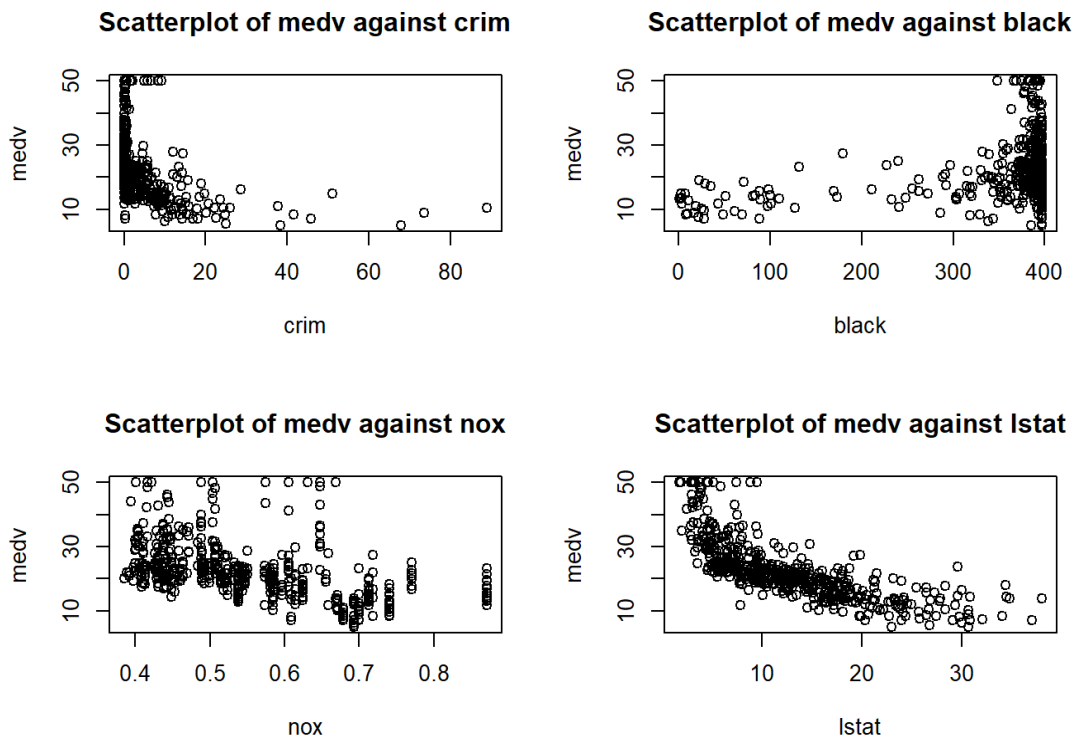

Figure 1: Scatterplot of medv against crim, black, nox and lstat

Interpretation:

medv vs crim -

There is a clear negative relationship. As crime rate increases, median house value generally decreases. Most high medv values are concentrated at very low crime rates, while high crime areas almost always have low medv. The relationship is nonlinear, with a sharp drop in medv even for moderate increases in crim, and a few extreme crime outliers.

medv vs black -

The points are highly scattered, with no clear linear pattern. While higher values of black are often associated with slightly higher median house values, the spread is large, indicating low strength of association. Overall, black has a weak and noisy relationship with medv compared to variables like lstat or nox.

medv vs nox -

There is a clear negative association. As nitrogen oxide concentration increases (worse air quality), median house value decreases. The relationship looks nonlinear, with medv dropping sharply beyond moderate nox levels. This suggests environmental quality strongly affects housing prices.

medv vs lstat -

This is the strongest and clearest relationship among the four. There is a strong negative, nonlinear relationship: as the percentage of lower-status population increases, median house value decreases sharply. High medv values occur almost exclusively at low lstat levels, while high lstat areas have consistently low medv.

Problem 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 predictors? Comment on your findings. Hint: Mention which percentile these values belong to.

```
#To find the lowest median value and values of the corresponding predictors
lowest.medv=df[medv==min(medv), ]
lowest.medv
```

```
##     medv    crim  black   nox lstat
## 399    5 38.3518 396.90 0.693 30.59
## 406    5 67.9208 384.97 0.693 22.98
```

The lowest median value comes out as 5000 dollars. And suburb 399 and suburb 406 have the lowest median value of owner-occupied homes which is 5000 dollars.

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

#For suburb 399
sapply(c("crim","nox","lstat","black"), function(v)
  percentile(df[[v]], lowest.medv[[v]][1])
)
```

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

```
#For suburb 406
sapply(c("crim","nox","lstat","black"), function(v)
  percentile(df[[v]], lowest.medv[[v]][2])
)
```

```
##     crim      nox    lstat    black
## 99.60474 85.77075 89.92095 34.98024
```

Comment:

Two suburbs share the lowest median house value (medv = 5). Both fall in the extreme upper percentiles for crime (≈99th) and high percentiles for nitrogen oxide concentration (nox ≈86th). The lower-status population (lstat) is also very high, ranging from about the 90th to 98th percentile. In contrast, black varies widely between the two suburbs (≈35th to 100th percentile), showing no consistent pattern.

Problem 4

Does any suburb of Boston stand out for having notably high crime rates, tax rates, or pupil–teacher ratios? Hint: Use a boxplot to detect any outliers. If so, identify the suburbs that show the outlier values.

```
#Boxplots
par(mfrow = c(1,3))
crim.out=boxplot(crim,main="Boxplot of Crime rate (crim)")$out
tax.out=boxplot(tax,main="Boxplot of Tax rate (tax)")$out
pt.out=boxplot(ptratio,main="Boxplot of Pupil-Teacher ratio (ptratio)")$out
```

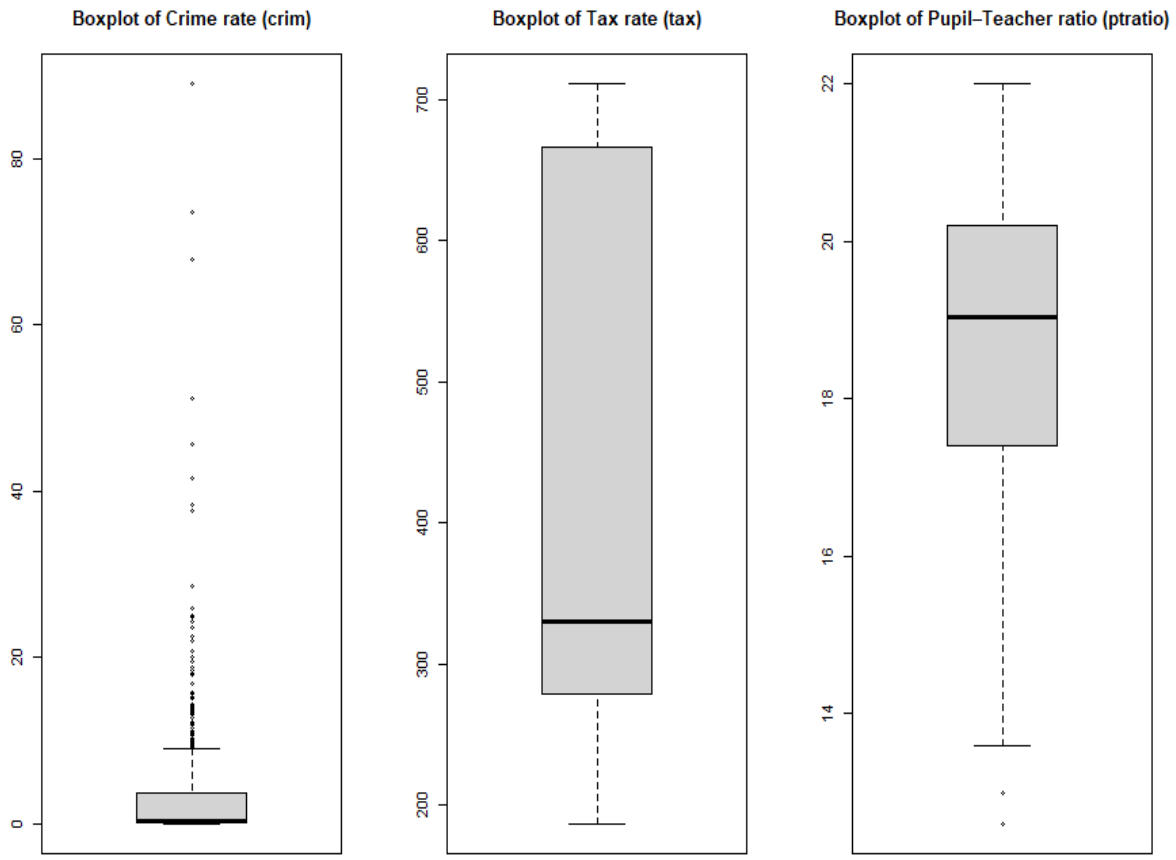

Figure 2: Boxplots of crim, tax and ptratio

Comment:

We can observe the presence of outliers for Crime rate and Pupil-Teacher ratio but not for the tax rates from figure 2.

Crime rate -

Yes, several suburbs clearly stand out as extreme outliers with very high crime rates. These values lie far beyond the upper whisker of the boxplot, indicating crime levels much higher than the majority of Boston suburbs.

Pupil-Teacher ratio -

The outliers occur on the lower end, indicating a few suburbs with unusually low pupil–teacher ratio compared to the rest.

Now we find out the suburbs which contains the respective outlier values.

```
#To find out which suburbs show the outlier values
#Suburbs with outliers in Pupil-Teacher Ratio
which(ptratio %in% pt.out)
```

```
##  [1] 197 198 199 258 259 260 261 262 263 264 265 266 267 268 269
```

```
#Suburbs with outliers in Crime Rate
which(crim %in% crim.out)
```

```
##  [1] 368 372 374 375 376 377 378 379 380 381 382 383 385 386 387 388 389
393 395
## [20] 399 400 401 402 403 404 405 406 407 408 410 411 412 413 414 415 416
417 418
## [39] 419 420 421 423 426 427 428 430 432 435 436 437 438 439 440 441 442
444 445
## [58] 446 448 449 455 469 470 478 479 480
```