PREDICTIVE PRACTICAL PS1

DEBKANTA GHOSH

2026-01-19

Problem 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?

```
library(MASS)
attach(Boston)
dim(Boston)
```

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

```
class(Boston)
```

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

Ans:-

Class of Boston is “data frame”.There are 506 rows and 14 columns.Rows represents 506 different neighborhood areas in Boston metropolitan area.Columns represents variables(13 independent variables/predictors and 1 study varible/response(medv-median value of owner-occupied homes in $1000s)).

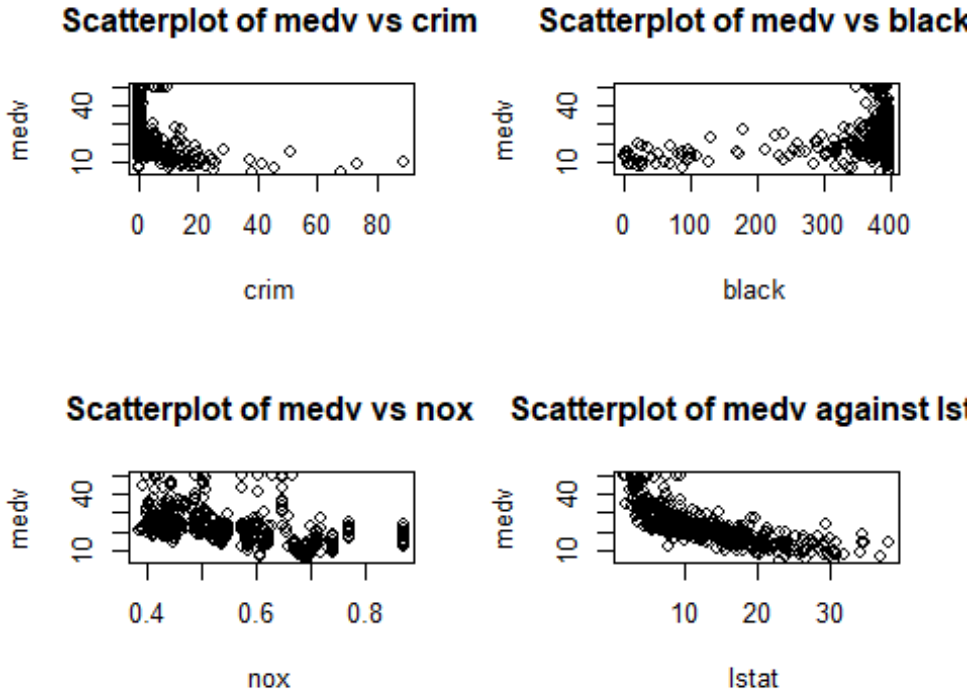
Problem 2

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.

```
df=Boston
df1=data.frame(medv,crim,nox,black,lstat)
head(df1)
```

```
##   medv    crim   nox  black 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(crim,medv,main="Scatterplot of medv vs crim")
plot(black,medv,main="Scatterplot of medv vs black")
plot(nox,medv,main="Scatterplot of medv vs nox")
plot(lstat,medv,main="Scatterplot of medv against lstat")
```



comment on finding:-

medv vs crim:- From the first plot we can clearly say that there is a negetive relation between medv and crim.As crim in increases medc=v decreases.

medv vs black:- Weak positive and nonlinear association between medv and black.Higher values of black are generally associated with higher house values, but the relationship is not strong. Black alone does not strongly explain house value variation.

medv vs nox:-There is negetive relation between medv and nox.As nitrogen oxide concentration increases medv decreases.Pollution negatively impacts housing prices.

medv vs lstat:- Clearly we can see from the fourth scatter plot there is strong negetive relation between lstat and medv.Neighborhoods with a higher percentage of lower-status population tend to have much lower house values.lstat is one of the strongest predictors of median house value.

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.

```
medv1=df1[medv==min(medv), ]
medv1
```

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

Suburb 399 and suburb 406 have the lowest median value of owner-occupied homes which is 5000 dollars.

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

sapply(c("crim","nox","lstat","black"), function(v)
 percentile(df[[v]], medv1[[v]][1])
)
```

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

```
sapply(c("crim","nox","lstat","black"), function(v)
 percentile(df[[v]], medv1[[v]][2])
)
```

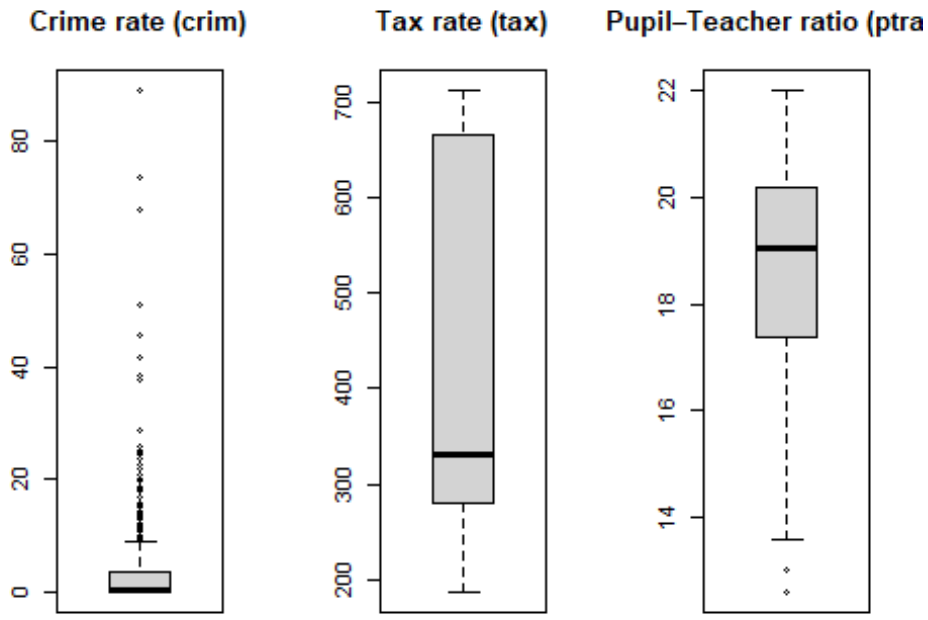
```
##     crim      nox    lstat    black
## 99.60474 85.77075 89.92095 34.98024
```

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.

```r
par(mfrow = c(1,3))
boxplot(crim,main="Crime rate (crim)")
boxplot(tax,main="Tax rate (tax)")
boxplot(ptratio,main="Pupil-Teacher ratio (ptratio)")
```



Comment:-

We can see there exist outlier for crim and pupil teacher ratio but not for tax rate.