Predictive practical set 1

Sayan Ganguly,Roll:714

2026-01-21

importing and viewing first 6 rows of the dataset

```
library(MASS)
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
```

Question-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?

Answer:

```
class(Boston)

## [1] "data.frame"

dim(Boston)

## [1] 506  14
```

So there are 506 rows and 14 columns in this dataset in which the rows represent the observations,i.e different suburbs in Boston.The columns represent the variables on which data is observed.

Question-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.

Answer:

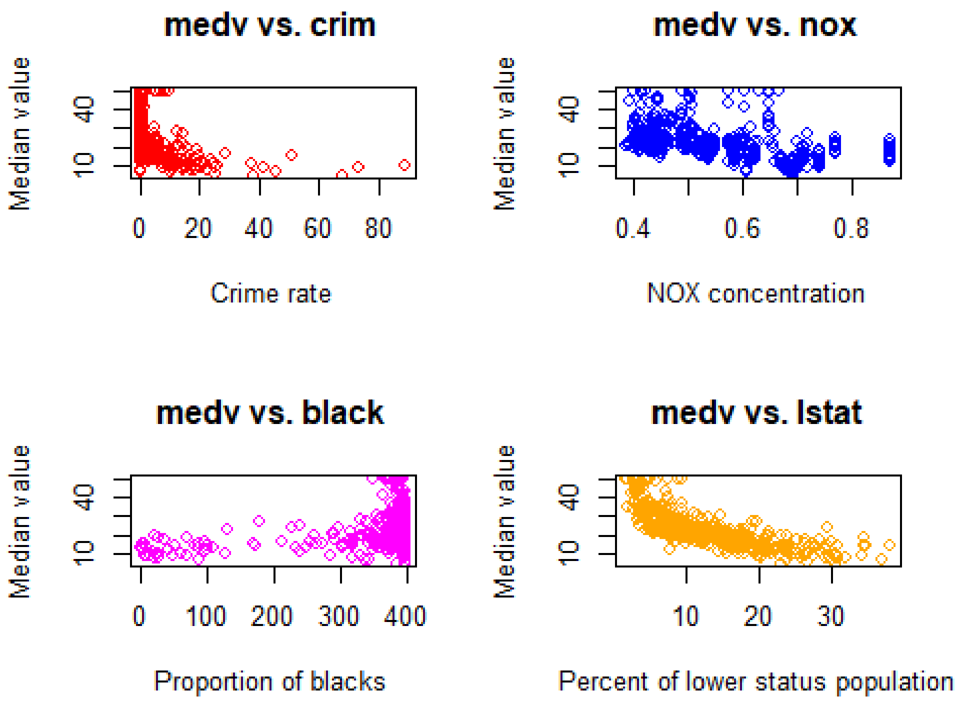
Smaller dataset with the given variables

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

##   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
```

scatter plots

```
par(mfrow=c(2,2))
plot(df$Boston.crim,df$Boston.medv,main = "medv vs. crim",xlab = "Crime
rate",ylab = "Median value",col="red",)
plot(df$Boston.nox,df$Boston.medv,main = "medv vs. nox",xlab = "NOX
concentration",ylab = "Median value",col ="blue")
plot(df$Boston.black,df$Boston.medv,main = "medv vs. black",xlab =
"Proportion of blacks",ylab = "Median value",col = "magenta")
plot(df$Boston.lstat,df$Boston.medv,main = "medv vs. lstat",xlab = "Percent
of lower status population",ylab = "Median value",col = "orange")
```



Comment: 1)Between 0-20 crime rate most of the observations are found and as the crime rate gradually increase the median prices tend to decrease which is quite obvious.So there is a negative correlation between the two variables.

2)As the nitrogen oxide concentration increases the median value of the prices decrease om the average.So there is a negative association between the two variables.There are a few clusters formed in the values of the median prices for different concentrations of nox.For NOX concentration values between 0.4 and 0.6 we observe that most of the price values are between 15000 and 35000 dollars.

3)We see that proportion of black people for most of the observations is very high(close to 400).The median values of homes are highly concentrated between 10000 and 35000 dollars.Since most observations are observed against a high number of black people we cannot make a suitable comment on the nature of association.For proportion of black people less than 300 we see that the prices are somewhat evenly distributed,i.e there is no such pattern.

4)We observe that as the percent of lower status population increases,median price decreases on the average and so we can conclude that there is a negative correlation between the two variables.

Question-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.

Answer:

lowest median value of owner-occupied homes

```
arr=c(Boston$medv)
which(arr==min(arr))

## [1] 399 406
```

So the lowest median value of owner occupied homes occur for the 399th and the 406th suburb.

```
Boston[c(399,406),c(1,5,12,13)]

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

Calculation of the percentiles

```
crim=c(sort(Boston$crim))
nox=c(sort(Boston$nox))
black=c(sort(Boston$black))
lstat=c(sort(Boston$lstat))
p_crim=c(length(which(crim<=38.3518))/506*100,length(which(crim<=67.9208))/50
6*100)
p_nox=length(which(nox<=0.693))/506*100
p_black=c(length(which(black<=396.90))/506*100,length(which(black<=384.97))/5
06*100)
p_lstat=c(length(which(lstat<=30.59))/506*100,length(which(lstat<=22.98))/506
*100)
p_nox

## [1] 85.77075

p_crim

## [1] 98.81423 99.60474

p_black

## [1] 100.00000  34.98024

p_lstat

## [1] 97.82609 89.92095
```

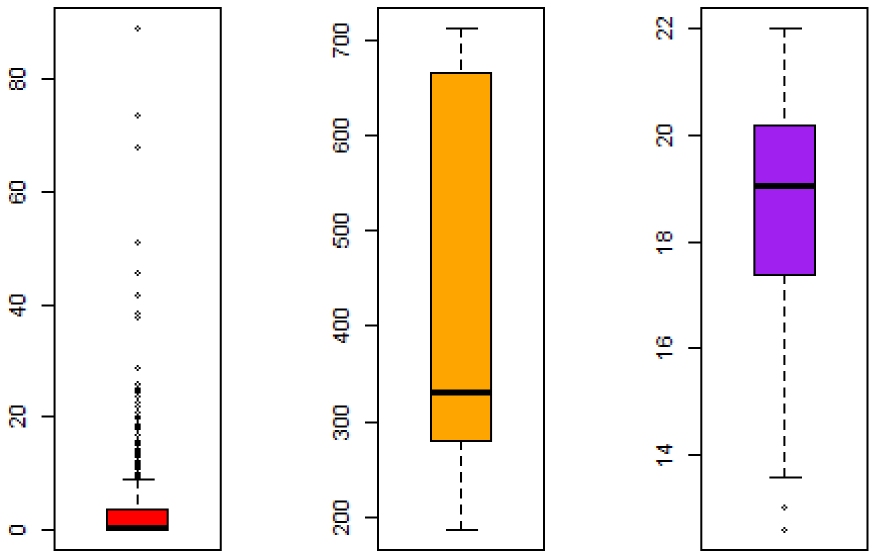
Comment: We observe that the values of the 4 variables under consideration corresponding to the lowest median price belong to a high percentile.So certainly the high values of crime rate,nox concentration,black population and percent of lower status of population contribute to the reduction of the median price.

Question-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.

Answer:

Box Plots

```
par(mfrow=c(1,3))
boxplot(Boston$crim,col="red")
boxplot(Boston$tax,col="orange")
boxplot(Boston$ptratio,col="purple")
```



Now we find out the suburbs which show the outlier values for each of the three variables seperately.

Suburbs showing outlier values for crim

```
crime_out=c(boxplot.stats(Boston$crim)$out)
outlier_crim=c()
for(i in 1:506)
  for(j in 1:length(crime_out))
```

```
    if(Boston$crim[i]==crime_out[j])
      outlier_crim=append(outlier_crim,i)
unique(outlier_crim)
```

```
##  [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
```

Suburbs showing outlier values for tax

```
print(boxplot.stats(Boston$tax)$out)
```

```
## numeric(0)
```

So there are no outlier values for tax.

Suburbs showing outlier values for ptratio

```
ptratio_out=c(boxplot.stats(Boston$ptratio)$out)
outlier_ptratio=c()
for(i in 1:506)
  for(j in 1:length(ptratio_out))
    if(Boston$ptratio[i]==ptratio_out[j])
      outlier_ptratio=append(outlier_ptratio,i)
unique(outlier_ptratio)
```

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

Comment: So there are a lot of outliers for “crim” and quite a few for “ptratio” also.