Boston Problem Set 1

Srijita Dey, Roll-720

2026-01-22

Predictive Analysis Problem Set 1

Download Boston housing data from MASS library in R.

```
library("MASS")
attach(Boston)
```

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"

nrow(Boston)

## [1] 506

ncol(Boston)

## [1] 14
```

The "class" of the dataset is Data Frame

Number of rows: 506

Number of columns: 14

The columns are the variables and the rows are the suburbs

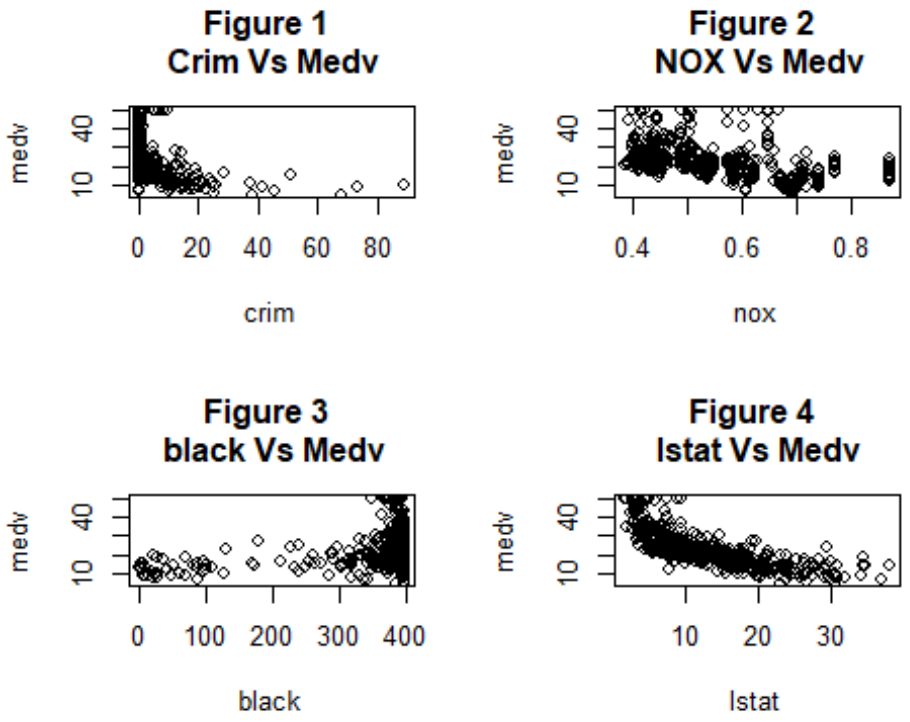
2. Create a smaller data set with the variables median value of owner-occupied homes, per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population. Choosing median value of owner occupied homes as the response and the rest as the predictors,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

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

##   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="Figure 1 \n Crim Vs Medv")
plot(nox,medv,main="Figure 2 \n NOX Vs Medv")
plot(black,medv,main="Figure 3 \n black Vs Medv")
plot(lstat,medv,main="Figure 4 \n lstat Vs Medv")
```



Comment:

Figure 1: Crim Vs Medv

There is a negative association between Crime rate and Median values.As the Crime rate increases, median values of the houses decrease

Figure 2: NOX vs Medv

When NOX level is ≥0.7, house values becomes low. So, as the nitrogen oxide level increases upto a certain level, the prices of the houses decrease.

Figure 3: black vs Medv

Here the values spread widely at higher levels of black, including high medv.

Figure 4: lstat vs Medv

A strong negative relationship; as lstat increases, medv decreases sharply. Lower-status neighborhoods have low home values.

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.

```
min_index = which.min(Boston$medv)
Boston[min_index, ]
```

```
##        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    20.2 396.9 30.59
##     medv
## 399    5
```

```
suburb=Boston[min_index, c("medv", "crim", "nox", "black", "lstat")]
suburb
```

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

```
percentile=function(x,value) ecdf(x)(value)*100

data.frame(
  Variable=c("crim", "nox", "black", "lstat"),
  Value=as.numeric(suburb[c("crim", "nox", "black", "lstat")]),
  Percentile=c(percentile(Boston$crim, suburb$crim), percentile(Boston$nox,
suburb$nox),percentile(Boston$black, suburb$black),percentile(Boston$lstat,
suburb$lstat)))
```

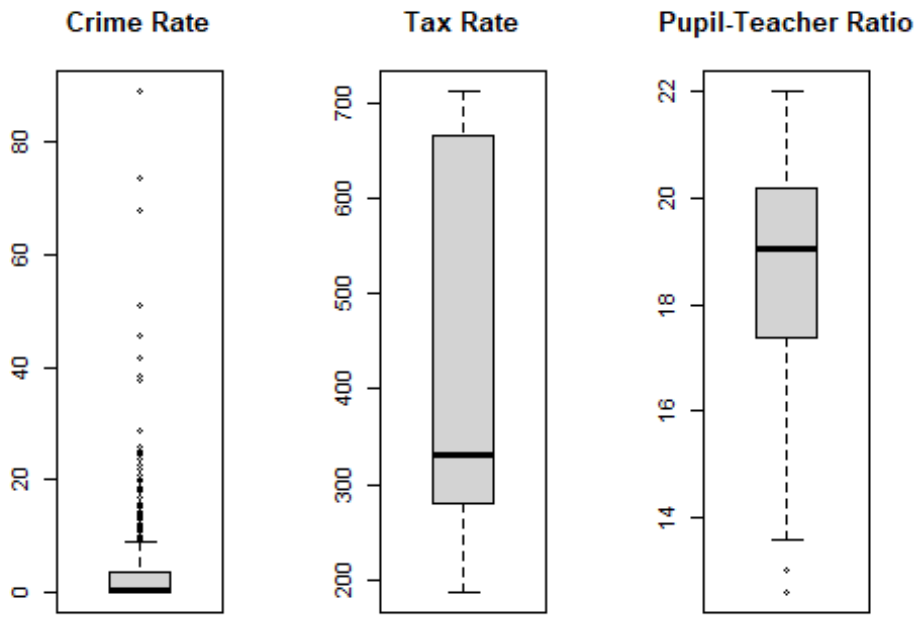
```
##   Variable    Value Percentile
## 1     crim  38.3518   98.81423
## 2      nox   0.6930   85.77075
## 3    black 396.9000  100.00000
## 4    lstat  30.5900   97.82609
```

Comment:

The suburb with the lowest median home value is characterized by extremely high black, crime, high pollution, and very high lower-status population.

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.

```
par(mfrow=c(1,3))
boxplot(Boston$crim, main="Crime Rate")
boxplot(Boston$tax, main="Tax Rate")
boxplot(Boston$ptratio, main="Pupil-Teacher Ratio")
```



Comment:

Using boxplots, several suburbs are detected as outliers for crime rate and pupil–teacher ratios.

```
get_outliers=function(x) {
  bp= boxplot.stats(x)
  which(x %in% bp$out)
}
```

The suburbs with Crime rate outlier values:

```
crim_outliers=get_outliers(Boston$crim);crim_outliers
```

```
##  [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
```

The suburbs with Pupil-teacher ratio outlier values:

```
ptratio_outliers=get_outliers(Boston$ptratio);ptratio_outliers
```

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