Predictive Analytics Practical Set 1

746_Subhodeep Bhattacharjee

2026-01-21

Download “Boston” housing data from MASS library in R

```
library(MASS)
data(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?

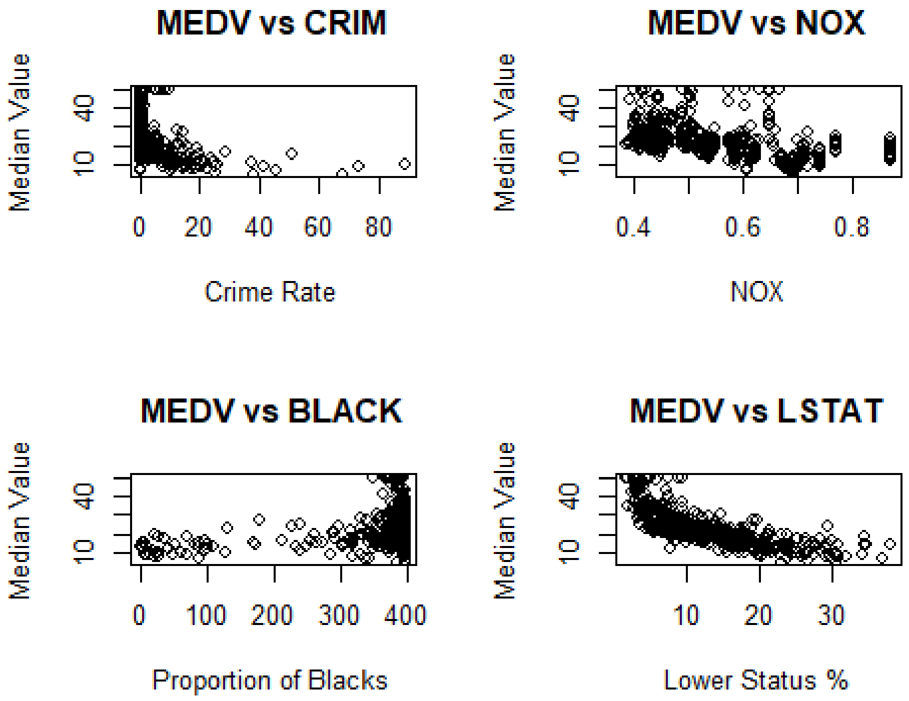
```
class(Boston)

## [1] "data.frame"

dim(Boston)

## [1] 506  14

 There are 506 rows and 14 columns

str(Boston)

## '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 ...
```

The **ROWS** represent the each entry of the Boston suburb.
The **COLUMNS** represent the housing or socio economic variables where there are 13 predictors and 1 response i.e medv-median value of owner-occupied homes in $1000s

---

***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
# Subset data
small_data=Boston[, c("medv", "crim", "nox", "black", "lstat")]

# Scatter plots in one panel
par(mfrow = c(2, 2))

plot(small_data$crim, small_data$medv,xlab="Crime Rate", ylab="Median Value",
main="MEDV vs CRIM")

plot(small_data$nox, small_data$medv,xlab="NOX", ylab="Median Value",
main="MEDV vs NOX")

plot(small_data$black, small_data$medv,xlab="Proportion of Blacks",
ylab="Median Value", main="MEDV vs BLACK")

plot(small_data$lstat, small_data$medv,xlab="Lower Status %", ylab="Median
Value", main="MEDV vs LSTAT")
```



We see that -
- medv vs crim has a negative relation
- medv vs nox has a strong negative - medv vs black has a weak positive relation
- medv vs lstat has a strong negative and non linear relationship

Hence, we can deduce that here lstat is the most strongest predictor as the lstat % increase the house price decreases and nox and crim negatively affect the house price.

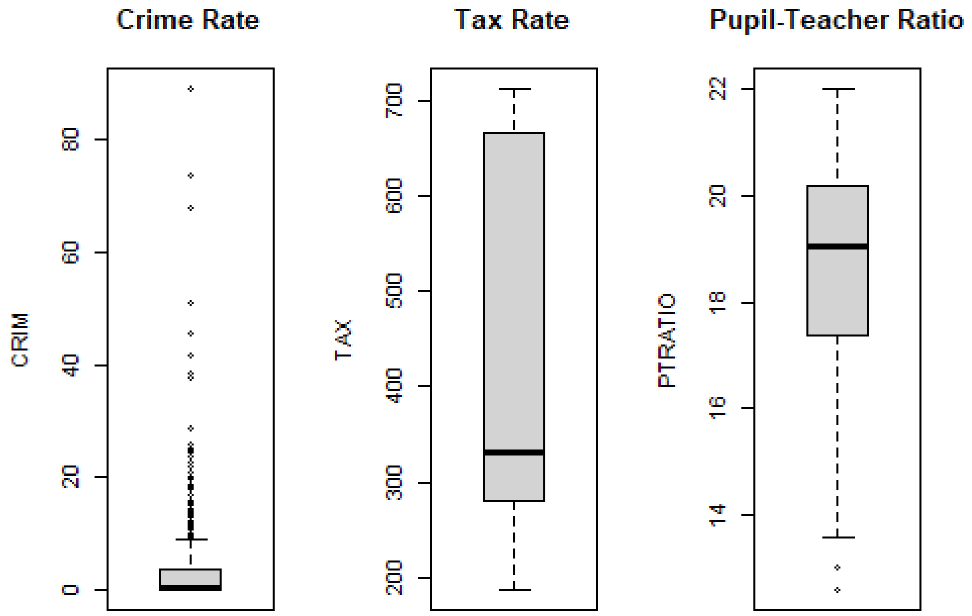
***3.** Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors mentioned in (2), for that suburb. How do these values compare to the overall ranges for those predictors? Comment on your findings. Hint: Mention which percentile these values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

```
# Suburb with Lowest MEDV
min_index=which.min(Boston$medv)
Boston[min_index, c("medv", "crim", "nox", "black", "lstat")]

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

vars=c("crim", "nox", "black", "lstat")

percentiles=sapply(vars, function(v) {ecdf(Boston[[v]])(Boston[min_index,
v])})
percentiles

##      crim       nox     black     lstat
## 0.9881423 0.8577075 1.0000000 0.9782609
```

The median value suburb shows that the crim is relative higher(higher percentile), nox shows high pollution, lstat is also high as it lies among high percentile and black is also showing a good proportion in the suburb.
Thus, we infer that the suburb lies in a socially and economically disadvantaged area which explains the low housing prices.

***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$crim, main="Crime Rate", ylab="CRIM")
boxplot(Boston$tax, main="Tax Rate", ylab="TAX")
boxplot(Boston$ptratio, main="Pupil-Teacher Ratio", ylab="PTRATIO")
```



```
# Identify outliers
which(Boston$crim %in% boxplot.stats(Boston$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
```

These suburbs show high crime rate, also high tax and pratio explains stranded public infrastructure.
These are the possible outliers observed.

```
which.max(Boston$crim)

## [1] 381
```

Amongst all the suburb 381 has notable high crime rate.