PREDICTIVE ANALYTICS
Problem Set 1

Amrita Nath

2026-01-20

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.

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)

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

data=Boston
class(data)

## [1] "data.frame"

nrow(data)

## [1] 506

ncol(data)

## [1] 14
```

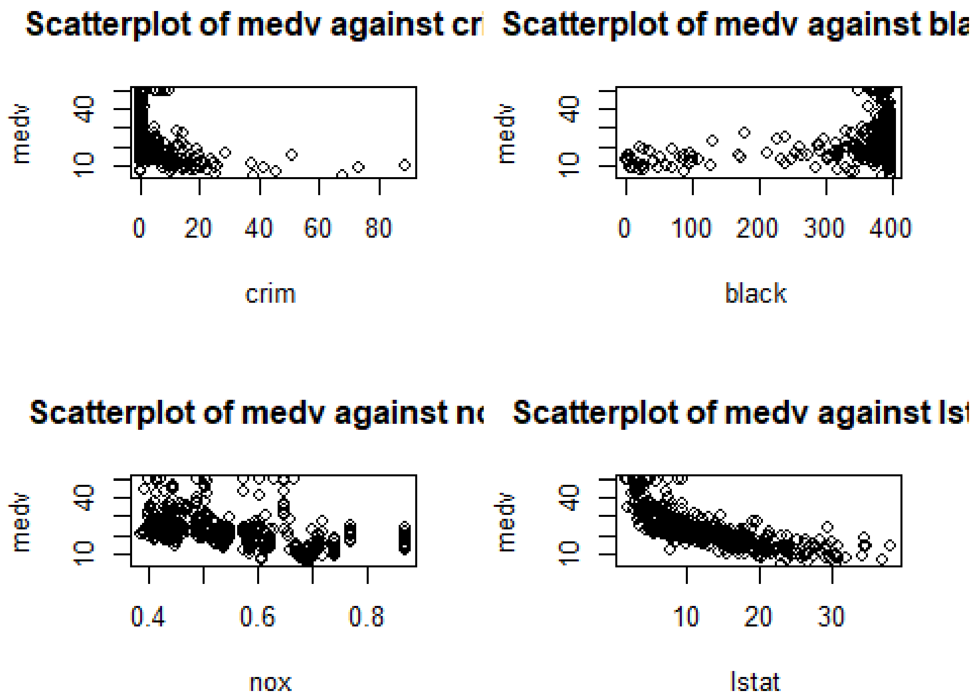
The class of the data set is “data.frame”. There are 50 rows and 14 columns in this data set. Rows represent 506 suburbs of Boston and columns represent 15 variables describing each suburbs.

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.

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

#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")
```



Interpretation

MEDV vs CRIM

-There is a negative relationship between crime rate and median house value.

-Higher median house values are concentrated in suburbs with very low crime rates.

-Suburbs with high crime rates seems to have low housing values.

-The relationship is nonlinear, with a sharp decline in medv even at moderate crime levels.

-A few extreme crime outliers are present.

MEDV vs BLACK

-The points are highly scattered pattern with no strong linear relationship.

-Higher values of black are sometimes associated with slightly higher median house values.

-The dispersion is wide which indicates low strength of association.

MEDV vs NOX

-There is a negative association between nox concentration and median house value.

-Median house values decrease as air pollution levels increase.

-The relationship seems nonlinear, with a sharp decline beyond moderate NOX levels.

-This indicates that environmental quality significantly influences housing prices.

MEDV vs LSTAT

-It has the strongest relationship among all predictors considered.

-Shows a strong, nonlinear negative association with median house value.

-With the increase in percentage of lower status population,median house value decreases sharply.

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.

```
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
```

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
}

#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
```

Interpretation

-The suburb with the lowest median house value (medv = 5) lies at the minimum of the housing price distribution.

-The crime rate (crim) for this suburb falls above the 98th percentile, indicating an exceptionally high crime level compared to most Boston suburbs.

-The nitrogen oxide concentration (nox) is also in the upper tail of the distribution (around 85th percentile), reflecting very poor air quality.

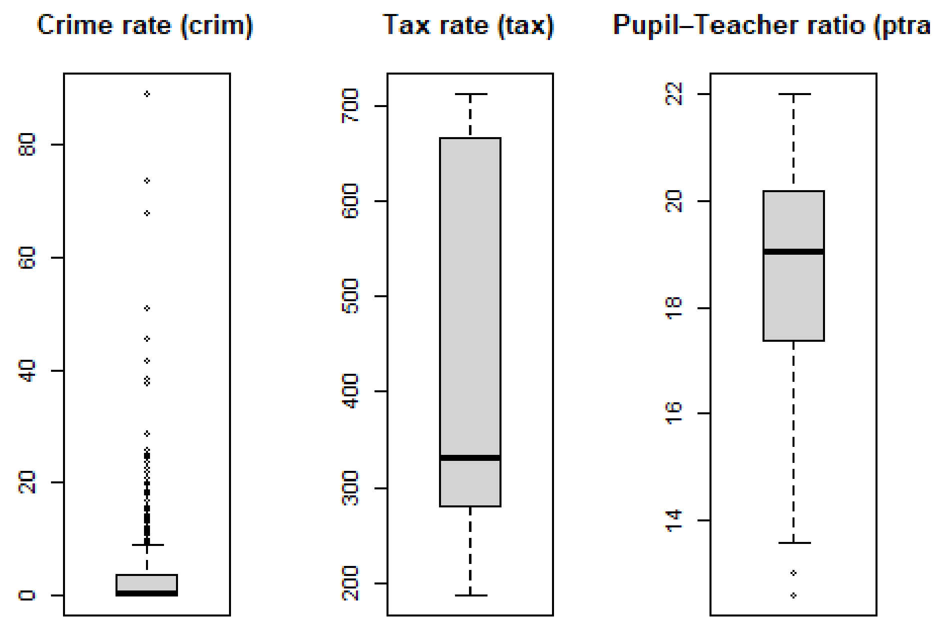
-The percentage of lower-status population (lstat) is among the highest observed values, lying above the 95th percentile.

-In contrast, the black varies widely between to suburbs indicating no consistent pattern.

-The key predictors (crim, nox, and lstat) for this suburb consistently fall in the upper percentiles of their distributions, which collectively explains the extremely low median value of owner-occupied homes.

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))
boxplot(crim,main="Crime rate (crim)")
boxplot(tax,main="Tax rate (tax)")
boxplot(ptratio,main="Pupil-Teacher ratio (ptratio)")
```



Interpretation -

Boxplot shows the presence of outliers in per capita crime rate (crim), indicating that a small number of suburbs experience exceptionally high crime levels compared to the majority.

-The pupil–teacher ratio (ptratio) shows outliers,indicating a few suburbs with unusually low pupil teacher ratio compared to rest.

Displaying the suburbs that show outlier values

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

#for crime rate
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 3
93 395
## [20] 399 400 401 402 403 404 405 406 407 408 410 411 412 413 414 415 416 4
17 418
## [39] 419 420 421 423 426 427 428 430 432 435 436 437 438 439 440 441 442 4
44 445
## [58] 446 448 449 455 469 470 478 479 480

#for pupil-teacher ratio
ptratio_outliers=get_outliers(Boston$ptratio)
ptratio_outliers
```

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