Predictive Analysis Assignment 1

SHRESTHA BAJAJ

ROLL N0. - 741

2026-01-19

```
library(MASS)

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

data(Boston)
```

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?

```
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 ...
```

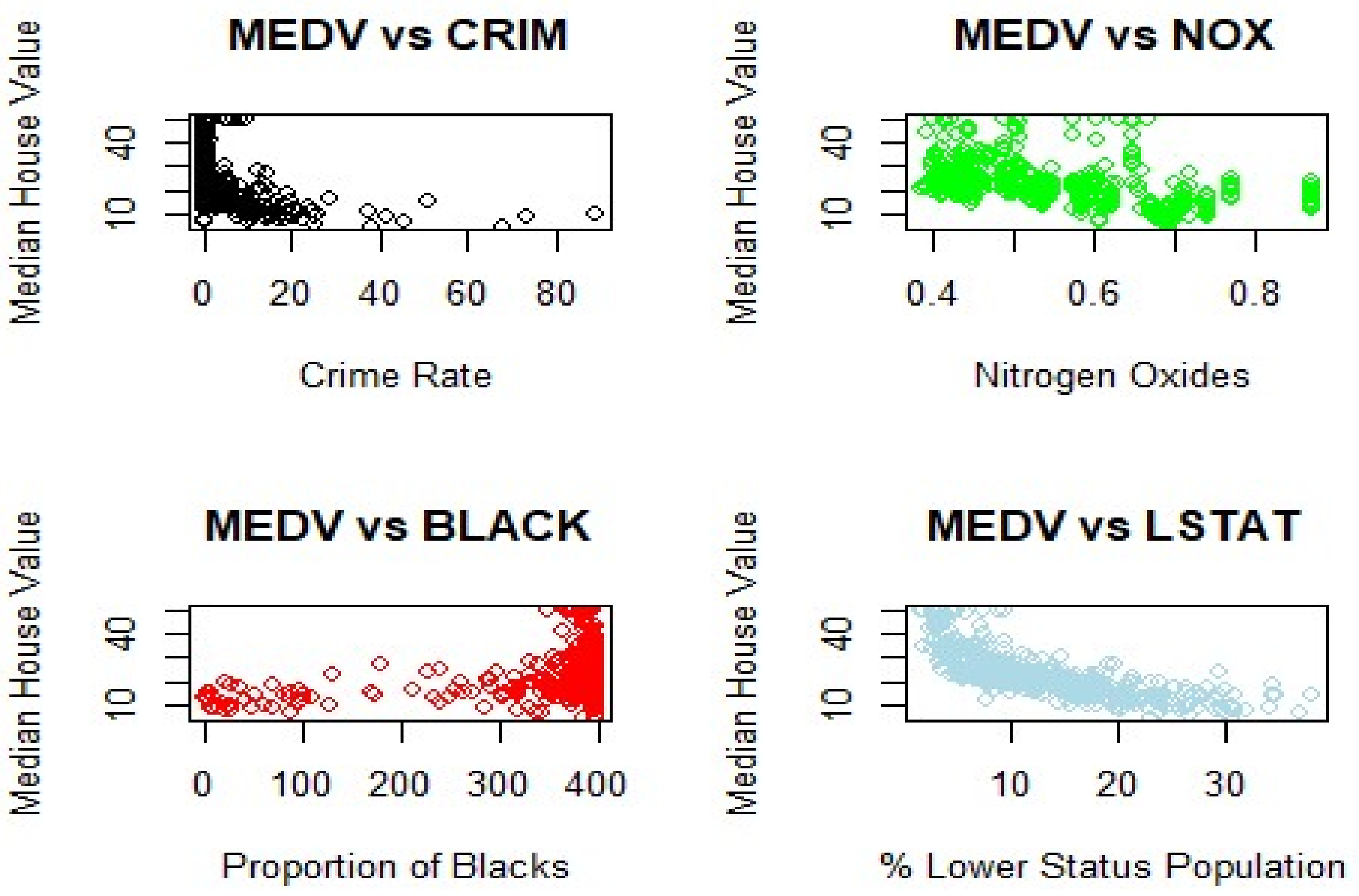
- The class of the data set “Boston” is data.frame
- The data set has 506 rows and 14 columns.
- 506 rows represent the number of observations (suburbs of Boston) in the data set and 14 columns represent the variables such as crime rate, nitric oxide concentration, tax rate and median house value which defines the characteristics of the observations.

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.

```r
# A smaller data set of Boston containing the variables mentioned
boston_small = Boston[, c("medv", "crim", "nox", "black", "lstat")]


# Scatter plot of median value of owner occupied home (response variable) against each
predictor
par(mfrow = c(2, 2))

plot(boston_small$crim, boston_small$medv, xlab = "Crime Rate", ylab = "Median House
Value", main = "MEDV vs CRIM",col="black")


plot(boston_small$nox, boston_small$medv, xlab = "Nitrogen Oxides", ylab = "Median House
Value", main = "MEDV vs NOX",col="green")


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


plot(boston_small$lstat, boston_small$medv, xlab = "% Lower Status Population", ylab =
"Median House Value", main = "MEDV vs LSTAT",col="lightblue")
```

-



Interpretation Of The Graphs -

- MEDV vs CRIM (Crime Rate):
  - The graph shows a strong negative association between the median value of house and per capita crime rate. It means as the crime rate increases the price of the house in that area falls.
  - Though for areas with low crime rates, the price of houses can vary from low to high.
- MEDV vs NOX (Nitrogen Oxides Concentration):
  - There is a clear negative relationship between nitrogen oxide concentration and median house value.
  - Suburbs with lower pollution levels tend to have higher house prices, while higher NOX levels are associated with significantly lower housing values.
- MEDV vs BLACK (Proportion of Blacks):
  - The relationship between median house value and the proportion of blacks is weak and non-linear.
  - The scatter plot is widely dispersed mostly clustered on the right side of the graph, suggesting that BLACK alone is not a strong direct predictor of housing prices.
- MEDV vs LSTAT (% Lower Status Population):
  - This plot displays a strong, non-linear negative relationship.
  - As the percentage of lower-status population increases, median house values decline sharply.

Question 3

Which suburb of Boston has the 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.

```
min_medianval = min(boston_small$medv)
min_medianval
```

```
## [1] 5
```

```
min_sub = boston_small[which.min(boston_small$medv),]
min_sub
```

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

```
vars = c("crim","nox","black","lstat")
perc =
sapply(vars,function(v){ecdf(Boston[[v]])(Boston[which.min(boston_small$medv),v])})
perc
```

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

Interpretation -

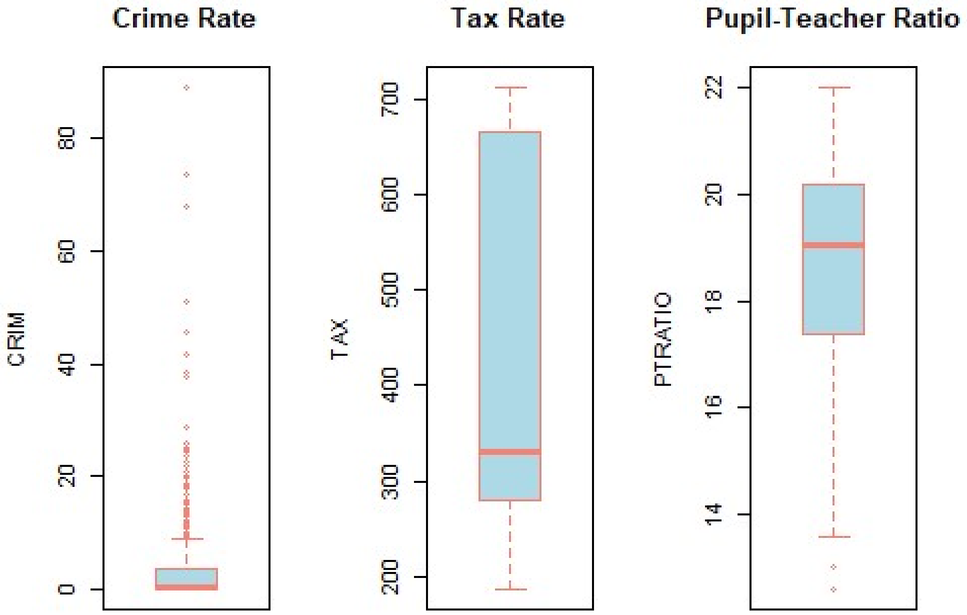
- The lowest median house value (MEDV) is 5.0.
- This suburb shows:
  - Very high crime rate (upper percentile of CRIM).
  - High nitrogen oxide concentration (upper percentile of NOX).
  - High percentage of lower-status population (upper percentile of LSTAT).
  - Low proportion of blacks compared to many suburbs.

These extreme values explain the very low housing prices. The suburb lies in the worst socio-economic and environmental percentiles, making it an outlier in terms of housing value.

Question 4

Does any suburb of Boston stand out for having notably high crime rates, tax rates, or pupil–teacher ratios? Use boxplots to detect any outliers. Identify the suburbs that show the outlier values.

```
par(mfrow = c(1, 3))


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



```
crim_outliers = which(Boston$crim %in% boxplot.stats(Boston$crim)$out)
tax_outliers = which(Boston$tax %in% boxplot.stats(Boston$tax)$out)
ptr_outliers = which(Boston$ptratio %in% boxplot.stats(Boston$ptratio)$out)


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

tax_outliers

## integer(0)

ptr_outliers

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

Interpretation -

- CRIM (Crime Rate):
    - The distribution is highly right-skewed.
    - A large number of extreme outliers appear above the upper whisker.
    - Some suburbs exhibit exceptionally high crime rates, far removed from the bulk of the data.
    - It means crime is unevenly distributed over Boston.
- TAX (Tax Rate):
    - The distribution shows slight right skewness.
    - There are no extreme outliers beyond the whiskers.
    - Several suburbs cluster near the upper end of the tax range, close to the maximum value.
    - Tax rates vary considerably among suburbs but remain within a reasonable and regulated range.
- PTRATIO (Pupil-Teacher Ratio):
    - The distribution is approximately symmetric.
    - A few low-end outliers exist, representing suburbs with exceptionally low pupil–teacher ratios.
    - No extreme high outliers are observed.
    - Educational inequality exists but is less pronounced compared to crime inequality.