Predictive Analysis Problem Set 1

Deeptarka Saha

2026-01-20

PROBLEM SET 1

BOSTON DATA SET

```
# Load MASS library
library(MASS)

# Load Boston dataset
data(Boston)

# View first few rows
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
```

QS 1

```
# Class of dataset
class(Boston)
## [1] "data.frame"
# Number of rows and columns
```

```
dim(Boston)
## [1] 506  14
# Structure of dataset
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 ...
```

Class: data.frame

Rows: 506 . Each row represents a suburb/town in Boston

Columns: 14. Each column represents a housing-related variable such as crime rate, pollution, tax, etc.

QS 2

```
# Response: medv (Median value of owner-occupied homes)
# Predictors: crim, nox, black, lstat
# CREATE SMALLER DATASETS
boston_small <- Boston[, c("medv", "crim", "nox", "black", "lstat")]
head(boston_small)
##   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
#Scatter plot in multiple panels
```

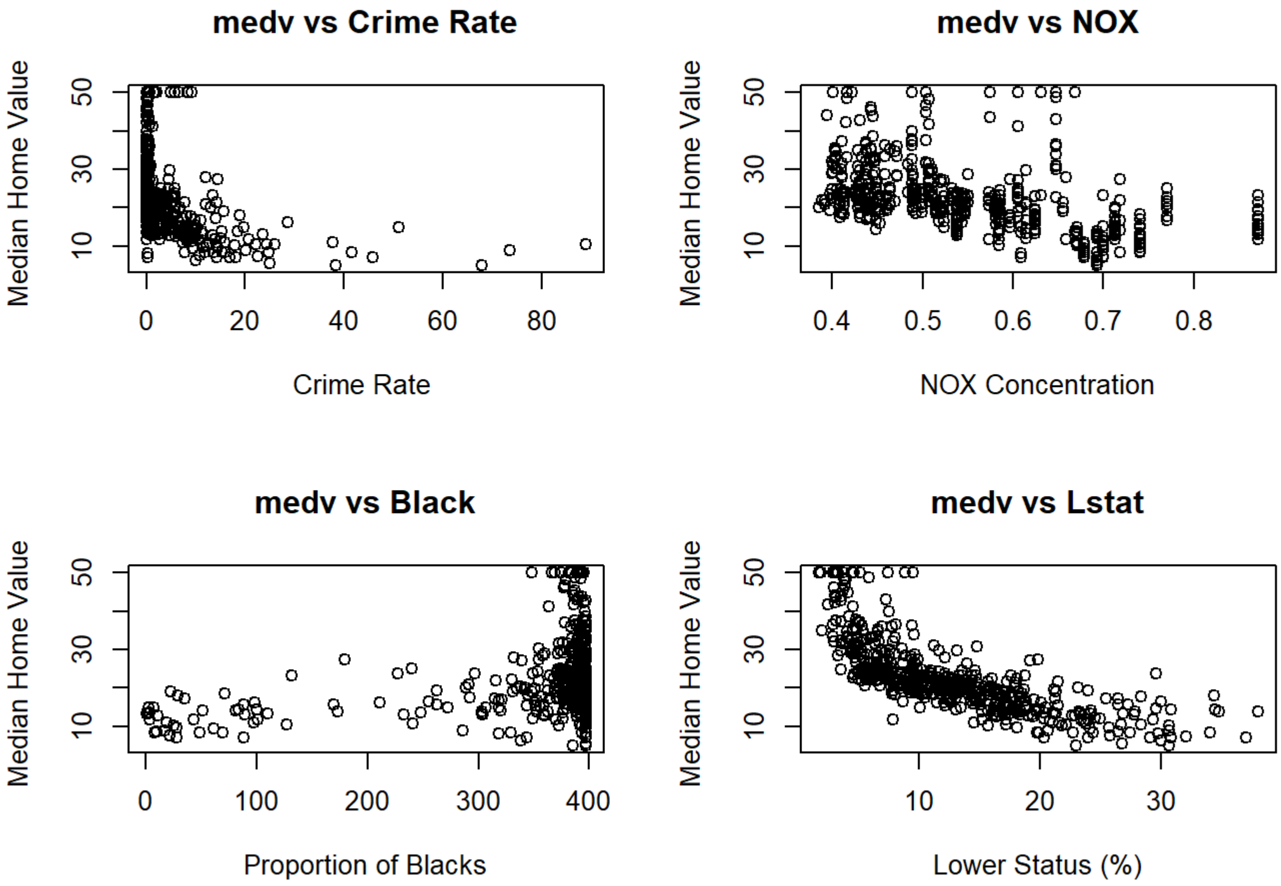
```
par(mfrow = c(2, 2))

# Scatter plots
plot(boston_small$crim, boston_small$medv,
     xlab = "Crime Rate ", ylab = "Median Home Value ",
     main = "medv vs Crime Rate")

plot(boston_small$nox, boston_small$medv,
     xlab = "NOX Concentration", ylab = "Median Home Value ",
     main = "medv vs NOX")

plot(boston_small$black, boston_small$medv,
     xlab = "Proportion of Blacks", ylab = "Median Home Value ",
     main = "medv vs Black")

plot(boston_small$lstat, boston_small$medv,
     xlab = "Lower Status (%)", ylab = "Median Home Value ",
     main = "medv vs Lstat")
```



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

Comments:

i. Crime rate : Higher crime → lower house prices
ii. NOX: Increased pollution → reduced house values
iii. Black: Weak positive relationship
iv. Lstat: Strong negative relationship

QS 3

```
# Find minimum median value
min_medv = min(Boston$medv);
min_medv
## [1] 5
# Suburb(s) with lowest median value
lowest_medv_suburb = Boston[Boston$medv == min_medv, ]
lowest_medv_suburb
##        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    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
selected_vars=lowest_medv_suburb[, c("crim", "nox", "black", "lstat", "medv")]
# Percentile Comparison
# Percentiles
# Function to compute percentile
percentile <- function(x, value) {
  mean(x <= value) * 100
}


# Calculate percentiles
percentiles <- data.frame(
  Variable = c("crim", "nox", "black", "lstat"),
  Value = c(selected_vars$crim,
            selected_vars$nox,
            selected_vars$black,
            selected_vars$lstat),
  Percentile = c(
    percentile(Boston$crim, selected_vars$crim),
    percentile(Boston$nox, selected_vars$nox),
    percentile(Boston$black, selected_vars$black),
    percentile(Boston$lstat, selected_vars$lstat)
  )
)
percentiles
##   Variable    Value Percentile
## 1     crim  38.3518   99.01186
## 2      nox  67.9208   85.77075
## 3    black   0.6930   66.00791
## 4    lstat   0.6930   94.07115
## 5     crim 396.9000   99.01186
## 6      nox 384.9700   85.77075
## 7    black  30.5900   66.00791
## 8    lstat  22.9800   94.07115
```

Comments:

i. The suburb with the lowest median home value (medv) represents one of the most economically disadvantaged areas.

ii. Its crime rate (crim) lies in a high percentile, indicating unusually high crime.

iii. Nitrogen oxide levels (nox) are also relatively high, suggesting environmental pollution.

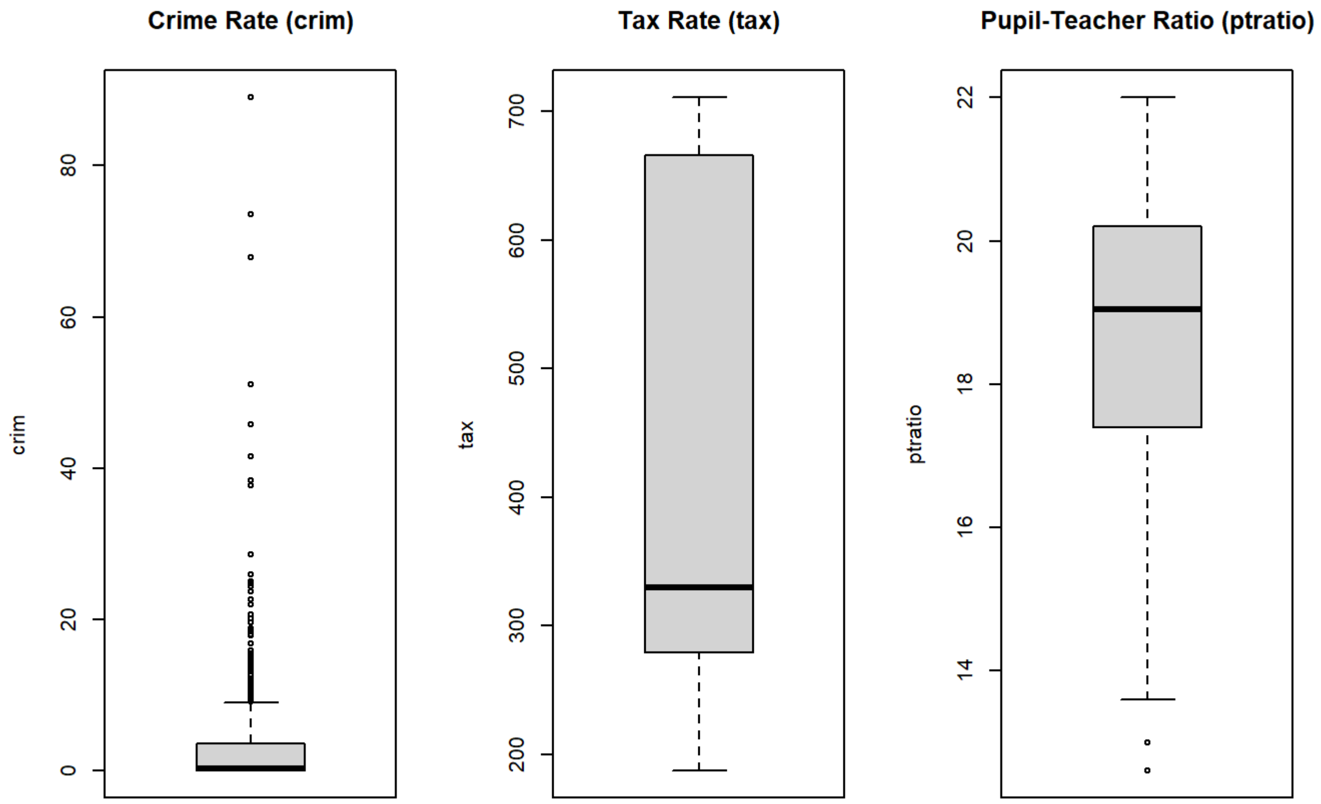
iv. Lower status population (lstat) is in the upper percentile, reinforcing socio-economic stress.

v. The black variable percentile indicates how this suburb compares demographically to others.

QS 4

```
# Set plotting area for 3 boxplots
par(mfrow = c(1, 3))

# Boxplots for detecting outliers
boxplot(Boston$crim, main = "Crime Rate (crim)", ylab = "crim")
boxplot(Boston$tax, main = "Tax Rate (tax)", ylab = "tax")
boxplot(Boston$ptratio, main = "Pupil-Teacher Ratio (ptratio)", ylab = "ptratio")
```



```
# Reset plotting layout
par(mfrow = c(1, 1))
# Find outliers for each variable
crim_outliers <- boxplot.stats(Boston$crim)$out
tax_outliers <- boxplot.stats(Boston$tax)$out
ptratio_outliers <- boxplot.stats(Boston$ptratio)$out
crim_outliers
##  [1] 13.52220  9.23230 11.10810 18.49820 19.60910 15.28800  9.82349 23.64820
##  [9] 17.86670 88.97620 15.87440  9.18702 20.08490 16.81180 24.39380 22.59710
## [17] 14.33370 11.57790 13.35980 38.35180  9.91655 25.04610 14.23620  9.59571
## [25] 24.80170 41.52920 67.92080 20.71620 11.95110 14.43830 51.13580 14.05070
## [33] 18.81100 28.65580 45.74610 18.08460 10.83420 25.94060 73.53410 11.81230
## [41] 11.08740 12.04820 15.86030 12.24720 37.66190  9.33889 10.06230 13.91340
## [49] 11.16040 14.42080 15.17720 13.67810  9.39063 22.05110  9.72418  9.96654
## [57] 12.80230 10.67180  9.92485  9.32909  9.51363 15.57570 13.07510 15.02340
## [65] 10.23300 14.33370
tax_outliers
## numeric(0)
```

```
ptratio_outliers

##  [1] 12.6 12.6 12.6 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0
```

Comments:

i. A few suburbs appear as extreme outliers, indicating unusually high crime compared to most Boston suburbs.

ii. Some suburbs have exceptionally high property tax rates, standing far above the upper quartile.

iii. One or two suburbs show high student–teacher ratios, suggesting possible strain on educational resources.