ASSIGNMENT 1

SAURABH MISHRA

2026-01-19

PREDICTIVE ANALYTICS

Problem Set 1: An Introduction

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.

```
library(MASS)

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

data(Boston)
```

About the Boston data set

Description of Variables in the Boston Housing Dataset

The Boston housing dataset contains information on housing, demographic, and environmental characteristics of suburbs in Boston. The variables included in the dataset are described below:

crim - Per capita crime rate by town.

zn - Proportion of residential land zoned for lots over 25,000 sq.ft.

indus- Proportion of non-retail business acres per town.

chas - Charles River dummy variable (1 if tract bounds river; 0 otherwise).

nox - Nitrogen oxides concentration (parts per 10 million).

rm - Average number of rooms per dwelling.

age - Proportion of owner-occupied units built prior to 1940.

dis - Weighted mean of distances to five Boston employment centres.

rad - Index of accessibility to radial highways.

tax - Full-value property-tax rate per $10,000.

ptratio - Pupil–teacher ratio by town.

black - Transformed measure of the proportion of Black residents by town: 1000*(Bk-0.63)^2, where Bk is the actual proportion.

lstat - Percentage of lower status population.

medv - Median value of owner-occupied homes in $1000s.

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
```

Each row represents a suburb of Boston.

Each column represents a variable describing housing, demographic, environmental, or socio-economic characteristics of the suburb.

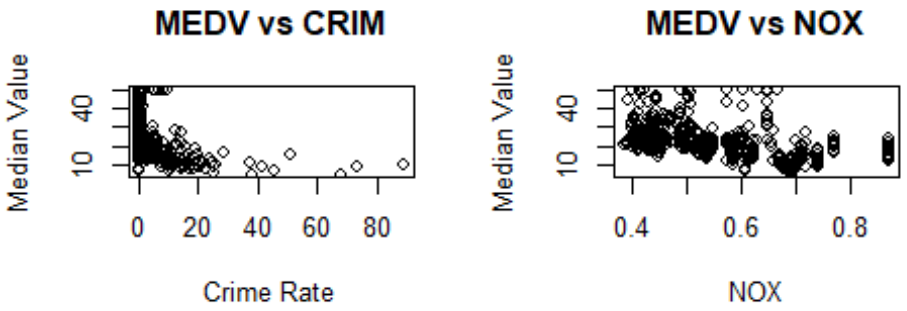
2. Create a smaller data set with the variables median value of owner-occupied homes, per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population. Choosing median value of owner occupied homes as the response and the rest as the pre- dictors,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

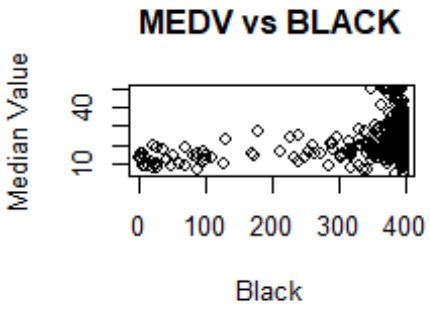
```
boston_small <- Boston[, c("medv", "crim", "nox", "black", "lstat")]
head(boston_small)
```

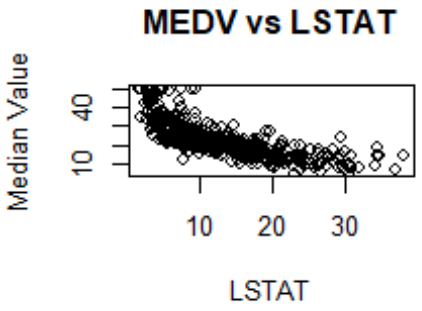
```
##   medv    crim   nox  black lstat
## 1 24.0 0.00632 0.538 396.90  4.98
## 2 21.6 0.02731 0.469 396.90  9.14
## 3 34.7 0.02729 0.469 392.83  4.03
## 4 33.4 0.03237 0.458 394.63  2.94
## 5 36.2 0.06905 0.458 396.90  5.33
## 6 28.7 0.02985 0.458 394.12  5.21
```

Scatter Plot

```
par(mfrow = c(2,2))
plot(boston_small$crim, boston_small$medv, xlab="Crime Rate", ylab="Median
Value", main="MEDV vs CRIM")
plot(boston_small$nox, boston_small$medv, xlab="NOX", ylab="Median Value",
main="MEDV vs NOX")
plot(boston_small$black, boston_small$medv, xlab="Black", ylab="Median
Value", main="MEDV vs BLACK")
plot(boston_small$lstat, boston_small$medv, xlab="LSTAT", ylab="Median
Value", main="MEDV vs LSTAT")
```







Findings

1. Median house value decreases as crime rate increases.
2. A strong negative relationship exists between medv and lstat.

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 pre- dictors? Comment on your findings. Hint: Mention which percentile these values belong to.

```
min_medv_index <- which.min(Boston$medv)
Boston[min_medv_index, c("medv", "crim", "nox", "black", "lstat")]
```

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

```
# Percentiles of predictors for that suburb
suburb_values <- Boston[min_medv_index, c("crim", "nox", "black", "lstat")]
percentiles <- sapply(
  c("crim", "nox", "black", "lstat"),
  function(var) ecdf(Boston[[var]])(suburb_values[[var]])
)

percentiles
```

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

Comments

The suburb corresponding to 399th row has lowest median value of owner-occupied homes.

The suburb with the lowest median house value lies in the upper percentiles of crime rate and lower-status population.

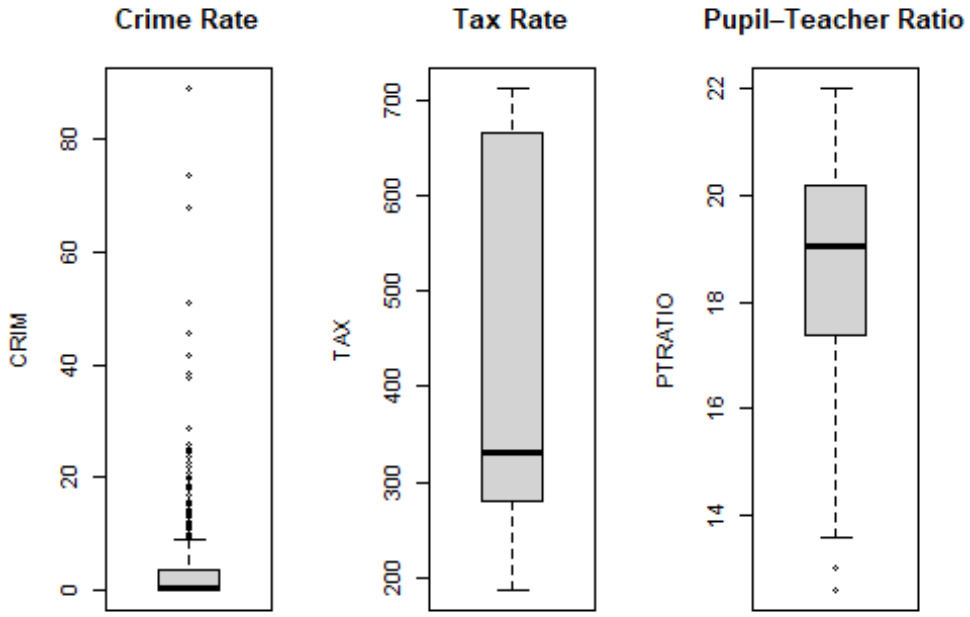
Pollution levels are also relatively high.

These characteristics explain the very low housing prices in this suburb.

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 suburbs with outliers
out_crim <- which(Boston$crim %in% boxplot.stats(Boston$crim)$out)
out_tax <- which(Boston$tax %in% boxplot.stats(Boston$tax)$out)
out_ptratio <- which(Boston$ptratio %in% boxplot.stats(Boston$ptratio)$out)
```

```
out_crim
```

```
##  [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
```

```
out_tax
```

```
## integer(0)
```

```
out_ptratio
```

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

Comments

A few suburbs exhibit extremely high crime rates.

Some suburbs also show unusually high tax rates and pupil–teacher ratios.

These outliers may strongly influence predictive models and should be handled carefully.