Predictive Practical 1

Srinjana Sen_735

2026-01-22

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)
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"
```

The Boston data set is of class “data.frame”.

```
dim(Boston)

## [1] 506  14
```

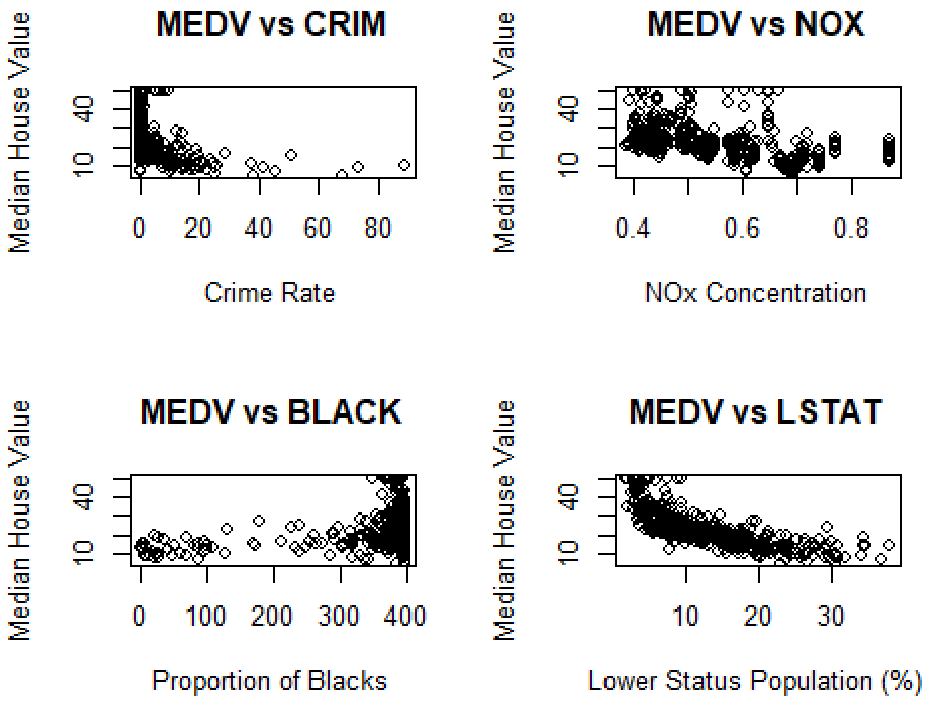
There are 506 rows and 14 columns in this data set.

Each of the 506 rows represent 506 suburbs (or census tracts) in the Boston area.
Each of the 14 columns represent 14 variables describing different characteristics of each suburb viz. crime rates, proportion of residential land, proportion of non-retail business, a Charles River dummy variable, nitrogen oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units, weighted mean of distances to five Boston employment centres, index of accessibility to radial highways, full-value property-tax rate, pupil-teacher ratio, proportion of blacks, lower status of the population, median value of owner-occupied homes.

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
smaller_data=Boston[, c("medv", "crim", "nox", "black", "lstat")]

par(mfrow = c(2, 2))

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

plot(smaller_data$nox, smaller_data$medv,
     xlab = "NOx Concentration",
     ylab = "Median House Value",
     main = "MEDV vs NOX")

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

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



the median values cluster around the zero crime rate zones. But it’s not a good predictor as the lower median value rates, too, cluster around the same area with some distinct outliers.

Median house value decreases as nitrogen oxide concentration increases, indicating that pollution negatively affects housing prices.

No proper correlation is seen. Higher values of black population cluster against higher median house values; thus it’s not a good predictor.

A strong negative relationship is obvious from the scatter plot. As the percentage of lower-status population increases, median house values decline sharply. This displays the strongest relationship among the predictors.

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.

```
min_medv_index=which(Boston$medv == min(Boston$medv))

Boston[min_medv_index, c("medv", "crim", "nox", "black", "lstat")]
```

```
##     medv    crim   nox  black lstat
## 399    5 38.3518 0.693 396.90 30.59
## 406    5 67.9208 0.693 384.97 22.98
```

Two suburbs, at indices 399 and 406, share the lowest median value of owner-occupied homes, with medv = 5.

```
get_percentile = function(value, column_data) {
  return(round(ecdf(column_data)(value) * 100, 2))
}
target_row = Boston[399, ]

  val_crim = target_row[["crim"]]
  perc_crim = get_percentile(val_crim, Boston[["crim"]])
  print(paste("crim", "=", val_crim, "| Percentile:", perc_crim, "%"))

## [1] "crim = 38.3518 | Percentile: 98.81 %"
```

Crime rate is 38.35, which lies in the 98.81st percentile, indicates that this suburb has one of the highest crime rates in the entire data set.

```
 val_nox = target_row[["nox"]]
  perc_nox = get_percentile(val_nox, Boston[["nox"]])
  print(paste("nox", "=", val_nox, "| Percentile:", perc_nox, "%"))

## [1] "nox = 0.693 | Percentile: 85.77 %"
```

Nitrogen oxides concentration is 0.693, corresponding to the 85.77th percentile, suggesting high levels of air pollution.

```
 val_black = target_row[["black"]]
  perc_black = get_percentile(val_black, Boston[["black"]])
  print(paste("black", "=", val_black, "| Percentile:", perc_black, "%"))

## [1] "black = 396.9 | Percentile: 100 %"
```

Proportion of blacks is 396.90, which falls at the maximum (100th percentile) of this variable.

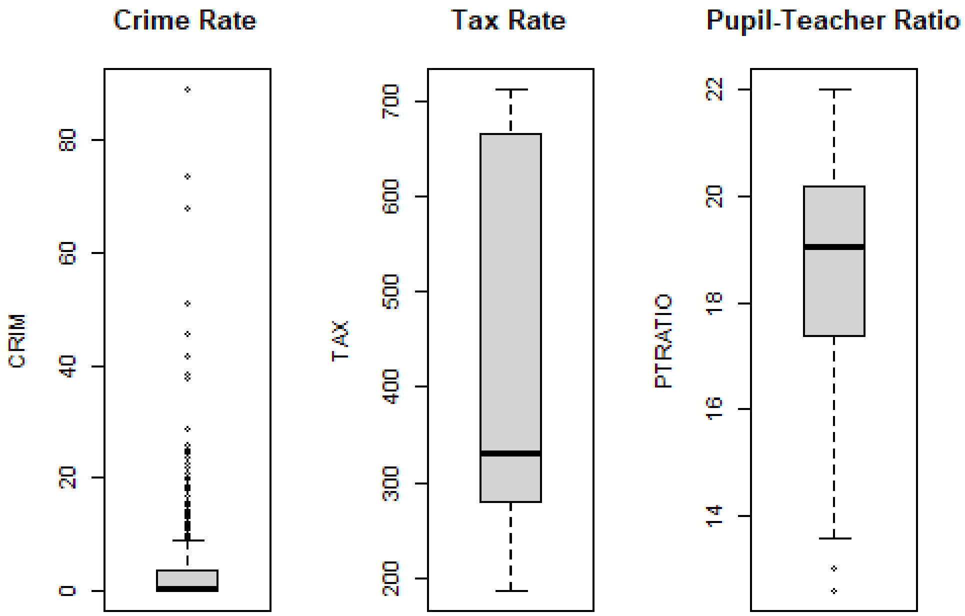
```
 val_lstat = target_row[["lstat"]]
  perc_lstat = get_percentile(val_lstat, Boston[["lstat"]])
  print(paste("lstat", "=", val_lstat, "| Percentile:", perc_lstat, "%"))

## [1] "lstat = 30.59 | Percentile: 97.83 %"
```

Percentage of lower-status population is 30.59, placing it in the 97.83rd percentile, indicating an extremely high level of socio-economic disadvantage.

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")
```



The distribution of crime rate appears to be right-tailed, negatively skewed and leptokurtic with many outliers. Several suburbs indicate unusually high crime rates appearing as extreme outliers.
The distribution of tax rate appears to be widely spread, right-tailed, negatively skewed with no visible outliers.
The distribution of pupil teacher ratio appears to be left-tailed, positively skewed with a moderate spread few outliers.