Assignment_1_710

PROTYAY CHATTERJEE

2026-01-21

```
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"
```

```
dim(Boston)
```

```
## [1] 506  14
```

The data set is of class data.frame.

There are 506 rows and 14 columns.

Each row represents a specific suburb or town within the Boston area.

Each column represents a specific variable associated with that 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 predictors, make scatter plots of the response versus each predictor. Presentthe scatter plots in different panels of the same graph. Comment on your findings.

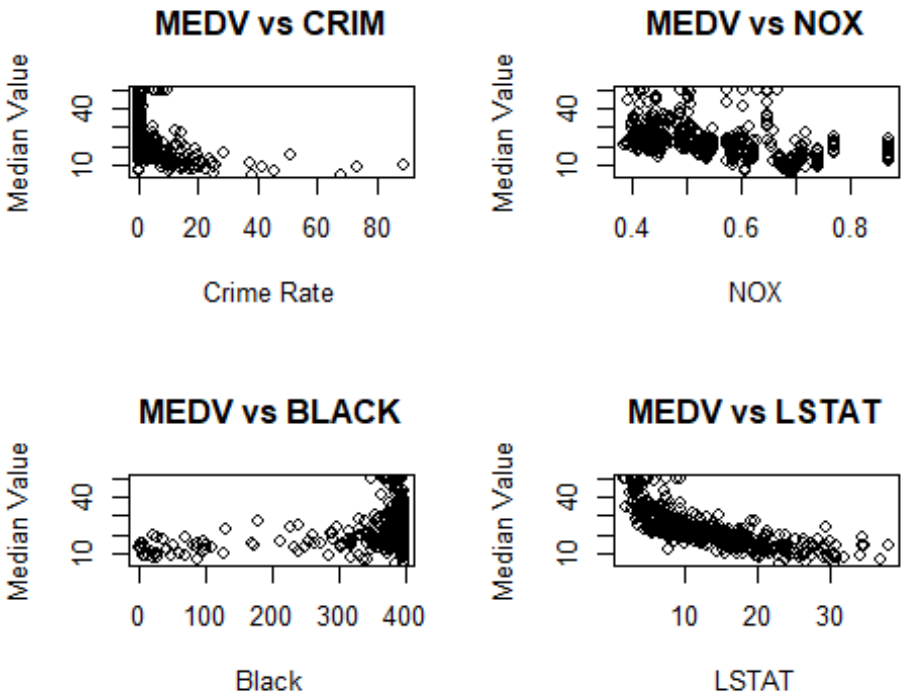
```
boston_trim <- Boston[, c("medv", "crim", "nox", "black", "lstat")]
head(boston_trim)
```

```
##   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
```

The Scatter Plots are as follows:

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

```
plot(boston_trim$lstat, boston_trim$medv, xlab="LSTAT", ylab="Median Value",
main="MEDV vs LSTAT")
```



My Findings :

Crime (crim): There is a negative relationship; as crime rates increase, the median value of homes tends to drop, though most data is clustered near zero crime.

Nitrogen Oxides (nox): There is a negative trend; higher pollution levels correlate with lower housing values.

Proportion of Blacks (black): The relationship is not strictly linear, but lower proportions often correspond with lower median values. There is a cluster of high median values where the proportion of blacks is maximized.

Lower Status Population (lstat): There is a strong, clear negative correlation. As the percentage of the lower-status population increases, the median home value decreases significantly.

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.min(boston_trim$medv)
lowest_suburb <- boston_trim[min_medv_index, ]
print(lowest_suburb)

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

```r
calc_percentile <- function(x, value) {
  mean(x <= value) * 100
}

crim_perc <- calc_percentile(boston_trim$crim, lowest_suburb$crim)
nox_perc <- calc_percentile(boston_trim$nox, lowest_suburb$nox)
black_perc <- calc_percentile(boston_trim$black, lowest_suburb$black)
lstat_perc <- calc_percentile(boston_trim$lstat, lowest_suburb$lstat)

cat("Percentiles:\n")
```

```
## Percentiles:
```

```r
cat("Crime:", crim_perc, "%\n")
```

```
## Crime: 98.81423 %
```

```r
cat("NOx:", nox_perc, "%\n")
```

```
## NOx: 85.77075 %
```

```r
cat("Black:", black_perc, "%\n")
```

```
## Black: 100 %
```

```r
cat("Lstat:", lstat_perc, "%\n")
```

```
## Lstat: 97.82609 %
```

```r
summary(boston_trim)
```

```
##       medv            crim              nox              black       
##  Min.   : 5.00   Min.   : 0.00632   Min.   :0.3850   Min.   :  0.32  
##  1st Qu.:17.02   1st Qu.: 0.08205   1st Qu.:0.4490   1st Qu.:375.38  
##  Median :21.20   Median : 0.25651   Median :0.5380   Median :391.44  
##  Mean   :22.53   Mean   : 3.61352   Mean   :0.5547   Mean   :356.67  
##  3rd Qu.:25.00   3rd Qu.: 3.67708   3rd Qu.:0.6240   3rd Qu.:396.23  
##  Max.   :50.00   Max.   :88.97620   Max.   :0.8710   Max.   :396.90  
##      lstat      
##  Min.   : 1.73  
##  1st Qu.: 6.95  
##  Median :11.36  
##  Mean   :12.65  
##  3rd Qu.:16.95  
##  Max.   :37.97
```

####$ Findings : i. Suburb with the Lowest Median Value The suburb with the lowest median value of owner-occupied homes ($5,000) is found at index 399

ii. Predictor Values for Suburb 399 For this specific suburb, the predictor values are:

Crime Rate (crim): 38.35

Nitrogen Oxides (nox): 0.693 parts per 10 million

Proportion of Blacks (black): 396.90

Lower Status Population (lstat): 30.59%

Crime: The value of 38.35 is extremely high, placing it in the 99th percentile. This is one of the most dangerous neighborhoods in the dataset.

NOx: A concentration of 0.693 is in the 85th percentile, indicating significantly higher pollution than the average Boston suburb.

Black: The value 396.90 is the maximum value possible in this dataset (100th percentile), a value shared by many suburbs.

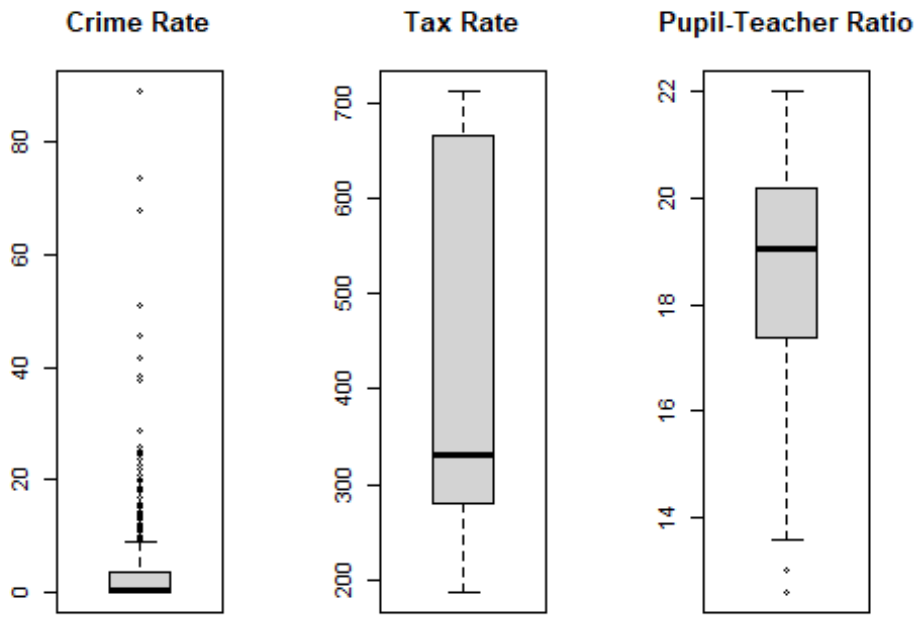
Lstat: A value of 30.59% is in the 97th percentile, indicating this suburb has one of the highest proportions of lower-status population in the entire area.

iv. The suburb with the lowest housing value is characterized by extreme signs of urban decay: it has exceptionally high crime rates, high pollution levels, and a very large population of lower-status residents. This confirms the strong negative correlations observed in the scatter plots.

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.

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

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




```
# Identify outliers
outliers_crim <- boxplot.stats(Boston$crim)$out
outliers_tax <- boxplot.stats(Boston$tax)$out
outliers_ptratio <- boxplot.stats(Boston$ptratio)$out

cat("Number of Crime outliers:", length(outliers_crim), "\n")

## Number of Crime outliers: 66

cat("Number of Tax outliers:  ", length(outliers_tax), "\n")

## Number of Tax outliers:   0

cat("Number of P-T outliers:  ", length(outliers_ptratio), "\n")

## Number of P-T outliers:   15

# Identify specific suburbs with high crime outliers (top 5 extreme)
high_crime_indices <- which(Boston$crim %in% outliers_crim)
print("Indices of suburbs with extreme crime rates (first 10):")

## [1] "Indices of suburbs with extreme crime rates (first 10):"

print(head(high_crime_indices, 10))

##  [1] 368 372 374 375 376 377 378 379 380 381
```

Findings : Crime: Yes, the boxplot shows significant outliers in the upper range. A small number of suburbs have drastically higher crime rates than the median.

Tax: There are no statistical outliers (points beyond the whiskers) in the standard boxplot, although the distribution is bimodal.

Pupil-Teacher Ratio: Shows very few outliers, mostly on the lower end.

Conclusion: The suburbs that “stand out” most are those with high crime rates.