Predictive Analysis - Problem Set 1 - Introduction

Priyanshu Dey 703

2026-01-20

Introduction

This report analyzes the “Boston” housing data from the MASS library in R.

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?

```r
data(Boston)

# Check class, dimensions
data_class = class(Boston)
data_dims = dim(Boston)

print(paste("Class of dataset:", data_class))
```

```
## [1] "Class of dataset: data.frame"
```

```r
print(paste("Number of rows:", data_dims[1]))
```

```
## [1] "Number of rows: 506"
```

```r
print(paste("Number of columns:", data_dims[2]))
```

```
## [1] "Number of columns: 14"
```

Answer: The `Boston` dataset is of class `data.frame`. It contains **506 rows** and **14 columns**.

Representation:

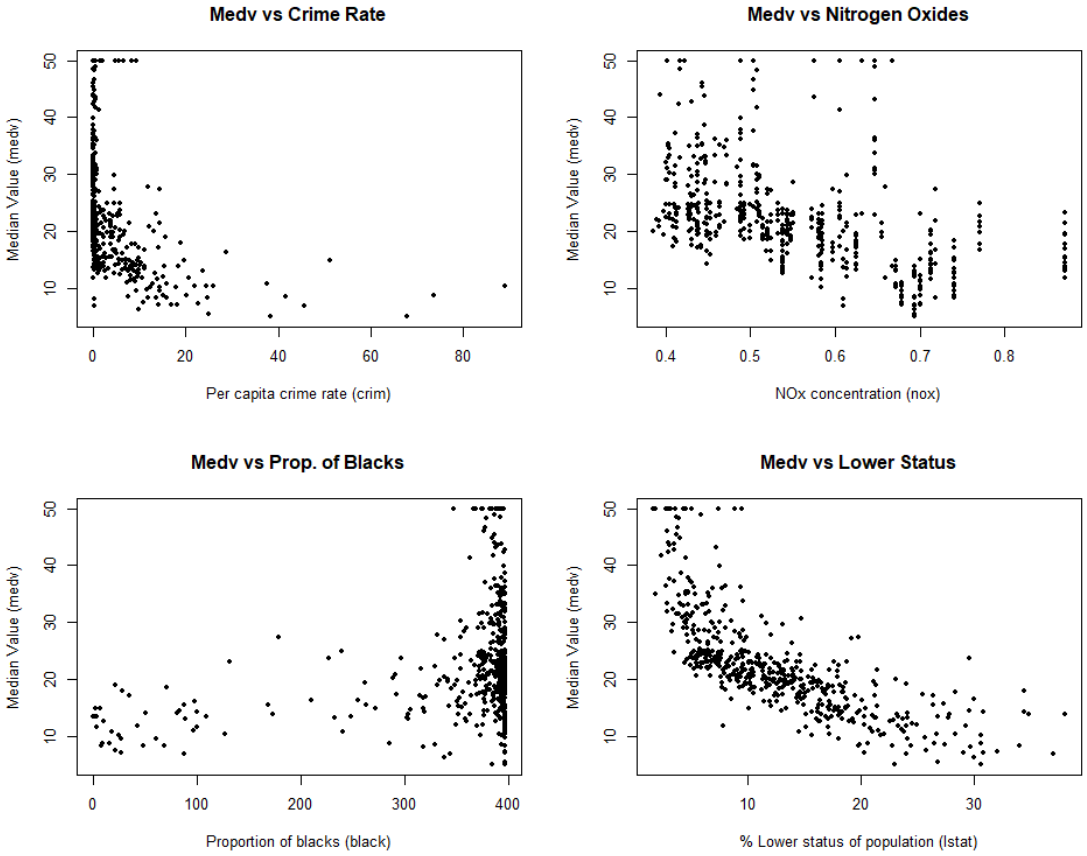
- **Rows:** Each row represents a specific suburb or town in the Boston area (census tract).
- **Columns:** Each column represents a specific attribute or variable associated with that suburb, such as crime rate, property tax rate, or median home value.

---

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 subset
# Variables: medv (response), crim, nox, black, lstat
subset_data = Boston[, c("medv", "crim", "nox", "black", "lstat")]

par(mfrow = c(2, 2))

# Plot medv vs crim
plot(subset_data$crim, subset_data$medv,
     main = "Medv vs Crime Rate",
     xlab = "Per capita crime rate (crim)",
     ylab = "Median Value (medv)", pch = 20)

# Plot medv vs nox
plot(subset_data$nox, subset_data$medv,
     main = "Medv vs Nitrogen Oxides",
     xlab = "NOx concentration (nox)",
     ylab = "Median Value (medv)", pch = 20)

# Plot medv vs black
plot(subset_data$black, subset_data$medv,
     main = "Medv vs Prop. of Blacks",
     xlab = "Proportion of blacks (black)",
     ylab = "Median Value (medv)", pch = 20)

# Plot medv vs lstat
plot(subset_data$lstat, subset_data$medv,
     main = "Medv vs Lower Status",
     xlab = "% Lower status of population (lstat)",
     ylab = "Median Value (medv)", pch = 20)
```



```r
# Reset
par(mfrow = c(1, 1))
```

Findings:

- **Crime (crim) vs Medv:** There is a negative relationship. Suburbs with very high crime rates invariably have lower median home values. However, low crime rates do not guarantee high home values (high variance at low x-values).
- **NOx (nox) vs Medv:** There is a general negative trend; as pollution (NOx) increases, median home values tend to decrease.
- **Black (black) vs Medv:** The relationship is not immediately linear or clear from the scatter plot alone, though many high-value homes are clustered at the higher end of the `black` variable.
- **Lower Status (lstat) vs Medv:** This shows a strong, non-linear negative relationship. As the percentage of lower-status population increases, the median home value drops 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.)

```r
# index of the minimum medv
min_medv_idx = which(Boston$medv == min(Boston$medv))
print(paste("Index of suburb with lowest medv:", min_medv_idx))
```

```
## [1] "Index of suburb with lowest medv: 399"
## [2] "Index of suburb with lowest medv: 406"
```

```r
# data for the suburbs
lowest_suburbs = Boston[min_medv_idx, c("medv", "crim", "nox", "black",
"lstat")]
print(lowest_suburbs)
```

```
##     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
```

```r
# Function to calculate percentile of a value within the whole dataset
get_percentile = function(value, column_data) {
  return(round(ecdf(column_data)(value) * 100, 2))
}

# Analyze the first suburb with the lowest value : 399
target_row = Boston[399, ]
predictors = c("crim", "nox", "black", "lstat")

print("Percentile Ranks for Suburb #399:")
```

```
## [1] "Percentile Ranks for Suburb #399:"
```

```r
for(pred in predictors) {
  val = target_row[[pred]]
  perc = get_percentile(val, Boston[[pred]])
  print(paste(pred, "=", val, "| Percentile:", perc, "%"))
}
```

```
## [1] "crim = 38.3518 | Percentile: 98.81 %"
## [1] "nox = 0.693 | Percentile: 85.77 %"
## [1] "black = 396.9 | Percentile: 100 %"
## [1] "lstat = 30.59 | Percentile: 97.83 %"
```

Answer: Two suburbs (Indices 399 and 406) are tied for the lowest median value (medv = 5).

- **Analyzing Suburb 399**

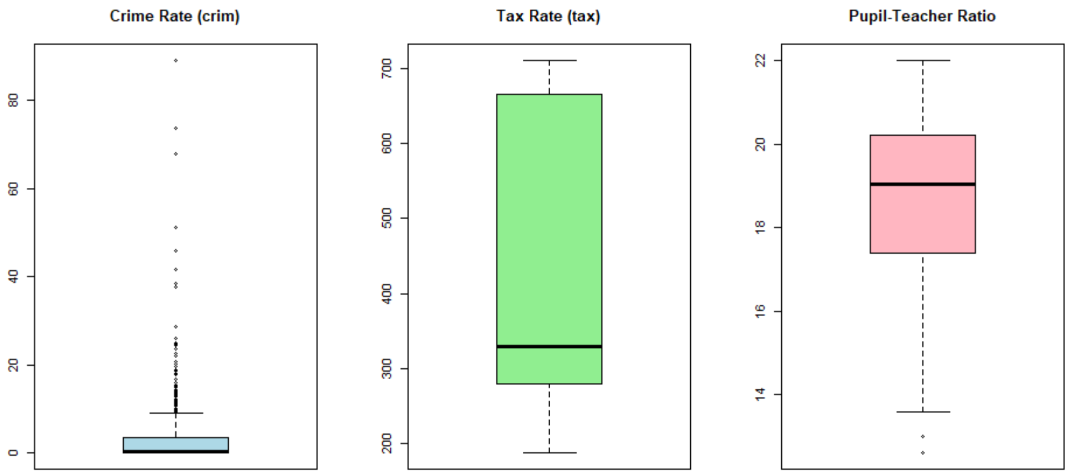
- **Crime (`crim`):** 38.35 (98.81 percentile). This suburb has one of the highest crime rates in the entire dataset.
- **NOx (`nox`):** 0.693 (85.77 percentile). The air pollution is in the upper quartile (high pollution).
- **Black (`black`):** 396.90 (100 percentile). This value is at the maximum range for this variable.
- **Lower Status (`lstat`):** 30.59 (97.83 percentile). The percentage of lower-status population is extremely high compared to other suburbs.

Comment: The suburb with the lowest housing value represents an extreme case. It suffers from the highest levels of crime and very high poverty levels (lstat).

---

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))

# Boxplot for Crime
boxplot(Boston$crim, main = "Crime Rate (crim)", col = "lightblue")
# Boxplot for Tax
boxplot(Boston$tax, main = "Tax Rate (tax)", col = "lightgreen")
# Boxplot for Pupil-Teacher Ratio
boxplot(Boston$ptratio, main = "Pupil-Teacher Ratio", col = "lightpink")
```



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

```
# Outliers for Crime
outliers_crim <- boxplot.stats(Boston$crim)$out
cat("Number of Crime outliers:", length(outliers_crim), "\n")

## Number of Crime outliers: 66

cat("Indices of suburbs with notably high crime (> 30):",
    which(Boston$crim > 30), "\n")

## Indices of suburbs with notably high crime (> 30): 381 399 405 406 411 415
419 428

# Outliers for Tax
outliers_tax <- boxplot.stats(Boston$tax)$out
cat("Number of Tax outliers:", length(outliers_tax), "\n")

## Number of Tax outliers: 0

# Identifying Outliers for PTRatio
outliers_ptratio <- boxplot.stats(Boston$ptratio)$out
cat("Number of PTRatio outliers:", length(outliers_ptratio), "\n")

## Number of PTRatio outliers: 15

cat("Indices of PTRatio outliers:", which(Boston$ptratio %in%
outliers_ptratio), "\n")

## Indices of PTRatio outliers: 197 198 199 258 259 260 261 262 263 264 265
266 267 268 269
```

Findings:

1. **Crime (`crim`):** Yes, there are significant outliers. The boxplot shows a very compressed distribution near zero with a long tail of extreme outliers reaching up to values of 80+. Suburbs like 381, 406, 419, and 428 have notably high crime rates.
2. **Tax (`tax`):** The boxplot does not show traditional “outliers” (points beyond the whiskers) in the standard definition, but there is a distinct bimodal distribution. There is a large cluster of suburbs with very high tax rates (near 666) compared to the median.
3. **Pupil-Teacher Ratio (`ptratio`):** There are a few outliers on the *low* end (indices 197, 198, 260), but not notably “high” outliers. The distribution is skewed but contained.