Predictive Analytics Assignment

Oindrila Chakraborty(728)

2026-01-19

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.

```
data1=read.csv("C:\\Users\\OINDRILA CHAKRABORTY\\Desktop\\Boston.csv")
head(data1)

##   X    crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
lstat
## 1 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
4.98
## 2 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
9.14
## 3 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
4.03
## 4 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
2.94
## 5 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
5.33
## 6 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
```

Question 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?

```
library(MASS)
data(Boston)
class(Boston)

## [1] "data.frame"

dim(Boston)
```

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

```
colnames(Boston)
```

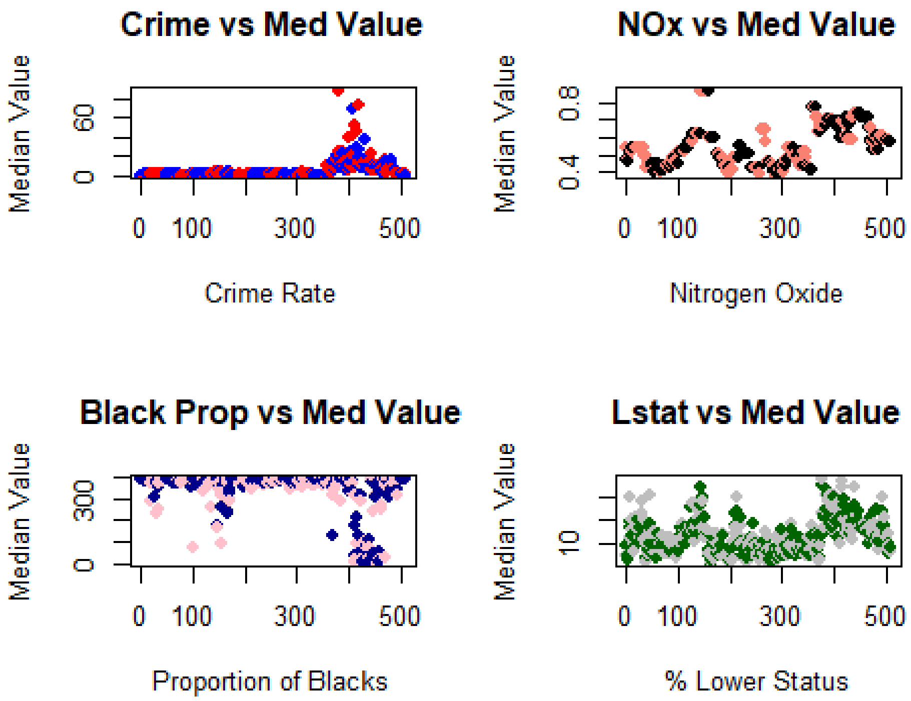
```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

Here,the dim() function gives the number of rows and columns.This data set contains 506 rows and 14 columns.Each row represents a specific suburb or town(given observations) in Boston and each column represents a specific socio-economic or environmental variable related to housing and demographics, such as crime rates, nitrogen oxide levels, and median home values.

Question 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.

```
data2=Boston[, c("crim","nox", "black", "lstat")]
par(mfrow = c(2, 2))
plot(data2$crim, data2$medv,col =c("red", "blue"), pch = 19, main="Crime vs
Med Value", xlab="Crime Rate", ylab="Median Value")
plot(data2$nox, data2$medv,col =c("salmon", "black"), pch = 19, main="NOx vs
Med Value", xlab="Nitrogen Oxide", ylab="Median Value")
plot(data2$black, data2$medv,col =c("pink", "darkblue"), pch = 19,main="Black
Prop vs Med Value", xlab="Proportion of Blacks", ylab="Median Value")
plot(data2$lstat, data2$medv,col =c("grey", "darkgreen"), pch =
19,main="Lstat vs Med Value", xlab="% Lower Status", ylab="Median Value")
```



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

For lstatvsmedv scatter plot we see,there is a clear negative correlation as the percentage of lower-status population increases, the median home value decreases significantly.

For crimvsmedv scatter plot we see,when crime rate is low there is high median values (blue points).As crime rates increase, the median values decreases.

For noxvsmedv,higher nitrogen oxide concentrations generally correspond to lower median home values, suggesting that industrial or high-traffic areas have cheaper housing.

For blackvsmedv,relationship between the proportion of blacks and home values shows a high concentration of data points at the higher end of the black variable scale. While expensive homes exist in these areas, the highest-priced outliers are mostly found in specific demographic clusters.

Question 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.

```
m1=which.min(Boston$medv)
m1

## [1] 399
```

```
data3=Boston[m1, ]
data3

##        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    20.2 396.9
30.59
##     medv
## 399    5

data3[, c("medv", "crim", "nox", "black", "lstat")]

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

crime=ecdf(Boston$crim)(data3$crim)
nox=ecdf(Boston$nox)(data3$nox)
lstat=ecdf(Boston$lstat)(data3$lstat)
print(paste("Crime Percentile:", round(crime*100, 2)))

## [1] "Crime Percentile: 98.81"

print(paste("NOx Percentile:", round(nox*100, 2)))

## [1] "NOx Percentile: 85.77"

print(paste("Lstat Percentile:", round(lstat*100, 2)))

## [1] "Lstat Percentile: 97.83"
```

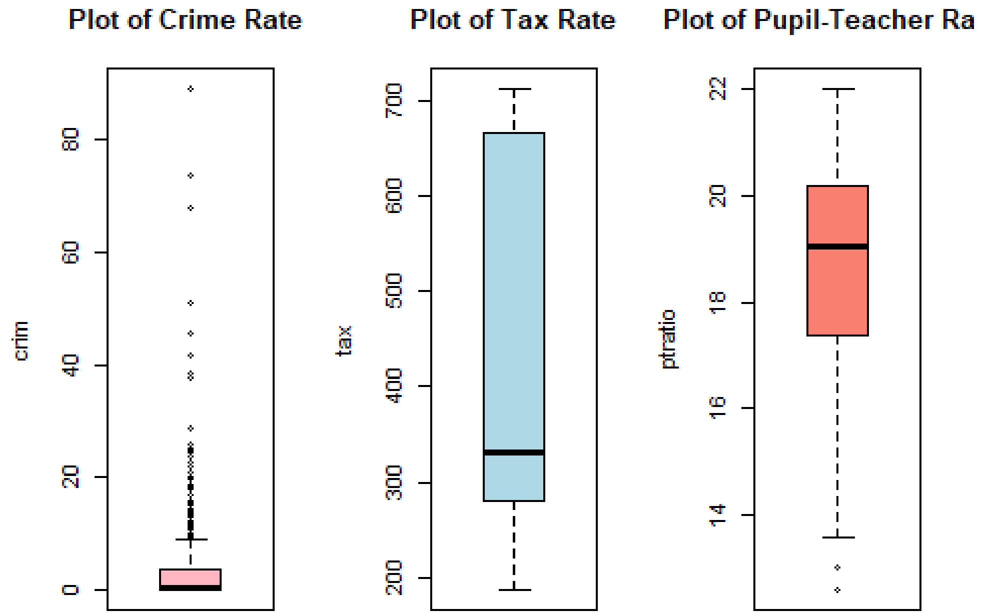
Thus, from here we can infer that this suburb falls into a very high percentile for crime(98.81 which exceeds the average).The concentration of nitrogen oxides is above the median(85.77), placing it in the upper percentiles for air pollution.This suburb has one of the highest percentages of lower-status residents in the entire data set.The value for the black variable is also high, indicating mixed ethinicities is common in the lower-priced housing tiers of this data set.

The suburb with the lowest median value(5.00 or $5,000) is characterized by extreme values in all predictor categories. It is an outlier not just in price, but in its high levels of pollution, crime, and socio-economic challenges.

Question 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$crim, main="Plot of Crime Rate ",col="lightpink", ylab="crim")
boxplot(Boston$tax, main="Plot of Tax Rate",col="lightblue",ylab="tax")
boxplot(Boston$ptratio, main="Plot of Pupil-Teacher Ratio",col="salmon",
ylab="ptratio")
```



```
outliers=function(x) {
  s=boxplot.stats(x)
  return(which(x %in% s$out))
}

outliers_found_in_crim=outliers(Boston$crim)
outliers_found_in_tax=outliers(Boston$tax)
outliers_found_in_ptratio=outliers(Boston$ptratio)

cat("Suburbs with outliers in crime:", length(outliers_found_in_crim), "\n")

## Suburbs with outliers in crime: 66

cat("Suburbs with outliers in tax:", length(outliers_found_in_tax), "\n")

## Suburbs with outliers in tax: 0

cat("Suburbs with outliers in pupil-teacher ratios:",
length(outliers_found_in_ptratio), "\n")

## Suburbs with outliers in pupil-teacher ratios: 15
```

The first boxplot reveals a large number of suburbs with high crime rates. These suburbs stand out as points far above the main box of the data, indicating that while most of Boston is relatively safe, certain areas have extreme levels of crime since a large amount of outliers is detected in this boxplot.

The second boxplot does not show outliers for tax rates because the data is skewed. However, there is a large group of suburbs with a very high tax rate, which stands out from the lower range (around 200-400).

In the third boxplot,there are a few suburbs that stand out, primarily on the lower end. These suburbs represent towns with exceptionally small class sizes compared to the rest of the dataset.

Yes, several suburbs stand out. The crime rate has the most extreme outliers, indicating a high level of inequality in safety across different Boston suburbs. The tax rate shows a distinct group of high-tax areas, and the pupil-teacher ratio identifies a few suburbs with significantly better (lower) student-to-teacher distributions than the rest of the dataset.