702_assignment

Sagnik Roy

2026-01-19

Predictive Analysis PS1

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)
class(Boston)

## [1] "data.frame"

dim(Boston)

## [1] 506  14
```

The class of this data set is data.frame.

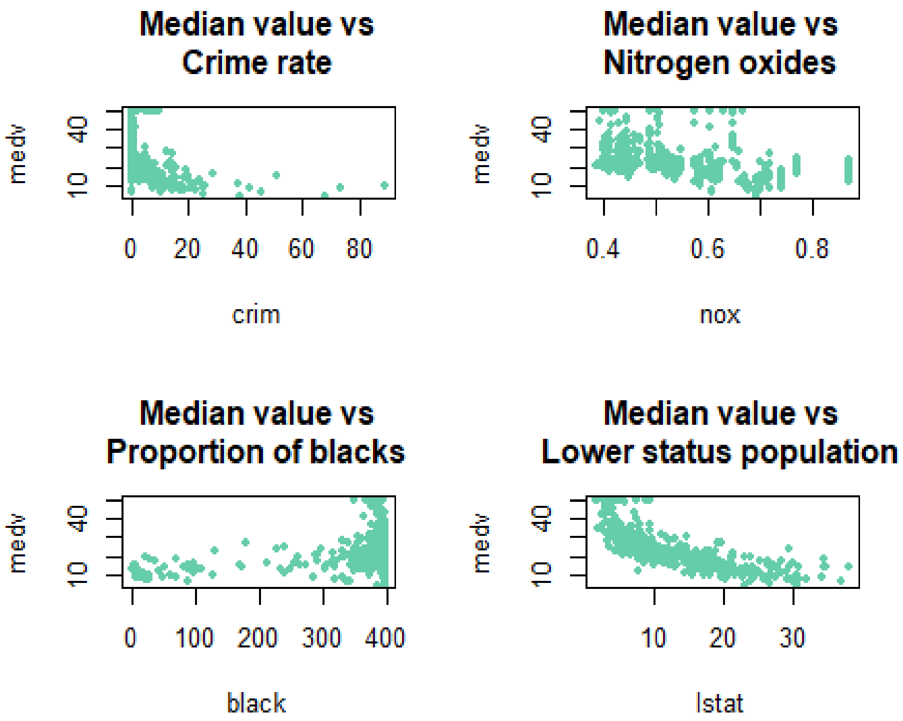
This data set has 14 columns, each representing a different housing related variable, and 506 rows, each representing a different suburb in Boston.

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.

```
Boston_sub=Boston[,c("medv","crim","nox","black","lstat")]
attach(Boston_sub)

par(mfrow=c(2,2))
plot(crim,medv,xlab="crim",ylab="medv",main="Median value vs\nCrime
rate",pch=20,col="mediumaquamarine")
plot(nox,medv,xlab="nox",ylab="medv",main="Median value vs\nNitrogen
oxides",pch=20,col="mediumaquamarine")
plot(black,medv,xlab="black",ylab="medv",main="Median value vs\nProportion of
blacks",pch=20,col="mediumaquamarine")
plot(lstat,medv,xlab="lstat",ylab="medv",main="Median value vs\nLower status
population",pch=20,col="mediumaquamarine")
```



```
detach(Boston_sub)
```

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.

```
min_medv=min(Boston$medv)
lowest_suburb=Boston[Boston$medv == min_medv,]
lowest_suburb

##        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    20.2 396.90
30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97
22.98
##     medv
## 399    5
## 406    5
```

We can see that the lowest value of medv (which is 5.0) is present for 2 different suburbs of Boston, i.e. suburb number 399 and 406.

```r
percentiles=sapply(names(Boston)[-which(names(Boston)=="medv")],
function(var){ecdf(Boston[[var]])(lowest_suburb[[var]])*100})
percentiles
```

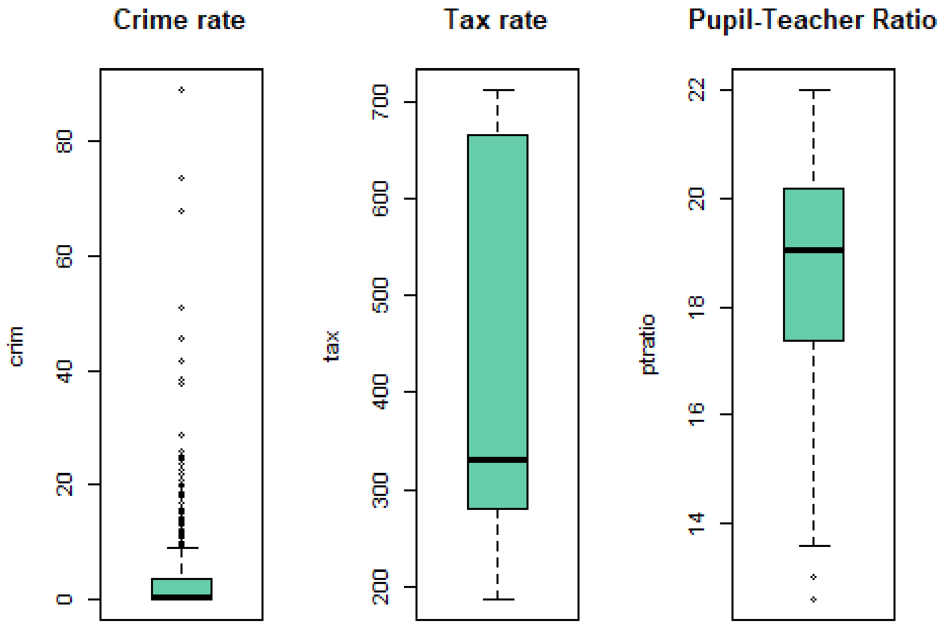
```
##          crim       zn    indus   chas      nox       rm age      dis rad
## [1,] 98.81423 73.51779 88.73518 93.083 85.77075  7.70751 100 5.731225 100
## [2,] 99.60474 73.51779 88.73518 93.083 85.77075 13.63636 100 4.150198 100
##           tax  ptratio     black    lstat
## [1,] 99.01186 88.93281 100.00000 97.82609
## [2,] 99.01186 88.93281  34.98024 89.92095
```

We can see that for most of the variables, both the suburbs have similar quantiles. We also observe that both the suburbs have very high crime rates, nitrous oxide and tax, hence resulting in the low price of the house.

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.

```r
par(mfrow=c(1,3))
boxplot(Boston$crim,main="Crime rate",ylab="crim",col="mediumaquamarine")
boxplot(Boston$tax,main="Tax rate",ylab="tax",col="mediumaquamarine")
boxplot(Boston$ptratio,main="Pupil-Teacher
Ratio",ylab="ptratio",col="mediumaquamarine")
```



The suburb with the lowest median home value lies in the extreme upper percentiles for crime rate and tax rate, indicating very high crime and property taxes relative to most suburbs. Its pupil–teacher ratio is also above the 75th percentile, suggesting larger class sizes. Overall, these predictors take unusually unfavorable values, which helps explain the exceptionally low housing prices in this suburb.

We do not need to find the outliers for the boxplot of tax because there are none.

The outliers of crime rates are given below:

```
boxplot.stats(Boston$crim)$out

##  [1] 13.52220  9.23230 11.10810 18.49820 19.60910 15.28800  9.82349
23.64820
##  [9] 17.86670 88.97620 15.87440  9.18702 20.08490 16.81180 24.39380
22.59710
## [17] 14.33370 11.57790 13.35980 38.35180  9.91655 25.04610 14.23620
9.59571
## [25] 24.80170 41.52920 67.92080 20.71620 11.95110 14.43830 51.13580
14.05070
## [33] 18.81100 28.65580 45.74610 18.08460 10.83420 25.94060 73.53410
11.81230
## [41] 11.08740 12.04820 15.86030 12.24720 37.66190  9.33889 10.06230
13.91340
## [49] 11.16040 14.42080 15.17720 13.67810  9.39063 22.05110  9.72418
9.96654
## [57] 12.80230 10.67180  9.92485  9.32909  9.51363 15.57570 13.07510
15.02340
## [65] 10.23300 14.33370
```

The outliers of pupil-teacher ratio are given below:

```
boxplot.stats(Boston$ptratio)$out

##  [1] 12.6 12.6 12.6 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0
13.0
```