Predictive Analysis Assignment 1

Sreshtha Chatterjee. Roll No: 724

2026-01-19

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

str(Boston)

## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524
0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Class of the data set: The data set is of class data.frame (pandas DataFrame equivalent).

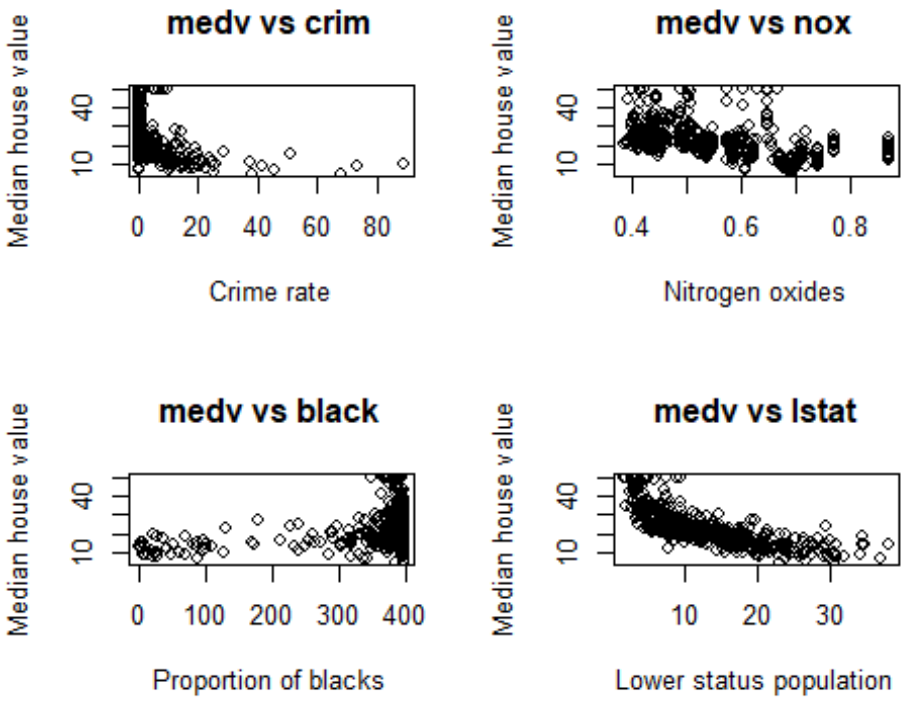
Dimensions: Number of rows: 506 Number of columns: 14

Interpretation: Rows represent 506 suburbs of Boston. Columns represent different socio-economic, environmental, and housing-related variables, such as crime rate, tax rate, pupil–teacher ratio, pollution levels, and 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 pre-dictors, 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")]
par(mfrow = c(2, 2))

plot(Boston_sub$crim, Boston_sub$medv,
     xlab = "Crime rate",
     ylab = "Median house value",
     main = "medv vs crim")

plot(Boston_sub$nox, Boston_sub$medv,
     xlab = "Nitrogen oxides",
     ylab = "Median house value",
     main = "medv vs nox")

plot(Boston_sub$black, Boston_sub$medv,
     xlab = "Proportion of blacks",
     ylab = "Median house value",
     main = "medv vs black")

plot(Boston_sub$lstat, Boston_sub$medv,
     xlab = "Lower status population",
     ylab = "Median house value",
     main = "medv vs lstat")
```



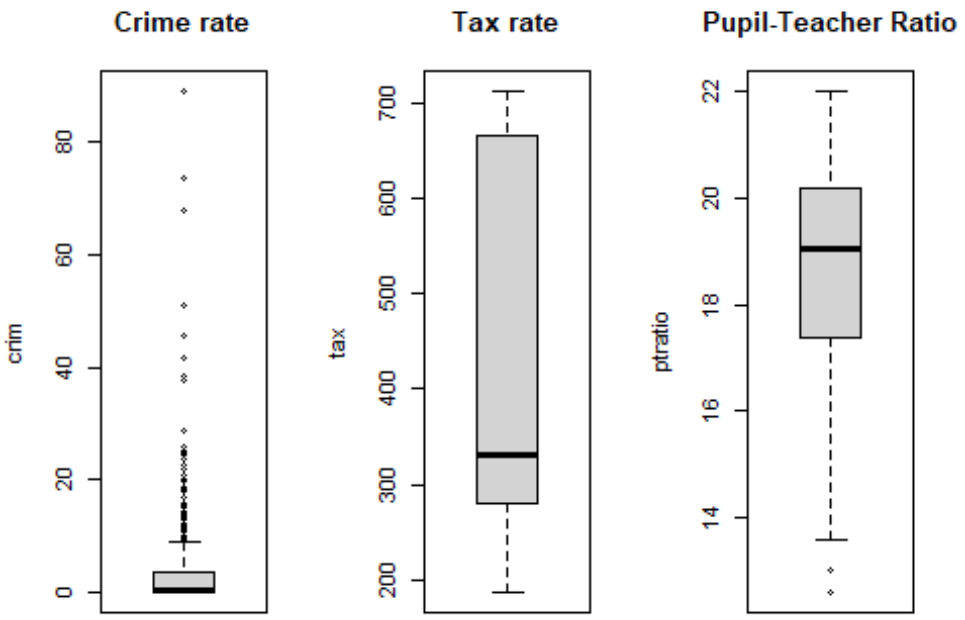
3. Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors mentioned in (2), for that suburb. How do these values compare to the overall ranges for those pre-dictors? Comment on your findings. Hint: Mention which percentile these values belong to.

```
min_medv=min(Boston$medv)
min_medv

## [1] 5

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

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
```

The percentage of lower-status residents and the crime rate are both above the 95th percentile in the suburb with the lowest median value of owner-occupied residences. This explains why the value of the house is so low.

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")
```




```
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
```

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

```
## numeric(0)
```

```
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
```

The boxplots show a number of extreme anomalies in the tax and crime rates. While the student-teacher ratio displays fewer and less extreme outliers, a few suburbs have abnormally high crime and tax values.