Exercises ISLR – Ch.2

Marcelo Previato Simoes Nº 2367070
18/08/2025

Conceptual

Exercises: 4, 7

Exercise 4a) Classification Problem

Prediction on Fraud Detection in Banking Transaction

Y = fraud classification {0 = legitimate, 1 = fraudulent}

X = multiple transaction information including:

a) **Transation features:** *Time, Amount, Amount Deviation, Velocity, Device, Device Deviation…*

b) **Customer features:** *Region, Region Mismatch, Age, Income, Regularity of transactions…*

c) **Recipient features:** *Region, New x Usual Recipient…*

Exercise 4b) Regression Problem

Inference on the Impact of Remote Work in Wealth

Y = wealth $ (in general, ln(wealth))

X = multiple information including:

a) Variable of Interest: *Remote Work (0 = No, 1 = Yes)* → inference about impact (causal inference)

b) Demographics Controls: *Region, State, Urban x Rural, Age, Age^2, Race, Gender…*

c) Experience Controls: *Years of Education, Degree, Course Category, Work Tenure, Public x Private Sector, Industry…*

Exercise 4c) Cluster Analysis

Prediction on Customer Segmentation in Grocery Purchase

Y = Customer Category (*e.g. Healthy Bulk, Healthy Light, Conventional Bulk, Conventional Light)*

X = multiple customer information:

a) Product Selections: *share of healthy food, share of processed food, diversity of SKUs, changes…*

b) Quantity Related: *average basket size, average basket value, proportion of large transactions, weight of purchase, % of large size itens, % of combo itens*

Exercise 7) Classification KNN

Obs.	X1	X2	X3	Y	D
1	0	3	0	Red	$\sqrt{9}$
2	2	0	0	Red	$\sqrt{4}$
3	0	1	3	Red	$\sqrt{10}$
4	0	1	2	Green	$\sqrt{5}$
5	-1	0	1	Green	$\sqrt{2}$
6	1	1	1	Red	$\sqrt{3}$

- K = 1 → obs 5

  P(Green) = 1 → Green

- K = 3 → obs 5,6,2

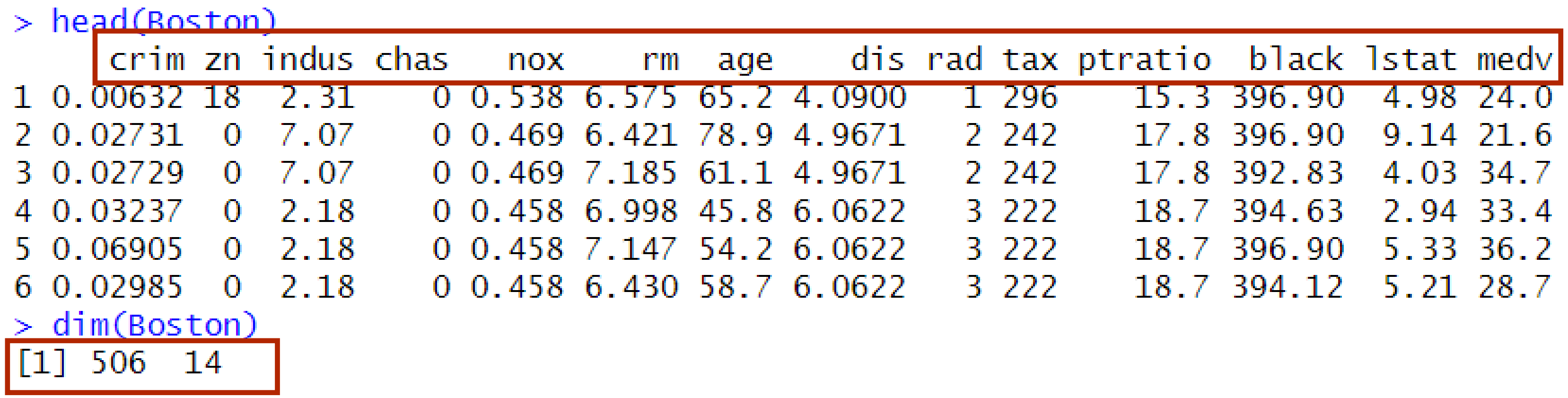
  P(Red) = 2/3 → Red

- **Non-linear → smaller K,** since it is **more flexible** and able to **capture the non linearity**

Applied

Exercises 10

Exercise 10a) Boston Suburbs Info



```
> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
> dim(Boston)
[1] 506  14
```

Greater Boston Residence Information

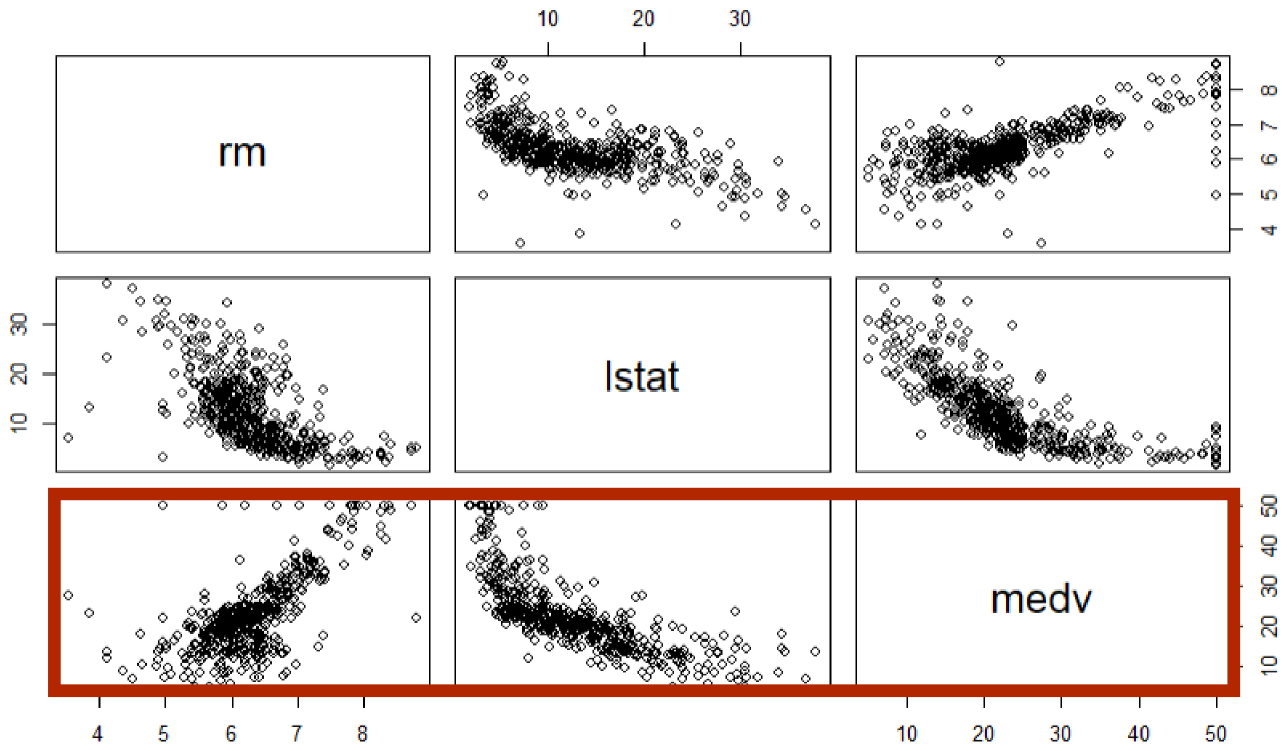
N = 506 observations

X = 14 variables (normally Medv is the dependent variable)

Exercise 10b) Covariates

X	Explanation
crim	per capita **crime rate** per town
zn	proportion of residential land zoned for **lots over 25,000 sq.ft.**
ind	proportion of **non-retail business** acres per town (**industrial**)
chas	**Charles river** bound
nox	**nitrogen oxides** concentration (**pollution**)
rm	average **number of rooms** per dwelling
age	proportion of owner-occupied units built **prior to 1940** (**antique**)
dis	weighted mean of **distances** to five Boston employment centres.
rad	index of **accessibility** to radial highways
tax	full-value **property-tax rate**
ptratio	**pupil-teacher** ratio by town.
black	Index based on **proportion of blacks** by town
lstat	**lower status** of the population
medv	median **value of owner-occupied** homes

Exercise 10c) Scatterplots



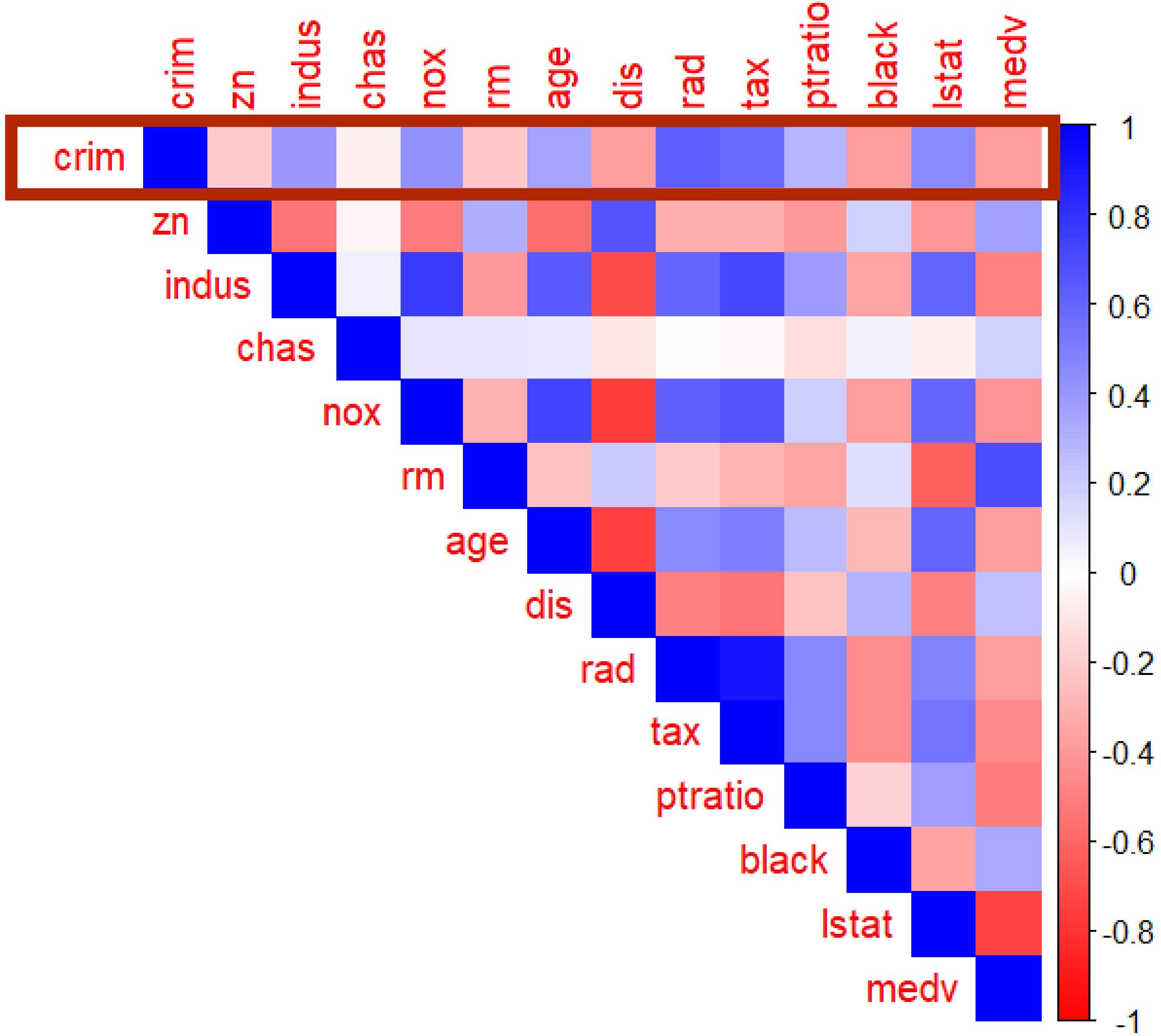
› **Cov(rm, medv) +**

The higher the number of rooms, the higher value of the residence

› **Cov(lstart, medv) -**

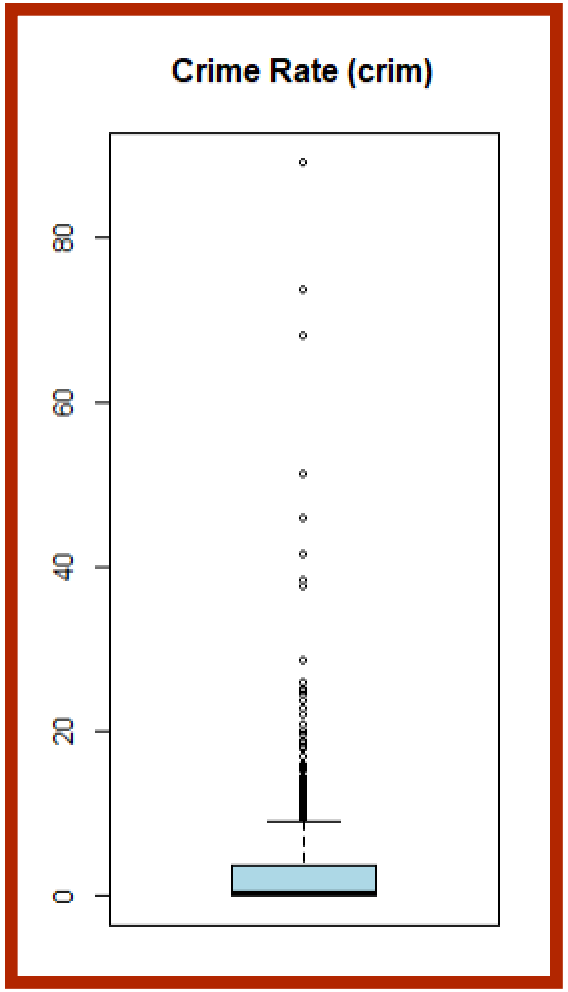
The higher the share of low status people, the lower the value of the residence

Exercise 10d) Correlations

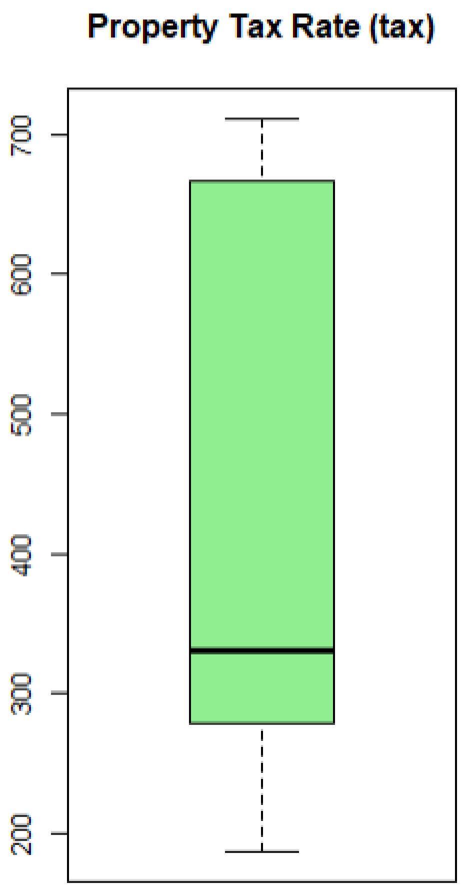


- **Positive:** indus, rad, nox → areas near industries and roads tend to have higher crime rate
- **Negative:** dis, black, medv → areas with less expensive residences, distant and with black residences tend to have lower crime rate

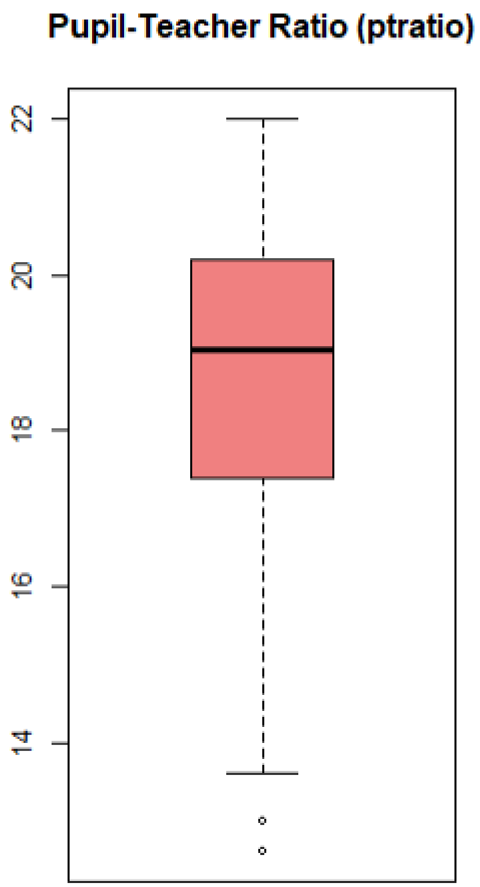
Exercise 10e) High values



Crime Rate (crim)

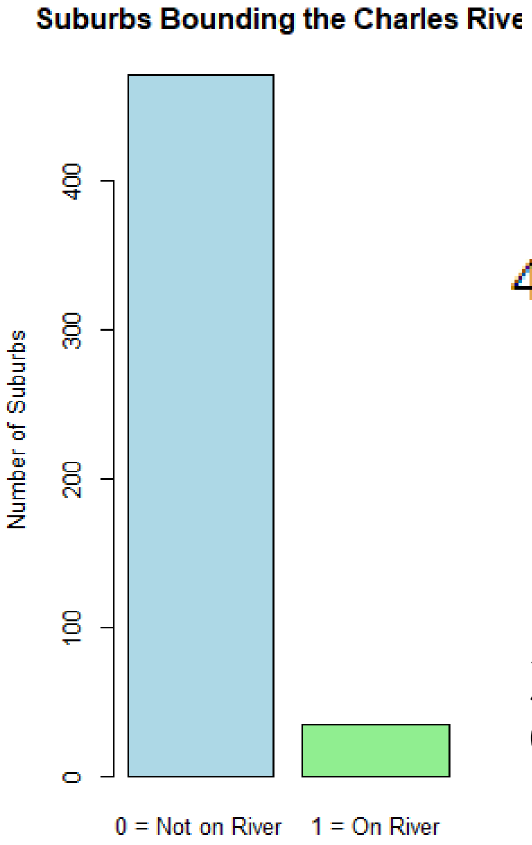


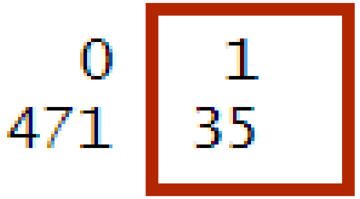
Property Tax Rate (tax)



Pupil-Teacher Ratio (ptratio)

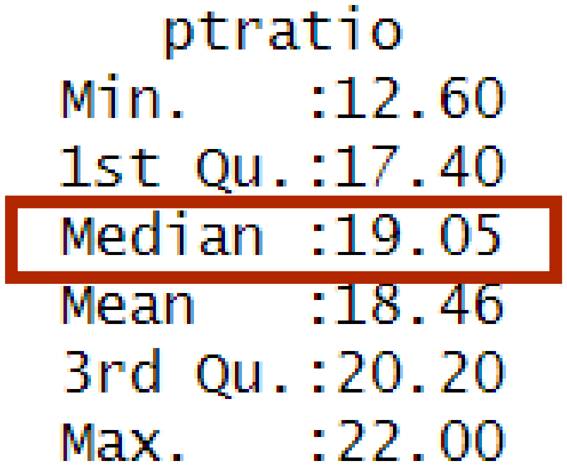
Exercise 10f,g) Descriptive





```
  0    1
471   35
```

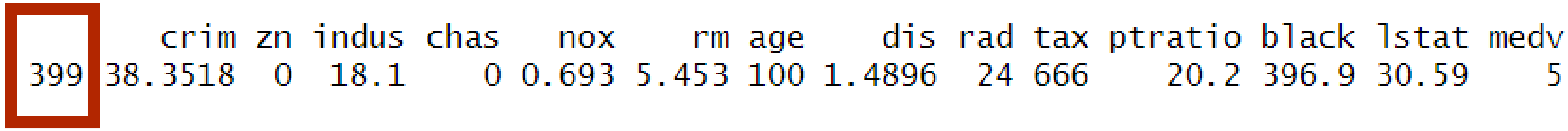
35 suburbs by
Charles river



```
    ptratio
 Min.   :12.60
 1st Qu.:17.40
 Median :19.05
 Mean   :18.46
 3rd Qu.:20.20
 Max.   :22.00
```

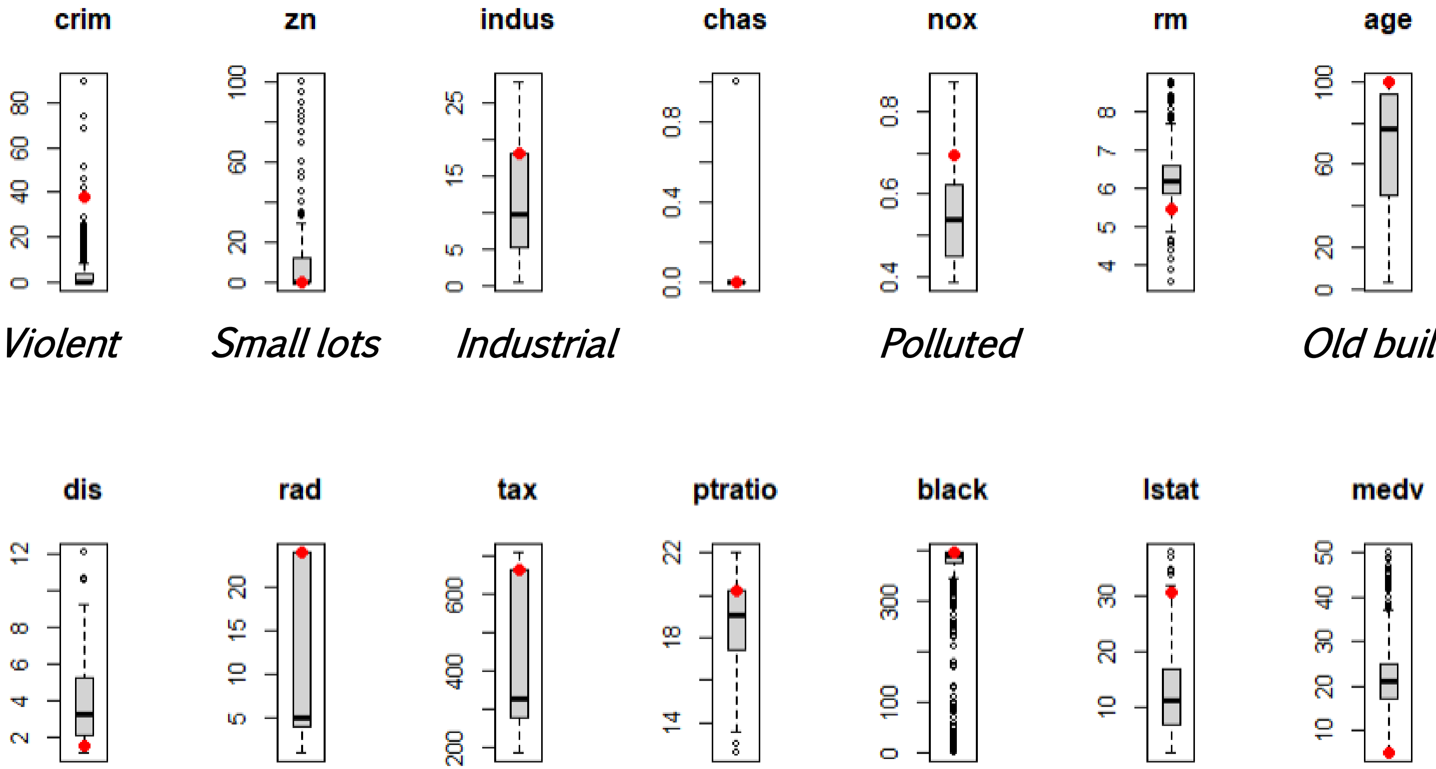
Median: 19.05

Exercise 10h) Lowest medv suburb



```
       crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59    5
```

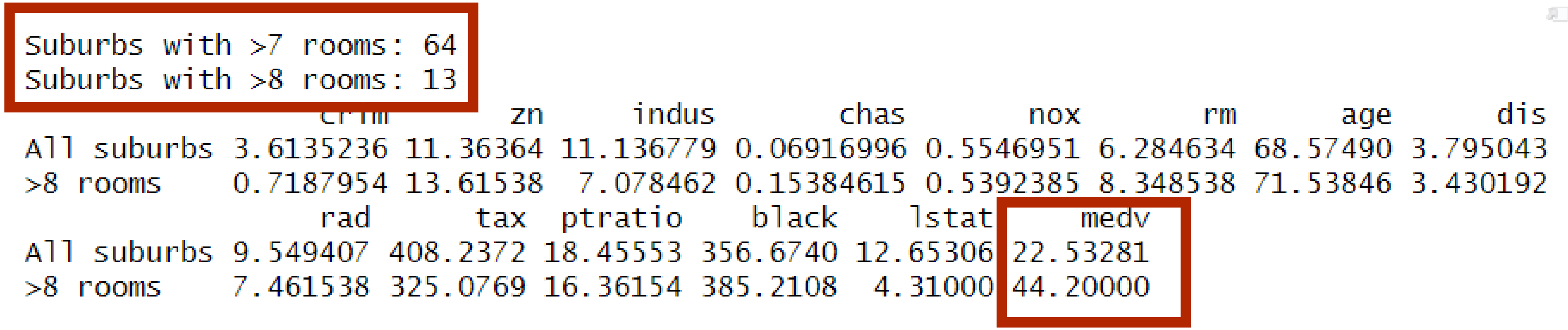
Suburb 399



Violent *Small lots* *Industrial* *Polluted* *Old buildings*

Near roads *High tax* *More black*

Exercise 10i) Larger residences



```
Suburbs with >7 rooms: 64
Suburbs with >8 rooms: 13
                 crim       zn     indus       chas       nox       rm      age      dis
All suburbs 3.6135236 11.36364 11.136779 0.06916996 0.5546951 6.284634 68.57490 3.795043
>8 rooms    0.7187954 13.61538  7.078462 0.15384615 0.5392385 8.348538 71.53846 3.430192
                 rad      tax  ptratio    black    lstat     medv
All suburbs 9.549407 408.2372 18.45553 356.6740 12.65306 22.53281
>8 rooms    7.461538 325.0769 16.36154 385.2108  4.31000 44.20000
```

Suburbs witn >8 rooms are twice as expensive as the entire database