Assignment 4: The Boston Housing Prices

Chang

5/8/2021

## The Business Goal

The dataset can be used to make useful economic predictions about the housing prices of a new house by observing the plot picture. The business goal is to predict the value of owner-occupied homes in the future.

## Source

The Boston Housing Dataset is retrieved from the U.S Census Service concerning housing in the area of Boston Mass. The StatLib archive collected the dataset at the website (<http://lib.stat.cmu.edu/datasets/boston>), and has extensively procedured the literature to benchmark algorithms. Conversely, they made these primarily done comparisons, which are somewhat suspect outside of Delve. Lastly, the dataset is in the small size with only 506 observations.

Harrison, D. and Rubinfeld, D.L originally published the data in `Hedonic prices and the demand for clean air’, J. Environ. Economics & Management, vol.5, 81-102, 1978.

## Code

The dataset is raw in HTML. It is saved in the boston information.txt file by using the Save As with the text format.

In Excel: Open the Excel spreadsheet where you want to save the data and click the Data tab. In the Get External Data group, click From Text. Select the TXT or CSV file you want to convert and click Import. Select “Delimited”. Click Next.

dataset <- read.csv(“Boston house price dataset.csv”, stringsAsFactors = FALSE)

The code, colnmaes(dataset) is to show what the 14 variables’ column names are. Another code, str(dataset) is to demonstrate the internal structure of a R object.

colnames(dataset)

## [1] "CRIM" "ZN" "INDUS" "CHAS" "NOX" "RM" "AGE"   
## [8] "DIS" "RAD" "TAX" "PTRATIO" "B" "LSTAT" "MEDV"

str(dataset)

## '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 : chr "65.2" "78.9" "61.1" "45.8" ...  
## $ 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 : int 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 ...  
## $ B : 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 ...

## Dataset Description

There are 14 attributes in each case of the natural 506-case dataset

CRIM = per capita crime rate by town

ZN = proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS = proportion of non-retail business acres per town

CHAS = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX = nitric oxides concentration (parts per 10 million)

RM = average number of rooms per dwelling

AGE = proportion of owner-occupied units built prior to 1940

DIS = weighted distances to five Boston employment centres

RAD = index of accessibility to radial highways

TAX = full-value property-tax rate per $10,000

PTRATIO = pupil-teacher ratio by town

B = 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT = % lower status of the population

MEDV = Median value of owner-occupied homes in $1000’s

The code, summary(dataset$MEDV), is to make the five-number summary in statistics to show the median value of owner-occupied homes in 1000’s.

summary(dataset$MEDV)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 17.02 21.20 22.53 25.00 50.00

The minimum of censoring price is 5,000 dollars. The median price is 21,200 dollars. The average price is 22,530 dollars. The maximum of censoring price is 50,000 dollars. There are 16 cases in the highest median price of exactly 50,000.

The code, sd(medv) tells what the standard deviation of the housing price is.

medv = dataset$MEDV  
sd(medv)

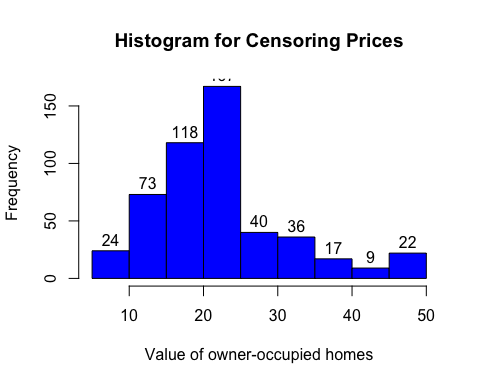
## [1] 9.197104

The standard deviation of the censoring price in the dataset is 9.20.

# Any data preparation and errors

The dataset is raw in the HTML. There are a few paragraphs like introduction and explanation at the top. The dataset of housing pricing is under the paragraphs. It is difficult to make the only dataset when the mix of paragraphs and dataset is imported into RStudio to show the errors. Removing and moving the paragraphed information is separated to save into another new .txt file. The one observation within the dataset has its own two rows. It is necessary to delete between the spaces back next to the end of the first row each time in 506 observations. The cleaned dataset is successfully imported into the RStudio by using the code of read.csv.

## Statistical Histogram in the modeling



## Output Discussion with any visuals

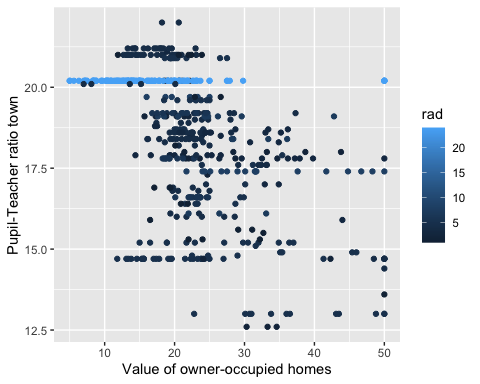
As for observing the statistical histogram, there are 506 observations in the housing values in Boston. The price between 40,000 and 45,000 is reported in the lowest 9 cases, while the one between 20,000 and 25,000 is in the highest 167 cases. According to the calculated summary, the median is 21,200 and mean is 22,530. They are very closed prices in the same histogram within the highest cases. The closed numbers around 23 cases have both the minimum and maximum of censoring prices in both the range of 5,000 to 10,000 and 45,000 to 50,000.

## Statistical Plot in the modeling

The summary is for the variable of PTRATIO, which is pupil-teacher ratio by town.

summary(dataset$PTRATIO)

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



## Output Discussion with any visuals

As for observing the statistical plots, there are more pupil-teacher relationships with the highest index of accessibility to radial highways at the number of 20.20, which has the highest index of accessibility to radial highways in the colorful light-blue plots in a row. The statistical plot shows that many parents would like to have their houses to be easily accessible to the good pupil-teacher school for their children’ good learning environment.