What is the complete guide to the Boston dataset in R?

Authored by

stats writer

June 25, 2024

The complete guide to the Boston dataset in R is a comprehensive overview of the widely used dataset in the field of statistics and data analysis. This guide provides detailed information on the structure, variables, and characteristics of the dataset, as well as step-by-step instructions on how to import, manipulate, and analyze the data using the R programming language. It also includes useful tips and techniques for data visualization and interpretation, making it a valuable resource for researchers, students, and professionals looking to gain a deeper understanding of the Boston dataset and its applications in statistical analysis.

A Complete Guide to the Boston Dataset in R

The Boston dataset from the MASS package in R contains information about various attributes for suburbs in Boston, Massachusetts.

This tutorial explains how to explore, summarize, and visualize the Boston dataset in R.

Load the Boston Dataset

Before we can view the Boston dataset, we must first load the MASS package:

library(MASS)

We can then use the head() function to view the first six rows of the dataset:

#view first six rows of Boston dataset

```
head(Boston)

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

To view a description of each variable in the dataset, we can type the following:

#view description of each variable in dataset
?Boston

This data frame contains the following columns:

'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' nitrogen 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 mean of 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.

'black' 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by
town.

'lstat' lower status of the population (percent).

'medv' median value of owner-occupied homes in $1000s.

Summarize the Boston Dataset

We can use the summary() function to quickly summarize each variable in the dataset:

```
#summarize Boston dataset
summary(Boston)
```

crim zn indus chas
Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000
1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000
Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000
Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917
3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000
Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000
nox rm age dis
Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130
1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100
Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207
Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795
3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188
Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127
rad tax ptratio black
Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32
1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38
Median : 5.000 Median :330.0 Median :19.05 Median

:391.44

Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67

3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23

Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90

lstat medv

Min. : 1.73 Min. : 5.00

1st Qu.: 6.95 1st Qu.:17.02

Median :11.36 Median :21.20

Mean :12.65 Mean :22.53

3rd Qu.:16.95 3rd Qu.:25.00

Max. :37.97 Max. :50.00

For each of the numeric variables we can see the following information:

Min: The minimum value.1st Qu: The value of the first quartile (25th percentile).Median: The median value.Mean: The mean value.3rd Qu: The value of the third quartile (75th percentile).Max: The maximum value.

We can use the dim() function to get the dimensions of the dataset in terms of number of rows and number of

columns:

#display rows and columns
dim(Boston)

506 14

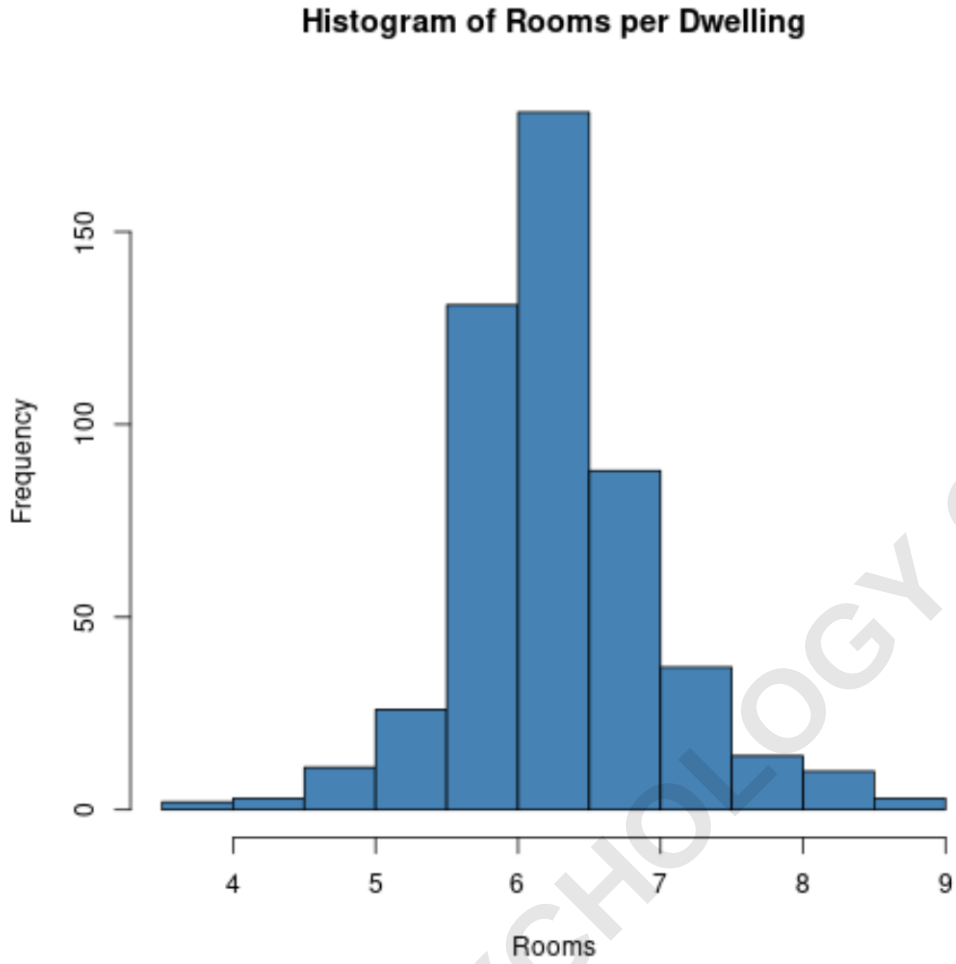
We can see that the dataset has 506 rows and 14 columns.

Visualize the Boston Dataset

We can also create some plots to visualize the values in the dataset.

For example, we can use the hist() function to create a histogram of the values for a certain variable:

```
#create histogram of values for 'rm' column
hist(Boston$rm,
col='steelblue',
main='Histogram of Rooms per Dwelling',
xlab='Rooms',
ylab='Frequency')
```



We can also use the plot() function to create a scatterplot of any pairwise combination of variables:

```
#create scatterplot of median home value vs crime rate
plot(Boston$medv, Boston$crime,
col='steelblue',
main='Median Home Value vs. Crime Rate',
xlab='Median Home Value',
ylab='Crime Rate',
```

pch=19)

We can create a similar scatterplot to visualize the relationship between any two variables in the dataset.

The following tutorials provide a complete guide to other popular datasets in R: