

# Banco de dados: Boston House Prices

Fernando Bispo, Jeff Caponero

# Sumário

<b>Introdução</b>	<b>2</b>
<b>Metodologia</b>	<b>3</b>
Sobre o conjunto de dados . . . . .	3
Variáveis a serem analisadas . . . . .	3
<b>Resultados</b>	<b>4</b>
Análise Descritiva dos Dados . . . . .	4
<!-- \usepackage{parskip} -->	
<!-- \setlength{\parsident}{1.25cm} -->	
<!-- \setlength{\parskip}{1.5cm} -->	

# Introdução

A busca pela moradia própria é o desejo da grande maioria das pessoas, contudo a conquista desse bem nos grandes centros não é tarefa fácil. Levando isso em consideração a procura por imóveis na região metropolitana torna-se uma opção viável economicamente, mesmo havendo penalizações no que diz respeito a distância e congestionamentos.

O objetivo deste relatório é trazer a luz as análises e conclusões acerca da utilização das técnicas de regressão linear a fim de determinar o preço das casas em Boston, baseado nos dados fornecidos pelo conjunto de dados obtido. Neste primeiro momento, em que se utilizará a regressão linear simples, se buscará determinar uma função que descreva a relação entre o Valor Médio dos imóveis e o Índice de Criminalidade.

Composto por 506 observações e 14 variáveis, o conjunto de dados, publicado no *Jornal of Environmental Economics & Management*, vol.5, 81-102, 1978.t, traz inúmeras características que servirão de parâmetros para resolução do seguinte questionamento: O valor médio dos imóveis é influenciado pelas diversas características externas observadas?

# Metodologia

## Sobre o conjunto de dados

Os dados de preços de 506 casas em Boston, publicados em Harrison, D. and Rubinfeld, D.L. *'Hedonic prices and the demand for clean air'*, J. Environ. Economics & Management, vol.5, 81-102, 1978, podem ser acessados na plataforma para aprendizado de ciência de dados [Kaggle](https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data) através do link:

<https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data>.

## Variáveis a serem analisadas

1. CRIM: Índice de criminalidade per capita por bairro.
2. ZN: Proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft.
3. INDUS: Proporção de hectares de negócios não varejistas por bairro.
4. CHAS: Margem do rio Charles (1 se o trecho margeia o rio; 0 caso contrário).
5. NOX: Concentração de óxidos nítricos (partes por 10 milhões).
6. RM: Número médio de cômodos por habitação.
7. AGE: Proporção de unidades próprias construídas antes de 1940.
8. DIS: Distâncias ponderadas para cinco centros de emprego de Boston.
9. RAD: Índice de acessibilidade às rodovias radiais.
10. TAX: Valor total do imposto predial por \$10.000.
11. PTRATIO: Proporção aluno-professor por bairro.
12. B: O resultado da equação  $B = 1000(Bk - 0,63)^2$  onde  $Bk$  é a proporção de negros por bairro.
13. LSTAT: % da população de “classe baixa”.

# Resultados

## Análise Descritiva dos Dados

Existem variáveis que em virtude da sua natureza e da falta de um dicionário, inviabilizam a interpretação e portanto serão excluídas das análises, sendo estas:

1. ZN: Proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft.
2. RAD: Índice de acessibilidade às rodovias radiais.
3. B: O resultado da equação  $B = 1000(Bk - 0,63)^2$  onde  $Bk$  é a proporção de negros por bairro.

Tabela 1: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
age	2,90	45,00	77,50	68,57	94,10	100,00	28,15	0,41	-0,60	-0,98
crim	0,01	0,08	0,26	3,61	3,68	88,98	8,60	2,38	5,19	36,60
dis	1,13	2,10	3,21	3,80	5,21	12,13	2,11	0,55	1,01	0,46
indus	0,46	5,19	9,69	11,14	18,10	27,74	6,86	0,62	0,29	-1,24
lstat	1,73	6,93	11,36	12,65	16,96	37,97	7,14	0,56	0,90	0,46
medv	5,00	17,00	21,20	22,53	25,00	50,00	9,20	0,41	1,10	1,45
nox	0,38	0,45	0,54	0,55	0,62	0,87	0,12	0,21	0,72	-0,09
ptratio	12,60	17,40	19,05	18,46	20,20	22,00	2,16	0,12	-0,80	-0,30
rm	3,56	5,88	6,21	6,28	6,62	8,78	0,70	0,11	0,40	1,84
tax	187,00	279,00	330,00	408,24	666,00	711,00	168,54	0,41	0,67	-1,15

Fonte: StatLib - Carnegie Mellon University

Legenda:

age: Proporção de unidades próprias construídas antes de 1940.

crim: Índice de criminalidade per capita por bairro.

dis: Distâncias ponderadas para cinco centros de emprego de Boston.

indus: Proporção de hectares de negócios não varejistas por bairro.

lstat: Percentual da população de "classe baixa".

mdev: Valor médio de residências ocupadas pelo proprietário.

nox: Concentração de óxidos nítricos (partes por 10 milhões).

ptratio: Proporção aluno-professor por bairro.

rm: Número médio de cômodos por habitação.

tax: Valor total do imposto predial por \$10.000.

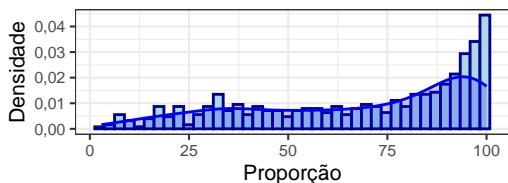
Analisando a Tabela 1 é possível identificar a presença de pontos atípicos para a variável **crim**, sendo esta a variável que mais se destaca pelos indicadores obtidos, gerando assim uma grande variabilidade, confirmada pelo Coeficiente de Variação - CV, sendo esta medida que avalia a dispersão dos dados em relação a média, além de verificar um elevado Coeficiente de Assimetria Positiva, indicando que a maioria dos dados são menores que a média e um comportamento Leptocúrtico, com base no Coeficiente de Curtose, ou seja, os dados possuem um comportamento mais alongado.

A Figura 1 traz os Histogramas das variáveis a fim de tentar identificar o comportamento das variáveis e uma possível associação a algum modelo de distribuição

Figura 1: Histogramas das variáveis em análise.

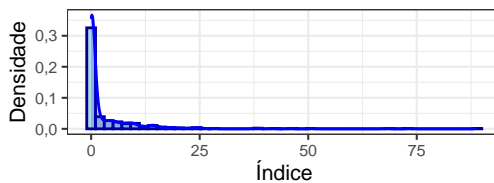
Sub Fig. A:

Unidades constuídas  
antes de 1940



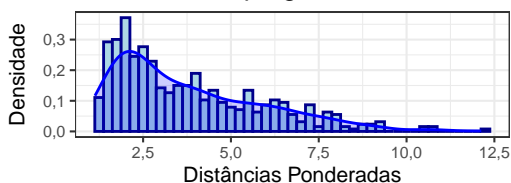
Sub Fig. B:

Índice de Criminalidade



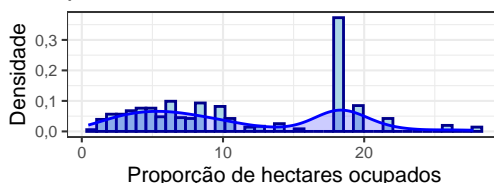
Sub Fig. C:

Distância para cinco  
centros de emprego.



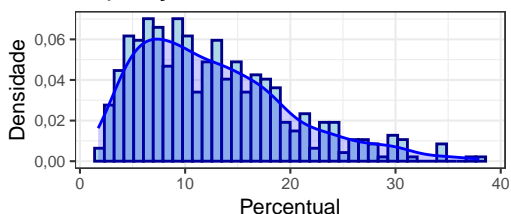
Sub Fig. D:

Negócios não varejistas  
por bairro



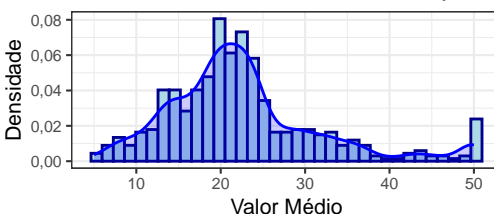
Sub Fig. A:

População de "classe baixa"



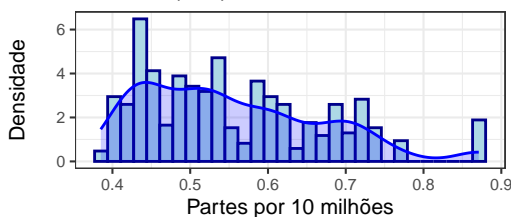
Sub Fig. B:

Valor médio de residências ocupadas



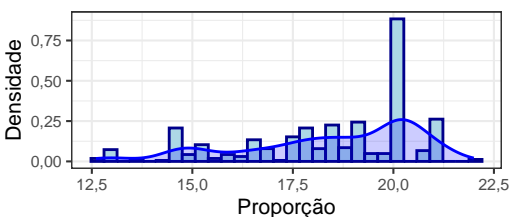
Sub Fig. C:

Concentração de Óxidos  
Nítricos (NO)

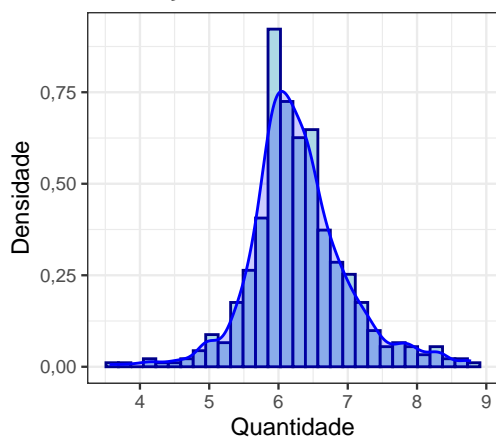


Sub Fig. D:

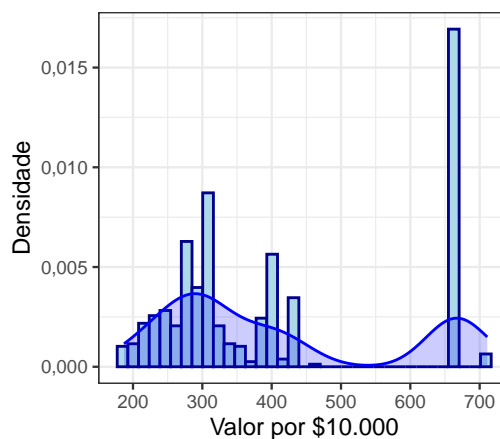
Aluno/Professor por bairro



Sub Fig. A:  
Número médio de cômodos por habitação



Sub Fig. B:  
Valor total do imposto predial



Fonte: StatLib – Carnegie Mellon University