

Análise Estatística do Boston Housing Dataset

Concentração, Distribuição, Moda e Correlação

Análise e Desenvolvimento de Sistemas

Data Science - 4º Semestre

João Pedro dos Santos

06 de Outubro de 2025

1. Introdução e Objetivos

Dataset Boston Housing:

- 506 observações (imóveis)
- 14 variáveis numéricas e categóricas

CrITÉrios de Análise:

1. **Concentração e Distribuição** de todas as colunas numéricas
2. **Moda** das colunas categóricas (CHAS e RAD)
3. **Correlação** entre todos os pares de colunas numéricas
4. **Análise de quartis** com boxplots
5. **Hipóteses comparativas** com valores dos imóveis

2. Análise Completa de Concentração e Distribuição

Script: `concentracao_distribuicao.py` - TODAS as 14 variáveis numéricas

Cobertura Expandida:

- **TODAS:** CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV
- **Alta Variabilidade ($CV > 50\%$):** CHAS (373%), CRIM (246%), ZN (214%), RAD (91%), INDUS (61%), DIS (56%), LSTAT (55%)
- **Baixa Variabilidade ($CV < 20\%$):** RM (11.2%), PTRATIO (11.7%)
- **Simétricas:** RM, INDUS | **Assimétricas à Direita:** Maioria

Visualização: Grade 4x4 com histogramas de **todas as variáveis** incluindo média e mediana.

3. Análise Executiva Completa - Visão 360°

Script: `analise.py` - Resumo executivo expandido de todas as variáveis

Estatísticas Executivas:

- **Dataset:** 506 imóveis, 14 variáveis numéricas, 120 valores ausentes tratados
- **Preço (MEDV):** Média \$22.53k, Mediana \$21.20k, CV 40.8%
- **Melhor Preditor:** LSTAT ($r=-0.723$) - Status socioeconômico

Correlações Categorizadas:

- **Fortes ($|r|>0.6$):** LSTAT (-0.723), RM (0.695)
- **Moderadas ($0.3-0.6$):** 9 variáveis
- **Fracas (≤ 0.3):** DIS, CHAS

Total Outliers: 289 registros (57.1%) - Indica heterogeneidade urbana moderada

4. Análise da Moda - Colunas Categóricas

Script: `moda_categorica.py` - Identificação automática de variáveis categóricas

CHAS (Acesso ao Rio Charles):

- **Moda:** 0.0 (sem acesso)
- **Frequência:** 472 (93.3%)
- **Interpretação:** Muito concentrada - maioria não tem acesso ao rio

RAD (Índice de Acessibilidade):

- **Moda:** 24 (alta acessibilidade)
- **Frequência:** 132 (26.1%)
- **Interpretação:** Dispersa - 9 categorias com distribuição mais equilibrada

Gráfico: Gráficos de barras com percentuais para visualização das frequências.

5. Análise de Correlação - Pares de Colunas Numéricas

Script: `correlacao.py` - Análise completa de todos os pares

Matriz Completa: 14x14 variáveis = 91 pares únicos analisados

Top 8 Correlações Mais Fortes:

- RAD ↔ TAX: 0.910 (Acessibilidade vs Impostos)
- NOX ↔ DIS: -0.769 (Poluição vs Distância do emprego)
- INDUS ↔ NOX: 0.738 (Indústria vs Poluição)
- LSTAT ↔ MEDV: -0.723 (Status vs Preço) ★

Gráfico: Heatmap de correlação completo com todas as relações.

6. Análise de Quartis com Boxplots

Script: `analise_quartis.py` - Análise de dispersão das 6 principais variáveis

Estatísticas dos Quartis:

- **LSTAT**: Q1=7.23, Q2=11.43, Q3=16.57, IQR=9.34
- **MEDV**: Q1=17.02, Q2=21.20, Q3=25.00, IQR=7.98
- **CRIM**: Q1=0.08, Q2=0.25, Q3=2.81, IQR=2.73

Outliers Identificados:

- **CRIM**: 81 outliers (16.0%) - Áreas com criminalidade extrema
- **MEDV**: 40 outliers (7.9%) - Imóveis com preços atípicos

Gráfico: Boxplots em grid 2x3 mostrando dispersão e valores extremos.

7. Hipótese Comparativa 1: Quartos vs Preço

Hipótese: Mais quartos (`RM`) = Maior preço (`MEDV`)?

Análise Comparativa por Categorias:

- ≤ 5.5 quartos: \$15.2k (n=42)
- 5.5-6.5 quartos: \$19.4k (n=312)
- > 6.5 quartos: \$31.1k (n=152)

Teste Estatístico:

- Correlação: $r = 0.695$ (forte positiva)
- P-valor: < 0.001 (significativo)

✓ Conclusão: Imóveis com mais quartos valem **104% mais**.

8. Hipótese Comparativa 2: Status Socioeconômico vs Preço

Hipótese: Melhor status (LSTAT baixo) = Maior preço (MEDV)?

LSTAT como Melhor Preditor: $r = -0.723$ (correlação mais forte)

Análise Comparativa:

- Status Alto (LSTAT baixo): Preços mais altos
- Status Baixo (LSTAT alto): Preços mais baixos

Teste Estatístico:

- Correlação: $r = -0.723$ (forte negativa)
- P-valor: < 0.001 (altamente significativo)

✓ Conclusão: Status socioeconômico é o fator mais determinante do preço.

9. Conclusões

✓ **Concentração/Distribuição:** Todas as 14 colunas numéricas analisadas

- CV, assimetria e formato de distribuição identificados

✓ **Moda Categórica:** CHAS (93.3% sem acesso) e RAD (disperso)

- Frequências e interpretações fornecidas

✓ **Correlação de Pares:** 91 pares analisados

- Top 5 correlações mais fortes identificadas, matriz completa gerada

✓ **Quartis:** Boxplots com IQR e outliers

- Dispersão e valores extremos quantificados

✓ **Hipóteses Comparativas:** Testes estatísticos rigorosos

- Correlações confirmadas com $p\text{-valor} < 0.001$

10. Resumo Final - Insights Expandidos

Descobertas Expandidas:

1. **ANÁLISE EXECUTIVA:** 57.1% dos registros têm outliers - heterogeneidade urbana moderada
2. **Status socioeconômico (LSTAT)** é o preditor mais forte ($r=-0.723$)
3. **VARIABILIDADE EXTREMA:** CHAS (373%), CRIM (246%), ZN (214%)
4. **ESTABILIDADE:** Apenas RM (11.2%) e PTRATIO (11.7%) são estáveis
5. **COBERTURA TOTAL:** 14/14 variáveis numéricas completamente analisadas
6. **CORRELAÇÕES CATEGORIZADAS:** 2 fortes, 9 moderadas, 2 fracas

Metodologia Simplificada: Análise executiva 360°, cobertura completa de todas as variáveis, visualizações expandidas (grade 4x4), classificações inteligentes, estrutura modular otimizada.