# Análise Exploratória do California Housing Dataset
## Introdução

Este relatório apresenta os resultados da análise exploratória do conjunto de dados "California Housing Dataset", que contém informações sobre residências e seus respectivos valores em diferentes regiões da Califórnia. A análise foi realizada utilizando as bibliotecas Python, Pandas, NumPy, Seaborn e Matplotlib, conforme especificado nos requisitos do trabalho.

## Metodologia

Base de Dados: O dataset utilizado contém 20.640 observações com 8 atributos cada:

MedInc: Renda média familiar na região

HouseAge: Idade média do imóvel na região

AveRooms: Número médio de cômodos no imóvel

AveBedrms: Número médio de quartos no imóvel

Population: População na região

AveOccup: Número médio de membros na família

Latitude: Latitude da região

Longitude: Longitude da região

A variável alvo (MedHouseVal) representa o valor do imóvel em múltiplos de US$ 100.000.

Ferramentas Utilizadas
Pandas para manipulação e análise de dados

NumPy para cálculos matemáticos

Seaborn e Matplotlib para visualizações

Scikit-learn para acesso ao dataset

Resultados por Requisito
Requisito 1: Estatísticas Descritivas Básicas

Observa-se uma grande variação nos valores, especialmente em AveRooms e AveOccup que apresentam valores máximos extremamente altos em comparação com a média, sugerindo a presença de outliers.

Requisito 2: Visualização Geográfica dos Imóveis
O gráfico de dispersão gerado com Latitude no eixo Y e Longitude no eixo X revela a distribuição geográfica dos imóveis na Califórnia. A visualização mostra claramente o formato geográfico do estado, com maior concentração de imóveis nas regiões costeiras, especialmente nas áreas metropolitanas de São Francisco e Los Angeles.

https://mapa_dispersao.png

Requisito 3: Métricas Estatísticas
Foram calculadas as seguintes métricas para as seis primeiras variáveis:

Média:

MedInc: 3.8707

HouseAge: 28.6395

AveRooms: 5.4290

AveBedrms: 1.0967

Population: 1425.4767

AveOccup: 3.0707

Mediana:

MedInc: 3.5348

HouseAge: 29.0000

AveRooms: 5.2291

AveBedrms: 1.0488

Population: 1166.0000

AveOccup: 2.8181

Moda:
Cada variável apresenta múltiplas modas, sendo os valores mais frequentes:

MedInc: 2.6354 (aparece 68 vezes)

HouseAge: 52.0000 (aparece 723 vezes)

AveRooms: 4.9012 (aparece 44 vezes)

AveBedrms: 1.0000 (aparece 1813 vezes)

Population: 322.0000 (aparece 42 vezes)

AveOccup: 2.5490 (aparece 40 vezes)

Variância:

MedInc: 3.6092

HouseAge: 158.4043

AveRooms: 6.1216

AveBedrms: 0.2246

Population: 1282467.0960

AveOccup: 107.8684

Desvio Padrão:

MedInc: 1.8998

HouseAge: 12.5856

AveRooms: 2.4742

AveBedrms: 0.4739

Population: 1132.4621

AveOccup: 10.3860

Quantis:

Q1 (25%):

MedInc: 2.5634

HouseAge: 18.0000

AveRooms: 4.4407

AveBedrms: 1.0061

Population: 787.0000

AveOccup: 2.4297

Q2/Mediana (50%): Valores já apresentados acima Q3

(75%):

MedInc: 4.7432

HouseAge: 37.0000

AveRooms: 6.0524

AveBedrms: 1.0995

Population: 1725.0000

AveOccup: 3.2823

IQR (Intervalo Interquartil):

MedInc: 2.1798

HouseAge: 19.0000

AveRooms: 1.6117

AveBedrms: 0.0934

Population: 938.0000

AveOccup: 0.8526

Requisito 4: Boxplots e Histogramas
Os boxplots e histogramas gerados para as seis variáveis revelam importantes características da distribuição dos dados:

MedInc (Renda Média):

Distribuição assimétrica à direita

Maior concentração entre 2 e 5 unidades

Presença de outliers na extremidade superior

HouseAge (Idade do Imóvel):

Distribuição aproximadamente normal

Valores concentrados entre 15 e 40 anos

Poucos outliers

AveRooms (Número Médio de Cômodos):

Distribuição extremamente assimétrica à direita

Muitos outliers com valores muito altos

Maioria dos valores entre 4 e 7 cômodos

AveBedrms (Número Médio de Quartos):

Distribuição similar à AveRooms, mas menos extrema

Valores tipicamente entre 1.0 e 1.2 quartos

Alguns outliers com valores muito altos

Population (População):

Distribuição muito assimétrica à direita

Grande quantidade de outliers com valores extremamente altos

Maioria das regiões tem população abaixo de 3000 habitantes

AveOccup (Número Médio de Membros na Família):

Distribuição extremamente assimétrica à direita

Muitos outliers com valores muito altos

Maioria dos valores entre 2.0 e 4.0 membros por família

Requisito 5: Identificação de Correlações
Foram identificados dois pares de variáveis com correlações significativas:

1. MedInc (Renda Média) e HouseAge (Idade do Imóvel) - Correlação Positiva

Regiões com maior renda média tendem a ter imóveis mais novos

Essa correlação sugere que áreas mais ricas possuem construções mais recentes ou renovações mais

frequentes

2. AveRooms (Número Médio de Cômodos) e AveBedrms (Número Médio de Quartos) - Correlação Positiva Forte

Existe uma relação quase linear entre o número de cômodos e quartos

Esta correlação era esperada, pois imóveis com mais cômodos geralmente têm mais quartos

Conclusão
A análise exploratória do California Housing Dataset revelou características importantes do mercado imobiliário da Califórnia. As distribuições das variáveis mostraram-se majoritariamente assimétricas, com presença significativa de outliers, especialmente nas variáveis relacionadas a tamanho de imóveis e população.

A visualização geográfica confirmou a concentração de imóveis nas áreas costeiras, particularmente nas regiões metropolitanas. As correlações identificadas sugerem relações esperadas entre as variáveis, como a relação entre renda e idade dos imóveis, e entre número de cômodos e quartos.

Este estudo fornece uma base sólida para análises mais aprofundadas, como modelagem preditiva de preços de imóveis com base nas características identificadas. A presença de outliers em várias variáveis indica a necessidade de tratamentos específicos para esses casos em análises futuras.

Anexos
Os códigos fonte utilizados para gerar estas análises estão disponíveis nos arquivos:

Trabalho_Requisito_01.py (Estatísticas descritivas)

Trabalho_Requisito_02.py (Visualização geográfica)

Trabalho_Requisito_03.py (Cálculo de métricas)

Trabalho_Requisito_04.py (Boxplots e histogramas)