UNIVERSIDADE FEDERAL DE MATO GROSSO
FACULDADE DE CIÊNCIAS DA COMPUTAÇÃO

Paulo Henrique Gonçalves Coelho
Carlos Eduardo da Silva Frazão

Análise exploratória de Dados:

BARRA DO GARÇAS - MT
2025

Introdução

Análise Exploratória do California Housing Dataset

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
MedHouseVal	A variável alvo, 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

Preparação do ambiente

Para a correta execução e reprodutibilidade deste projeto, é fundamental a devida configuração do ambiente de desenvolvimento. A análise de dados, a modelagem estatística e a visualização dos resultados foram realizadas utilizando a linguagem de programação Python, juntamente com um conjunto de bibliotecas especializadas que fornecem as ferramentas necessárias para manipulação e interpretação de dados.

Para instalar as bibliotecas necessárias, é preciso ter o Python e o seu gerenciador de pacotes, o *pip*, devidamente instalados em seu sistema.

Instalação no Windows

1. **Instale as Bibliotecas:** Execute o seguinte comando para instalar todas as dependências de uma só vez. O pip se encarregará de baixar e instalar cada uma delas.

```
pip install pandas numpy seaborn matplotlib scikit-learn
```

2. **Aguarde a Instalação:** O processo de download e instalação pode levar alguns minutos, dependendo da sua conexão com a internet. Ao final, as bibliotecas estarão prontas para serem utilizadas.

Instalação no Linux (Debian, Ubuntu e derivados)

1. **Instale as Bibliotecas:** Utilize o pip para instalar todas as dependências necessárias com um único comando.

```
pip3 install pandas numpy seaborn matplotlib scikit-learn
```

2. **Conclusão**: Após a execução do comando, o pip fará o download e a instalação de todas as bibliotecas listadas. Ao final do processo, seu ambiente estará pronto para executar o projeto.

Resultados por Requisito

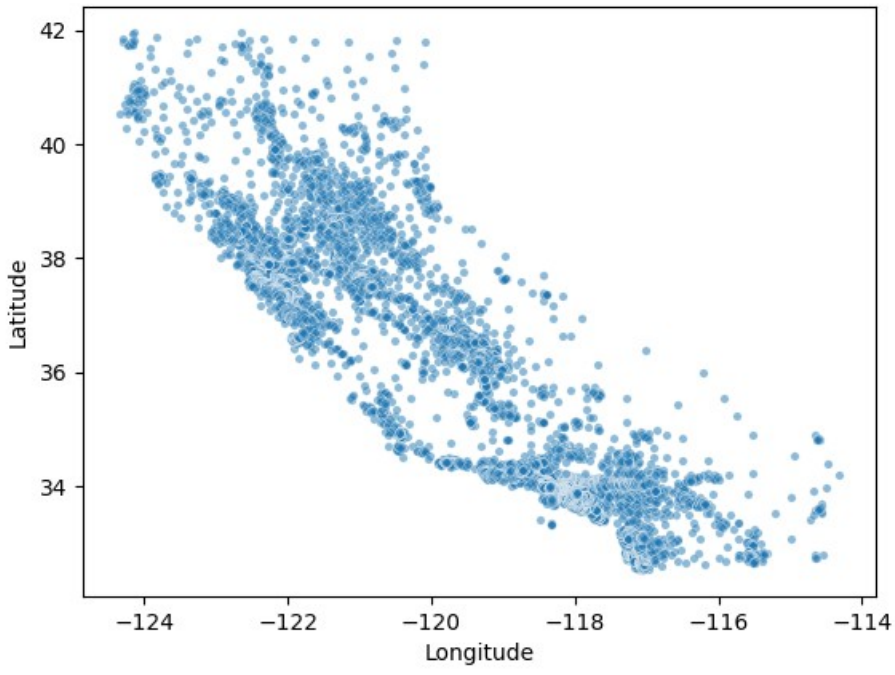
Requisito 1: Estatísticas Descritivas Básicas

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.

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
count	20640	20640	20640	20640	20640	20640	20640	20640
mean	3.87	28.63	5.42	1.09	1425.47	3.07	35.63	-119.56
std	1.89	12.5	2.47	0.47	1132.46	10.38	2.13	2.00
min	0.49	1.00	0.84	0.33	3.00	0.69	32.54	-124.35
25%	2.56	18.00	4.44	1.00	787.00	2.42	33.93	-121.80
50%	3.53	29.00	5.22	1.04	1166.00	2.81	34.26	-118.49
75%	4.74	37.00	6.05	1.09	1725.00	3.28	37.71	-118.01
max	15.00	52.00	141.90	34.06	35682.00	1243.33	41.95	-114.31

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.

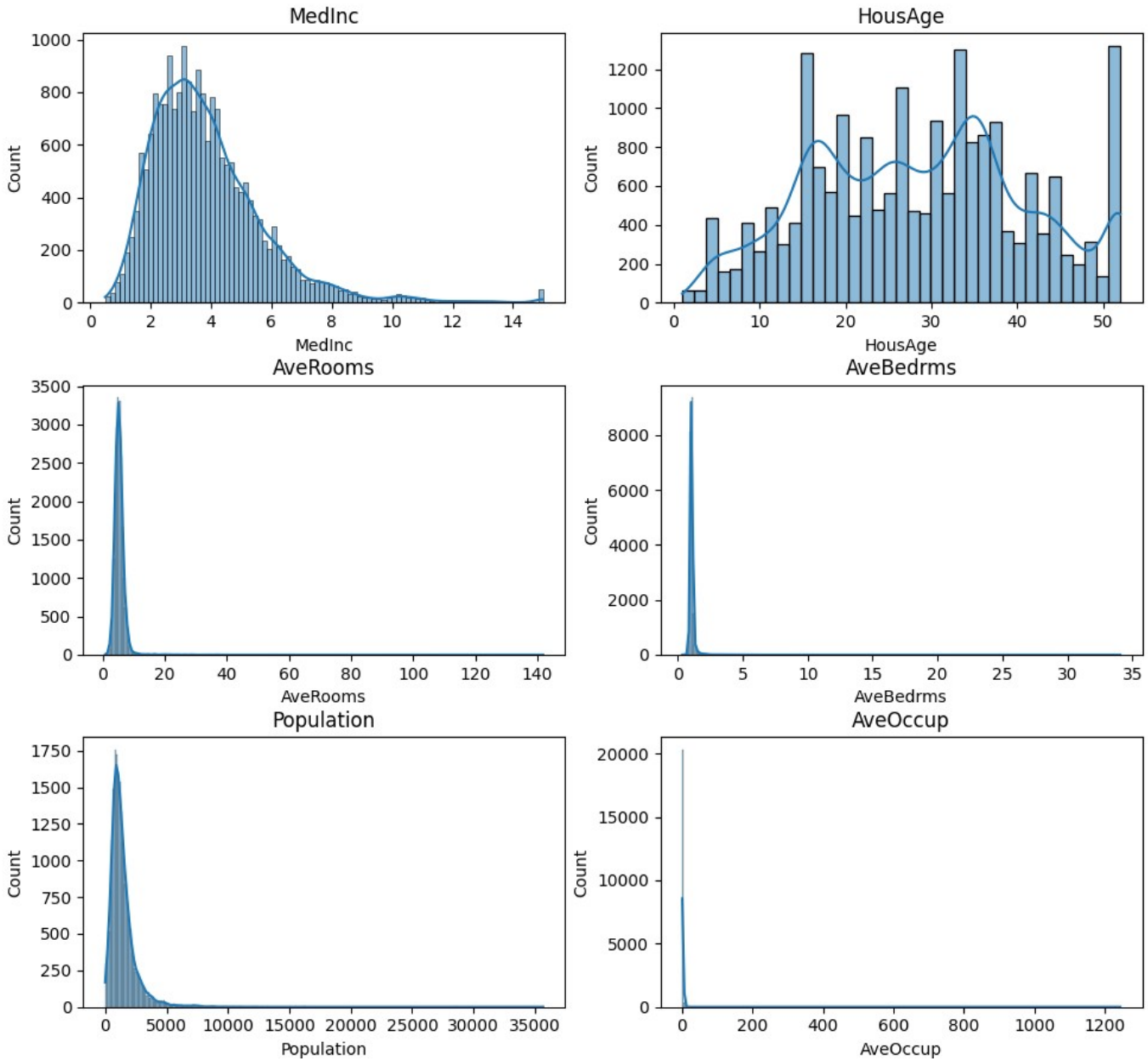


Requisito 3: Métricas Estatísticas

A execução deste requisito gera um volume considerável de dados como saída. Por questões de brevidade e para manter a clareza do documento, optou-se por não exibir esses resultados diretamente aqui. No entanto, a saída completa pode ser consultada e verificada através da execução do script `Trabalho_Requisito_03.py`

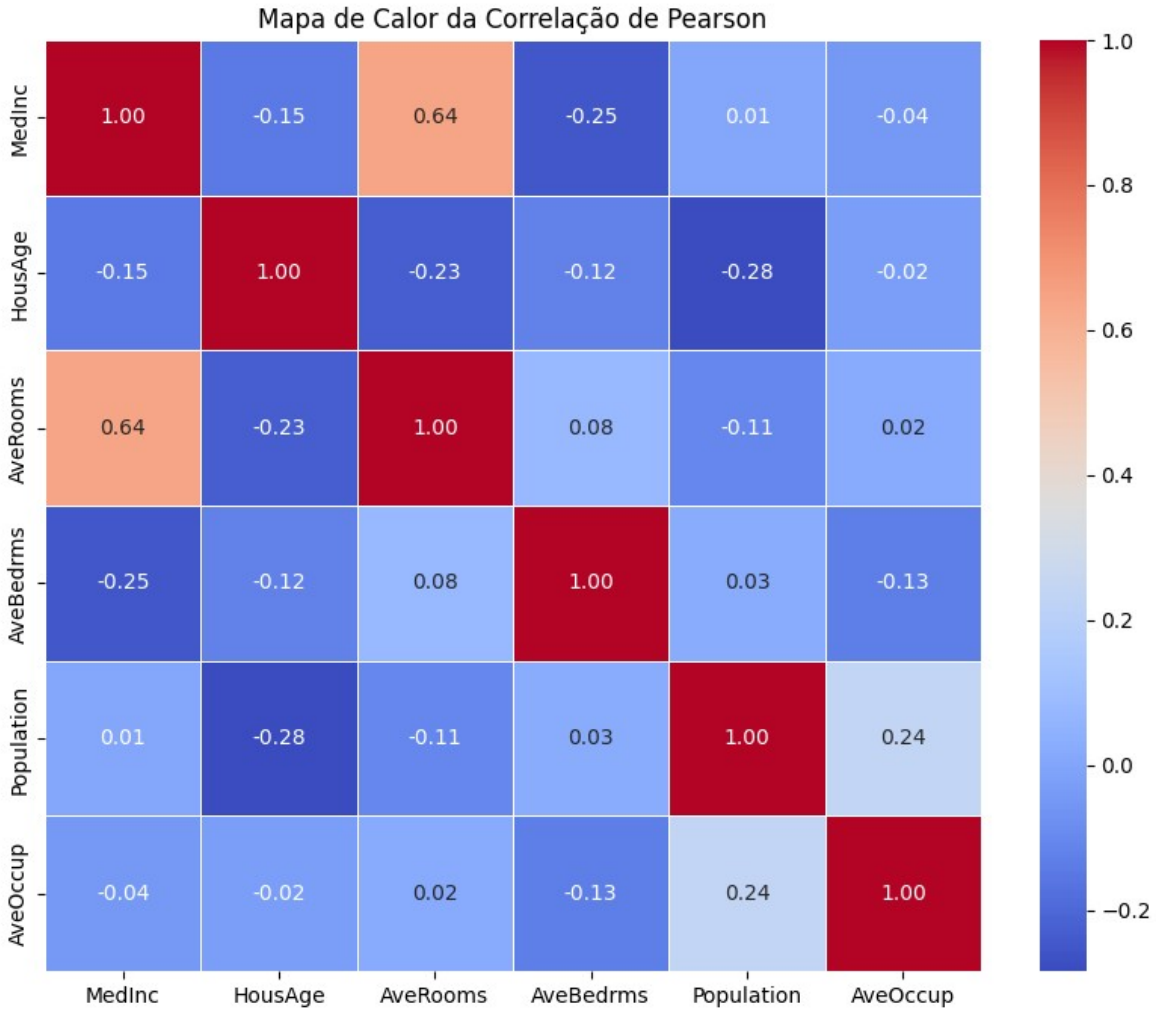
Requisito 4: Boxplots e Histogramas

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



Requisito 5: Identificação de Correlações

Neste requisito vamos trabalhar com a Correlação de Spearman (ρ_s)**.** Ela mede a relação monotônica entre duas variáveis. Uma relação é monotônica quando as variáveis tendem a se mover na mesma direção, mas não necessariamente a uma taxa constante. É uma medida não paramétrica, o que a torna adequada para dados ordinais ou quando a relação não é linear.



1. **Correlação Positiva Mais Forte**

- `MedInc` (Renda Média) e `AveRooms` (Média de Quartos): +0.64
    - Esta é a correlação mais forte e positiva no gráfico.
    - **Interpretação:** Há uma forte tendência de que, em áreas onde a renda média (`MedInc`) é maior, as casas também possuam, em média, um número maior de quartos (`AveRooms`). Isso faz sentido, pois pessoas com maior poder aquisitivo podem comprar casas maiores.

2. **Correlações Negativas Moderadas**

- `HousAge` (Idade do Imóvel) e `Population` (População): -0.28
    - **Interpretação:** Existe uma correlação negativa fraca a moderada. Isso sugere que áreas com casas mais antigas (`HousAge`) tendem a ter uma população ligeiramente menor. Isso pode ocorrer porque bairros mais novos costumam ser planejados para uma maior densidade populacional.
- `MedInc` (Renda Média) e `AveBedrms` (Média de Quartos de Dormir): -0.25
    - **Interpretação:** Esta é uma correlação negativa interessante e talvez contra-intuitiva. Sugere que em locais com renda média mais alta, a média de quartos de dormir (`AveBedrms`) é ligeiramente menor.
- `HousAge` (Idade do Imóvel) e `AveRooms` (Média de Quartos): -0.23
    - **Interpretação:** Há uma leve tendência de que imóveis mais antigos (`HousAge`) tenham, em média, menos quartos (`AveRooms`). Isso pode indicar que construções mais recentes são, em geral, maiores.

3. Correlações Fracas (Próximas de Zero)

Muitos pares de variáveis mostram uma correlação muito fraca, indicando que não há uma relação linear clara entre eles.

- `AveBedrms` (Média de Quartos de Dormir) e `AveRooms` (Média de Quartos): +0.08
    - **Detalhe:** A correlação entre o número médio de quartos e o número médio de quartos de dormir é quase nula. Isso reforça a ideia de que o aumento no total de quartos (`AveRooms`) não se deve necessariamente a um aumento no número de quartos de dormir,
- `MedInc` (Renda Média) e `Population` (População): +0.01
    - **Interpretação:** Não há praticamente nenhuma relação linear entre a renda média de uma área e sua população total. Uma área rica não é necessariamente mais ou menos populosa.
- `Population` (População) e `AveBedrms` (Média de Quartos de Dormir): +0.03
    - **Interpretação:** A população de uma área e a média de quartos de dormir por casa não estão linearmente relacionadas.

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.