



UNIVERSIDADE FEDERAL DO MARANHÃO  
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA  
BACHARELADO INTERDISCIPLINAR EM CIÊNCIA E  
TECNOLOGIA  
ENGENHARIA DA COMPUTAÇÃO

# **PREDIÇÃO DE PREÇOS DE IMÓVEIS: MINERAÇÃO DE DADOS E MACHINE LEARNING**

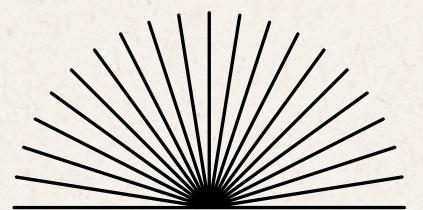
**Análise e Modelagem com House Prices Dataset**

**DISCENTE:**

Luís Guilherme Freitas de  
Almeida Silva

**DOCENTE:**

Prof. Dr. Thales Levi Azevedo  
Valente



# Agenda

03	<b>Introdução</b>
04	<b>Objetivos</b>
05	<b>Materiais e Métodos</b>
16	<b>Resultados e Discussões</b>
18	<b>Conclusão</b>
19	<b>Agradecimentos</b>

**01** O mercado imobiliário é altamente influenciado por diversos fatores, como localização, tamanho, infraestrutura e demanda. A correta precificação de imóveis é essencial para compradores, vendedores e investidores tomarem decisões estratégicas.

**02** Desenvolver um modelo preditivo de preços de imóveis utilizando Mineração de Dados e Técnicas de Machine Learning com o House Prices Dataset (Kaggle). O objetivo é analisar quais características mais influenciam o valor de um imóvel e criar um modelo capaz de estimar preços com alta precisão.

**03** A precificação de imóveis envolve desafios como a presença de outliers, dados incompletos e forte correlação entre variáveis. Este estudo busca entender esses padrões e construir um modelo robusto para auxiliar na tomada de decisões imobiliárias.



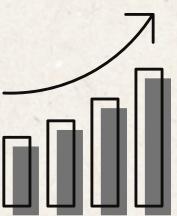
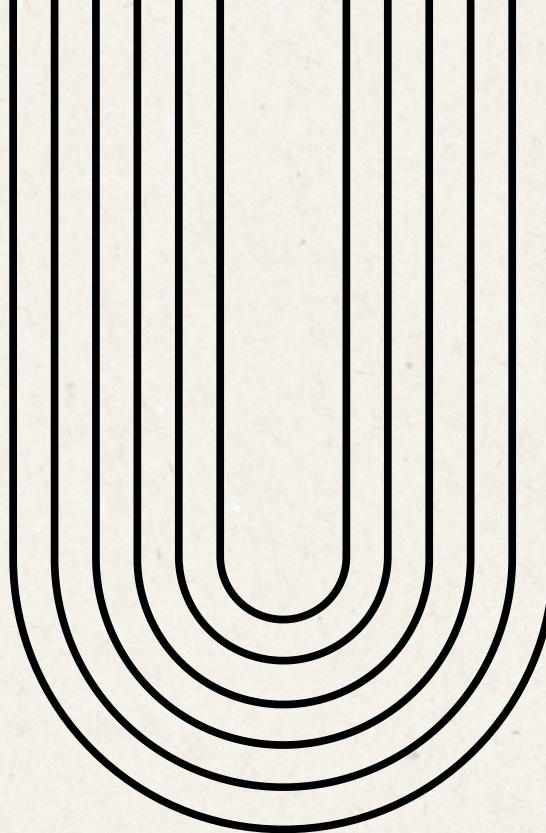
# Introdução

Predição de Preços de Imóveis: Motivação e Objetivos

fonte: Unsplash (<https://unsplash.com/pt-br/fotografias/fotografia-aerea-de-casas-brancas-com-telhados-alaranjados-durante-o-dia-zDFVEGflVFA>)

# Objetivos

Desenvolver um modelo de predição de preços de imóveis utilizando Mineração de Dados e Machine Learning, explorando padrões no dataset House Prices para auxiliar na tomada de decisões imobiliárias.



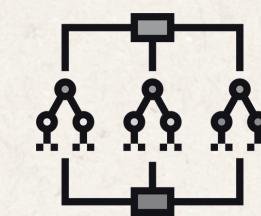
## Identificar os principais fatores que influenciam o preço dos imóveis

Analizar a correlação entre variáveis e o preço final para entender quais características impactam mais o valor de mercado.



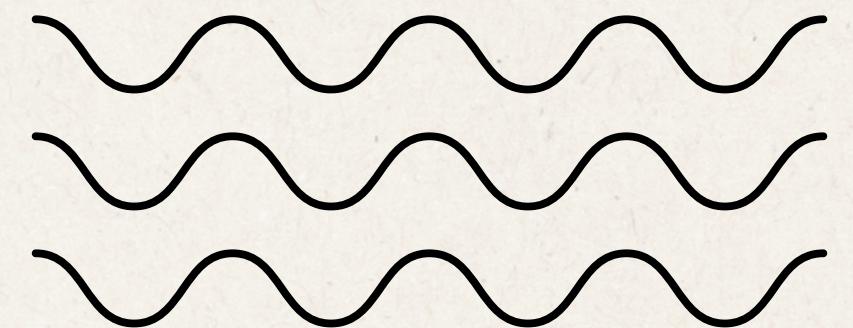
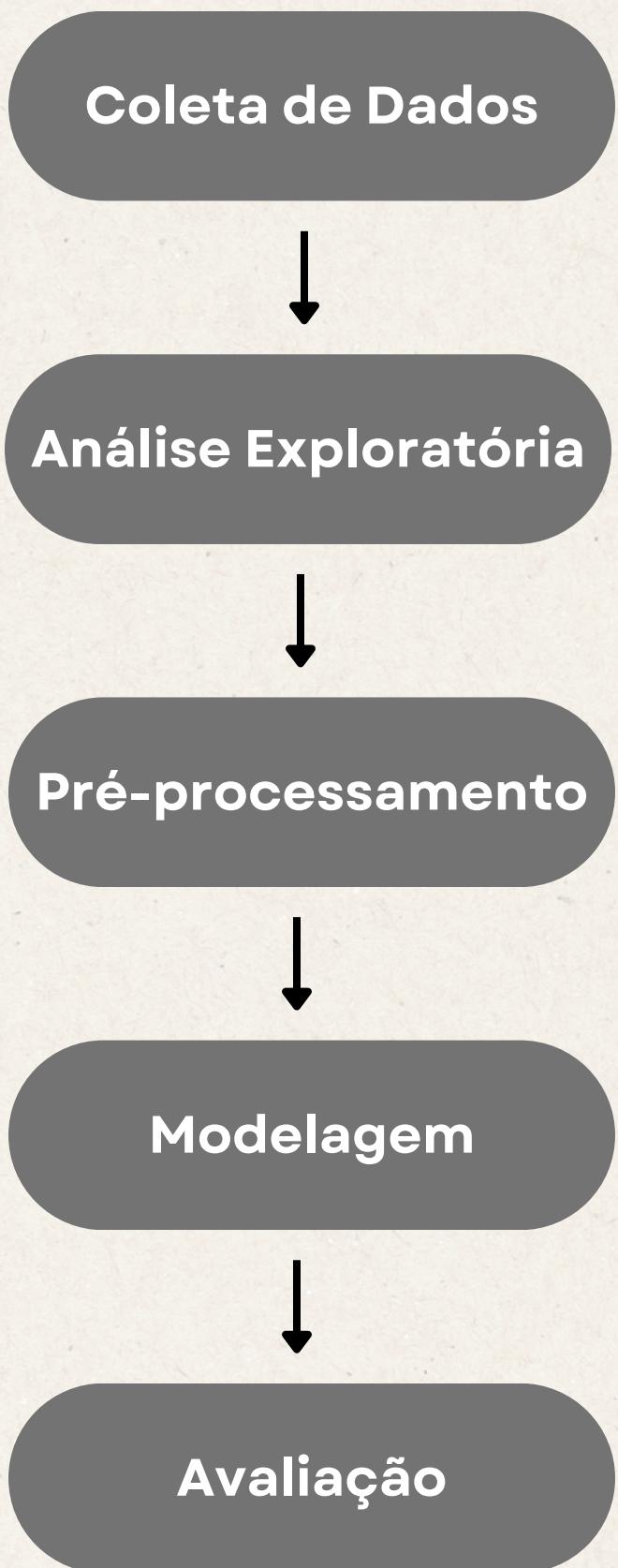
## Aplicar técnicas de pré-processamento para melhorar a qualidade dos dados

Realizar limpeza, normalização, engenharia de características e seleção de variáveis para otimizar a modelagem.



## Testar e avaliar diferentes modelos de regressão

Comparar modelos como Regressão Linear, Ridge, Lasso, Random Forest e XGBoost, ajustando hiperparâmetros para obter a melhor performance.



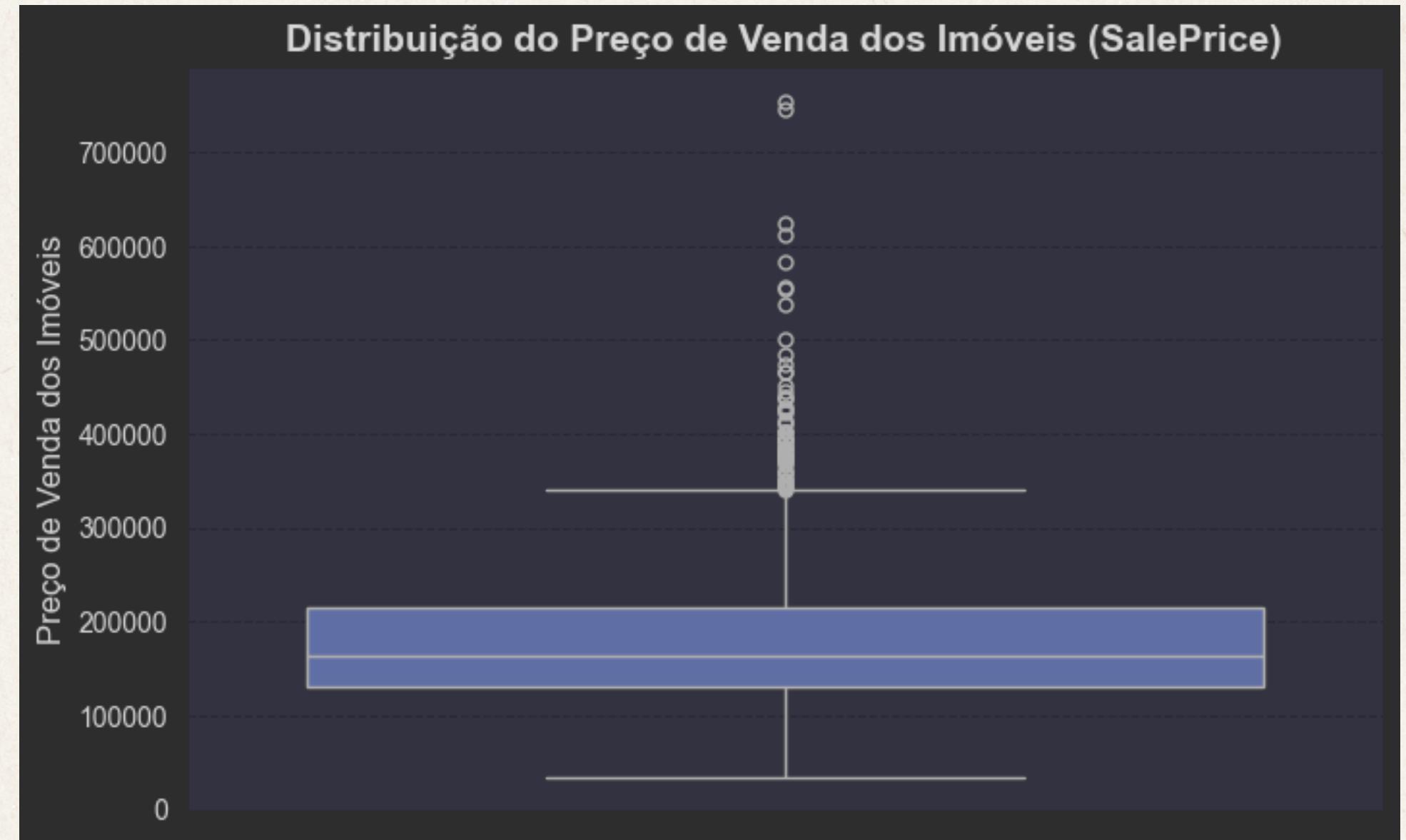
## Materiais e Métodos

Nesta seção, apresentamos os materiais utilizados e os métodos empregados para desenvolver o modelo preditivo de preços de imóveis. O fluxo abrange desde a coleta dos dados até a avaliação dos modelos treinados.

# Base de Dados

- ORIGEM: KAGGLE – HOUSE PRICES DATASET.
- REGISTROS: 1.460 AMOSTRAS.
- VARIÁVEIS: 81 COLUNAS.
- VARIÁVEL-ALVO: SALEPRICE (PREÇO DE VENDA DO IMÓVEL).
- TIPOS DE DADOS: NUMÉRICOS E CATEGÓRICOS.

Boxplot mostrando a distribuição da variável SalePrice na base de dados House Prices



fonte: Captura de tela do autor

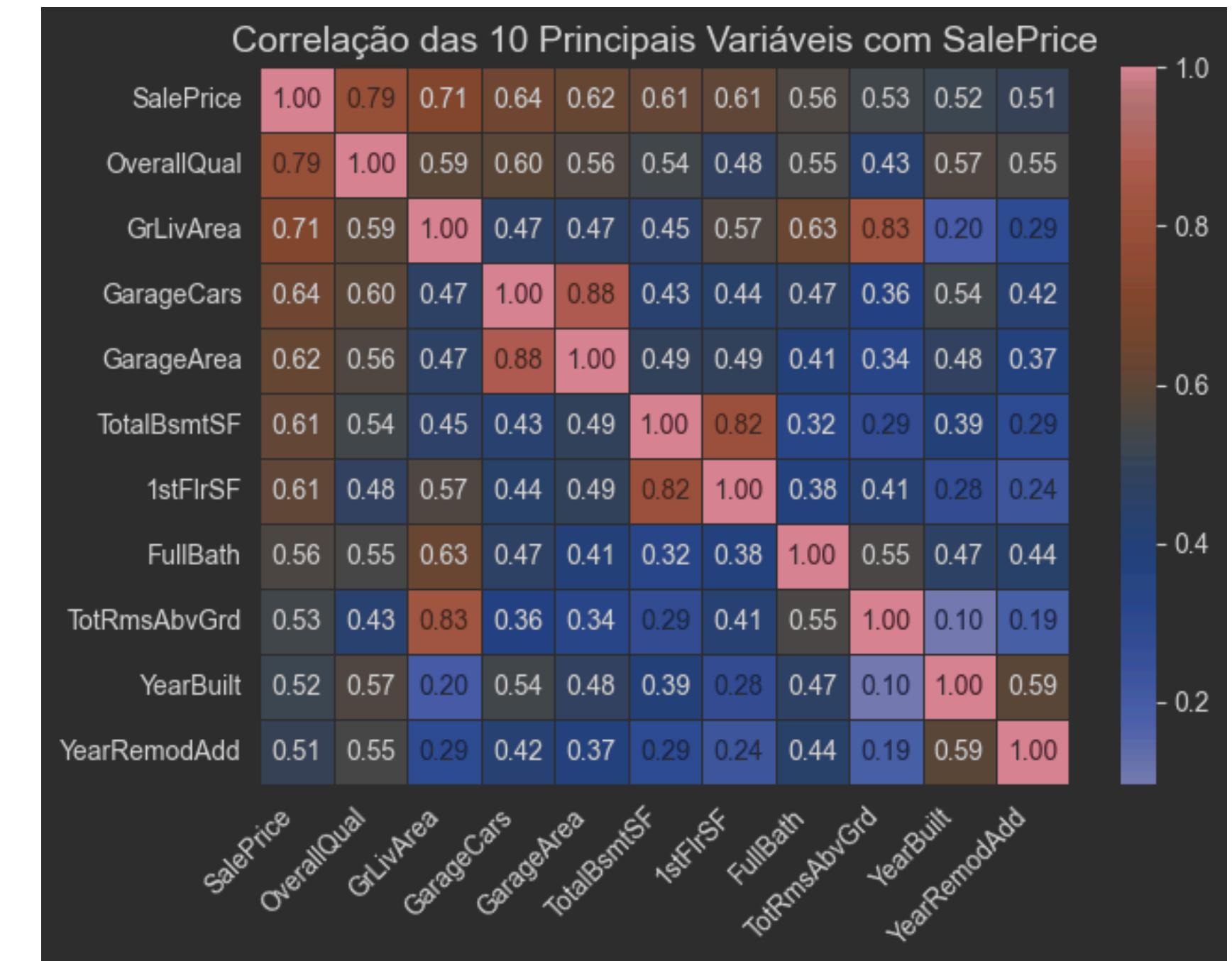
# Análise Exploratória

Nesta etapa, exploramos a base de dados para identificar padrões, relações entre variáveis e possíveis outliers que possam impactar a modelagem preditiva.

## Correlação entre Variáveis

- A matriz de correlação exibe as relações entre as principais variáveis da base de dados.
- A variável SalePrice é nossa variável-alvo e buscamos entender quais variáveis mais influenciam seu valor.
- A cor indica o grau de correlação:
- ● Correlação positiva (próximo de 1): indica que um aumento nessa variável tende a aumentar SalePrice.
- ● Correlação negativa (próximo de -1): indica que um aumento nessa variável tende a diminuir SalePrice.
- As variáveis com maior correlação com o preço dos imóveis incluem:
  - OverallQual (qualidade geral da construção)
  - GrLivArea (área útil acima do solo)
  - GarageCars (número de vagas na garagem)

Matriz de Correlação das 10 principais variáveis com SalePrice.



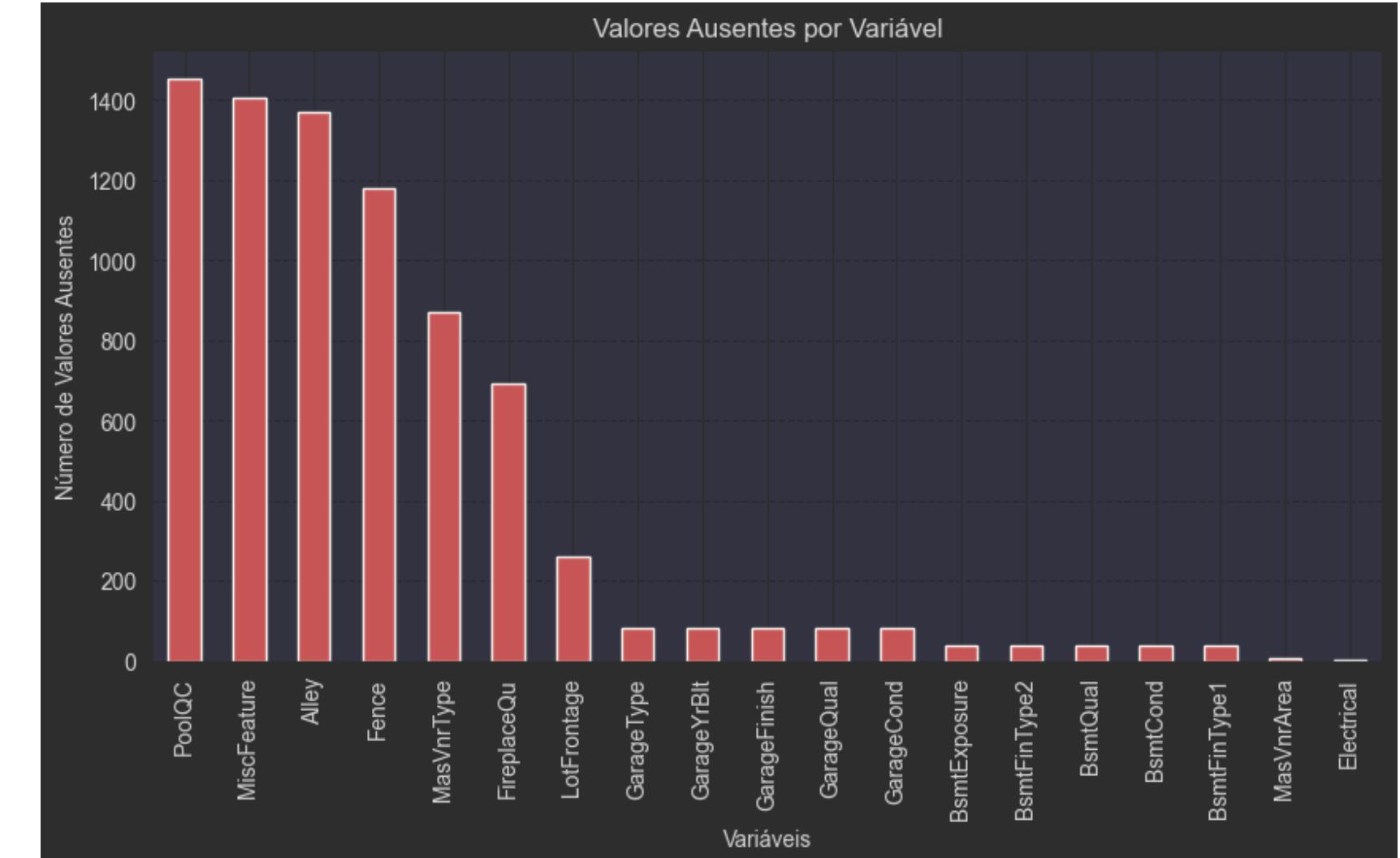
fonte: Captura de tela do autor.

# Análise Exploratória

## Verificação de Valores Nulos/Ausentes

- Antes de construir os modelos, é essencial verificar valores ausentes na base de dados.
- Valores nulos podem impactar a análise e a modelagem, exigindo tratamento adequado, como remoção ou imputação (substituição por um valor estimado).
- A seguir, apresentamos a quantidade de valores ausentes em cada variável do dataset.

Gráfico de barras mostrando as variáveis que possuem valores ausentes e suas respectivas quantidades



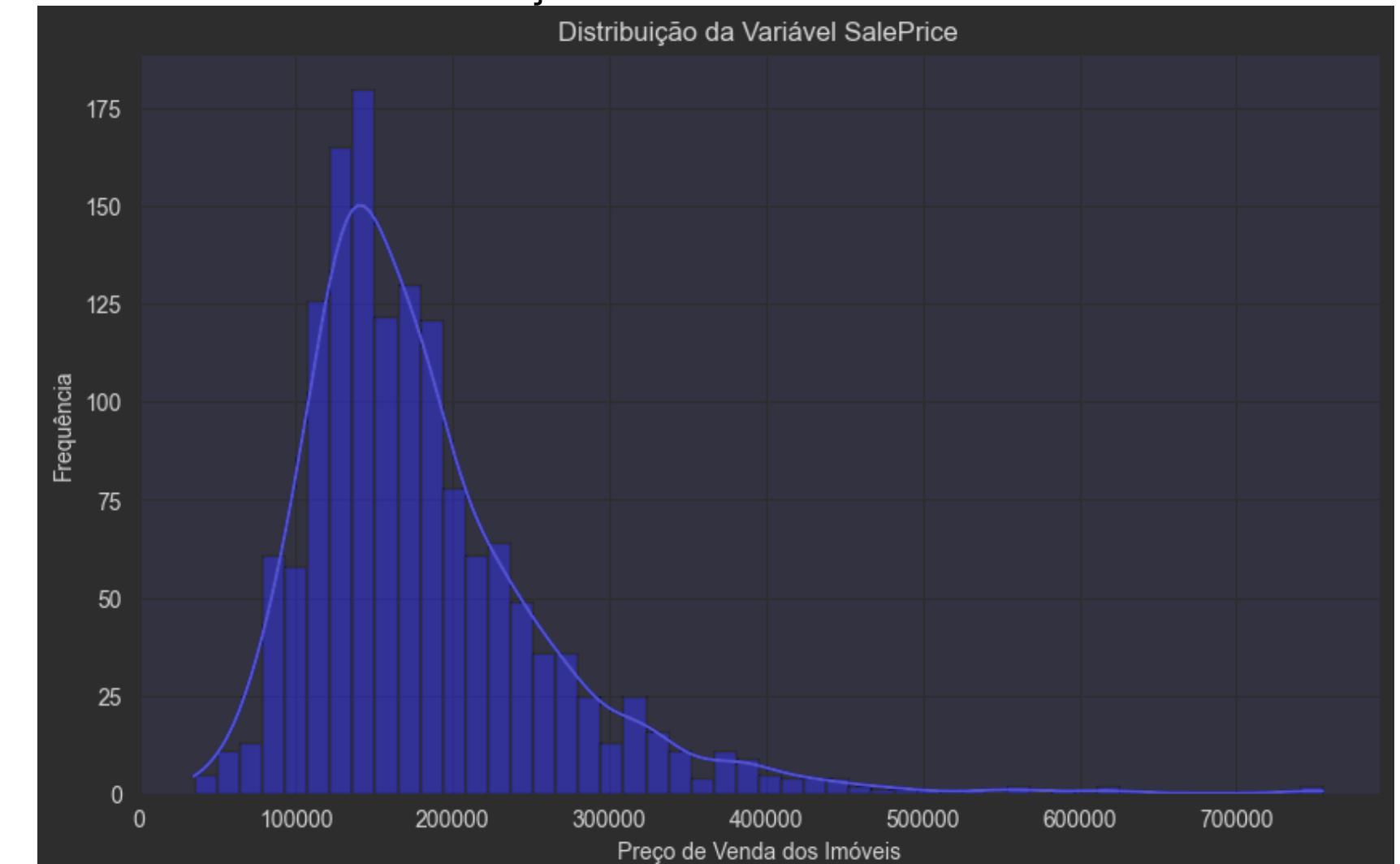
fonte: Captura de tela do autor

# Análise Exploratória

## Distribuição do Preço de Venda dos Imóveis

- O histograma ao lado mostra a distribuição da variável SalePrice, que representa o preço de venda dos imóveis na base de dados.
- Podemos observar que a distribuição não é simétrica, apresentando uma cauda longa à direita (assimetria positiva).
- Isso indica que a maioria dos imóveis está concentrada em valores menores, com alguns poucos imóveis sendo vendidos por preços elevados.
- Essa assimetria pode impactar modelos de regressão, pois valores extremos (outliers) podem distorcer as previsões.
- Uma possível solução é aplicar transformações nos dados, como a transformação logarítmica, para tornar a distribuição mais próxima de uma normal.

Distribuição da Variável SalePrice



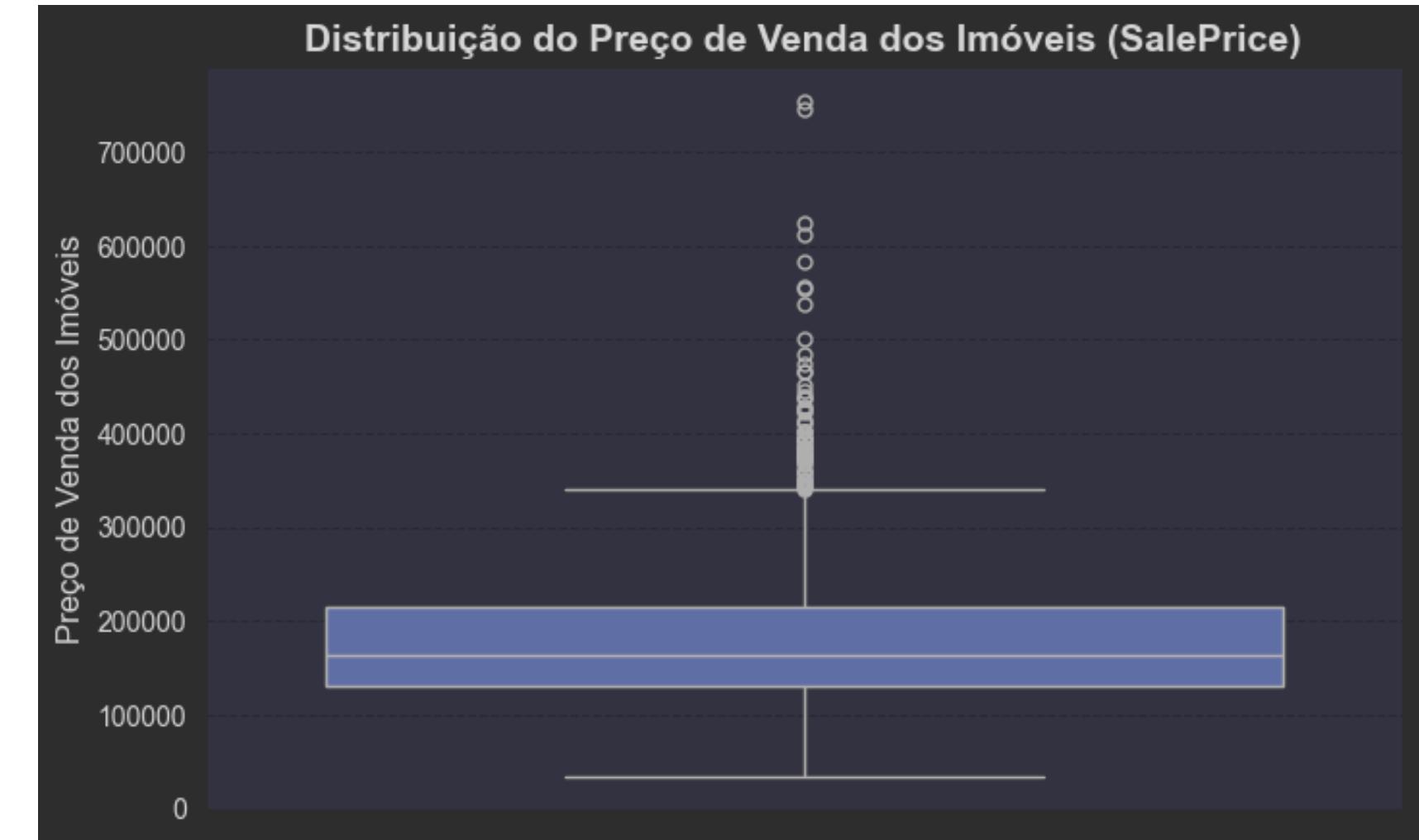
fonte: Captura de tela do autor

# Análise Exploratória

## Identificação de Outliers em SalePrice

- O boxplot ao lado representa a distribuição do preço de venda dos imóveis (SalePrice) e permite identificar a presença de outliers.
- Os círculos fora das "extremidades" do boxplot indicam valores considerados outliers, ou seja, preços significativamente diferentes do padrão da maioria dos imóveis.
- Nota-se que há vários valores extremos acima de 500.000, representando imóveis de alto padrão ou com características diferenciadas.
- A presença de outliers pode afetar modelos preditivos, pois distorcem estatísticas como a média e aumentam a variância.
- Em alguns casos, a remoção ou transformação desses outliers pode melhorar a qualidade do modelo, mas isso deve ser avaliado caso a caso.
- Para este estudo, a remoção já foi realizada durante o pré-processamento, garantindo que a análise esteja ajustada para os modelos.

Boxplot mostrando a distribuição da variável SalePrice na base de dados House Prices



fonte: Captura de tela do autor

# Pré-processamento

Antes de aplicar qualquer modelo de Machine Learning, é essencial tratar os valores ausentes. O gráfico ao lado exibe as variáveis com maior quantidade de valores faltantes na base de dados House Prices.

## Técnicas Utilizadas

### Preenchimento com "Nenhum":

- Variáveis categóricas que representam características opcionais foram preenchidas com "Nenhum", indicando ausência da característica.
- Exemplos: Alley, FireplaceQu, PoolQC, Fence, BsmtQual, GarageType, MasVnrType.

### Preenchimento com 0:

- Variáveis numéricas onde o valor ausente significa ausência da característica receberam 0.
- Exemplos: MasVnrArea (Área de revestimento de alvenaria), GarageYrBlt (Ano de construção da garagem).

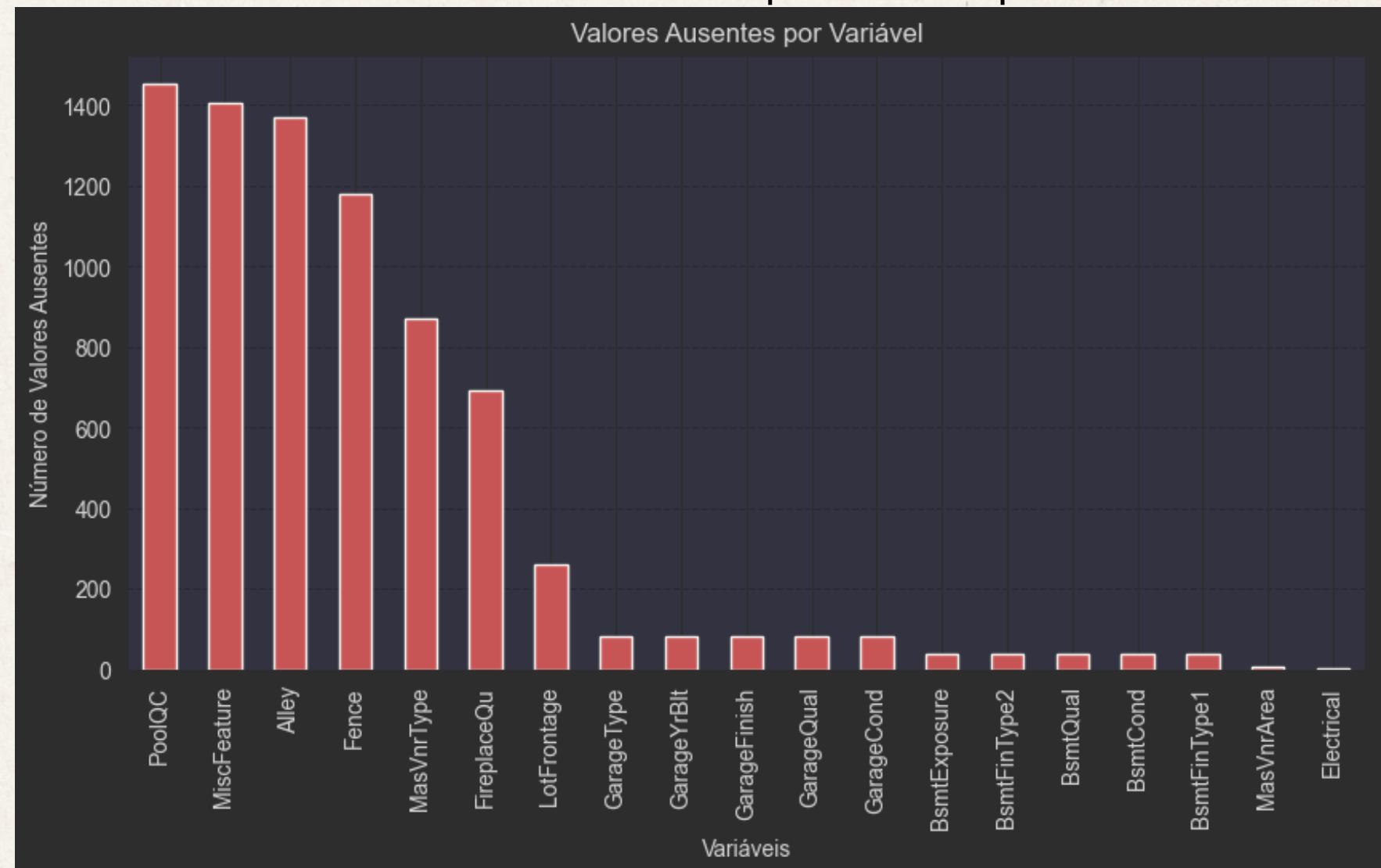
### Preenchimento com a Mediana do Bairro:

- A variável LotFrontage (largura da fachada do lote) foi preenchida com a mediana do bairro correspondente.

### Preenchimento com a Moda:

- A variável Electrical, que possui apenas um valor ausente, foi preenchida com o valor mais frequente (moda).

Gráfico de barras mostrando as variáveis que possuem valores ausentes e suas respectivas quantidades



fonte: Captura de tela do autor

# Pré-processamento

A variável alvo SalePrice não segue uma distribuição normal, o que pode impactar o desempenho de alguns modelos de Machine Learning. Portanto, aplicamos a Transformação Logarítmica para estabilizar a variância e aproximar a distribuição de uma normal.

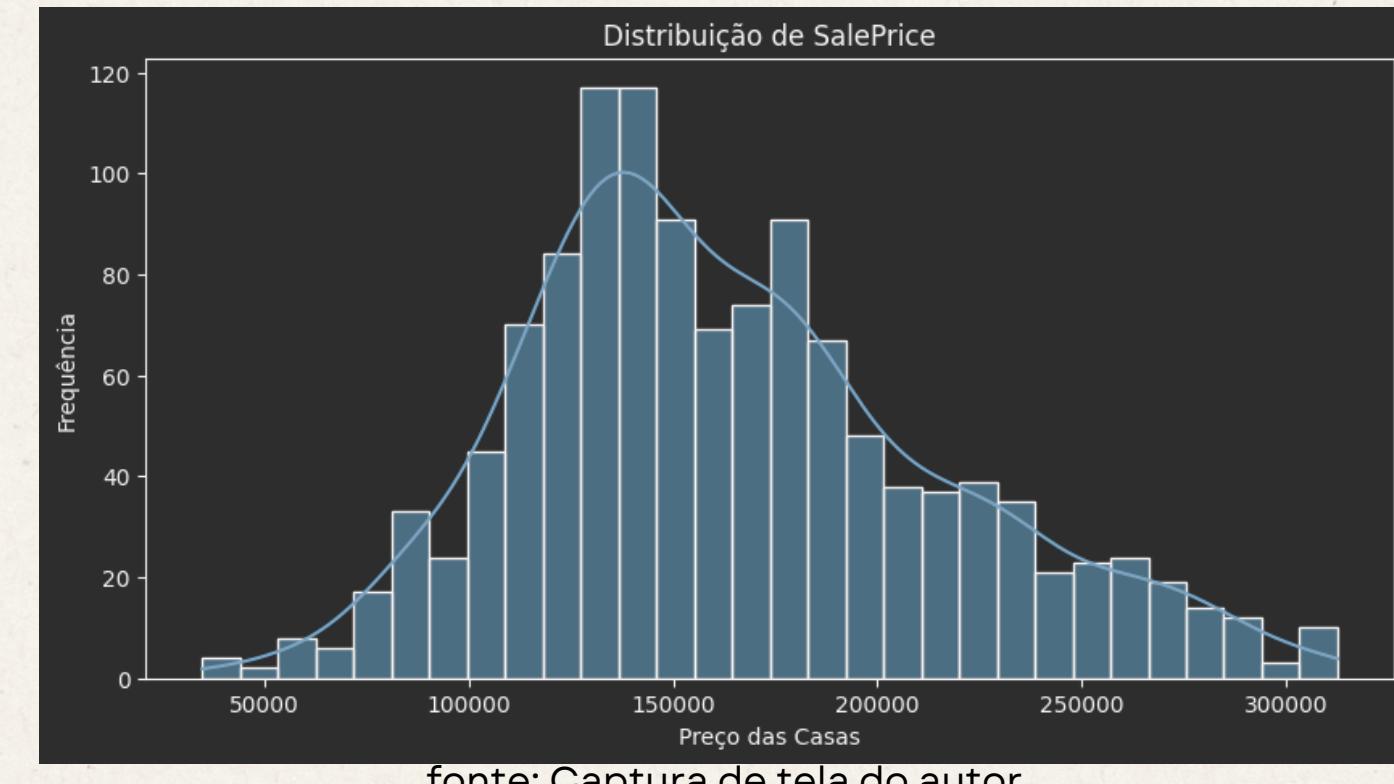
## Análise Inicial

- O histograma inicial mostra uma distribuição assimétrica à direita (skewed right), com uma longa cauda de valores altos.
- O Teste de Normalidade de Shapiro-Wilk confirmou que a variável SalePrice não segue uma distribuição normal ( $p\text{-valor} < 0.05$ ).

## Transformação Aplicada

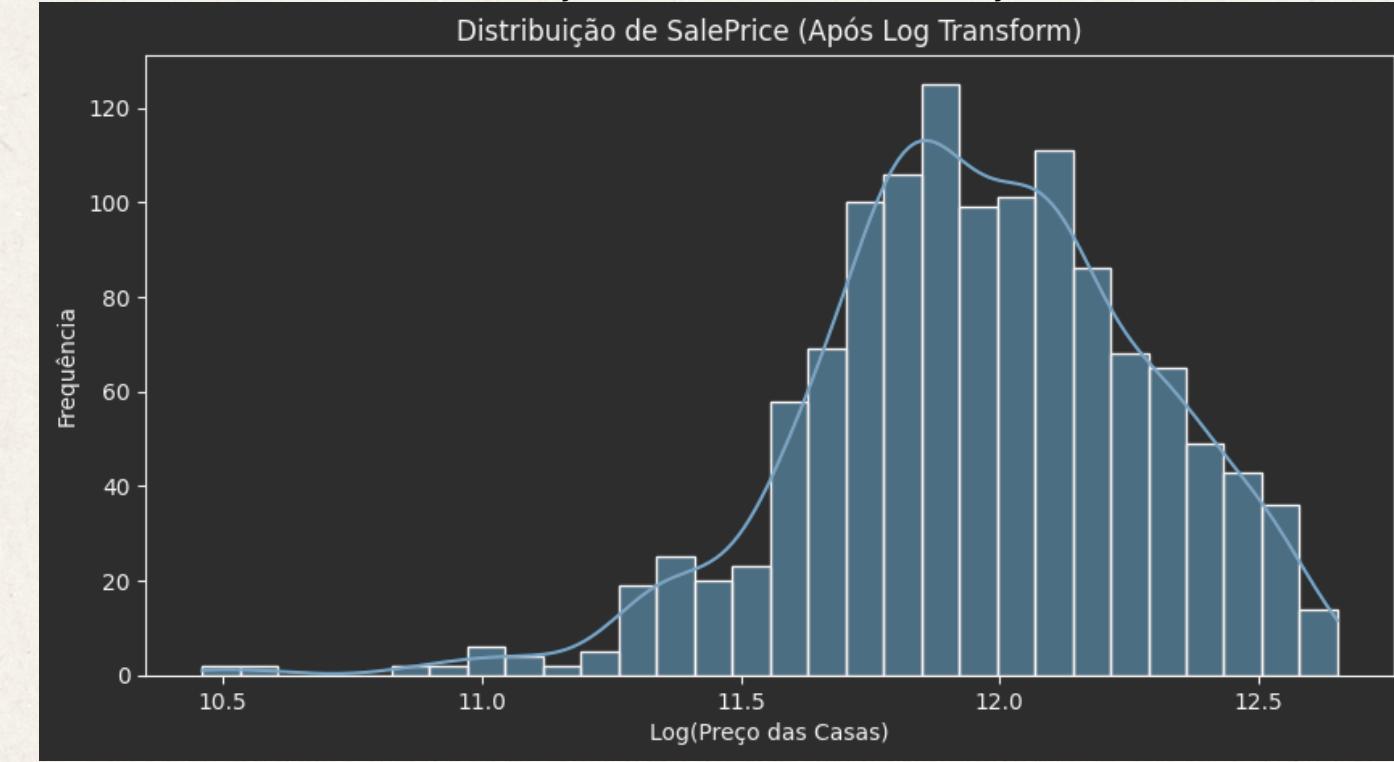
- Para corrigir essa assimetria, aplicamos a Transformação Logarítmica utilizando  $\log1p(\text{SalePrice})$ , que ajuda a normalizar a distribuição.
- O segundo histograma mostra que, após a transformação, a distribuição se aproxima de uma normal, tornando os dados mais adequados para modelos que assumem normalidade.

Distribuição assimétrica à direita, com uma longa cauda de valores altos



fonte: Captura de tela do autor

Distribuição após a transformação



fonte: Captura de tela do autor

# Pré-processamento

Após a transformação da variável SalePrice, aplicamos normalização nas variáveis numéricas para evitar que atributos com escalas diferentes influenciem desproporcionalmente o modelo.

## Método Utilizado

Utilizamos StandardScaler do sklearn.preprocessing, que:

- Centraliza os dados subtraindo a média.
- Escala para que a variância seja 1.

## Passos Executados

- Selecioneamos as colunas numéricas, excluindo SalePrice para evitar vazamento de dados.
- Aplicamos StandardScaler para normalizar os dados.
- Salvamos os dados normalizados para uso nos modelos

## Por que Normalizar?

### Regressão Linear:

- Modelos de regressão linear são sensíveis a escalas diferentes. Se uma variável tem valores muito altos em comparação com outra, ela pode dominar os coeficientes do modelo.
- A normalização melhora a estabilidade numérica e evita que variáveis de grande magnitude dominem os resultados.

### Regressão de Lasso:

- Lasso adiciona uma penalização (L1) que pode zerar coeficientes irrelevantes.
- Se as variáveis não estiverem normalizadas, a penalização pode ser injusta, favorecendo variáveis com escalas menores.
- A normalização garante que o modelo penalize todas as variáveis de forma justa e equilibrada.

### Random Forest:

- Modelos baseados em árvores, como Random Forest, não são diretamente afetados por normalização, pois não dependem de distância entre os dados.
- No entanto, se os dados forem muito desbalanceados ou com grande variância, a normalização pode ajudar na interpretação dos gráficos de importância de variáveis.

# Pré-processamento

Transformar as variáveis categóricas em um formato adequado para os modelos de machine learning, garantindo que possam ser interpretadas corretamente por algoritmos baseados em regressão e árvores de decisão.

## Método Utilizado

Utilizamos One-Hot Encoding (`pd.get_dummies()`), que:

- Cria uma nova coluna para cada categoria existente.
- Representa a presença/ausência de cada categoria com 0 ou 1.
- Evita problemas de interpretação de grandezas numéricas associadas a categorias.

## Passos Executados

- Identificamos as colunas categóricas presentes no dataset.
- Aplicamos One-Hot Encoding para transformar essas colunas em variáveis binárias.
- Utilizamos `drop_first=True` para evitar multicolinearidade entre categorias.
- Salvamos os dados finais prontos para modelagem no arquivo

## Por que Converter Variáveis Categóricas?

### Regressão Linear e Regressão de Lasso:

- Esses modelos assumem que as variáveis são numéricas e não podem lidar com dados categóricos diretamente.
- Se passássemos as categorias como texto, o modelo não saberia como utilizá-las.
- A conversão permite que cada categoria seja tratada como uma variável independente.

### Random Forest:

- Diferente da regressão, Random Forest pode lidar com variáveis categóricas sem necessidade de encoding explícito.
- Porém, ao utilizar One-Hot Encoding, garantimos que a informação categórica seja considerada corretamente em todos os modelos.

# Modelagem Preditiva

Aplicar algoritmos de aprendizado de máquina para prever o preço de venda dos imóveis.

## Algoritmos Utilizados

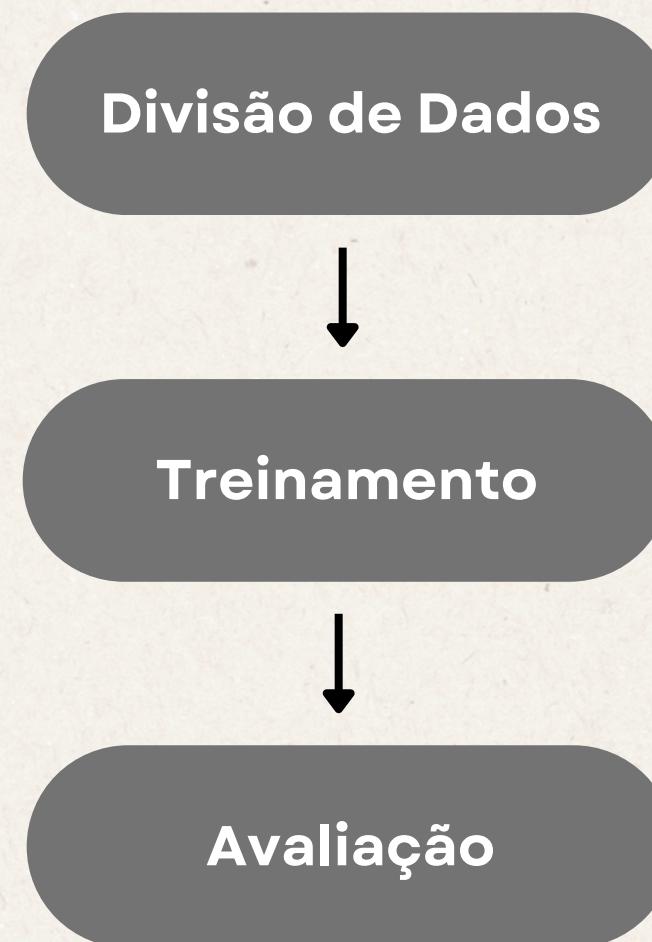
- Regressão Linear
- Regressão Linear com Feature Selection
- Ridge-Lasso Regression
- Random Forest
- XGBoost

Cada modelo foi treinado e avaliado utilizando dados normalizados e variáveis transformadas.

## Métricas de Avaliação

- Erro Absoluto Médio (MAE) – Média dos erros em valor absoluto.
- Raiz do Erro Quadrático Médio (RMSE) – Penaliza erros maiores.
- Coeficiente de Determinação ( $R^2$ ) – Mede o quanto o modelo explica a variância da variável SalePrice.

Os resultados serão analisados na seção de Resultados e Discussão.



# Resultados e Discussões

## PRINCIPAIS CONCLUSÕES:

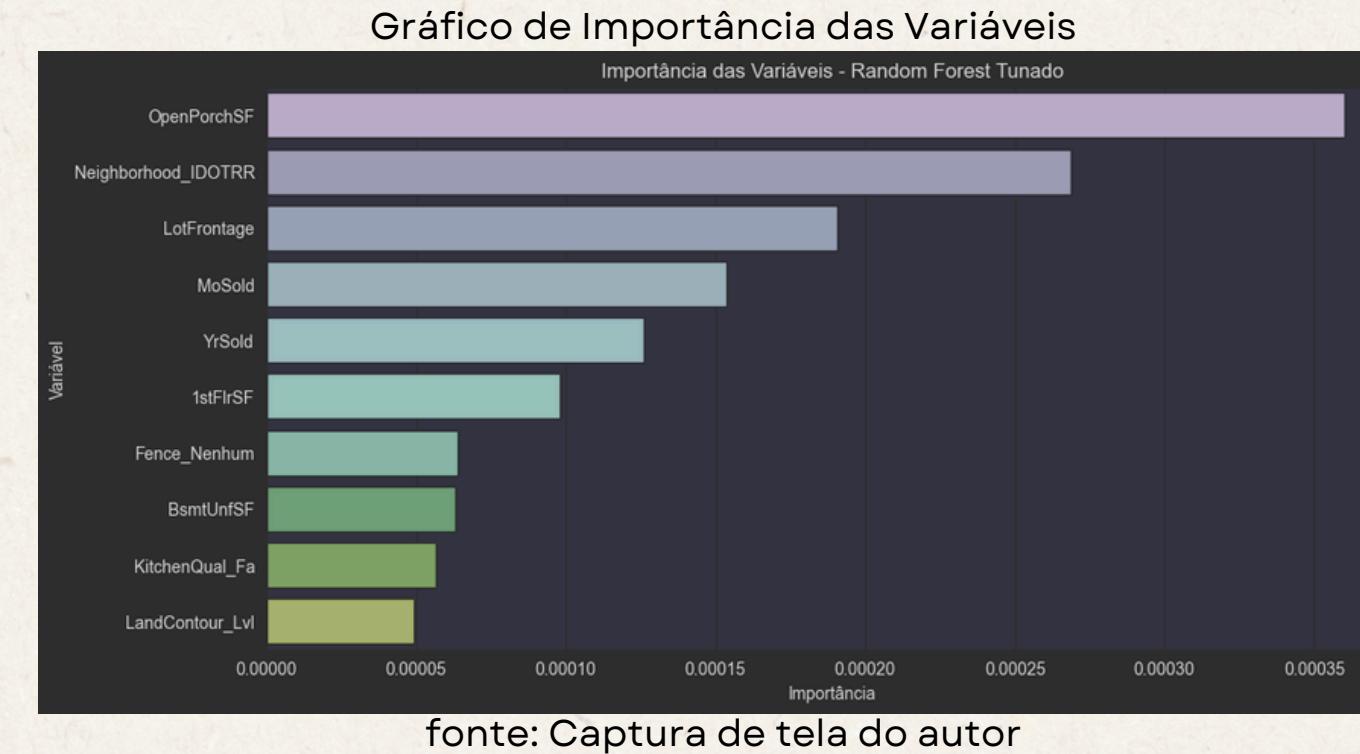
- Random Forest Tunado obteve o melhor desempenho, com  $R^2 = 0.9989$ , indicando uma excelente capacidade de prever o preço dos imóveis.
- XGBoost também teve um bom desempenho, mas com erro ligeiramente maior que o Random Forest.
- Modelos lineares tiveram desempenho inferior, pois não capturaram relações não lineares complexas entre as variáveis.
- Feature Selection na Regressão Linear não trouxe melhora significativa, indicando que a regressão linear não estava sobrecarregada com variáveis irrelevantes.

Modelo	MAE	RMSE	$R^2$
Regressão Linear	0,1429	0,2525	0,9378
Regressão Linear com Feature Selection	0,1340	0,2110	0,9566
Ridge Regression	0,1286	0,2163	0,9544
Lasso Regression	0,1321	0,2097	0,9571
Lasso Tunado	0,1202	0,0403	0,9607
Random Forest	0,0066	0,0012	0,9988
Random Forest Tunado	0,0060	0,0011	0,9989
XGBoost	0,0159	0,0057	0,9945

# Gráficos de Avaliação

## IMPORTÂNCIA DAS VARIÁVEIS:

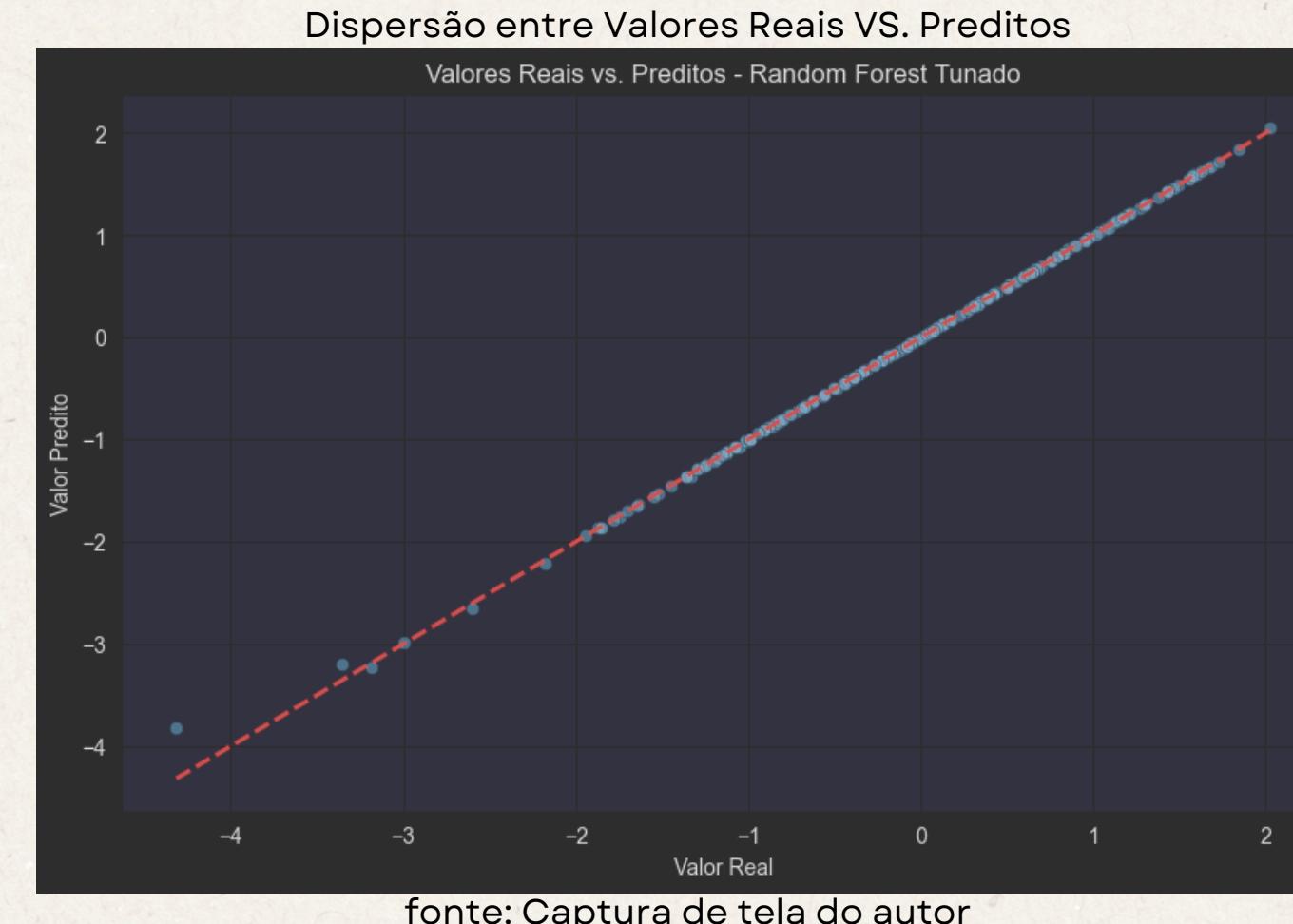
- O modelo identificou as variáveis OpenPorchSF, Neighborhood\_IDOTRR e LotFrontage como as mais relevantes para prever SalePrice.



## DISPERSÃO ENTRE VALORES REAIS VS.

## PREDITOS:

- O gráfico de dispersão confirma que os valores preditos estão alinhados com os valores reais, indicando alta precisão do modelo.



# Conclusão

## Resumo dos principais achados:

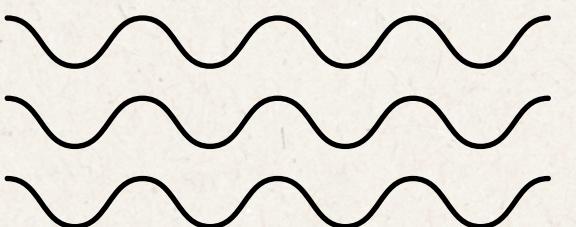
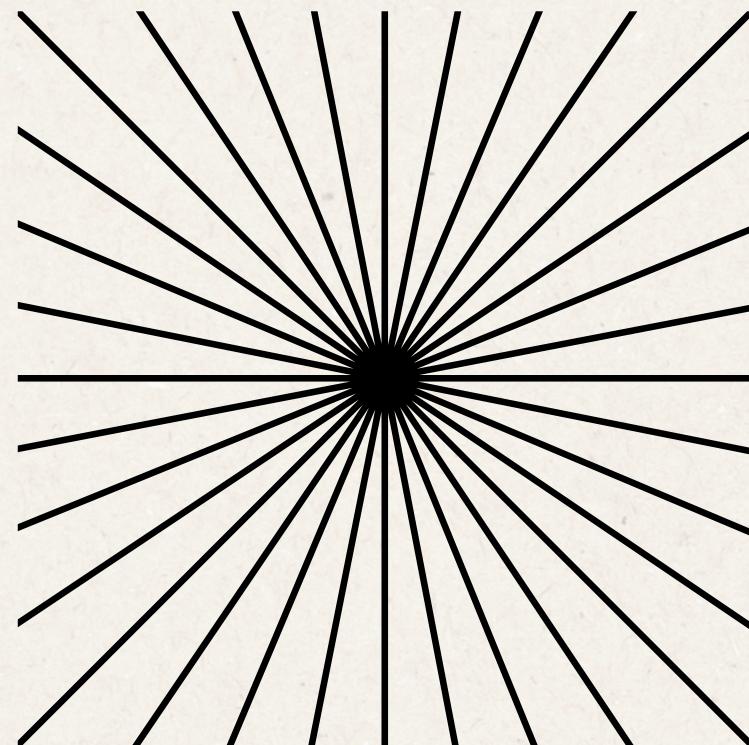
- O Random Forest Tunado foi o modelo com melhor desempenho, alcançando  $R^2 = 0.9989$ .
- As variáveis mais influentes na previsão do preço dos imóveis foram OpenPorchSF, Neighborhood\_IDOTRR e LotFrontage.
- A normalização e o tratamento de outliers melhoraram significativamente a performance dos modelos.

## Desafios encontrados:

- Alto número de variáveis categóricas exigiu um trabalho intensivo de engenharia de features.
- Algumas variáveis apresentaram muitos valores nulos, necessitando técnicas diferentes de imputação.

## Trabalhos futuros:

- Explorar modelos mais avançados, como Deep Learning.
- Investigar técnicas adicionais de seleção de variáveis para reduzir a complexidade do modelo.
- Criar uma API para integrar o modelo em um sistema de previsão de preços de imóveis.



# Muito obrigado!

## CONTACT US

E-mail            luis.gfas@discente.ufma.br

**Github do  
Projeto**        <https://github.com/lgfaf/house-prices-data-mining>