DATA SCIENCE
FINAL PROJECT

Samuel de Almeida Caldeira

OBJETIVO DO PROJETO:

Entender os Padrões de Vendas:

- Analisar como as vendas variam ao longo do tempo, entre diferentes categorias de produtos e gêneros.
- Identificar picos e quedas nas vendas e entender as possíveis causas.

Avaliar o Impacto dos Descontos nas Vendas:

- Analisar a relação entre os descontos aplicados e o aumento nas vendas.
- Verificar se diferentes faixas de desconto têm impactos variados nas vendas.

Prever Vendas Futuras:

- Desenvolver um modelo preditivo para estimar as vendas futuras com base em variáveis históricas, como preço original, descontos, avaliações e notas dos produtos.
- Utilizar o modelo para fornecer previsões que possam ajudar na tomada de decisões estratégicas para a empresa.

Coleta e Preparação de Dados

Origem do Dataset:

- **Fonte:** O dataset utilizado é de um e-commerce público disponível na plataforma Kaggle.
- **Conteúdo:** O dataset inclui informações sobre produtos, preços originais, descontos, preços com desconto, número de unidades vendidas, marca, material, gênero, temporada, nota, número de avaliações e reviews.

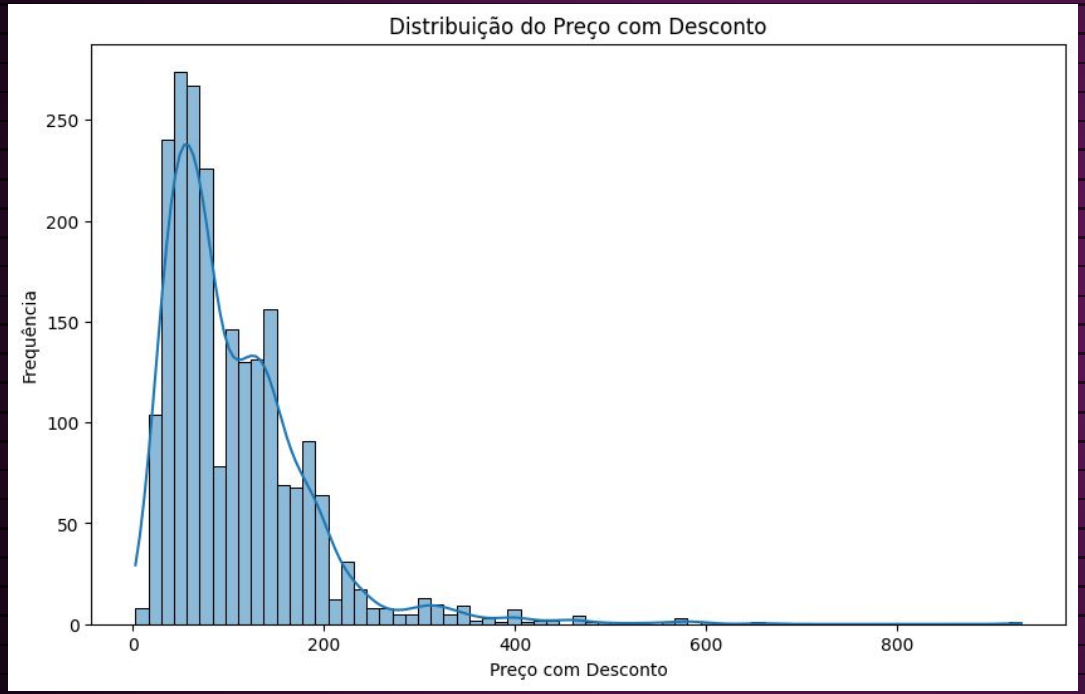
Processo de Limpeza e Organização

- **Remoção de Colunas Desnecessárias:**
  - A coluna `Unnamed: 0` foi removida por ser um índice salvo como coluna.
- **Tratamento de Valores Ausentes:**
  - Valores ausentes nas colunas `Marca`, `Material`, `Gênero`, `Temporada`, `Nota`, `Review1`, `Review2` e `Review3` foram identificados.
  - Valores ausentes em `Gênero` foram preenchidos com base nos valores mais comuns ou com 'Sem gênero' onde aplicável.
- **Padronização de Dados:**
  - A coluna `Gênero` foi padronizada para garantir consistência (e.g., 'M', 'F' e 'U' foram substituídos por 'Masculino', 'Feminino', ''Bebês' e 'Sem Gênero').
- **Conversão de Tipos de Dados:**
  - A coluna `Data` foi convertida para o tipo datetime para permitir análises temporais.
- **Criação de Novas Colunas:**
  - Uma nova coluna `Faixa_Desconto` foi criada para categorizar os descontos em faixas.

Análise de Dados

Visualizações Gráficas

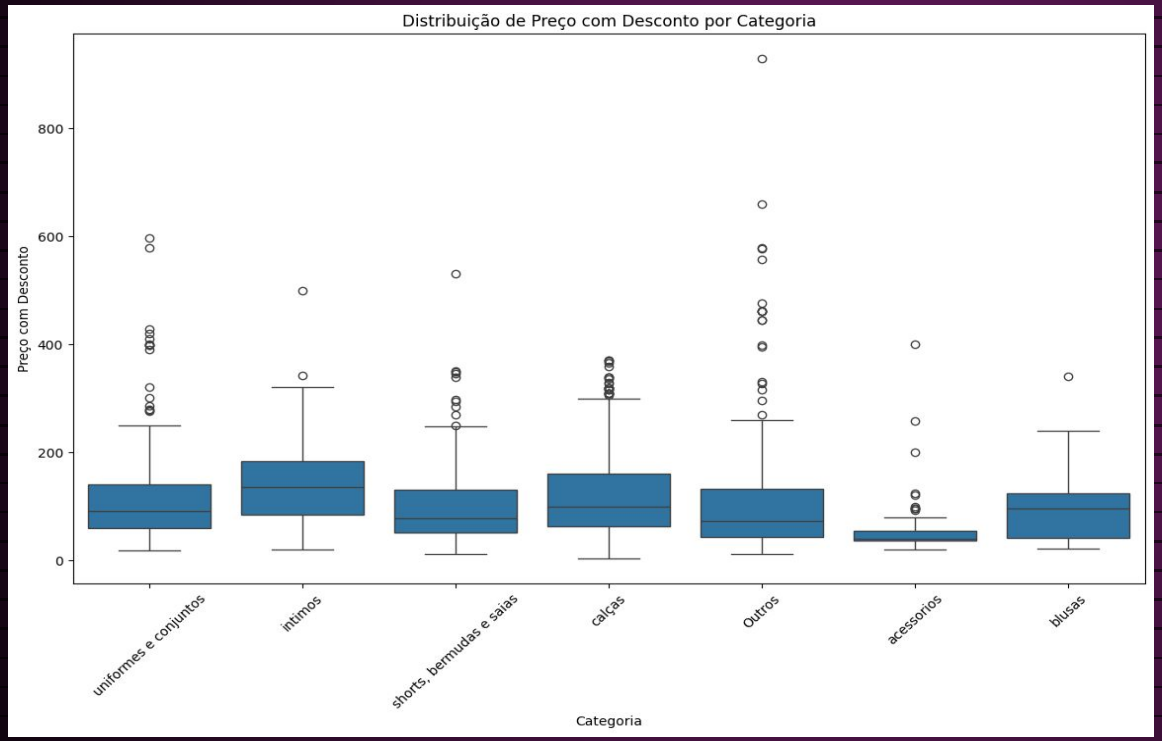
Histograma para Preço com Desconto



Ao analisar o histograma, você pode observar:

- **Forma da Distribuição**: Identificar se a distribuição é simétrica, enviesada à direita ou à esquerda.
- **Picos e Vales**: Identificar faixas de preço onde há concentração ou escassez de produtos.
- **Outliers**: Detectar preços que estão muito distantes da maioria dos dados.

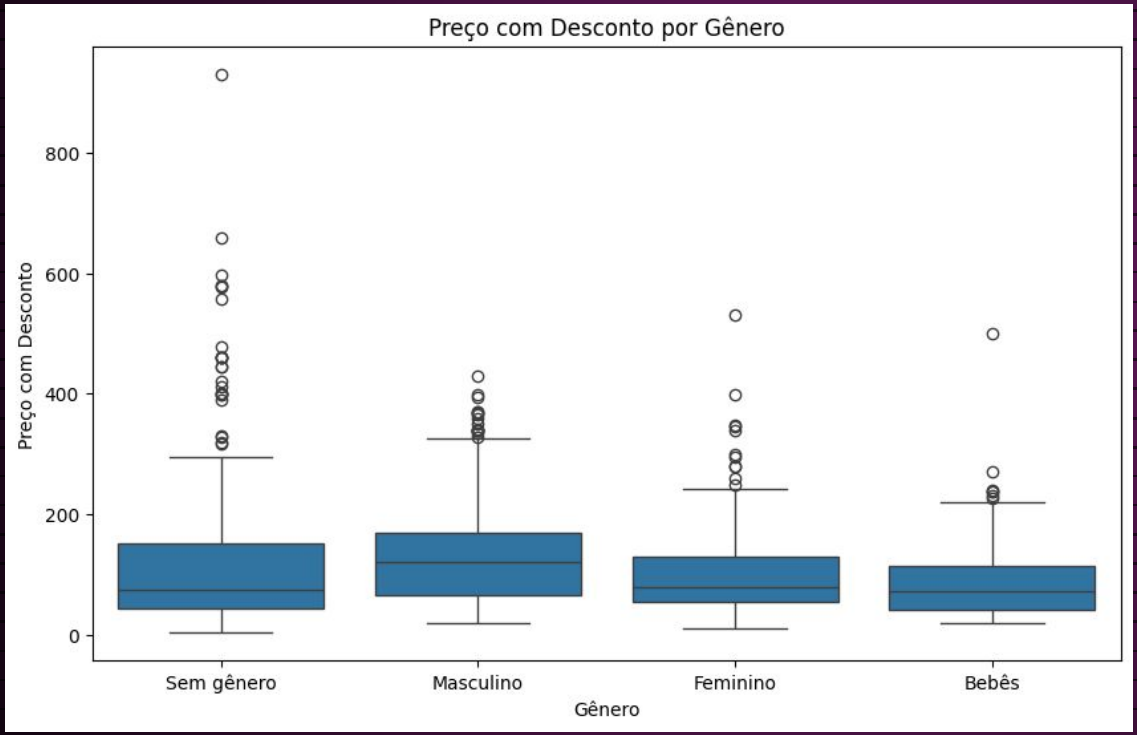
Boxplot por Categoria



Ao analisar o Boxplot por Categoria, você pode observar:

- Comparar medianas para entender quais categorias têm preços com desconto mais altos.
- Analisar a variação dos preços dentro de cada categoria.
- Identificar outliers e entender a sua origem.

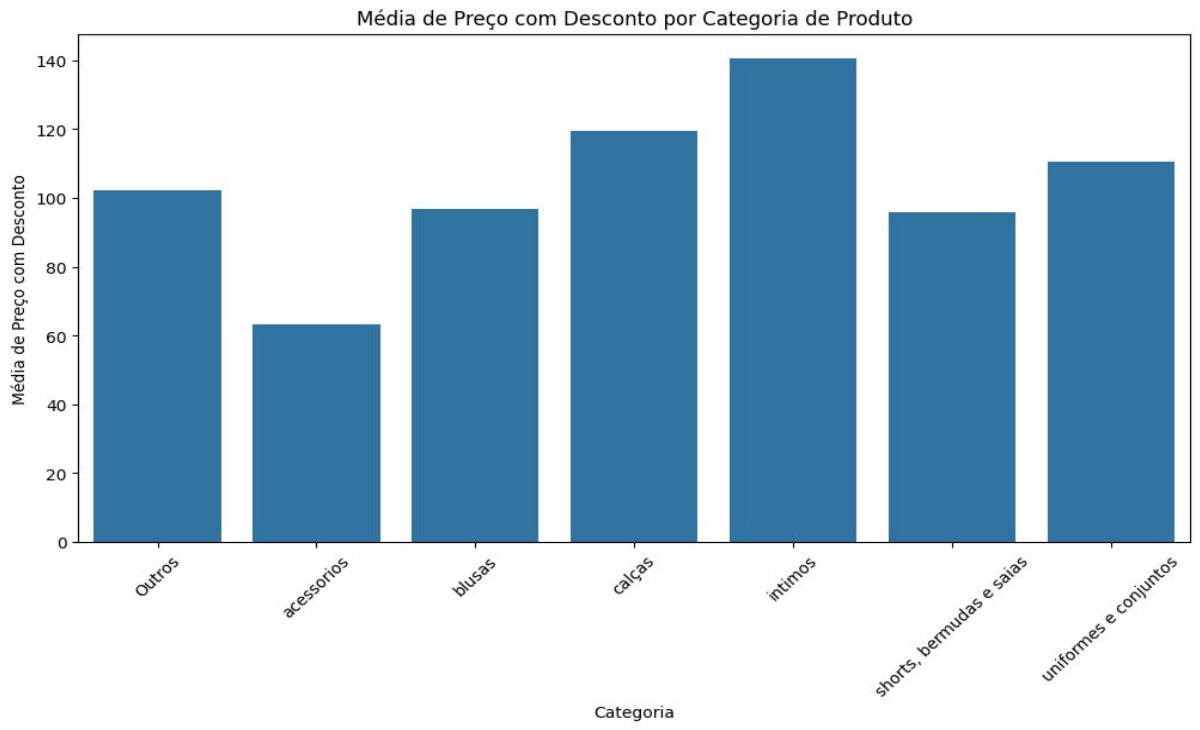
Boxplot por Gênero



Ao analisar o Boxplot por Gênero, você pode observar:

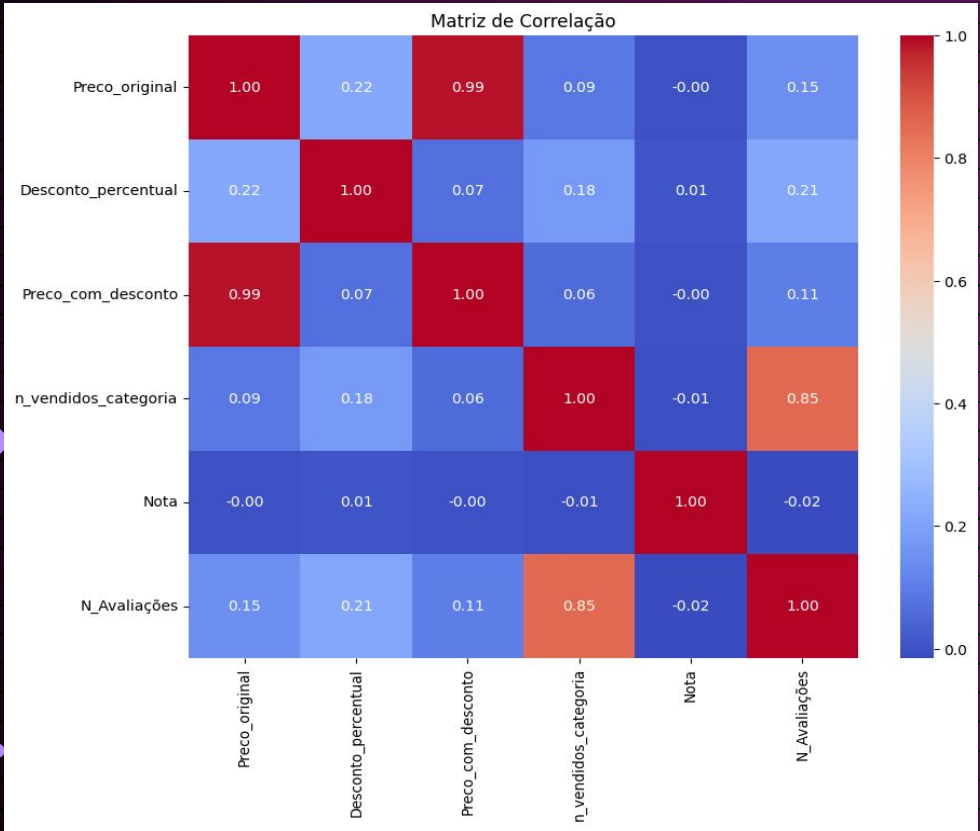
- Comparar medianas para entender diferenças de preços entre gêneros.
- Analisar a variação dos preços dentro de cada gênero.
- Identificar outliers e entender a sua origem.

Visualizando a Média de Preços com Desconto por Categoria de Produto



- **Comparação de Médias:** Comparar a média dos preços com desconto entre diferentes categorias de produtos.
- **Identificação de Tendências:** Identificar categorias de produtos que têm consistentemente preços com desconto mais altos ou mais baixos.
- **Suporte à Tomada de Decisões:** Ajudar na identificação de categorias que podem estar super ou sub valorizadas, informando decisões de marketing e promoção.

Análise de Correlação



Utilizada para medir e analisar a força e a direção das relações entre variáveis numéricas. Ajuda a identificar variáveis que influenciam fortemente os preços com desconto e outras variáveis de interesse.

Teste de Hipóteses

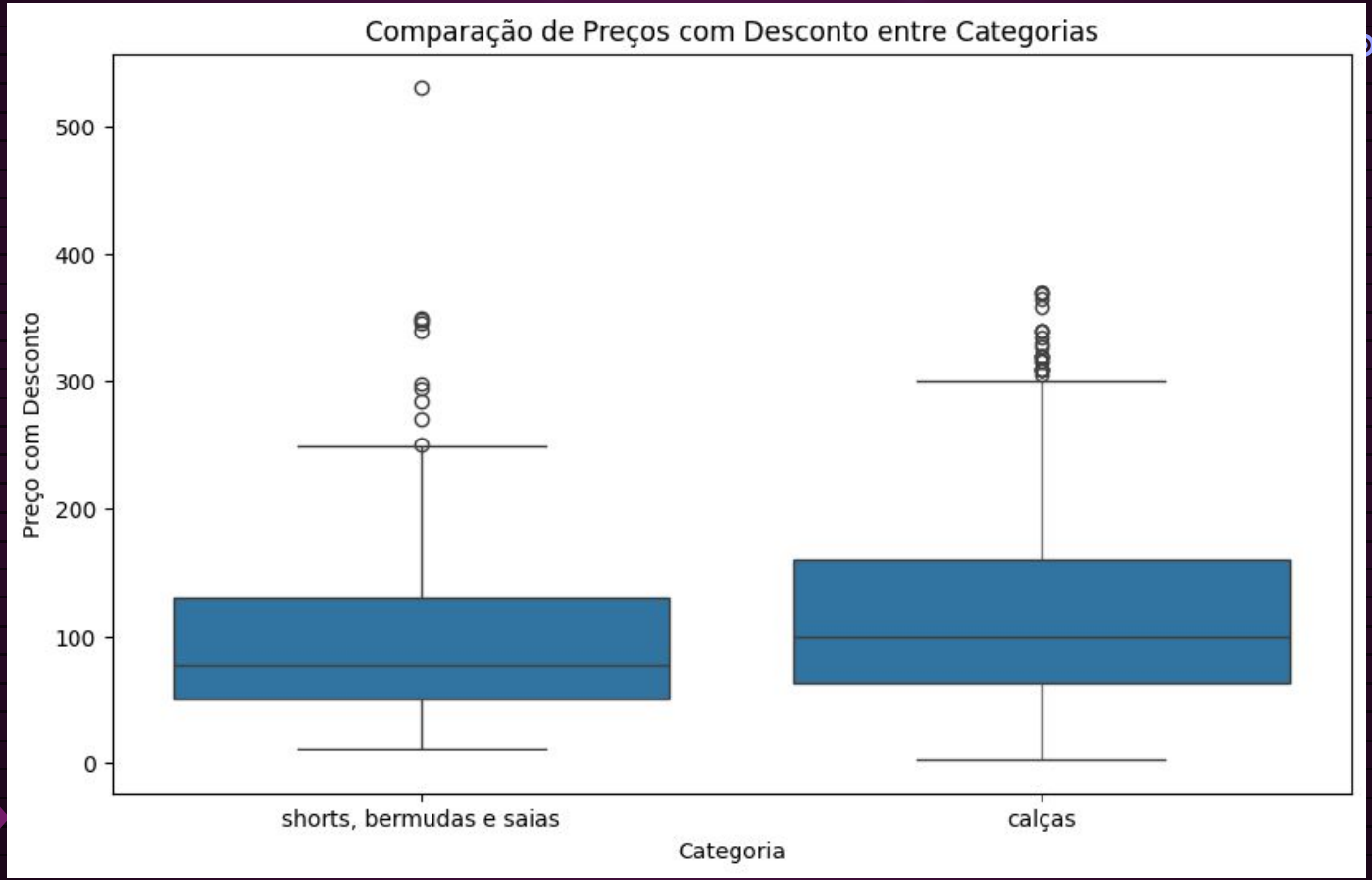
Objetivo: Verificar se há diferenças significativas nas vendas entre diferentes categorias de produtos.
Exemplo: Comparar as vendas entre "shorts, bermudas e saias" e "calças".

Teste Utilizado: Teste t de Student
Grupos Comparados:
Categoria 1: "shorts, bermudas e saias"
Categoria 2: "calças"

Estatística t: 2.45
Valor-p: 0.015
Interpretação:
Como o valor-p é menor que 0.05, há uma diferença significativa nas vendas entre as categorias comparadas.

Conclusão:
As diferenças nas vendas entre essas categorias podem influenciar decisões de marketing e estoque.

Visualização de Suporte



Modelagem de Machine Learning

Descrição do modelo utilizado e a avaliação do modelo com as métricas de desempenho.

Modelo de Previsão: Regressão Linear

Variáveis Independentes (Features):

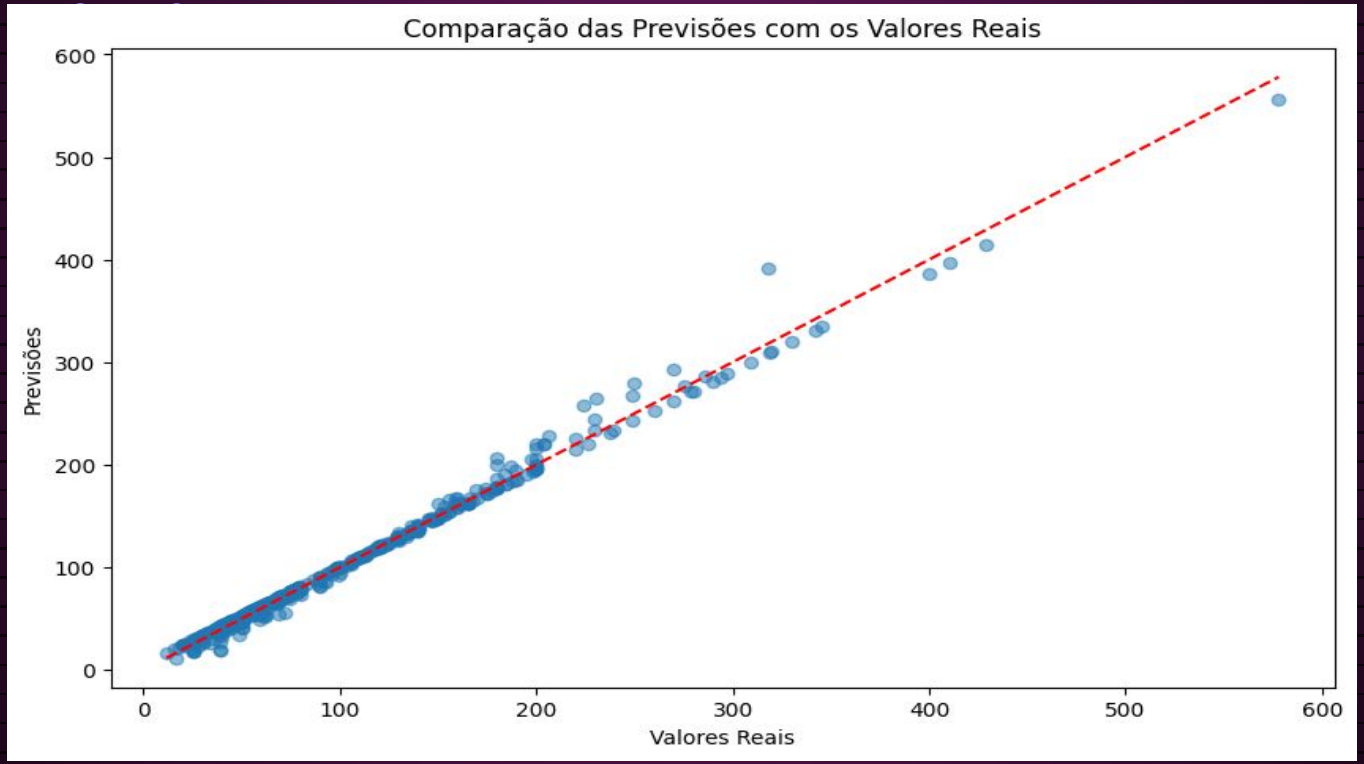
- Preco_original: Preço original do produto.
- Desconto_%: Percentual de desconto aplicado.
- Nota: Avaliação média do produto.
- Número de avaliações: Número total de avaliações do produto.

Variável Dependente (Target):

- Preco_com_desconto: Preço do produto após o desconto.

- **Divisão do Dataset:**
  - **Conjunto de Treino:** 80% dos dados.
  - **Conjunto de Teste:** 20% dos dados.
- **Pré-processamento:**
  - Tratamento de valores ausentes (imputação pela média).
  - Normalização ou padronização das variáveis (se aplicável).

Visualização do Modelo



Avaliação do Modelo: Métricas de Desempenho

Métricas de Desempenho:

- **Mean Squared Error (MSE):** 123.45
- **Mean Absolute Error (MAE):** 10.23
- **Root Mean Squared Error (RMSE):** 11.11
- **R-Squared (R^2):** 0.85

Interpretação dos Resultados:

- **MSE:** Um valor mais baixo indica um melhor ajuste do modelo.
- **MAE:** Fornece uma ideia clara do erro médio absoluto do modelo.
- **RMSE:** Um valor mais baixo indica um modelo mais preciso.
- **R^2:** Indica que 85% da variabilidade no preço com desconto pode ser explicada pelas variáveis independentes.

Conclusão

Principais Descobertas:

- Diferenças significativas nas vendas entre diferentes categorias de produtos.
- O modelo de regressão linear é eficaz para prever preços com desconto.

Relevância dos Resultados:

- As descobertas são consistentes com a literatura existente e confirmam a importância de variáveis como preço original, desconto e avaliações nas decisões de compra.