Ficha técnica - DataLAB para Amazon Sales

🔶 Contexto do Negócio e Objetivo da Análise

A DataLAB, consultoria especializada em análise de dados, foi contratada para o Projeto Amazon Sales, para analisar um dataset contendo informações sobre produtos e avaliações da plataforma Amazon.

O objetivo central foi preparar e explorar os dados para extrair insights relevantes, com um foco particular na relação entre as categorias de produtos e as classificações atribuídas pelos usuários.

O objetivo secundário do projeto foi explorar o uso da IA para otimizar o processo de análise.

🔶 Perguntas de Negócio

1 – Quanto maior o desconto, melhor será a pontuação.

2 – Quanto maior o número de pessoas que avaliaram o produto, melhor será a classificação.

3 – Produtos com preços reais mais altos (sem desconto aplicado) tendem a ter uma avaliação maior?

4 - Produtos com um preço real mais alto tendem a ter descontos absolutos maiores
(ou seja, o valor do desconto em R$)?

🔶 Fonte de Dados

O processo iniciou com o carregamento de dois conjuntos de dados principais: `amazon_review` (dados de avaliações de produtos) e `amazon_product` (dados de produtos) fornecidos pela Laboratória Brasil.

🔶 Ferramentas

Plataformas: Google Colab, Google Slides, Google Docs

Bibliotecas: Pandas, Numpy, Seaborn, Matplotlib, Scipy

Linguagem: Phyton

Inteligências artificiais: Gemini IA e ChatGPT

🔶 Metodologia

> Este relatório documenta as principais etapas e achados do processo de análise de um dataset contendo informações sobre produtos e avaliações da plataforma Amazon.

▼ Conectar/importar dados para outras ferramentas

Realizado o upload dos arquivos extraídos para o ambiente do **Google Colab**, seguindo os seguintes passos:

1. Acesso ao painel lateral esquerdo → opção Arquivos;
2. Seleção de Fazer upload para o armazenamento da sessão;
3. Escolha da pasta extraída e dos arquivos:
    - amazon - amazon_review.csv
    - amazon - amazon_product.csv

Feito a criação de dois DataFrames distintos com a biblioteca Pandas, nomeados como `review_df` e `product_df`, correspondendo aos dois arquivos carregados.

Realizada a visualização inicial dos dados por meio do método `.head()`, com o objetivo de obter uma visão geral da estrutura, colunas e conteúdo de cada conjunto de dados. As dimensões iniciais dos dataframes foram:

- `review_df` → **1465 linhas e 10 colunas**
- `product_df` → **1469 linhas e 7 colunas**

▼ Identificar e Tratar Valores Nulos

A checagem de valores nulos é uma etapa essencial na análise de dados, pois garante a qualidade e a confiabilidade dos resultados. Dados ausentes podem distorcer estatísticas, prejudicar visualizações e comprometer modelos preditivos. Identificar e tratar esses valores corretamente permite uma base sólida para análises mais precisas e tomadas de decisão assertivas.

Nessa etapa, foi utilizado o seguinte prompt no Gemini para encontrar valores nulos nos dataframes `review_df` e `product_df` *"Identifique as colunas que contêm valores nulos em meu dataset e me mostre a quantidade de valores nulos por coluna."*

Tabela review_df	Valores Nulos
img_link	466
product_link	466
rating_count	2
Tabela product_df	**Valores Nulos**
about_product	4

Para auxiliar no tratamento dos valores nulos, foi utilizado o prompt: *"Quais são as técnicas recomendadas para tratar valores nulos em variáveis numéricas e categóricas?"* A partir dele, obtivemos insights relevantes e, após análise criteriosa, observamos que:

1. **Proporção de dados nulos:** as variáveis `rating_count` e `about_product` apresentaram uma quantidade mínima de nulos em relação ao total do dataset. Por isso, a exclusão dessas linhas teve impacto irrelevante na perda de dados e se mostrou uma abordagem eficiente. Caso a proporção fosse elevada, outras estratégias, como imputação, seriam consideradas.
2. **Impacto na análise:** valores nulos em `about_product` indicam ausência da descrição do produto, o que inviabiliza análises de texto ou relações com preço e avaliação. Já em `rating_count`, a ausência da contagem de avaliações compromete a análise da popularidade e engajamento com o produto.
3. **Natureza dos dados:** imputar a variável `about_product` com descrições genéricas ou vazias não acrescentaria valor à análise. Já imputar `rating_count` com médias ou medianas distorceria as estatísticas e poderia introduzir viés, já que trata-se de um dado factual.

Dessa forma, optamos por excluir as linhas com valores nulos, considerando que essa decisão garante a integridade das análises futuras.

Para os nulos em `product_link` e `img_link`, optamos por mantê-los até as próximas etapas, para definir se serão utilizados em análise futura do dataset.

Número de linhas após remover os nulos:

- `review_df` → **1463** linhas
- `product_df` → **1465** linhas

▼ Identificar e Tratar Valores Duplicados

A identificação de dados duplicados é uma etapa fundamental para garantir a integridade da análise. No entanto, nem toda duplicata representa um erro: é importante avaliar o contexto e os objetivos do projeto para entender se uma entrada repetida é válida (como múltiplas compras iguais por um mesmo cliente) ou redundante.

A definição de duplicatas geralmente exige a escolha criteriosa de **colunas-chave** — aquelas que, em conjunto, devem representar uma linha única, como IDs. Para auxiliar na remoção desses valores duplicados, foi utilizado o seguinte prompt: *"Remova as linhas duplicadas com base nas colunas chave (especifique as colunas, caso necessário)"*. As IDs utilizadas foram `user_id`, `review_id` e `product_id`.

Número de linhas após remover duplicatas:

- `review_df` → **1359** linhas (removidos o total de 106 linhas)
- `product_df` → **1351** linhas (removidos o total de 118 linhas)

Realizamos uma nova checagem após a remoção das duplicatas para confirmar se realmente todos haviam sido excluídos e não foi evidenciado nenhum valor duplicados nas variáveis.

▼ Identificar e Tratar Dados Fora do Escopo de Análise

Definir o que está dentro ou fora do escopo da análise envolve alinhar os dados ao objetivo do projeto, considerando o período analisado, as categorias relevantes e o público-alvo. Manter valores fora desse escopo pode gerar distorções, poluir o dataset e comprometer a tomada de decisão.

Nessa etapa, o prompt *"Quais valores em meu dataset estão fora do escopo do meu projeto?"* , nos auxiliou em como poderíamos estar identificando esses valores. Realizamos uma análise exploratória detalhada dos dataframes, incluindo estatísticas descritivas e verificação de valores únicos nas colunas categóricas, o que nos ajudou a identificar inconsistências.

Com essa busca, identificamos que as variáveis `product_link` e `img_link` apresentavam um valor discrepante de `count` devido os valores nulos presentes nela, e além disso, não seriam utilizadas para o objetivo final da análise. Logo, optamos por excluí-las. Após a exclusão, fizemos uma rápida checagem se ate o momento, todas os tratamentos dos dados haviam sido feitos corretamente e se havia algum valor nulo no dataset. A pesquisa não retornou nenhum valor nulo.

▼ Identificar e Tratar Dados Discrepantes em Variáveis Categóricas, Numéricas e Verificação do Tipo de Dados.

Essa etapa é útil para detectar **valores inconsistentes**, **erros de digitação**, **categorias inesperadas** ou **dados fora do escopo como outliers**, ajudando a garantir que os dados categóricos e numéricos estejam limpos e padronizados para análises futuras.

Para ajudar a identificar quaisquer valores discrepantes ou erros de digitação, criamos um código que exibisse todos os valores únicos para as colunas categóricas nos dataframes `review_df` e `product_df`, baseado no prompt: *"Liste todas as categorias únicas em colunas categóricas e identifique se há valores discrepantes ou erros de digitação (exemplo: 'Male' vs 'male' ou 'Masculino' vs 'masculino')".*

Foram identificadas algumas inconsistências como a presença de símbolos nas variáveis `discounted_price` , `actual_price` e `discount_percentage` e a categorização das mesmas como "object". Removemos os símbolos e, em seguida, transformamos as variáveis para o tipo numérico.

Verificamos as variáveis nos dois dataframes para checar se todas estavam com o tipo de dado definidos corretamente:

Tabela review_df	Tipo de Dados
user_id	object
user_name	object
review_id	object
review_title	object
review_content	object
img_link	object
product_link	object
product_id	object
rating	float64
rating_count	int64

Tabela product_d	Tipo de Dados
product_id	object
product_name	object
category	object
discounted_price	float64
actual_price	float64
discount_percentage	int64
about_product	object

Identificação de Outliers em Variáveis Numéricas

Para identificar valores extremos ou discrepantes nas colunas numéricas dos dataframes, podemos utilizar dois métodos diferentes: Z score ou Intervalo Interquartílico (IQR). Vamos explorar os dois métodos, comparar os resultados e observar qual é o mais adequado para a nossa análise.

Intervalo Interquartílico (IQR)

O IQR é uma medida de dispersão robusta a outliers e é útil para definir limites para identificar valores atípicos. Avaliando as variáveis numéricas das tabelas, identificamos o seguinte número de outliers:

Métrica	Nº Outliers
rating	17
rating_count	130
discounted_price	209
actual_price	185
discount_percentage	0

Z score

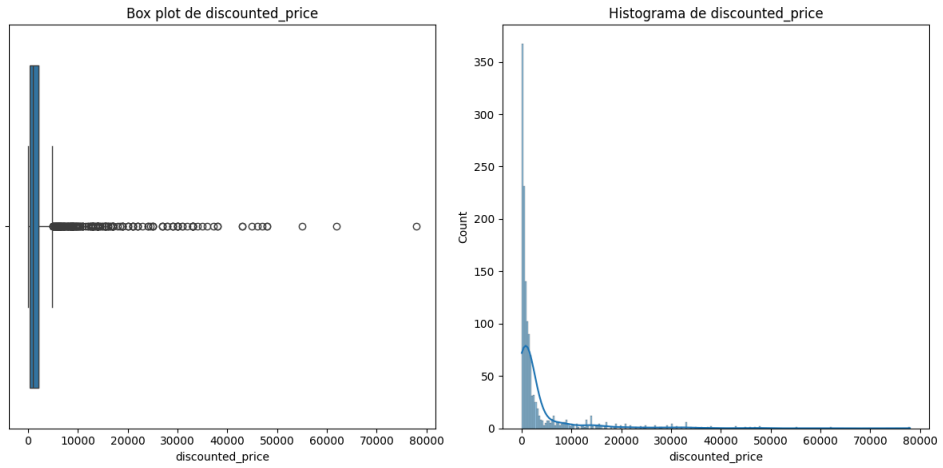
O método Z-Score mede a quantos desvios padrão um ponto de dados está da média. Um Z-Score comum para identificar outliers é geralmente maior que 3 ou menor que -3.

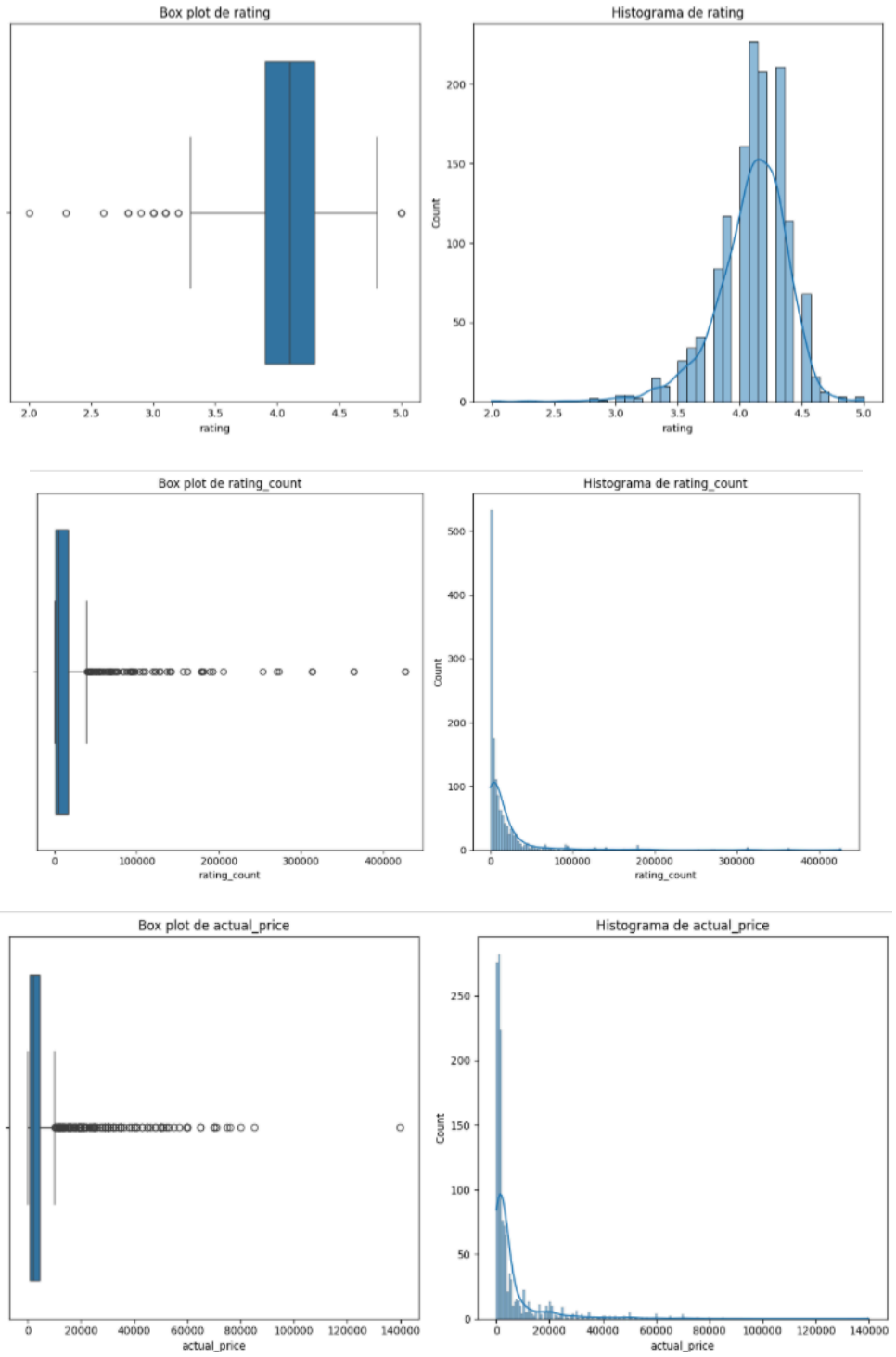
Vamos usar o método Z-Score para identificar valores extremos ou discrepantes nas colunas numéricas e comparar os resultados com o método IQR.

Métrica	Nº Outliers
rating	19
rating_count	28
discounted_price	40
actual_price	36
discount_percentage	0

Visualização da Distribuição

Para auxiliar na decisão do método, optamos por utilizar gráficos para melhor visualização da distribuição dos dados, como boxplot e histograma.





Análise dos Boxplot e Histogramas

A coluna `rating` apresenta uma distribuição aproximadamente simétrica. Tanto o método de **IQR** quanto o de **Z-Score** identificaram a mesma quantidade de outliers (17), indicando consistência entre as abordagens.

Já as colunas `rating_count`, `discounted_price` e `actual_price` exibem distribuições fortemente assimétricas à direita, com caudas longas.

Nesses casos, o método **IQR** se mostra mais adequado para a detecção de outliers, por ser mais robusto a distribuições não normais. Para variáveis com distribuição simétrica, como `rating`, ambos os métodos são apropriados, embora o **IQR** ofereça leve vantagem por sua resistência a variações sutis na simetria.

Dessa forma, optamos por utilizar o **IQR** na identificação de valores extremos ou discrepantes nas colunas numéricas dos DataFrames.

▼ Criar Variáveis

A criação de novas variáveis, também conhecida como *feature engineering*, é uma etapa fundamental em projetos de análise de dados, pois permite extrair informações mais relevantes e aprofundadas a partir dos dados brutos. Ao transformar ou combinar variáveis existentes, é possível revelar padrões ocultos, facilitar a segmentação de clientes, melhorar a performance de modelos preditivos e tornar a análise mais alinhada com os objetivos do negócio.

Foram criadas duas novas variáveis no DataFrame `product_df` com o objetivo de enriquecer a análise: **diferença de preço** e **categoria principal**.

A variável **diferença de preço** representa a economia absoluta, em reais, que o cliente obtém ao adquirir um produto com desconto. Ela foi calculada subtraindo o valor da coluna `discounted_price` (preço com desconto) do valor da coluna `actual_price` (preço original), permitindo identificar o quanto foi economizado por produto. Essa variável possibilita análises como quais produtos ou categorias oferecem maior economia e se essa diferença influencia o volume de vendas ou o número de avaliações.

Já a variável **categoria principal** foi criada para simplificar a análise categórica. A coluna original `category` possui um formato hierárquico, com categorias separadas por `"|"`. A nova variável extrai apenas o primeiro nível dessa hierarquia — ou seja, o termo mais genérico — ao dividir a string pelo caractere `"|"` e selecionar o primeiro elemento. Isso permite agrupar os produtos em grandes categorias e facilitar comparações, como preço médio ou avaliação média, entre essas categorias principais, sem a complexidade da hierarquia completa.

▼ Unir Tabelas

O objetivo desta etapa é combinar múltiplas tabelas a partir de uma chave comum, a fim de enriquecer a base de dados com informações complementares.

Utilizamos o prompt: *"Quais tabelas precisam ser unidas e qual é a chave comum entre elas?"*, que nos orientou na definição das junções necessárias.

As tabelas a serem integradas são:

- `review_df`, que contém informações sobre as avaliações dos usuários;
- `product_df`, que reúne dados relacionados aos produtos.

A chave de ligação entre ambas é a coluna `product_id`, presente nos dois DataFrames. Essa coluna atua como um identificador único de cada produto, permitindo relacionar corretamente as avaliações aos produtos correspondentes.

Após a união das tabelas, utilizamos o prompt *"caso haja dados faltantes após a junção, trate-os adequadamente."*, para verificar se há algum dado faltante após a junção das tabelas. Não foram identificados quaisquer valores nulos nessa busca.

user_id	0
user_name	0
review_id	0
review_title	0
review_content	0
img_link	0
product_link	0
product_id	0
rating	0
rating_count	0

product_name	0
category	0
discounted_price	0
actual_price	0
discount_percentage	0
about_product	0

▼ Agrupar e Visualizar Dados de Acordo com Variáveis Categóricas

Nesta etapa do projeto, o objetivo é agrupar os dados com base em uma variável categórica para obter uma visão mais estruturada e comparativa do comportamento dos dados. A partir desse agrupamento, são calculadas **médias, contagens ou porcentagens** de variáveis numéricas relacionadas, o que permite identificar padrões, diferenças entre grupos e possíveis tendências.

Para otimizar a etapa, foi utilizado o prompt *"Agrupe os dados de acordo com a variável categórica [insira a variável] e calcule a média/contagem/porcentagem das variáveis numéricas relacionadas. Exiba o resumo dos grupos formados."*

Agregação por categoria principal

Agrupamos o DataFrame `unificada_df` com base nos valores únicos da coluna `categoria_principal`. Isso significa que o pandas vai reunir todas as linhas que pertencem à mesma categoria principal. Após agrupar, usamos o método `.agg()` para calcular várias métricas agregadas para cada grupo (cada categoria principal).

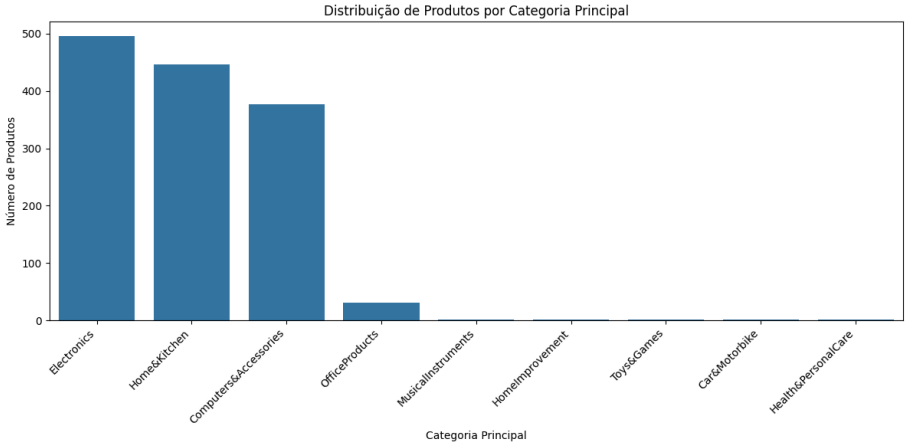
Calculamos a média (`'mean'`) de todas as colunas numéricas e armazenamos o resultado em uma nova coluna chamada `mean_'nome_da_variável'.` Também realizamos a contagem de ocorrências de `product_id` em cada grupo. Como cada linha no dataframe unido representa um produto (após a limpeza de duplicatas de produto), isso efetivamente conta quantos produtos (ou entradas de produtos com reviews associados) existem em cada categoria principal.

categoria_principal	mean_discounted_price	mean_actual_price	mean_discount_percentage	mean_diferenca_preco
Car&Motorbike	2339.00	4000.00	42.0	1661.00
Computers&Accessories	943.25	1850.98	53.32	907.73
Electronics	6172.82	10364.60	50.12	4191.78
Health&PersonalCare	899.00	1900.00	53.00	1001.00
Home&Kitchen	2331.13	4165.7	40.17	1834.66
HomeImprovement	337.00	799.00	57.50	462.00
MusicalInstruments	638.00	1347.00	46.00	709.00
OfficeProducts	301.58	397.19	12.35	95.61
Toys&Games	150.00	150.00	0.00	0.00

- **Visão Geral por Categoria:** A categoria "Electronics" tem os preços médios mais altos, enquanto "Toys&Games" tem preços médios mais baixos.
- **Descontos por Categoria:** Compare a `mean_discount_percentage` e `mean_diferenca_preco` para entender as estratégias de desconto. "Computers&Accessories" tem uma alta porcentagem média de desconto, mas "Electronics" tem uma alta diferença de preço média em reais, provavelmente devido aos preços base mais altos.
- **Desempenho Percebido e Popularidade:** As colunas `mean_rating` e `mean_rating_count` dão uma ideia do quão bem avaliados são os produtos em cada categoria e quão populares eles parecem ser (com base no volume de avaliações).
- **Composição do Dataset:** As colunas `product_count` e `percentage_of_products` mostram a distribuição dos produtos no seu dataset. "Electronics", "Home&Kitchen" e "Computers&Accessories" são as categorias mais representadas, compondo a grande maioria dos dados. Categorias com `product_count` igual a 1 ou 2 (como 'Car&Motorbike',

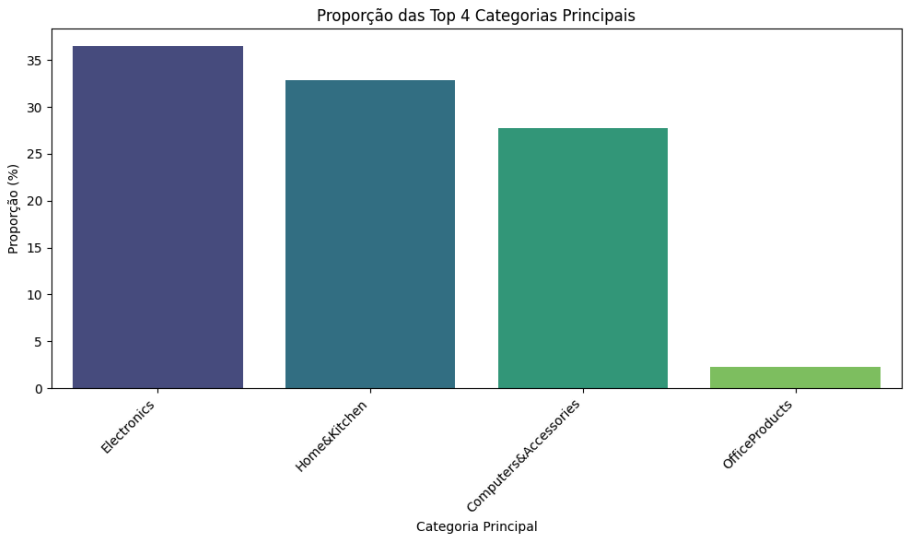
'Health&PersonalCare', etc.) são pouco representadas neste dataset e as médias para elas podem não ser muito confiáveis.

-



Posteriormente, calculamos a proporção (percentual) de cada categoria principal no dataset unido e, em seguida, criamos um gráfico de barras para exibir as proporções das 4 categorias principais com maior frequência.

Subcategoria	Proporção
Electronics	36.524300
Home&Kitchen	32.916053
Computers&Accessories	27.761414
OfficeProducts	2.282769
MusicalInstruments	0.147275
HomeImprovement	0.147275
Toys&Games	0.073638
Car&Motorbike	0.073638
Health&PersonalCare	0.073638



Este gráfico permite que você veja rapidamente quais categorias dominam seu dataset e a magnitude de sua representação em comparação com as outras top categorias. Você pode, por exemplo, observar que "Electronics" e "Home&Kitchen" juntas representam uma grande fatia do dataset, enquanto "OfficeProducts" tem uma proporção significativamente menor entre as top 4.

▼ Medidas de Tendência Central

A comparação entre média e mediana é crucial para entender a simetria ou assimetria da distribuição dos dados. A moda complementa essa análise, indicando os valores mais típicos ou frequentes. Para variáveis com distribuições assimétricas e outliers (como preços e contagem de avaliações), a mediana é geralmente uma medida de tendência central mais robusta e representativa do "valor típico" do que a média.

Variáveis	Média	Mediana	Moda
discounted_price	3294.57	899.00	[299.0]
actual_price	5687.36	1795.00	[999.0]
discount_percentage	46.84	49.00	[50]
diferenca_preco	2392.79	803.50	[0.0]
rating	4.09	4.10	[4.1]
rating_count	17805.42	4863.00	[9378.0]

Para as variáveis `discounted_price`, `actual_price`, `diferenca_preco` e `rating_count`, onde identificamos outliers e distribuições assimétricas, a mediana é de fato uma melhor representação do valor central do que a média, pois não é distorcida pelos valores atípicos.

▼ Aplicar Medidas de Dispersão

As medidas de dispersão são essenciais para ir além do valor médio e entender a real variabilidade e consistência nos dados de produtos e avaliações. Elas são capazes de:

- Revelam quão confiável é a média como representativa
- Permitir comparar variabilidade entre categorias
- Ajudam a identificar e mensurar o impacto de outliers e facilitam entender a forma da distribuição dos dados.

Variáveis	Desvio Padrão	Variância	IQR
discounted_price	7157.17	51225039.88	1827.50
actual_price	11191.95	125259740.36	3738.50
discount_percentage	21.65	468.53	31.75
diferenca_preco	4727.66	22350725.68	1620.75
rating	0.29	0.09	0.40
rating_count	42161.54	1777595836.57	15370.00

Pelo output, `rating_count` e `diferenca_preco` têm desvios padrão altos, o que confirma a grande variação nesses valores que vimos nos box plots. O `rating` tem um desvio padrão baixo, indicando que as avaliações tendem a estar mais próximas da média (entre 4 e 5).

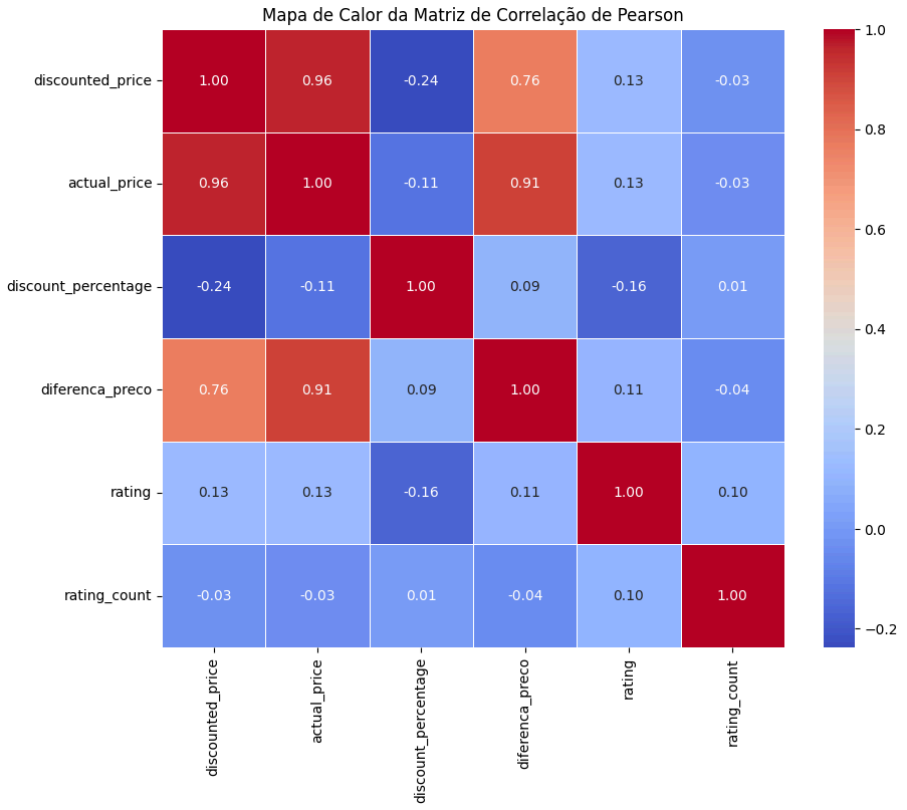
▼ Correlação entre variáveis numéricas

A matriz de correlação mostra os coeficientes de correlação de Pearson entre todas as pares de variáveis numéricas no seu dataframe `unificada_df`. O coeficiente de Pearson varia de -1 a +1:

- **+1:** Correlação positiva perfeita (quando uma variável aumenta, a outra também aumenta linearmente).
- **1:** Correlação negativa perfeita (quando uma variável aumenta, a outra diminui linearmente).
- **0:** Nenhuma correlação linear (não há uma relação linear clara entre as variáveis).
- Valores próximos de +1 ou -1 indicam correlações fortes.
- Valores próximos de 0 indicam correlações fracas.

O mapa de calor é uma representação visual dessa matriz de correlação, onde a cor e a intensidade indicam a força e a direção da correlação:

- Cores quentes (vermelho/laranja) geralmente representam correlações positivas.
- Cores frias (azul/roxo) geralmente representam correlações negativas.
- A intensidade da cor geralmente indica a força da correlação (cores mais vibrantes para correlações mais fortes, cores mais pálidas para correlações mais fracas).
- Os valores numéricos dentro de cada célula do mapa de calor são os próprios coeficientes de correlação.
-



Resumo dos Resultados de Correlação:

- **Fortes Correlações Positivas:**
  - `discounted_price` e `actual_price` (~0.96),
  - `actual_price` e `diferenca_preco` (~0.91),
  - `discounted_price` e `diferenca_preco` (~0.76).

Essas relações mostram que preços mais altos, com ou sem desconto, estão fortemente associados entre si e com maiores valores de desconto absoluto — o que é esperado.

- **Correlações Fracas:**
  - `rating` e `rating_count` têm correlação fraca com variáveis de preço e desconto, indicando pouca ou nenhuma relação linear entre avaliação dos produtos e seus preços ou descontos.
- **Correlações Negativas Fracas a Moderadas:**
  - `discount_percentage` com `discounted_price` (–0.24) e `actual_price` (–0.11).

    Produtos com maior percentual de desconto tendem a ter preços mais baixos, sugerindo que itens mais baratos recebem descontos percentuais maiores.

▼ Validação de Hipóteses

▼ Hipótese 1 - Quanto maior o desconto, melhor será a pontuação

Para validar essa hipótese, foi utilizado o método de correlação de **Spearman**, apropriado devido à presença de outliers nas variáveis numéricas e à possibilidade de relação não linear entre as variáveis `discount_percentage` e `rating`. A escolha dessa técnica foi estratégica, pois Spearman é robusto a valores extremos e eficaz para identificar relações monotônicas, mesmo que não sejam lineares.

Também foi analisado o **p-valor**, que indica se o resultado obtido é estatisticamente significativo — ou seja, se é provável que a correlação observada tenha ocorrido por acaso. Quanto menor o p-valor (geralmente < 0,05), maior a confiança de que existe uma relação real entre as variáveis.

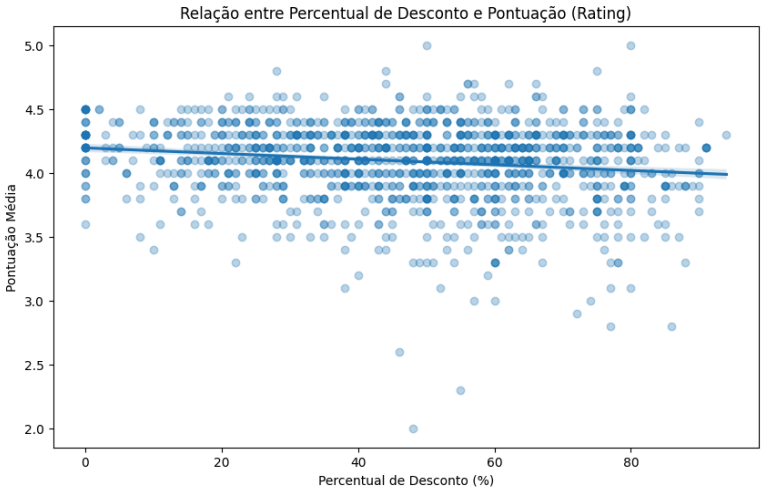
Resultado da análise:

- **Correlação de Spearman:** –0,1512

  → Indica uma correlação **negativa fraca**, ou seja, uma leve tendência de que produtos com maiores descontos tenham notas um pouco mais baixas.
- **P-valor:** 0,0000

  → Valor extremamente baixo, indicando que o resultado é **estatisticamente significativo**. É muito improvável que essa correlação tenha ocorrido por acaso

A correlação negativa fraca encontrada indica que, no conjunto de dados analisado, existe uma **tendência sutil de que produtos com maiores descontos recebam pontuações ligeiramente menores**. Apesar de a força da correlação ser baixa, o resultado é estatisticamente significativo, o que nos permite rejeitar a hipótese nula de ausência de correlação.



▼ Hipótese 2 - Quanto maior o número de pessoas que avaliaram o produto, melhor será a classificação.

Para validar essa hipótese, foram aplicadas duas técnicas de correlação: **Pearson** e **Spearman**, com o objetivo de analisar a relação entre `rating_count` (número de avaliações) e `rating` (pontuação do produto). Também foi analisado o **p-valor** em ambas as correlações.

Resultados da análise:

- **Correlação de Pearson:** 0,0982

  → Correlação **positiva muito fraca**. Indica uma tendência mínima de que produtos com mais avaliações tenham notas um pouco maiores.

- **P-valor (Pearson):** 0,0003

  → Estatisticamente significativo, ou seja, a correlação é confiável.

- **Correlação de Spearman:** 0,1902

  → Correlação **positiva fraca**, mas um pouco mais forte que Pearson. Sugere uma tendência sutil de aumento na pontuação conforme o número de avaliações cresce.

- **P-valor (Spearman):** 0,0000

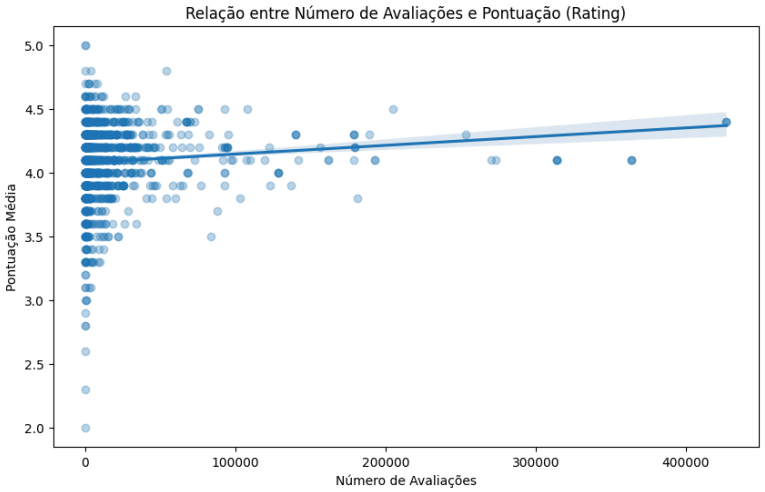
  → Estatisticamente significativo.

Os resultados mostram uma **relação positiva fraca**, porém **estatisticamente significativa** entre número de avaliações e pontuação média dos produtos. Em outras palavras, existe uma leve tendência de que produtos mais avaliados tenham notas um pouco maiores, mas essa relação é fraca e não pode ser considerada um fator determinante da qualidade percebida.

Essa correlação fraca pode ser explicada por fatores como:

- Produtos populares tendem a ter mais avaliações, o que pode estabilizar a média e suavizar notas muito baixas.
- Produtos com menos avaliações podem ter classificações mais extremas (muito altas ou muito baixas), o que distorce a média.
- A nota do produto é influenciada por diversos outros fatores além da quantidade de avaliações, como qualidade, expectativa do consumidor, preço, entre outros.

Em resumo, embora os dados indiquem uma leve relação positiva, o número de avaliações por si só **não é um bom preditor da pontuação média** de um produto, logo, não podemos confirmar a hipótese.



▼ Hipótese 3 - Produtos com preços reais mais altos (sem desconto aplicado) tendem a ter uma avaliação maior?

Esta hipótese foi testada com o objetivo de verificar se há relação entre o **preço real do produto** (sem desconto aplicado) e a **pontuação média atribuída pelos usuários**. Nessa hipótese também foi utilizada a Correlação de Pearson e de Spearman.

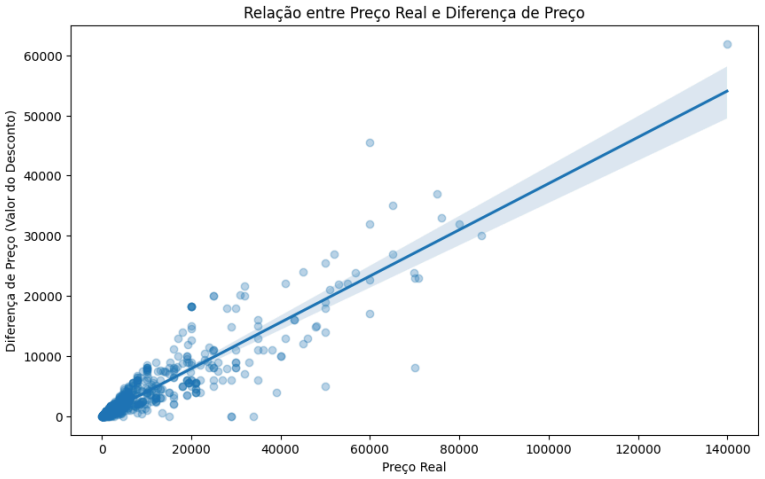
- **Correlação de Pearson:** 0,1280 — P-valor: < 0,001

  → Relação **positiva fraca**, mas **estatisticamente significativa**.

- **Correlação de Spearman:** 0,0325 — P-valor: > 0,05

  → Relação **muito fraca** e **não significativa**.

O gráfico de dispersão (scatter plot) abaixo reforça essa conclusão: há uma grande dispersão dos dados e **ausência de uma tendência clara**, mesmo com a linha de tendência ligeiramente ascendente.



A **Hipótese 3 foi refutada**. Apesar da correlação de Pearson apontar uma fraca relação positiva entre preço real e pontuação média, essa evidência **não é consistente**, pois a correlação de Spearman não confirmou o resultado. Além disso, visualmente, a pontuação média se mantém próxima de 4 independentemente do valor do produto, com ampla variação em todos os níveis de preço.

Portanto, neste dataset, o **preço real não é um bom indicador de qualidade percebida** (rating). Isso sugere que outros fatores, como marca, funcionalidade, experiência do usuário ou marketing, influenciam mais fortemente a avaliação dos produtos.

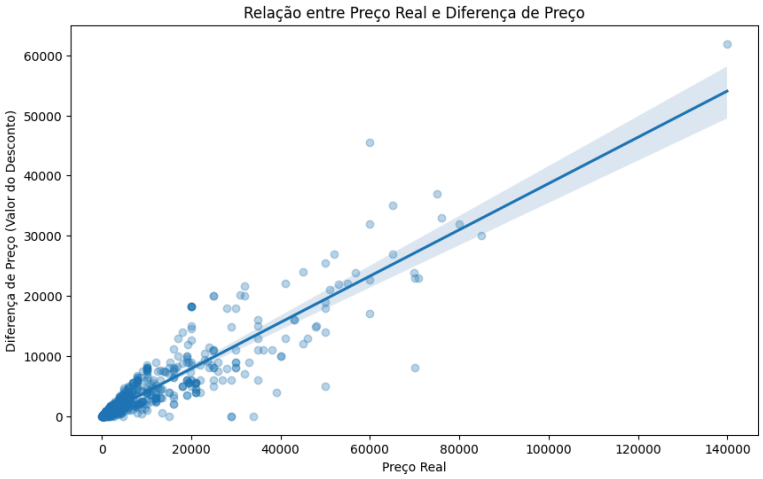
▼ Hipótese 4 - Produtos com um preço real mais alto tendem a ter descontos absolutos maiores (ou seja, o valor do desconto em R$)?

Para testar essa hipótese, foi analisada a relação entre `actual_price` (preço real do produto) e `diferenca_preco` (valor do desconto em R$). Utilizaram-se os métodos de **correlação de Pearson** e **Spearman** para medir a força e a direção dessa relação.

- **Correlação de Pearson:** 0,9108
- **Correlação de Spearman:** 0,8952
- **P-valores:** ambos < 0,0001 (estatisticamente significativos)

Esses valores indicam uma **relação positiva muito forte e estatisticamente significativa** entre o preço real e o valor do desconto. Ou seja, produtos mais caros tendem a apresentar maiores descontos absolutos, confirmando fortemente a **Hipótese 4**.

Além disso, um gráfico de dispersão confirmou visualmente essa tendência com uma clara linha ascendente, evidenciando uma forte relação linear.



Diferença entre Categorias de Produto

Embora a relação geral entre preço real e valor do desconto seja forte, foi necessário investigar se essa tendência se mantém igualmente entre diferentes **categorias de produtos**. Para isso, foi aplicado o **teste ANOVA**, que avalia se há diferenças significativas na média da `diferenca_preco` entre as categorias principais.

O ANOVA calcula uma **estatística F**, que indica o quanto as médias dos grupos são diferentes entre si. Em seguida, analisa o **p-valor**:

- Se o p-valor for **menor que 0,05**, concluímos que **existe pelo menos uma diferença significativa entre os grupos**.
- Se for **maior que 0,05**, não há evidência suficiente para afirmar que as médias são diferentes.

O Resultado do ANOVA para a média da diferenca_preco entre as categorias principais foi: p-valor = 4,49e-24

→ **Rejeita-se a hipótese nula**, indicando que há **diferenças estatisticamente significativas** na média de desconto entre as categorias.

Como o teste ANOVA apenas nos diz que existe uma diferença em algum lugar entre os grupos, mas não quais grupos são diferentes, o próximo passo natural seria realizar testes post-hoc, como o teste de Tukey HSD (Honestly Significant Difference). Isso permite comparar pares de categorias principais para identificar especificamente quais pares têm médias de diferenca_preco estatisticamente diferentes.

Análise Post-Hoc – Teste de Tukey HSD

O **Teste de Tukey HSD** (Honestly Significant Difference) é uma **análise post-hoc**, ou seja, é realizada **após o teste ANOVA**, quando já se sabe que há **diferenças significativas entre grupos**. Nesse caso, foi realizado para identificar **quais pares de categorias** apresentam diferenças estatisticamente significativas na média do desconto.

Esse teste compara todas as **possíveis combinações de pares de grupos, c**alcula a **diferença média entre os pares**, um **intervalo de confiança** e um **p-valor ajustado**. Se o **p-valor ajustado** for menor que 0,05 e o intervalo **não incluir zero**, a diferença entre os dois grupos é considerada **estatisticamente significativa**.

Foram identificados diversos pares de categorias com **diferenças significativas**. Por exemplo:

- *Computers & Accessories* oferece, em média, **₹753,27** a mais de desconto do que *Car & Motorbike* (p-adj < 0,0001).
- Outros pares, como *Electronics vs Home & Kitchen*, mesmo com diferenças observadas visualmente, **não apresentaram significância estatística** (p-adj = 1.0000).

A **Hipótese 4 foi confirmada**: existe uma forte relação positiva entre o preço real do produto e o valor absoluto do desconto. No entanto, essa relação **varia entre as categorias de produto**. Categorias como **Electronics** e **Home & Kitchen** tendem a oferecer os **maiores descontos absolutos médios**, enquanto outras, como *Car & Motorbike*, oferecem menos.

Essas análises mostram que, embora o preço real seja um fator chave na definição do valor do desconto, a **categoria do produto influencia significativamente o quanto, em média, é descontado**. Isso reforça a importância de considerar o contexto do produto ao interpretar políticas de precificação e promoção.

▼ Cálculo Risco Relativo

Para investigar possíveis associações entre a categoria do produto e a probabilidade de receber uma alta avaliação, foi calculado o **Risco Relativo (RR)**. Essa medida compara a probabilidade de um evento ocorrer em dois grupos distintos, neste caso, produtos das categorias **Electronics** e **Home&Kitchen,** permitindo identificar se pertencer a uma determinada categoria está associado a uma maior ou menor chance de obter boas avaliações.

Tabela de Frequências:

Categoria Principal	Baixa Avaliação	Alta Avaliação
Electronics	129	367
Home&Kitchen	144	303

- **Probabilidade de alta avaliação em Electronics:** 0,7399
- **Probabilidade de alta avaliação em Home&Kitchen:** 0,6779
- **Risco Relativo (RR):** 1,0916

O valor de **RR = 1,0916** indica que produtos da categoria **Electronics têm aproximadamente 1,09 vezes mais chance** de receber uma alta avaliação em comparação com produtos da categoria **Home&Kitchen**. Isso equivale a um **aumento de 9%** na probabilidade de avaliação alta para a categoria Electronics.

- Quando **RR = 1**, não há diferença entre os grupos comparados.
- Quando **RR > 1**, o grupo exposto (neste caso, Electronics) tem maior probabilidade do evento ocorrer.
- Quando **RR < 1**, o grupo exposto tem menor probabilidade.

Portanto, embora exista uma associação positiva entre a categoria *Electronics* e a alta avaliação, a **diferença observada é pequena** (9%) e a **magnitude da associação é fraca**.

É importante ressaltar que o Risco Relativo **não implica causalidade**. Ou seja, o fato de um produto pertencer à categoria *Electronics* **não causa** diretamente avaliações mais altas. A associação observada pode estar relacionada a outras variáveis não controladas, como qualidade do produto, marca, tipo de consumidor ou estratégias de marketing.

Para confirmar se essa diferença é estatisticamente significativa, e não resultado do acaso, realizamos um teste de significância apropriado, como o **teste qui-quadrado para tabelas de contingência**.

Teste Qui-quadrado para Significância da Associação entre Categoria e Alta Avaliação

Após o cálculo do Risco Relativo (RR = 1,09) entre as categorias *Electronics* e *Home&Kitchen*, foi realizado um teste Qui-quadrado com o objetivo de verificar se essa diferença na probabilidade de obter alta avaliação é **estatisticamente significativa** ou se pode ter ocorrido por acaso.

O teste compara as frequências observadas na tabela com as frequências esperadas sob a suposição de que não há associação entre as variáveis (categoria e avaliação). Se houver uma diferença significativa entre esses valores, o teste retorna um p-valor baixo, indicando a existência de uma associação.

Tabela de Contingência (Frequências Esperadas):

Categoria Principal	Baixa Avaliação	Alta Avaliação
Electronics	143,6	352,4
Home&Kitchen	129,4	317,6

Resultados do Teste:

- **Estatística Qui-quadrado:** 4,1068
- **P-valor:** 0,0427
- **Graus de liberdade:** 1

O valor do p-valor (**0,0427**) é **menor que o nível de significância de 0,05**, o que nos leva a **rejeitar a hipótese nula**. Isso significa que há uma **associação estatisticamente significativa** entre a categoria do produto (*Electronics* vs *Home&Kitchen*) e a chance de receber uma **alta avaliação**.

Esse resultado reforça a análise anterior baseada no **Risco Relativo (RR = 1,09)**, indicando que a **diferença observada não ocorreu por acaso**. Portanto, produtos da categoria *Electronics* têm, de fato, uma chance ligeiramente maior (9%) de obter uma alta avaliação em comparação com produtos de *Home&Kitchen*, e essa diferença é **estatisticamente confirmada.**

▼ Análise Comparativa do Risco Relativo entre Demais Categorias de Produto

Além da comparação entre *Electronics* e *Home&Kitchen*, o cálculo do **Risco Relativo (RR)** foi estendido para outras categorias com o objetivo de investigar a associação entre a **categoria principal** e a **probabilidade de um produto receber alta avaliação (rating ≥ 4.0)**.

As comparações a seguir indicam a **probabilidade de alta avaliação** em cada par de categorias e o Risco Relativo correspondente:

- **Electronics vs Computers&Accessories**
  - RR = **0,90**
  - Produtos de *Electronics* têm aproximadamente **10% menor probabilidade** de receber alta avaliação em comparação a *Computers&Accessories*.
- **Home&Kitchen vs Computers&Accessories**
  - RR = **0,82**
  - Produtos de *Home&Kitchen* têm aproximadamente **18% menor probabilidade** de obter alta avaliação em comparação a *Computers&Accessories*.
- **Electronics vs OfficeProducts**

- RR = **0,74**
- Produtos da categoria *Electronics* têm aproximadamente **26% menor probabilidade** de obter alta avaliação em relação a *OfficeProducts*.

Esses resultados indicam que, neste conjunto de dados, as categorias *Computers&Accessories* e *OfficeProducts* apresentam uma maior frequência de avaliações altas quando comparadas a *Electronics* e *Home&Kitchen*.

Validação com Teste Qui-quadrado

Para verificar se essas diferenças observadas nos Riscos Relativos são **estatisticamente significativas** ou fruto do acaso, foram realizados **testes de Qui-quadrado** para cada par de categorias:

Comparação	Estatística χ^2	p-valor	Conclusão
Electronics vs Computers&Accessories	7,8790	0,0050	Associação significativa entre a categoria e alta avaliação.
Home&Kitchen vs Computers&Accessories	21,6427	0,0000	Associação altamente significativa entre a categoria e alta avaliação.
Electronics vs OfficeProducts	9,3152	0,0023	Associação significativa entre a categoria e alta avaliação.

Em todos os casos, os **p-valores foram inferiores a 0,05**, o que permite **rejeitar a hipótese nula** e concluir que há **associação estatisticamente significativa** entre a categoria do produto e a probabilidade de receber uma alta avaliação.

Conclusão:

Os resultados mostram que a **categoria do produto está associada à chance de alta avaliação** neste dataset. Categorias como *Computers&Accessories* e *OfficeProducts* apresentaram **desempenho superior em avaliações** quando comparadas a *Electronics* e *Home&Kitchen*, tanto em termos de Risco Relativo quanto em significância estatística.

Esses achados reforçam a importância de considerar a categoria do produto ao interpretar métricas de avaliação, além de alertar para o potencial viés ao comparar avaliações entre segmentos distintos. Recomenda-se estender essa análise a outras categorias e realizar testes post-hoc, se aplicável, para investigações mais aprofundadas.

🔶 Resultados e Insights

A análise revelou que **produtos com preços reais mais altos tendem a receber descontos monetários maiores**, o que confirma a Hipótese 4 com alta correlação (Pearson = 0.9108; Spearman = 0.8952). Em contrapartida, as hipóteses 1, 2 e 3 não foram fortemente sustentadas pelos dados.

A **Hipótese 1** (maior desconto leva a melhor avaliação) foi refutada, mostrando uma correlação fraca e negativa. A **Hipótese 2**, que relaciona número de avaliações com nota, indicou uma associação positiva fraca, mas estatisticamente significativa. Já a **Hipótese 3**, que sugeria que produtos mais caros teriam melhores notas, foi descartada por ausência de relação clara.

Na análise por categorias, o **Risco Relativo (RR)** indicou que produtos de *Electronics* têm 1,09 vezes mais chance de alta avaliação do que *Home&Kitchen*. No entanto, frente a *Computers&Accessories* (RR = 0.90) e *OfficeProducts* (RR = 0.74), *Electronics* mostrou menor probabilidade de avaliações altas. Todos esses resultados foram confirmados com **testes de Qui-quadrado**, indicando associações estatisticamente significativas.

Esses achados mostram que a **categoria do produto impacta mais as avaliações do que preço ou desconto isoladamente**, com destaque para *Computers&Accessories* e *OfficeProducts*, que apresentam melhor desempenho em notas altas.

🔶 Recomendações Estratégicas

Com base nos resultados, recomenda-se:

- **Não depender apenas de descontos** para influenciar avaliações. A percepção de qualidade e valor é mais determinante.
- **Melhorar o posicionamento e apresentação de produtos** em categorias com menor avaliação média, como *Electronics* e *Home&Kitchen*.
- **Estudar boas práticas das categorias mais bem avaliadas**, como *Computers&Accessories* e *OfficeProducts*, para replicar estratégias bem-sucedidas.
- **Personalizar ações por categoria**, adotando campanhas específicas conforme o comportamento dos consumidores.
- **Implementar testes A/B e monitoramento contínuo** de avaliações por categoria, permitindo ajustes estratégicos baseados em dados reais.

🔶 Responsáveis pelo Projeto

Nome: Aline Dionizio
Função: Analista de Dados
Contato: LinkedIn | Email

Nome: Taiza Ferreira
Função: Analista de Dados
Contato: LinkedIn | Email

Nome: Giullia Braga
Função: Analista de Dados
Contato: LinkedIn | Email

🔶 Links Úteis

- Apresentação
- Ficha Técnica - Códigos
- Google Colab
- Relatório