

# Relatório

## Análise de dados DataLab Amazon Sales

**Preparado por:**

Aline Dionizio | Giulia Braga | Taiza Ferreira

## 1. Introdução

### Objetivo:

Este relatório apresenta uma análise exploratória e estatística dos dados de produtos e avaliações da Amazon, com o objetivo de identificar padrões, relações e insights relevantes para entender o comportamento dos produtos e as percepções dos clientes. O conjunto de dados utilizado compreende informações detalhadas sobre produtos (preços, descontos, categorias) e avaliações (pontuação, contagem de avaliações).

## 2. 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.

## 3. Preparação e Limpeza de Dados

As seguintes etapas foram realizadas para limpar e preparar os dados para análise:

- **Carregamento dos Dados:** Os datasets `amazon_product.csv` e `amazon_review.csv` foram carregados em DataFrames pandas.
- 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:
  1. `review_df` **1465 linhas e 10 colunas**
  2. `product_df` **1469 linhas e 7 colunas**

### 3.1. Identificação e tratamento de valores Nulos

Valores nulos foram identificados nas colunas `'about_product'` do dataframe de produtos e `'rating_count'`, `'img_link'` e `'product_link'` do dataframe de avaliações. Linhas com nulos em `'about_product'` e `'rating_count'` foram removidas. As colunas `'img_link'` e `'product_link'` foram removidas devido ao alto número de valores ausentes e sua menor relevância para os objetivos iniciais da análise. Após a união dos dataframes, um valor nulo remanescente na coluna `'rating'` também foi removido.

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

### 3.2. Identificação e tratamento de valores Duplicados

Linhas duplicadas foram removidas com base em 'product\_id' no dataframe de produtos e em combinações de 'product\_id', 'user\_id' e 'review\_id' no dataframe de avaliações.

**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)

### 3.3. Identificação e tratamento de dados fora do escopo

Nesta etapa, focamos em alinhar os dados aos objetivos do projeto, identificando valores que não seriam relevantes ou poderiam causar distorções. Uma análise exploratória, incluindo estatísticas descritivas e verificação de valores únicos, foi realizada. Identificamos que as colunas product\_link e img\_link continham muitos valores nulos e não eram essenciais para a análise final. Optamos por excluí-las. Após essa remoção, confirmamos que não havia outros valores nulos remanescentes no dataset, garantindo que os tratamentos de dados anteriores foram eficazes.

### 3.4. Verificação e alteração dos tipos de dados e tratamento de dados discrepantes em variáveis categóricas e numéricas.

Colunas com informações numéricas contendo símbolos (como '₹' e '%') e vírgulas foram limpas e convertidas para tipos numéricos (float64 ou int64) para permitir cálculos e análises quantitativas. A coluna 'rating' também foi convertida para numérico.

### 3.5. Criação de novas variáveis

Duas novas variáveis foram criadas no dataframe de produtos:

- `diferenca_preco`: Calculada como `actual_price - discounted_price`, representando a economia absoluta em R\$.
- `categoria_principal`: Extraída da coluna 'category', representando o nível mais alto da hierarquia de categorias.

### 3.6. União dos DataFrames

Os dataframes de produtos e avaliações foram unidos usando a coluna 'product\_id' como chave, resultando no dataframe `unificada_df`.

Após estas etapas, o dataframe `unificada_df` está pronto para a análise exploratória e estatística.

## 4. Análise Exploratória de Dados (AED)

A Análise Exploratória de Dados foi conduzida para entender a distribuição e as características das variáveis no conjunto de dados unido.

### Estatísticas Descritivas

As medidas de tendência central (média, mediana, moda) e dispersão (desvio padrão, variância, IQR) foram calculadas para as variáveis numéricas relevantes: `discounted_price`, `actual_price`, `discount_percentage`, `diferenca_preco`, `rating` e `rating_count`.

## 4.1. Medidas de Tendência Central

Variáveis	Média	Mediana	Moda
discounted_price	3294.57	899.00	[2.99]
actual_price	5687.36	1795.00	[999.00]
discount_percentage	4684	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]

### Interpretação das Medidas de Tendência Central:

A comparação entre a média e a mediana revela a assimetria nas distribuições de discounted\_price, actual\_price, diferenca\_preco e rating\_count, indicando a presença de valores extremos (outliers) que puxam a média. A mediana, sendo menos sensível a outliers, fornece uma medida mais robusta do valor típico para essas variáveis. A moda identifica os valores mais frequentes. Para rating e discount\_percentage, a média e a mediana são mais próximas, sugerindo distribuições mais simétricas.

## 4.2. Medidas de Dispersão

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	4688.53	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

### Interpretação das Medidas de Dispersão:

As medidas de dispersão quantificam a variabilidade dos dados. Um alto desvio padrão e variância indicam que os dados estão mais espalhados (como em `rating_count` e `diferenca_preco`), enquanto valores baixos sugerem maior concentração dos dados (como em `rating`). O Intervalo Interquartil (IQR) fornece uma medida robusta da dispersão dos 50% centrais dos dados, sendo menos afetado por outliers do que o desvio padrão e a variância. Essas medidas complementam as medidas de tendência central, oferecendo uma visão mais completa da distribuição das variáveis.

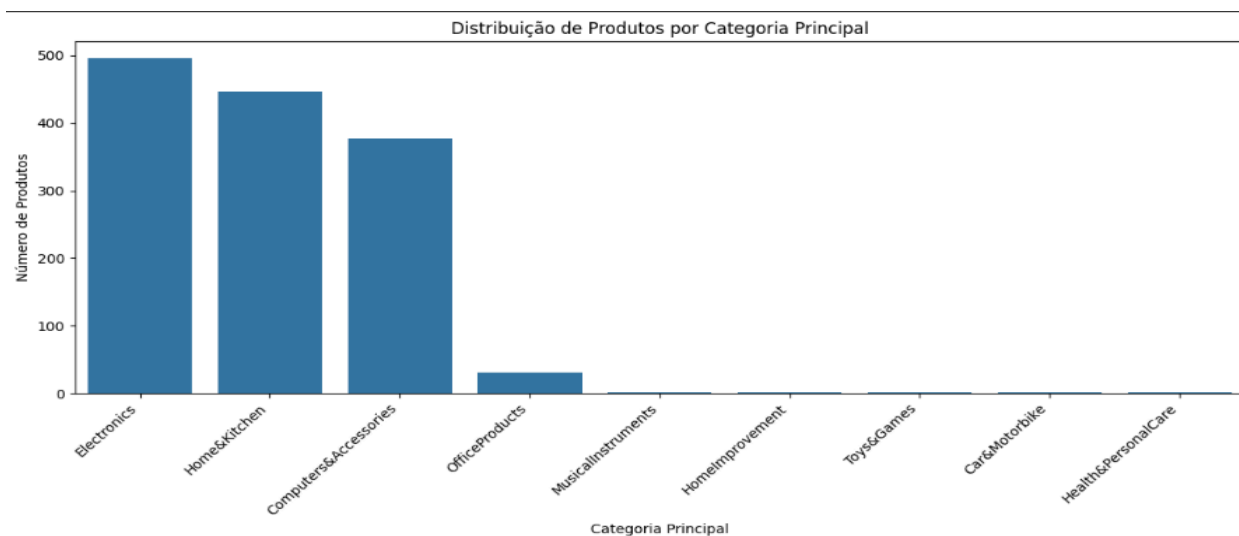
## 4.3. Agrupar e ver dados de acordo com as variáveis categóricas

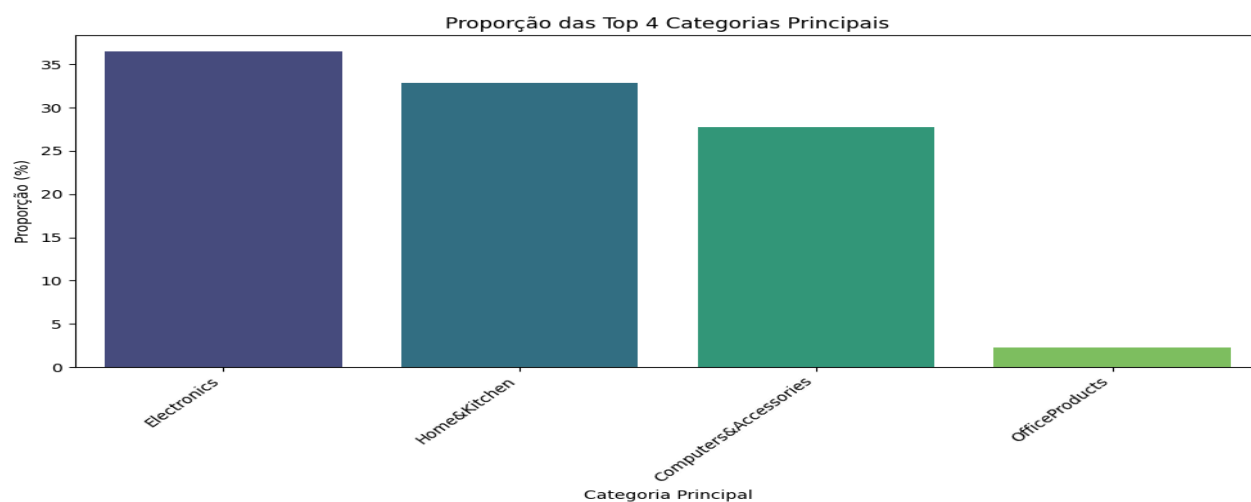
A distribuição dos produtos pelas `categoria_principal` e `category` (subcategorias) foi analisada para entender a composição do catálogo no dataset.

### Interpretação da Distribuição por Categoria Principal:

O gráfico de barras e a tabela de proporção mostram que o dataset é dominado por algumas categorias principais, como Eletrônicos, Casa&Cozinha e Computadores&Acessórios. Essas categorias representam a vasta maioria dos produtos no conjunto de dados, enquanto outras categorias têm uma representação muito menor. Essa concentração em poucas categorias é um fator importante a ser considerado em análises mais detalhadas e na generalização de resultados.

### Visualizações:





#### Proporção de cada Categoria Principal:

categoria_principal	
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

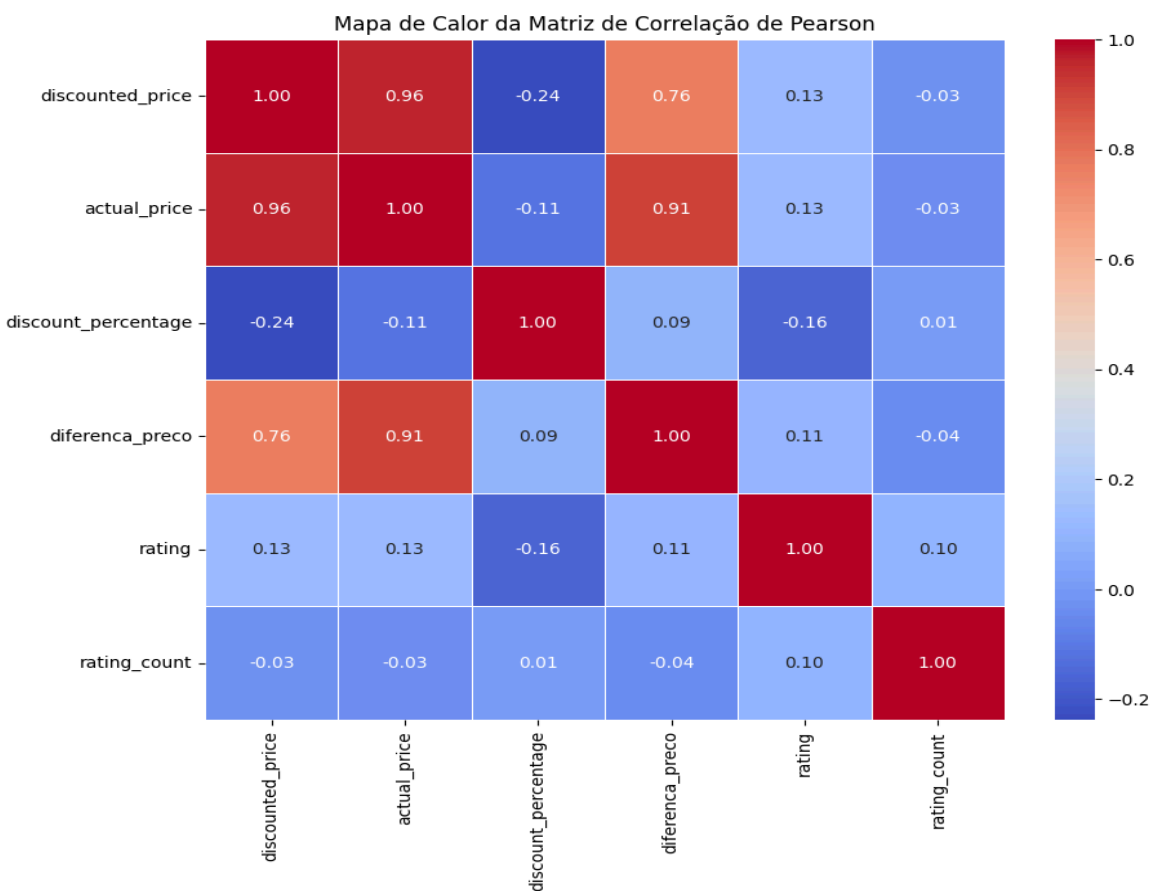
## 4.4. Correlação entre variáveis

A correlação entre as variáveis numéricas foi calculada usando o método de Pearson para identificar a força e a direção das relações lineares entre elas.

### Matriz de Correlação de Pearson entre as variáveis numéricas:

	discounted price	actual price	discount percentage	diferenca preco	rating	rating count
discounted price	1.000000	0.962103	-0.237639	0.763728	0.127307	-0.025049
actual price	0.962103	1.000000	-0.112198	0.910814	0.128014	-0.034842
discount percentage	-0.237639	-0.112198	1.000000	0.094150	-0.162386	0.007621
diferenca preco	0.763728	0.0910814	0.094150	1.000000	0.110324	-0.044561
rating	0.127307	0.128014	-0.162386	0.110324	1.000000	0.098156
rating count	-0.025049	-0.034842	0.007621	-0.044561	0.098156	1.000000

### Mapa de calor da Matriz de correlação de Pearson





### Interpretação da Análise de Correlação:

O mapa de calor e a matriz de correlação revelam as relações entre as variáveis numéricas:

- **Fortes Correlações Positivas:** `discounted_price`, `actual_price` e `diferenca_preco` apresentam fortes correlações positivas entre si. Isso é esperado, pois a diferença de preço é derivada dos preços original e com desconto, e produtos com preços originais mais altos tendem a ter preços com desconto e diferenças de preço maiores.
- **Correlações Fracas:** Variáveis como `rating` e `rating_count` mostram correlações fracas com as variáveis de preço e desconto. Isso sugere que a pontuação e o volume de avaliações não estão fortemente ligados linearmente aos aspectos de preço e desconto neste dataset.
- **Correlação Negativa Fraca a Moderada:** Há uma correlação negativa fraca a moderada entre `discount_percentage` e os preços (`discounted_price` e `actual_price`). Isso indica que produtos com preços mais altos tendem a ter percentuais de desconto ligeiramente menores, ou vice-versa.

As fortes correlações entre as variáveis de preço indicam multicolinearidade potencial em modelos preditivos, sugerindo que nem todas essas variáveis devem ser usadas simultaneamente como preditoras. As correlações fracas envolvendo `rating` e `rating_count` indicam que outros fatores além de preço e desconto podem ser mais influentes na avaliação dos produtos.

## 5. Validação de Hipóteses

Para investigar relações específicas nos dados, quatro hipóteses foram formuladas e testadas estatisticamente.

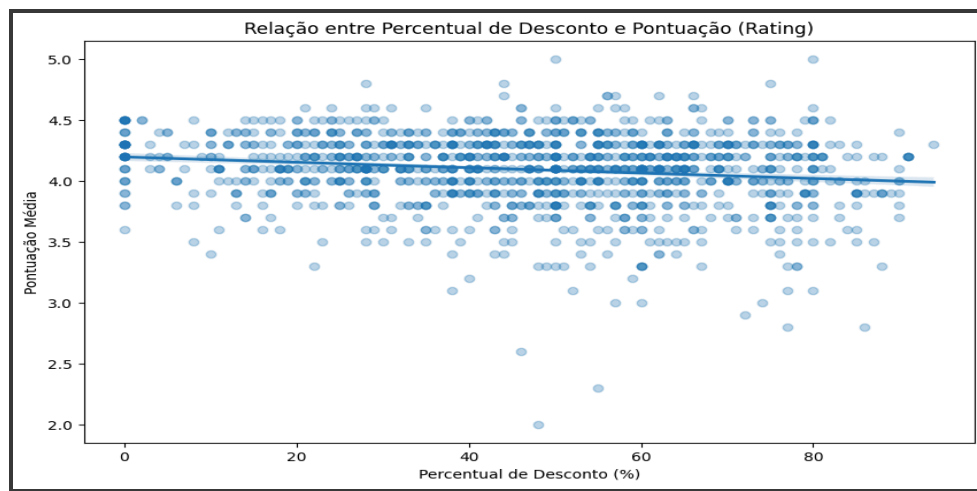
### Hipótese 1: Quanto maior o desconto, melhor será a pontuação?

Análise da relação entre o percentual de desconto (`discount_percentage`) e a pontuação do produto (`rating`).

#### Interpretação da Hipótese 1:

A correlação de Spearman (-0.1512) e o p-valor (0.0000) indicam uma correlação negativa muito fraca e estatisticamente significativa entre o percentual de desconto e a pontuação do produto. Embora a relação seja estatisticamente suportada, sua magnitude é pequena, sugerindo que um percentual de desconto maior não está associado a avaliações significativamente melhores neste dataset.

### Visualização:



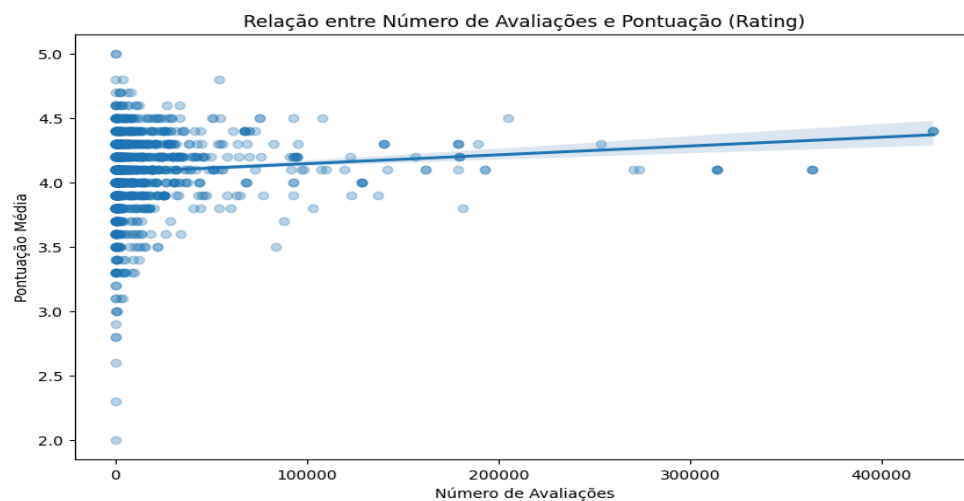
### Hipótese 2: Quanto maior o número de pessoas que avaliaram o produto, melhor será a classificação?

Análise da relação entre a contagem de avaliações (rating\_count) e a pontuação do produto (rating).

#### Interpretação da Hipótese 2:

As correlações de Pearson (0.0982) e Spearman (0.1902), ambas com p-valores próximos de 0.0000, indicam uma associação positiva fraca, mas estatisticamente significativa, entre o número de avaliações e a pontuação. Produtos com mais avaliações tendem a ter pontuações ligeiramente maiores, mas essa relação é fraca e não um fator determinante na classificação do produto.

### Visualização:



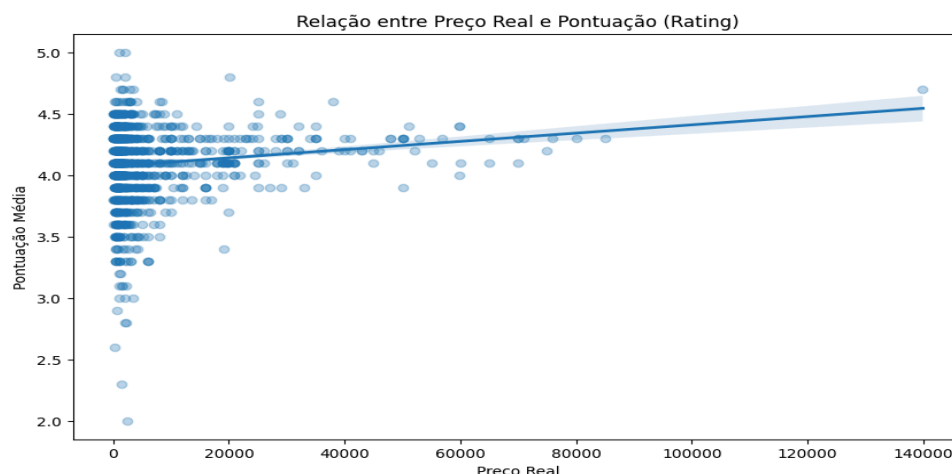
### Hipótese 3: Produtos com preços reais mais altos (sem desconto aplicado) tendem a ter uma avaliação maior?

Análise da relação entre o preço real (actual\_price) e a pontuação do produto (rating).

#### Interpretação da Hipótese 3:

A Correlação de Pearson (0.1280) sugere uma associação positiva fraca e estatisticamente significativa ( $p < 0.001$ ), mas a Correlação de Spearman (0.0325) não é estatisticamente significativa ( $p > 0.05$ ). A evidência para uma relação linear forte entre preço real e pontuação é fraca e inconsistente neste dataset. O preço real, por si só, não parece ser um forte preditor da avaliação do produto.

#### Visualização:



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

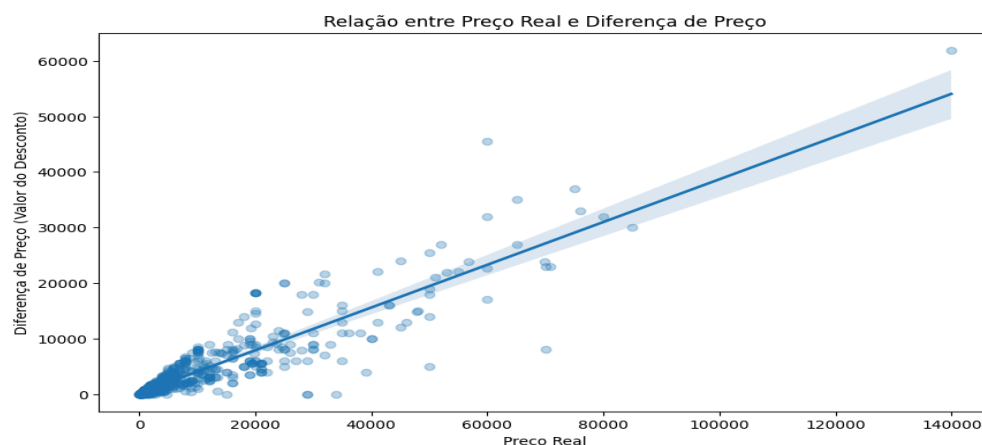
Análise da relação entre o preço real (actual\_price) e a diferença de preço (diferenca\_preco).

#### Interpretação da Hipótese 4:

As correlações de Pearson (0.9108) e Spearman (0.8952), ambas com valores próximos de 0.0000, indicam uma associação positiva muito forte e estatisticamente significativa entre o preço real e a diferença de preço. Isso confirma que produtos com preços originais mais altos tendem a ter descontos absolutos (em valor monetário) muito maiores.

Para complementar a análise da Hipótese 4, investigamos se a média da diferença de preço varia significativamente entre as categorias principais.

## Visualização:



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.

### Teste ANOVA e interpretação para Diferença de Preço por Categoria Principal

O p-valor extremamente baixo (4.489483e-24, essencialmente 0) confirma que há uma diferença estatisticamente significativa na média da diferença de preço entre pelo menos uma das categorias principais. Isso significa que a magnitude dos descontos em termos monetários varia dependendo da categoria do produto.

Para identificar quais pares de categorias têm médias de diferença de preço significativamente diferentes, realizamos um teste post-hoc de Tukey HSD.

### Teste 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, calcula 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:

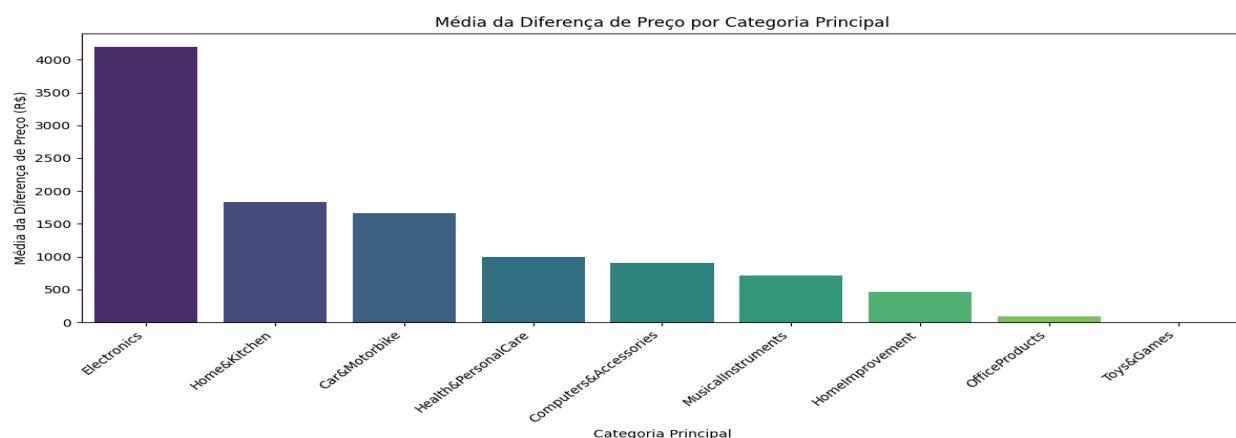
- *Computers & Accessories* oferece, em média, ₹753,27 a mais de desconto do que *Car & Motorbike* (p-adj < 0,0001).
- Outros pares, como *Electronics* vs *Home & Kitchen*, mesmo com diferenças observadas visualmente, não apresentaram significância estatística (p-adj = 1.0000).

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.

Para visualizar essa variação, geramos um gráfico de barras da média da diferença de preço por categoria principal.

### Média da Diferença de Preço por Categoria Principal:



### Interpretação da Média da Diferença de Preço por Categoria Principal:

O gráfico de barras visualmente suporta os resultados do ANOVA e Tukey HSD, mostrando claramente que a média da diferença de preço varia consideravelmente entre as categorias principais. Categorias como Eletrônicos e Casa&Cozinha tendem a ter descontos absolutos médios maiores, enquanto outras, como Office Products, têm descontos absolutos médios menores.

## 6. Análise de Risco Relativo

Para avaliar a probabilidade de um evento (ter uma alta avaliação, definida como rating  $\geq 4.0$ ) ocorrer em diferentes categorias principais, calculamos o Risco Relativo (RR) para pares de categorias relevantes e utilizamos o teste Qui-quadrado para avaliar a significância estatística dessa associação.

Primeiro, criamos uma variável binária alta\_avaliacao e focamos na comparação entre as categorias "Electronics" e "Home&Kitchen".

Cálculo do Risco Relativo e Teste Qui-quadrado: Electronics vs Home&Kitchen

### Tabela de Frequências:

Categoria Principal	Baixa Avaliação	Alta Avaliação
Eletrônicos	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
Eletronics	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 ( $\text{rating} \geq 4.0$ ).

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

Resultados e Interpretação para Pares Adicionais:

- Electronics vs Computers&Accessories:
  - Risco Relativo: 0.8998
  - P-valor do Teste Qui-quadrado: 0.0050
  - Interpretação: Há uma associação estatisticamente significativa ( $p < 0.05$ ). Produtos de Electronics têm aproximadamente 0.90 vezes a probabilidade de ter uma alta avaliação em comparação com Computers&Accessories (ou seja, menor probabilidade).
  -
- Home&Kitchen vs Computers&Accessories:
  - Risco Relativo: 0.8244
  - P-valor do Teste Qui-quadrado: 0.0000
  - Interpretação: Há uma associação estatisticamente significativa ( $p < 0.05$ ). Produtos de Home&Kitchen têm aproximadamente 0.82 vezes a probabilidade de ter uma alta avaliação em comparação com Computers&Accessories (ou seja, menor probabilidade).
  -
- Electronics vs OfficeProducts:
  - Risco Relativo: 0.7399
  - P-valor do Teste Qui-quadrado: 0.0023
  - Interpretação: Há uma associação estatisticamente significativa ( $p < 0.05$ ). Produtos de Electronics têm aproximadamente 0.74 vezes a probabilidade de ter uma alta avaliação em comparação com Office Products (ou seja, menor probabilidade).

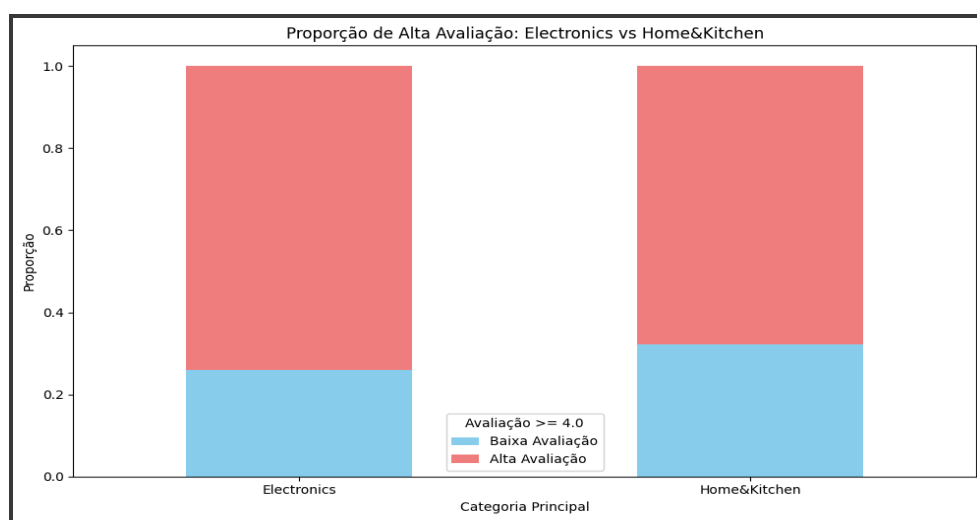


Em todos os pares analisados, os testes Qui-quadrado confirmaram que as diferenças observadas na proporção de altas avaliações são estatisticamente significativas.

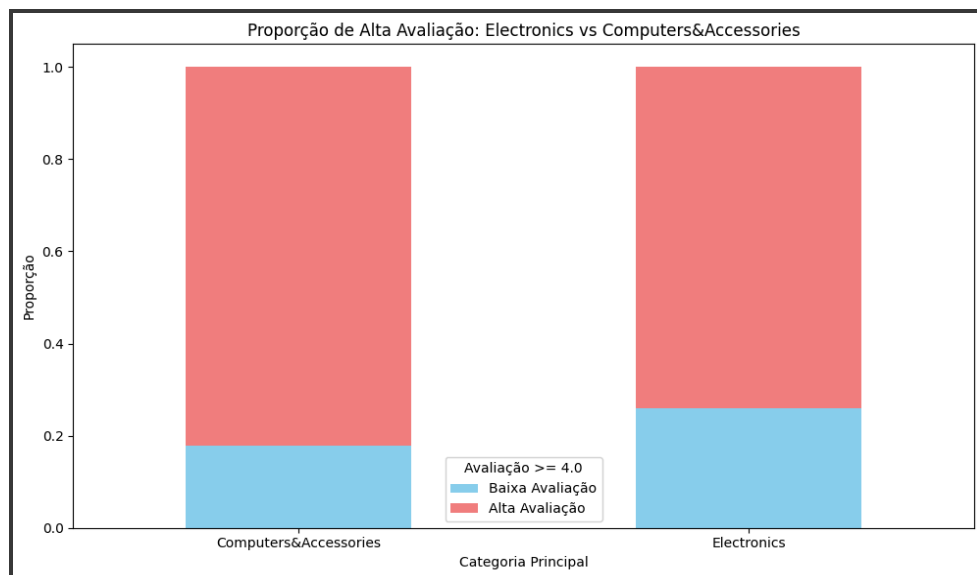
Para visualizar a proporção de alta avaliação em cada um desses pares, geramos gráficos de barras empilhadas.

## Visualização da Proporção de Alta Avaliação por Categoria Principal (Pares Comparados)

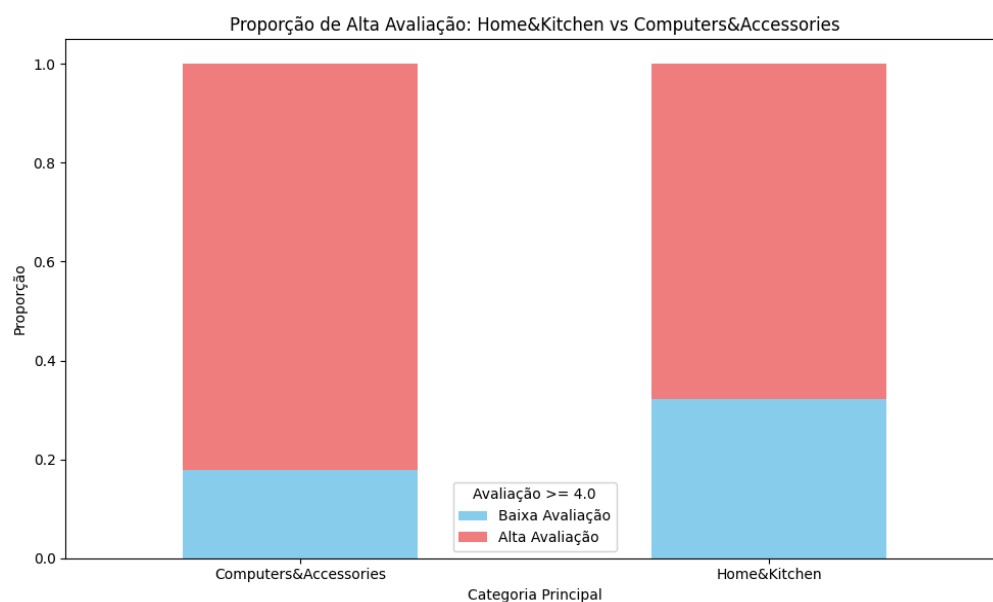
### 1 - Eletrônicos vs Home&Kitchen



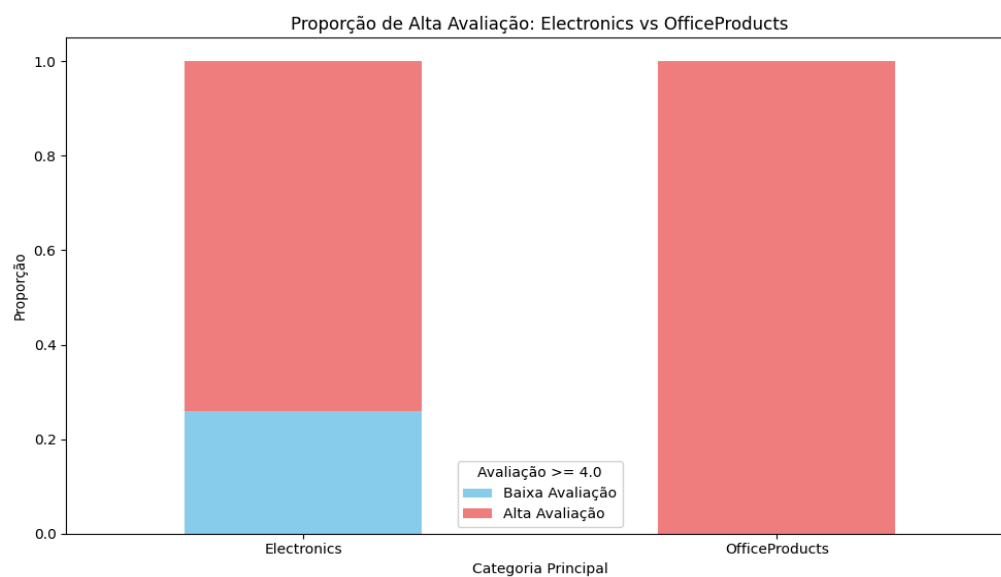
### 2 - Eletrônicos vs Computers&Accessories



### 3 - Home&Kitchen vs Computers&Accessories



### 4 - Electronics vs OfficeProducts



## 7. 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.

## 8. 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.

## 9. Recomendações

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.