# Projeto: Amazon Sales

Neste projeto foi aplicado a segmentação de produto , e a validação de hipóteses. Ambas metodologias são fundamentais na tomada de decisões, visto que são baseadas em evidências e não apenas em crenças ou opiniões. Na análise de dados, é comum formular suposições ou hipóteses sobre relações, tendências ou diferenças entre as variáveis dos dados disponíveis. A validação dessas hipóteses (confirmar ou refutar) é alcançada com técnicas e métodos projetados para determinar se os resultados observados nos dados são estatisticamente significativos ou se podem ser atribuídos ao acaso. Neste projeto, exploramos este aspecto da análise de dados, destacando como ele ajuda a melhorar a compreensão dos fenômenos, apoiar a pesquisa e tomar decisões informadas.

# Contexto

Entender o comportamento de compra do seu cliente é um tesouro para qualquer empresa. Evita o desperdício de tempo e dinheiro. Por isso, cada vez mais, as organizações estão em busca de ferramentas que ajudem a prever o comportamento de compra do seu cliente. Pensando nisso, o objetivo desse projeto é **segmentar os produtos por categoria**, identificar os **produtos mais avaliados e os menos avaliados**, a fim de auxiliar na tomada de decisão do gerente de vendas. Esta categorização permitirá diferenciar:

- Quais são os produtos com maior preço de venda ?
- Quais são as categorias com maior faturamento ?
- Qual o ticket médio por categoria?
- Qual a média de desconto por categoria?
- Qual o produto com maior desconto?
- Quais são as categorias mais e menos avaliadas?
- Qual a melhor categoria no ranking?

Validar hipótese:

- Quanto maior o desconto, melhor será a pontuação?
- Quanto maior o número de pessoas que avaliaram o produto, melhor será a classificação?

Esta informação oferece o atendimento da demanda de clientes, procurando disponibilizar o produto na prateleira, sem que haja excesso.

**Projeto Individual:** Nathalia Guimarães

## Ferramentas e Tecnologias

1. Bigquery
2. Power BI
3. Canva Apresentações
4. Canva Vídeo
5. GitHub

### Linguagens

1. linguagem SQL no BigQuery

### Links Ferramentas Aplicadas no Projeto

Link Apresentação do Projeto > Projeto Amazon
Link Relatório de Resultado Power BI > Relatório Amazon

# Processamento e Análises

- **Banco de dados**: O conjunto de dados do projeto estava dividido em duas tabelas. A tabela Amazon product e a tabela Amazon review. A tabela Amazon Product contém informações: ID produto, nome do produto, categoria do produto, os preços do produto com desconto e sem desconto, o percentual de desconto aplicado ao produto e a descrição do produto. Na tabela Amazon review informações como ID do usuário que avaliou o produto, nome do usuário que escreveu a avaliação, tinha avaliação breve e completa do usuário sobre o produto, link com imagem, link com informações do produto, ID produto, classificação do produto no ranking, e número de usuários que avaliaram o produto.
- **Limpeza dos dados**: Após o upload do conjunto de dados na Bigquery. Realizada união das tabelas para agrupar os dados Amazon product e Amazon review. A união ocorreu a partir de uma coluna em comum de ambas as tabelas ID do produto. O código SQL aplicado para Unir as tabelas:

```
SELECT P.*,A.user_id,A.user_name,A.review_id, A.review_title, A.review_content, A.img_link, A.product_link, A.rating, A.rating_count FROM `amazon-430517.dataset.Tabela_Amazon_Produto` AS P JOIN `amazon-430517.dataset.Tabela_Amazon_Avaliacao` AS A ON P.product_id = A.product_id;
```

1. Após unir as tabelas foi verificado o número maior de linhas resultando no total de 1741. As tabelas originais tinham as seguintes quantidades de linhas: Avaliação 1465 e Produto 1469. Foram identificados 486 Product_id duplicados (Lista de Product_ID Duplicados). Diante disso foi realizado essa consulta através do código abaixo retorna os product_id sem duplicidade. Resultado = Tabela_Unidas com total de 1255 linhas.

```
WITH avaliacao_semduplicadas AS ( SELECT * FROM `amazon-430517.dataset.Tabela_Amazon_Avaliacao` WHERE product_id NOT IN ( SELECT product_id FROM `amazon-430517.dataset.Tabela_Amazon_Avaliacao` GROUP BY product_id HAVING COUNT(*) > 1 ) ), produto_semduplicados AS ( SELECT * FROM `amazon-430517.dataset.Tabela_Amazon_Produto` WHERE product_id NOT IN ( SELECT
```

product_id FROM `amazon-430517.dataset.Tabela_Amazon_Produto` GROUP BY product_id HAVING COUNT(*) > 1 ) ) SELECT P.*, A.user_id, A.user_name, A.review_id, A.review_title, A.review_content, A.img_link, A.product_link, A.rating, A.rating_count FROM produto_semduplicados AS P JOIN avaliacao_semduplicadas AS A ON P.product_id = A.product_id;

2. Após retirar os ID duplicados foi realizada a consulta abaixo para verificar se ainda constava algum ID duplicado.

SELECT product_id AS OCORRENCIAS FROM `amazon-430517.dataset.Tabelas_Unidas` GROUP BY product_id HAVING count(*) > 1

3. Para verificar as células nulas nas tabelas. As colunas Img-link e Product_link apresentaram 456 células nulas, e a rating_count apresentou 2 células vazias, foi decidido retirar da análise apenas as duas células nulas da coluna rating count, devido ser um indicador relevante de análise, as demais permanecem. Abaixo a consulta realizada para retirar as células nulas.

SELECT * FROM `amazon-430517.dataset.Tabelas_Unidas` WHERE rating_count IS NOT NULL - células sem nulos

4. Retirada da célula com dados inconsistentes de análise, o Id produto 'B08L12N5H1' foi excluído do banco de dados por não apresentar posição no ranking de produto. Como esta é uma variável importante para a análise, o ID foi retirado. Abaixo a consulta para retirada da célula.

CREATE OR REPLACE TABLE `amazon-430517.dataset.Tabela_Unidas_Final`AS SELECT * FROM `amazon-430517.dataset.Tabela_Unidas_Final` WHERE product_id NOT IN ('B08L12N5H1');

5. Foi necessário usar a função FLOAT64 para considerar os números decimais e aplicar o quartil, posteriormente. A coluna estava como String e não conseguia aplicar a função Quartil. Abaixo consulta realizada para a alteração.

CREATE OR REPLACE TABLE `amazon-430517.dataset.Tabela_Unidas_Final` AS ( WITH ranking AS ( SELECT *, CAST(rating AS FLOAT64) AS rating_ FROM `amazon-430517.dataset.Tabela_Unidas_Final` ) SELECT product_id, product_name, category, discounted_price, actual_price, discount_percentage, about_product, user_id, user_name, review_id, review_title, review_content, img_link, product_link, ranking.rating_,rating_count FROM ranking )

- **Processamento dos dados**:

1. Após a alteração da célula foi possível Criar os Quartis, e as criar Categorias e Unir tabelas. O quartil foi calculado a partir da média da classificação e as categorias foram divididas em 50% para Melhor e Menos Avaliados. Abaixo consulta realizada:

```
CREATE OR REPLACE TABLE `amazon-430517.dataset.Tabela_Unidas_Final` AS ( WITH Avaliacao_Media AS ( SELECT product_id, AVG(CAST(rating_ AS FLOAT64)) AS media_avaliacao FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id ), Ranking_Quartil AS ( SELECT product_id, media_avaliacao, NTILE(4) OVER (ORDER BY media_avaliacao DESC) AS quartil_avaliacao FROM Avaliacao_Media ) SELECT f.*, r.media_avaliacao, r.quartil_avaliacao, CASE WHEN r.quartil_avaliacao = 1 THEN 'Melhor Avaliados' WHEN r.quartil_avaliacao = 2 THEN 'Melhor Avaliados' WHEN r.quartil_avaliacao = 3 THEN 'Menos Avaliados' WHEN r.quartil_avaliacao = 4 THEN 'Menos Avaliados' END AS categoria_avaliacao FROM Ranking_Quartil r JOIN `amazon-430517.dataset.Tabela_Unidas_Final` f ON r.product_id = f.product_id ORDER BY r.quartil_avaliacao, r.media_avaliacao DESC )
```

2. Criada tabela auxiliar traz informações sobre o produto com maior receita, de qual categoria ele faz parte e sua posição no ranking. Esta consulta ajuda a identificar qual produto está impulsionando o negócio. Avaliando a Categoria , olhar para a segmentação, permite identificar qual categoria dá mais lucro e direcionar esforços para impulsionar as vendas da categoria e similares. Quando olhamos para o ranking, a identificação do produto melhor avaliado nos permite comparar com demais produtos e identificar as vantagens competitivas deste produto. Ainda, a posição do produto de maior receita no ranking geral, indica a força do produto em relação aos demais.

```
WITH MaiorReceita AS ( SELECT product_id, discounted_price AS receita_total FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id,discounted_price ORDER BY receita_total DESC ), CategoriaProduto AS ( SELECT product_id, category AS categoria FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, category ), PontuacaoProduto AS ( SELECT product_id, rating_ AS pontuacao FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, rating_ ) SELECT m.product_id, m.receita_total, c.categoria, p.pontuacao FROM MaiorReceita m JOIN CategoriaProduto c ON m.product_id = c.product_id JOIN PontuacaoProduto p ON c.product_id = p.product_id
```

3. Tabela auxiliar para Identificar o Produto Maior Desconto e Pontuação. Nesta consulta o objetivo é verificar se o desconto tem relação direta com a pontuação.

```
WITH ProdutoMAiorDesconto AS ( SELECT product_id,discount_percentage AS desconto FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, discount_percentage ORDER BY desconto DESC ), MaiorPontuacao AS ( SELECT product_id,rating_ AS pontuacao FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, rating_ ) SELECT p.product_id, p.desconto, m.pontuacao FROM ProdutoMAiorDesconto p JOIN MaiorPontuacao m ON p.product_id = m.product_id
```

4. Tabela auxiliar identificar os produtos mais avaliados e Pontuação.Nesta consulta o objetivo é verificar se o produto com mais número de avaliação tem relação direta com a pontuação.

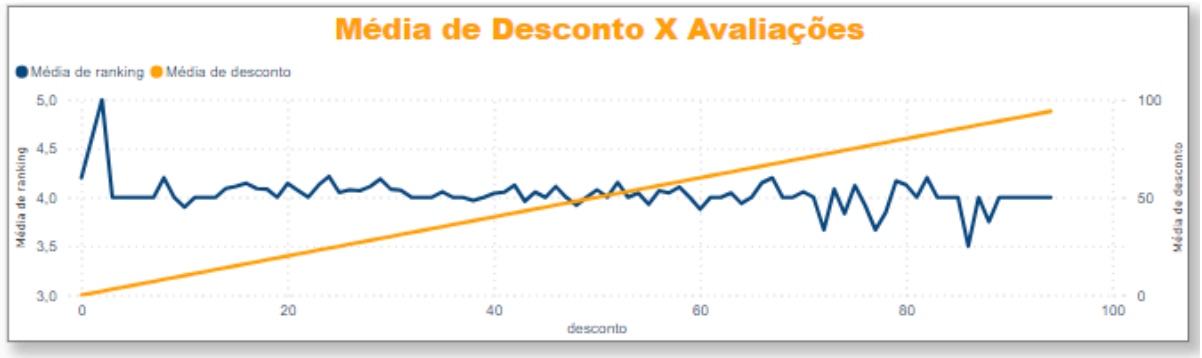
```
WITH ProdutoMaisAvaliados AS ( SELECT product_id,rating_count AS avaliados FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, rating_count ORDER BY avaliados DESC ), MaiorPontuacao AS ( SELECT product_id,rating_ AS pontuacao FROM `amazon-430517.dataset.Tabela_Unidas_Final` GROUP BY product_id, rating_ ) SELECT
```

p.product_id, p.avaliados, m.pontuacao FROM ProdutoMAisAvaliados p JOIN MaiorPontuacao m ON p.product_id = m.product_id

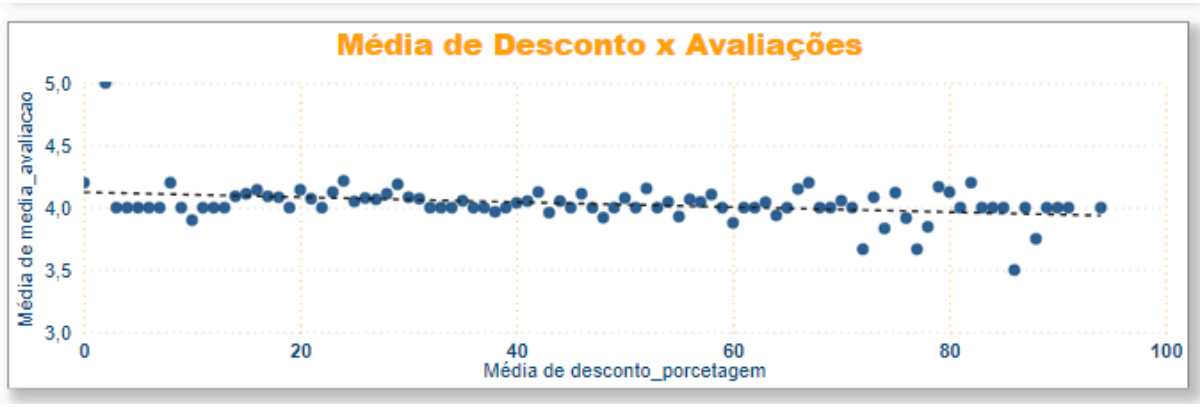
# Hipóteses à confirmar

- Quanto maior o desconto, melhor será a pontuação?

O gráfico de linha apresenta uma tendência de crescimento gradual conforme o desconto aumenta. No entanto, a linha azul (média de ranking) apresenta uma variação considerável, sem seguir uma tendência clara de aumento ou diminuição. Isso indica que, embora o desconto tenda a aumentar, a pontuação não acompanha essa tendência de forma consistente.



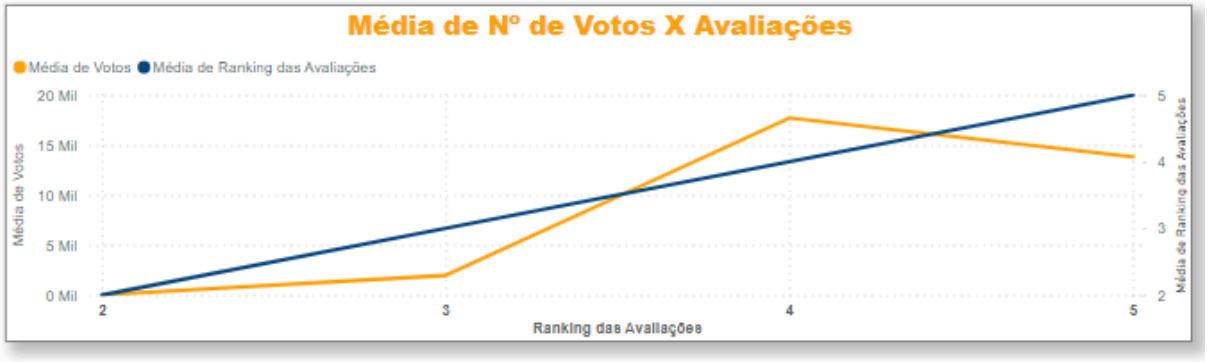

O gráfico de dispersão mostra uma dispersão dos dados indicando que não há uma correlação linear forte entre a média do desconto e a média da avaliação. A linha pontilhada horizontal representa uma média constante, sugerindo que, em média, a avaliação se mantém relativamente estável, independentemente do valor do desconto.



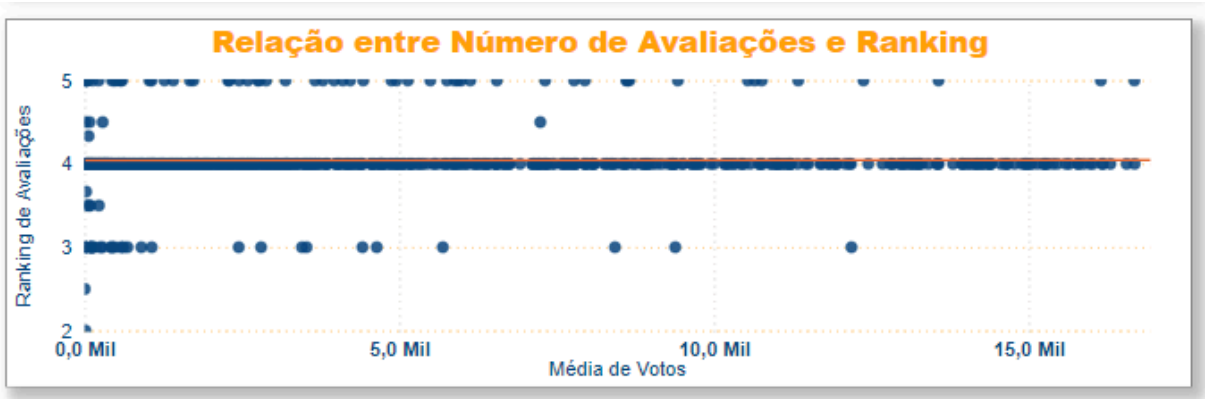

Os gráficos sugerem que o desconto é apenas um dos muitos fatores que influenciam a pontuação. Não se pode estabelecer uma relação de causa e efeito direta entre um maior desconto e uma pontuação mais alta. Por isso, outros fatores podem ser determinantes para aumento da pontuação, como, por exemplo, qualidade do produto, tempo de entrega e a experiência do cliente. Diante dos dados apresentados a hipótese foi refutada.

- Quanto maior o número de pessoas que avaliaram o produto, melhor será a classificação?

O gráfico de linha abaixo, ambos os eixos (número de votos e ranking das avaliações) apresentam uma tendência de crescimento. No entanto, a taxa de crescimento não é exatamente a mesma para ambas as linhas. Isso sugere que o aumento no número de votos não garante um aumento proporcional na classificação.




O gráfico de dispersão mostra uma dispersão dos dados, indicando que não há uma correlação linear forte entre o número de avaliações e o ranking. Embora exista uma tendência geral de aumento no ranking conforme o número de votos aumenta, há muitos pontos fora dessa tendência, o que sugere que outros fatores também influenciam a classificação.

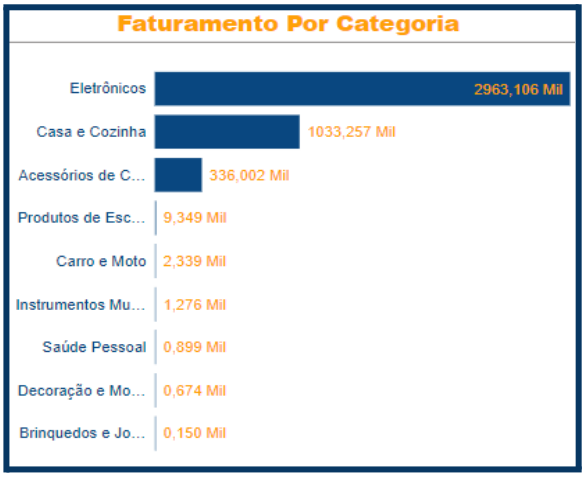



Com base nos gráficos apresentados, não podemos afirmar categoricamente que quanto maior o número de pessoas que avaliaram um produto, melhor será sua classificação.
Os gráficos sugerem que o número de avaliações é apenas um dos muitos fatores que influenciam a classificação de um produto, como a qualidade do produto, o preço, a marca, a experiência do cliente, etc. Não se pode estabelecer uma relação de causa e efeito direta entre um maior número de avaliações e uma classificação mais alta. A hipótese inicial não foi totalmente refutada, mas também não foi confirmada de forma conclusiva. Os dados sugerem que existe uma correlação entre o número de avaliações e a classificação, mas essa relação é mais complexa do que uma simples relação de causa e efeito.

# Resultados e Conclusões

- Quais são as categorias com maior faturamento?

A maior parte do faturamento da empresa está concentrada nas categorias de eletrônicos e casa e cozinha. Embora as categorias de eletrônicos e casa e cozinha sejam as mais lucrativas no momento, outras categorias podem apresentar potencial de crescimento, dependendo de fatores como tendências de mercado, investimentos em novos produtos e estratégias de marketing. Analisando o desempenho de cada categoria, a empresa pode identificar oportunidades para otimizar seus recursos, como investir em produtos com maior margem de lucro ou reduzir custos em categorias menos lucrativas.




- Quais são os produtos com maior preço de venda?

A categoria de eletrônicos conta com os três televisores de maior preço, que variam de R$ 55.000 a R$ 78.000. O que sugere uma variação de preços onde a empresa busca atender a diferentes segmentos de mercado, com produtos posicionados em faixas de preço distintas. A diferença de preço entre as televisões pode estar relacionada a diversos fatores, como:Tamanho da tela, Tecnologia, Marca, Design, Resolução e etc.


- Qual o ticket médio por categoria?

A categoria de eletrônicos possui o maior ticket médio, sugerindo que os consumidores estão dispostos a pagar mais por produtos eletrônicos, como televisores, smartphones e computadores. As demais categorias apresentam ticket médios menores, o que pode indicar que os produtos nesses segmentos são geralmente mais acessíveis ou que os consumidores costumam comprar quantidades maiores de produtos por compra.Analisando o ticket médio de cada categoria, a empresa pode identificar oportunidades de crescimento, como aumentar o ticket médio em categorias com potencial, ou desenvolver novos produtos para atender a segmentos de mercado com maior poder aquisitivo.

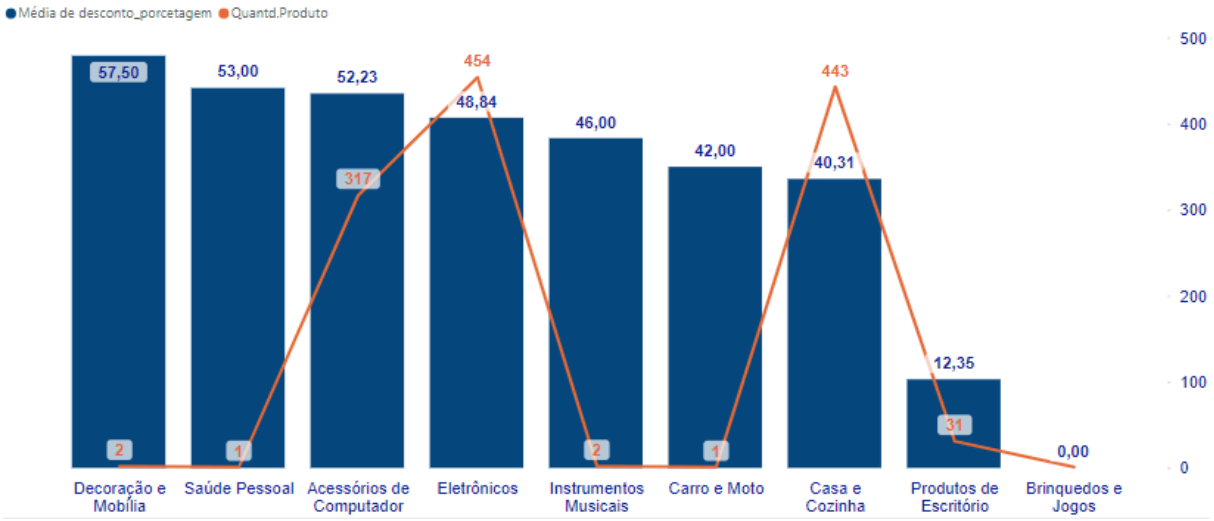
**Visão Geral Categoria de Produtos**

| Categoria | Total_Produto | Total_Desconto | Total_Vendas_Categoria | Ticket Médio |
|---|---|---|---|---|
| Eletrônicos | 454 | 1920 Mil | 2963,106 Mil | 6.526,67 |
| Carro e Moto | 1 | 2 Mil | 2,339 Mil | 2.339,00 |
| Casa e Cozinha | 443 | 817 Mil | 1033,257 Mil | 2.332,41 |
| Acessórios de Computador | 317 | 310 Mil | 336,002 Mil | 1.059,94 |
| Saúde Pessoal | 1 | 1 Mil | 0,899 Mil | 899,00 |
| Instrumentos Musicais | 2 | 1 Mil | 1,276 Mil | 638,00 |
| Decoração e Mobília | 2 | 1 Mil | 0,674 Mil | 337,00 |
| Produtos de Escritório | 31 | 3 Mil | 9,349 Mil | 301,58 |
| Brinquedos e Jogos | 1 | 0 Mil | 0,15 Mil | 150,00 |

- Qual a média de desconto por categoria?

A primeira observação que chama a atenção é que não há uma correlação direta entre a média de desconto e a quantidade de produtos. Significa dizer que as categorias com maior quantidade de produtos não necessariamente possuem menores médias de desconto, e vice-versa.  A média de desconto varia significativamente entre as categorias, indicando que a empresa adota estratégias de preços diferentes para cada segmento. A média geral de

desconto é de 45,75%, o que indica que, em média, os produtos da empresa são vendidos com um desconto de quase metade do preço original.  É necessário, uma análise mais aprofundada, considerando os dados específicos de cada categoria e outros fatores, como, por exemplo, sazonalidade, promoção, concorrência e margem de contribuição para entender as dinâmicas de preços da empresa e tomada de decisões estratégicas.



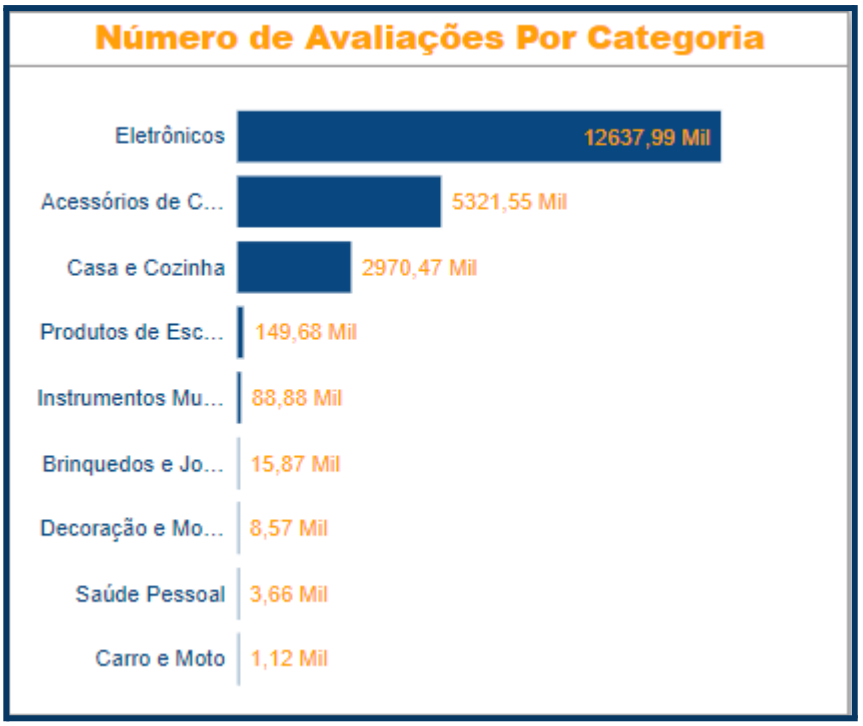

- Qual o produto com maior desconto?

O produto com o maior desconto é o Mini USB da categoria Acessórios de  Computador, com um desconto de 94%. Em seguida, a categoria "Eletrônicos" possui dois produtos idênticos (Smart Watch) com o mesmo desconto de 91%. Podemos concluir que as categorias que apresentam maior faturamento também são as categorias que apresentam maior desconto.


- Quais são as categorias mais e menos avaliadas?

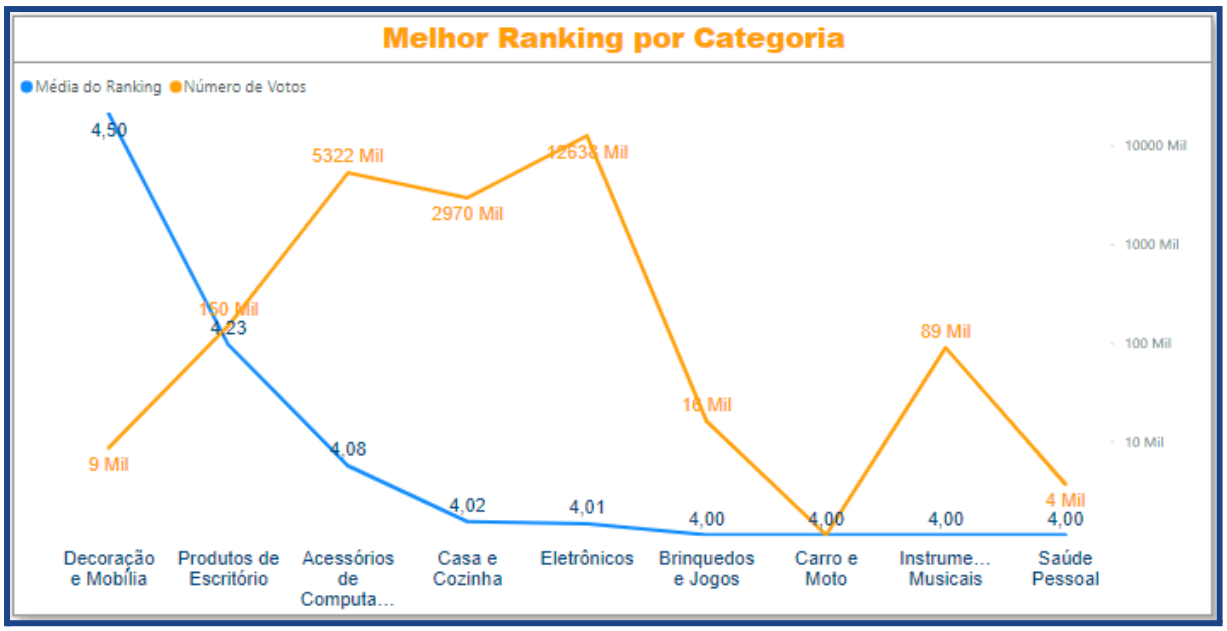
A categoria de eletrônicos é a mais avaliada, com mais de 12 milhões de avaliações,o que indica um grande interesse dos consumidores por este tipo de produto e uma alta taxa de engajamento. A segunda categoria mais avaliada é a de acessórios de computador, com mais de 5 milhões de avaliações. Isso mostra que os consumidores também estão bastante interessados em personalizar seus computadores e dispositivos. A análise das avaliações

por categoria revela um grande interesse dos consumidores por produtos eletrônicos e acessórios de computador. Por outro lado, categorias como carro e moto e saúde pessoal apresentam um número significativamente menor de avaliações. Essa variação pode ser explicada por diversos fatores, como a popularidade dos produtos, o preço, a facilidade de avaliação e o número de produtos disponíveis.




- Qual a melhor categoria no ranking?

É importante analisar a relação entre a média de ranking e o número de votos. Uma categoria com uma média alta e um número baixo de votos pode indicar uma avaliação muito positiva, mas baseada em um número menor de opiniões. Por outro lado, uma categoria com uma média um pouco menor, mas com um número muito alto de votos, pode indicar uma avaliação mais consistente e representativa. Se considerarmos a popularidade como critério para a melhor categoria no Ranking, a categoria de **Eletrônicos** é claramente a mais popular, com milhões de votos. Uma análise mais detalhada dos dados numéricos seria necessária para determinar qual categoria possui a maior média de ranking. No entanto, para a análise em questão foi atribuída a popularidade como critério relevante para identificar a melhor categoria no ranking.




# Limitações

A análise preliminar dos dados revelou a necessidade de aprimorar o banco de dados para obter insights mais precisos. A ausência de informações como quantidade de produtos em estoque, tempo de estoque e margem de contribuição limita a capacidade de avaliar o desempenho dos produtos e identificar oportunidades de otimização. Recomenda-se a inclusão dessas informações no banco de dados para permitir uma análise mais completa e estratégica.

Para uma análise mais aprofundada dos dados, seria necessário incluir informações adicionais, como:

- Quantidade de produtos em estoque: para avaliar o giro de estoque e a demanda por cada produto.
- Tempo de estoque: para identificar produtos encalhados ou com alta rotatividade.
- Data de compra e data da avaliação: para correlacionar as avaliações com o tempo de uso do produto e identificar possíveis problemas de qualidade ao longo do tempo.
- Margem de contribuição: para avaliar a rentabilidade de cada produto e categoria.
- Critérios de qualidade das avaliações: para entender o significado das pontuações de 1 a 5 e garantir a comparabilidade entre as avaliações.

Essas informações permitiriam uma melhor compreensão do desempenho dos produtos, identificação de oportunidades de melhoria e tomada de decisões mais estratégicas.

# Descrição Variável

## Amazon sales

Este conjunto de dados contém dados de mais de 1.000 classificações e análises de produtos disponíveis para venda na Amazon. Amazon é uma empresa americana de tecnologia com operações multinacional, cujos interesses comerciais incluem comércio eletrônico, para o qual compram e armazenam estoque, e cuidam de todo o processo, desde a precificação até o envio, atendimento ao cliente e devoluções.

Os dados vêm de um repositório no Kaggle.

**Descrição das variáveis:**

**Tabela amazon_product**

- product_id: ID do produto
- product_name: nome do produto
- category: categoria do produto
- discounted_price: preço com desconto do produto
- actual_price: preço real do produto
- discount_percentage: porcentagem de desconto do produto
- about_product: Descrição sobre o produto

**Tabela amazon_review**

- user_id: ID do usuário que escreveu a avaliação do produto
- user_name: nome do usuário que escreveu a avaliação do produto
- Review_id: ID da avaliação do usuário
- Review_title: Breve avaliação do usuário
- Review_content: Avaliação completa do usuário
- Img_link: link da imagem do produto
- Product_link: Link para o site oficial do produto
- Product_id: ID do produto
- Rating: Classificação do produto
- Rating_count: número de pessoas que votaram na classificação da Amazon.