Ficha Técnica: Projeto de Análise de Dados

Título do Projeto: Avaliações da Amazon

Contexto: A Amazon é uma plataforma de e-commerce que oferece milhares de produtos em diversas categorias. Nosso dataset contém informações numéricas de preço (actual_price), preço com desconto (discounted_price), porcentagem de desconto (discount_percentage), avaliação do cliente (rating) e volume de engajamento (rating_count).

Objetivo: Analisar, de forma exploratória e estatística, o impacto de preço e desconto nas avaliações (rating) e no engajamento (rating_count) dos clientes da Amazon, identificando correlações e padrões por categoria para apoiar decisões de precificação e campanhas promocionais.

Perguntas Importantes:

- Existe correlação entre percentual de desconto e nota média do produto?
- Faixas de desconto maiores levam a um maior engajamento (rating_count)?
- Qual a relação entre preço real (actual_price) e avaliação média?
- Que categorias se destacam em termos de nota média e desconto médio?

Público Alvo: Stakeholders: gerentes de pricing e marketing da Amazon, interessados em otimizar estratégias promocionais com base em métricas de satisfação e engajamento.

Processamento e análises:

1. **Limpeza dos Dados**

   Após análise dos dados, foram identificados os seguintes resultados relativos à contagem de valores nulos:

   - Na tabela de produtos, a coluna about_product apresentou 4 valores nulos, o que representa 0.27% da base.
   - Na tabela de avaliações, as colunas product_link e img_link apresentaram 466 valores nulos. Quando comparado, foi analisado que os nulos eram mútuos.
   - Decisão: Foi verificado se a ausência dessas informações poderia impactar significativamente as avaliações dos consumidores. Foram realizados os seguintes

testes:

a. Comparação da média de avaliação (rating) entre produtos com e sem link/imagem.
    i. Produtos com link/imagem tiveram média de 4,12.
    ii. Produtos sem link/imagem tiveram média de 4,04.
b. Comparação da média de avaliação entre produtos com e sem descrição (about_product).
    i. Produtos com descrição tiveram média de 4,11.
    ii. Produtos sem descrição tiveram média de 4,00.

Apesar dessas diferenças existirem, os valores encontrados foram considerados pequenos (variação de aproximadamente 0,08 a 0,11 pontos na média de avaliações). Além disso, tratam-se de variáveis bastante subjetivas e inconsistentes, pois podem depender da forma como o cadastro foi feito e não necessariamente do produto em si.

Optou-se por não considerar essas variáveis nas análises principais, pois além da diferença ser pequena, não são informações objetivas sobre os produtos ou sobre a experiência de compra que justifiquem alterar conclusões ou direcionar estratégias com base nesses resultados.

Na base de produtos,foram identificados 118 registros duplicados com base no product_id, que deveria ser único para cada produto. Ao investigar essas duplicatas, observamos que, apesar de o nome e a categoria dos produtos serem os mesmos, havia pequenas variações em campos como preço e na presença ou ausência da descrição (about_product).

Como nosso foco está em análises quantitativas e essas diferenças eram pontuais e de baixo impacto, decidimos remover as duplicatas, priorizando os registros que possuíam o campo about_product preenchido. Assim, garantimos uma base limpa e consistente, sem prejuízo para as análises planejadas.

Identificamos que a base de avaliações possuía 271 registros com review_id duplicados, indicando avaliações repetidas. Ao examinar essas duplicatas, percebemos que a principal diferença entre os registros era a presença ou ausência de links para imagens e produtos.

Para garantir a qualidade dos dados, adotamos o critério de manter apenas a versão da avaliação que contém os links, descartando as duplicatas sem esses elementos. Aplicamos essa filtragem priorizando as avaliações com links e removendo as outras duplicatas pelo review_id.

Como resultado, a base de avaliações foi reduzida de 1465 para 1194 registros únicos, mantendo as informações mais completas e evitando perdas de dados relevantes.

Decidimos focar na análise utilizando apenas variáveis estruturadas, objetivas e com dados mais completos, como product_id, category, preços (discounted_price, actual_price, discount_percentage) para os produtos, review_id, user_id, product_id, rating e rating_count para as avaliações. Optamos por excluir variáveis como about_product, product_name, user_name, review_content, além dos links (img_link e product_link), pois são dados textuais muito subjetivos, incompletos ou que não agregam valor direto à análise quantitativa. Além disso, analisar textos sem um processamento de linguagem natural avançado é complexo e está fora do escopo atual do projeto. Essa decisão visa garantir maior consistência, evitar ruído e focar em informações que realmente impactam nos resultados, otimizando o trabalho e a robustez do estudo.

Extraímos apenas os dois primeiros níveis da hierarquia da variável category, para simplificar a análise e reduzir a granularidade excessiva. Em seguida, identificamos as categorias resultantes e agrupamos aquelas com 5 ou menos ocorrências na base sob o rótulo "Outros", de forma a evitar a dispersão de categorias pouco representativas e tornar as análises mais objetivas e interpretáveis.

Realizamos a identificação de outliers nas variáveis numéricas dos produtos (actual_price, discounted_price e discount_percentage) utilizando a metodologia do IQR (Intervalo Interquartil). Os resultados indicaram que não havia outliers estatísticos nessas colunas, o que sugere uma distribuição de preços e descontos dentro de uma faixa consistente e esperada para o contexto analisado.

Em relação à variável rating, aplicamos o mesmo método e encontramos 74 valores estatisticamente classificados como outliers. No entanto, ao inspecioná-los, verificamos que esses valores não se desviavam significativamente do padrão geral de notas observadas (variando entre 2.8 e 5.0). Considerando a natureza subjetiva das avaliações e o comportamento usual em bases de review de produtos, optamos por manter esses registros na base, evitando o descarte de informações potencialmente relevantes para a análise.

Realizamos a conversão dos tipos de dados para garantir a coerência e precisão nas análises. As variáveis numéricas relacionadas aos preços (actual_price, discounted_price e discount_percentage) foram mantidas no formato float64 para preservar valores decimais. A coluna rating também foi mantida como float64, considerando que as avaliações podem conter notas com casas decimais. Já a variável rating_count, que representa contagem de avaliações, foi convertida para o tipo inteiro (Int64) para refletir corretamente seu caráter discreto. As variáveis categóricas e textuais permaneceram como objetos, e as flags booleanas foram definidas como bool. Essa etapa assegura que as operações estatísticas e

agrupamentos sejam realizados corretamente.

No processo, corrigimos as conversões das colunas numéricas relacionadas a preços dos produtos (actual_price, discounted_price e discount_percentage), removendo caracteres não numéricos e convertendo os valores para tipos numéricos adequados.

Na base de avaliações, excluímos uma linha que apresentava nota de avaliação (rating) ausente, dado que esta variável é essencial para a análise da satisfação do cliente. Também identificamos 2 avaliações que não contavam com o número de avaliações, mas adicionamos esse número.

Além disso, realizamos uma conferência cruzada entre os produtos e suas avaliações, identificando que alguns produtos (162) não possuíam avaliações associadas ou tinham número de avaliações nulo. Esses produtos não serão considerados na unificação.

Essa revisão final foi fundamental para assegurar que as bases de dados estejam limpas, consistentes e alinhadas para a etapa de união, proporcionando uma base robusta para análises confiáveis e insights precisos.

Para a etapa final de preparação dos dados, decidimos realizar a junção das tabelas de produtos e avaliações_limpa utilizando a chave product_id. Nessa integração, optamos por trabalhar apenas com variáveis estruturadas e objetivas, excluindo informações textuais e subjetivas que não contribuem diretamente para a análise quantitativa do projeto.

- Variáveis mantidas:
    - Produtos:

product_id, category_simplificada, actual_price, discounted_price, discount_percentage

    - Avaliações:

review_id, user_id, product_id, rating, rating_count

- Variáveis excluídas:

about_product, product_name, user_name, review_content, img_link, product_link

Essas exclusões se justificam por se tratarem de dados textuais subjetivos, incompletos ou fora do escopo atual, que busca concentrar-se em informações estruturadas e diretamente relevantes para a análise quantitativa de preços, categorias e avaliações de produtos.

Essa decisão visa garantir maior consistência no dataset, reduzir ruído e focar em variáveis que impactam diretamente nos resultados, otimizando o trabalho e a robustez do estudo.

2. **Análise Exploratória**
    a. **Análise por categoria: agrupamento e resumo estatístico**

Nesta etapa da Análise Exploratória, agrupamos os dados da base unificada pela variável **category_simplificada** e, para cada categoria, calculamos:

- A média de:
    - **Preço cheio (actual_price)**
    - **Preço com desconto (discounted_price)**
    - **Percentual de desconto (discount_percentage)**
    - **Nota média de avaliação (rating)**
- A soma de:
    - **Quantidade de avaliações (rating_count)**

Essa operação permitiu sintetizar e organizar o comportamento dos produtos de cada categoria em relação às métricas-chave do projeto.

- **Principais Insights**

Categorias com maiores preços médios:

- **Electronics|HomeTheater,TV&Video** e **Electronics|Mobiles&Accessories** possuem os preços médios mais altos (acima de **₹ 14.000 e ₹ 8.000**, respectivamente).

Categorias com maiores percentuais médios de desconto:

- **Electronics|WearableTechnology** (66,20%)
- **Electronics|Accessories** (62,57%)
- **Electronics|Headphones,Earbuds&Accessories** (60,45%)

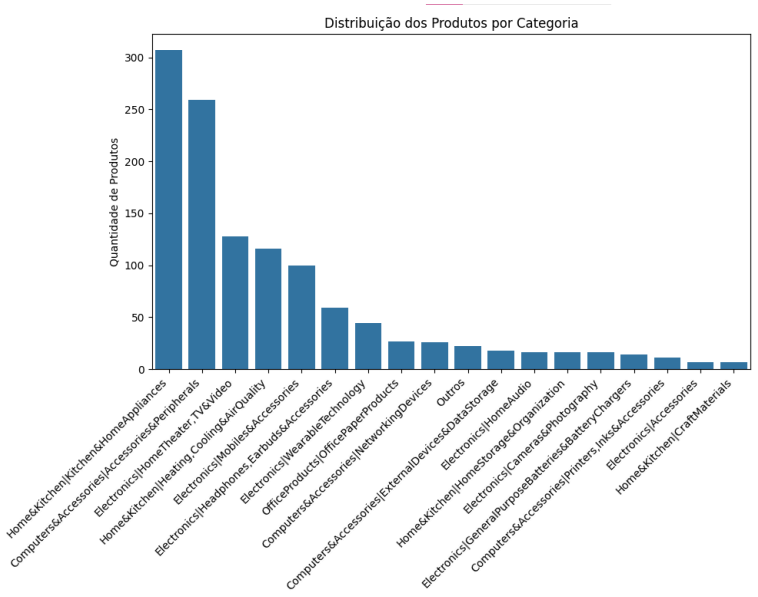
Categorias com melhores médias de avaliação:

- **Electronics|GeneralPurposeBatteries&BatteryChargers** (4,35)
- **Electronics|Accessories** (4,34)
- **Home&Kitchen|CraftMaterials** (4,34)
- Mostrando que algumas categorias com preços médios mais acessíveis podem ter melhor avaliação.

Categorias com maior volume de avaliações:

- **Electronics|Headphones,Earbuds&Accessories** com **3.596.849 avaliações**.
- **Computers&Accessories|Accessories&Peripherals** e **Home&Kitchen|Kitchen&HomeAppliances** também se destacam com volumes acima de **2 milhões**.

b. Quantidade de produtos e gráfico



Durante a análise exploratória, levantamos a quantidade de produtos disponíveis em cada categoria da base. Os resultados mostraram que as categorias com maior número de itens são *Home&Kitchen | Kitchen&HomeAppliances*, com 307 produtos, seguida por *Computers&Accessories | Accessories&Peripherals*, com 259, *Electronics | HomeTheater,TV&Video*, com 128, *Home&Kitchen | Heating,Cooling&AirQuality*, com 116, e *Electronics | Mobiles&Accessories*, com 100. Esse levantamento evidencia que a maior parte da oferta de produtos está concentrada em categorias ligadas a eletrodomésticos de cozinha, acessórios de informática e produtos de áudio e vídeo para o lar.

c. Medidas de Tendência Central

Nesta etapa da análise exploratória, nosso objetivo foi calcular as medidas de tendência central (média, mediana e moda) para as variáveis numéricas da base unificada, a fim de resumir o comportamento geral dos dados e identificar possíveis padrões ou distorções.

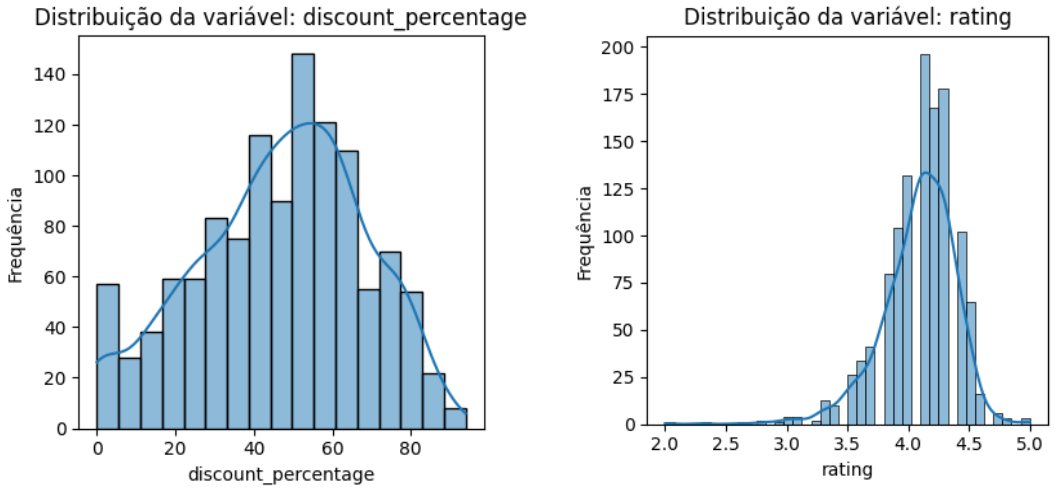
Como a função mode() do pandas pode retornar mais de um valor (quando há empate) ou até nenhum (se a série estiver vazia), criamos a função safe_mode() para capturar de forma segura apenas o primeiro valor da moda de cada variável ou None caso não houvesse moda. Assim, evitamos erros no processo de agregação.

Em seguida, aplicamos as funções de média, mediana e nossa safe_mode() às variáveis:

- actual_price (preço original)
- discounted_price (preço com desconto)
- discount_percentage (percentual de desconto)
- rating (nota de avaliação)
- rating_count (quantidade de avaliações)

Principais resultados obtidos:

- O preço médio dos produtos foi ₹ 4.843,65, enquanto a mediana foi ₹ 1.699,00, evidenciando assimetria positiva (alguns preços muito altos elevam a média).
- O percentual de desconto médio ficou em 46,56%, e a mediana em 49%, mostrando que a maior parte dos produtos recebe descontos próximos de 50%.
- A nota média das avaliações foi 4,08, com a mediana e a moda em 4,1, indicando avaliações concentradas em notas altas, com leve assimetria negativa.
- A quantidade média de avaliações por produto foi de 14.306, mas a mediana foi 3.842, novamente sugerindo a presença de produtos com altíssimo volume de reviews puxando a média para cima. A moda ficou em 4 avaliações, reforçando essa dispersão.



d. Medidas de Dispersão

Selecionamos as variáveis numéricas da base (actual_price, discounted_price, discount_percentage, rating e rating_count) e calculamos:

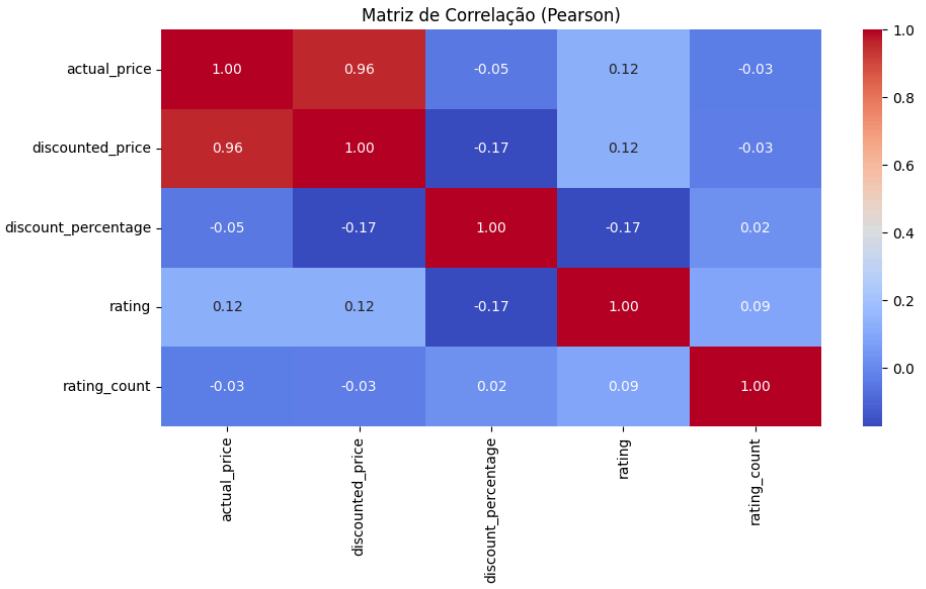
- Desvio padrão (std) → mede, em média, quanto os valores se afastam da média.
- Variância (var) → é o quadrado do desvio padrão; representa a dispersão total dos dados.
- Intervalo interquartílico (IQR) → diferença entre o 3º quartil (Q3) e o 1º quartil (Q1), ou seja, o intervalo onde estão os 50% valores centrais da distribuição. Essa métrica é resistente a outliers e excelente para identificar a dispersão central.

Principais achados:

- **Preços (actual_price e discounted_price)** apresentaram **altíssima dispersão**.
  → O desvio padrão do preço cheio foi de aproximadamente **₹ 10.125**, e o preço com desconto, **₹ 6.122**.
  → Isso confirma a existência de produtos com preços muito acima da média, elevando a variabilidade geral.
- **Percentual de desconto (discount_percentage)** apresentou **variação moderada**, com desvio padrão de **21,46%** e intervalo interquartílico (IQR) de **30%**, indicando diversidade nas promoções aplicadas, mas sem exageros extremos.
- **Notas de avaliação (rating)** mostraram-se **muito homogêneas**.
  → O desvio padrão foi de apenas **0,31**, e o IQR de **0,40**, indicando que a maior parte das avaliações dos produtos se concentra na faixa entre 4 e 4,5.
- **Quantidade de avaliações (rating_count)** apresentou **enorme dispersão**.
  → O desvio padrão foi de **33.160** e o IQR de **12.309**, revelando que alguns produtos têm pouquíssimas avaliações, enquanto outros acumulam dezenas de milhares.

e. Matriz de Correlação

Nesta etapa da análise exploratória, calculamos a **matriz de correlação de Pearson** para as principais variáveis numéricas do dataset: preço cheio (actual_price), preço com desconto (discounted_price), percentual de desconto (discount_percentage), nota média de avaliação (rating) e quantidade de avaliações (rating_count). O objetivo foi identificar o grau de associação linear entre essas variáveis.



Principais achados:

- Houve uma correlação muito alta e positiva entre o preço cheio e o preço com desconto (r ≈ 0,96), indicando que produtos com preço cheio elevado tendem a ter também preço com desconto alto, refletindo consistência no posicionamento de preços.
- O percentual de desconto apresentou correlação negativa moderada com os preços, especialmente com o preço com desconto (r ≈ -0,17), sugerindo que produtos com preços mais altos tendem a ter descontos percentuais menores, ou seja, promoções menos agressivas.
- A correlação entre o percentual de desconto e a nota média das avaliações foi negativa, com coeficiente aproximadamente -0,17, indicando uma tendência leve de que maiores descontos estejam associados a notas médias ligeiramente menores. Isso pode sugerir que produtos com descontos maiores nem sempre apresentam melhor avaliação dos clientes.
- Entre si, a nota média e a quantidade de avaliações mostraram correlação positiva fraca (r ≈ 0,09), indicando que produtos mais avaliados tendem a ter avaliações ligeiramente melhores, mas essa associação é bastante tênue.

Os resultados indicam que o preço cheio e o preço com desconto estão muito relacionados, o que é esperado, dado que o desconto é aplicado sobre o preço cheio. Já o percentual de desconto varia independentemente do preço em maior grau, mostrando uma política de desconto heterogênea.

A correlação negativa leve entre percentual de desconto e nota média sugere que maiores descontos não garantem necessariamente melhores avaliações, o que pode apontar para estratégias promocionais que não impactam diretamente a satisfação do cliente ou para características específicas dos produtos em promoção.
A baixa correlação das variáveis de avaliação com preços e descontos sugere que fatores como preço e desconto não influenciam diretamente as avaliações dos consumidores, ou que essa relação não é linear.

3. Análise

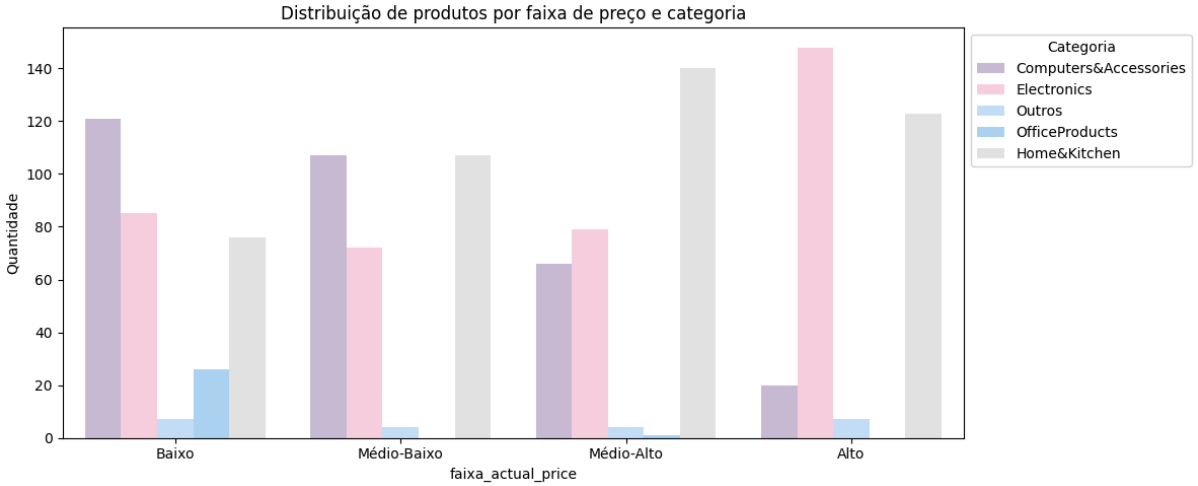
a. Segmentação

Para facilitar as análises segmentadas e tornar os dados mais interpretáveis, criamos faixas (categorias) para variáveis numéricas. Assim conseguimos, por exemplo, analisar o comportamento das avaliações e descontos entre produtos de preço baixo, médio e alto, ou verificar se produtos com mais avaliações têm notas melhores.
A seguir, o detalhamento das faixas que definimos:

- Preço Original (actual_price)
    - Critério: Dividimos os preços originais dos produtos em quartis (25%, 50%, 75%) para equilibrar as quantidades em cada faixa.
    - Faixas criadas:
        - Baixo
        - Médio-Baixo
        - Médio-Alto
        - Alto
    - Método: pd.cut() com os quartis da variável.
- Preço com Desconto (discounted_price)
    - Critério: Mesmo método aplicado para o preço original, usando os quartis da variável discounted_price.
    - Faixas criadas:
        - Baixo
        - Médio-Baixo
        - Médio-Alto
        - Alto

    - Método: pd.cut() com os quartis da variável.
- Porcentagem de Desconto (discount_percentage)
    - Critério: Criamos faixas fixas e interpretáveis de acordo com intervalos percentuais comuns de desconto.
    - Faixas criadas:
        - 0-10%
        - 10-30%
        - 30-50%
        - 50-70%
        - 70-100%
    - Método: pd.cut() com intervalos definidos manualmente.
- Nota da Avaliação (rating)
    - Critério: Criamos intervalos de nota que abrangem intervalos decimais.
    - Faixas criadas:
        - Ruim (1-2) → de 1 até 2.4
        - Médio (3) → de 2.5 até 3.4
        - Bom (4) → de 3.5 até 4.4
        - Excelente (5) → de 4.5 até 5
    - Método: Função apply() com regras condicionais usando if/elif.
- Quantidade de Avaliações (rating_count)
    - Critério: Como essa variável possui grande variação de valores, usamos os quartis para dividi-la em quatro grupos proporcionais.
    - Faixas criadas:
        - Poucas Avaliações
        - Avaliações Moderadas
        - Muitas Avaliações
        - Muitas +
        - Método: pd.cut() com os quartis da variável.

b. Distribuição de Produtos por Faixa de Preço e Categoria



➔ Faixa de preço "Baixo"

Computers & Accessories lidera com 121 produtos, seguido de Electronics (85) e Home & Kitchen (76).

Mostra que há uma boa oferta de produtos de informática e eletrônicos mais baratos na base.

Office Products e Outros são muito pequenos aqui.

➔ Faixa de preço "Médio-Baixo"

Empate técnico entre Computers & Accessories (107) e Home & Kitchen (107).

Electronics com 72 produtos — ainda forte.

Office Products some (0 produtos), o que indica que produtos de escritório tendem a ficar ou muito baratos ou muito caros.

➔ Faixa de preço "Médio-Alto"

Home & Kitchen dispara com 140 produtos, muito acima de outras categorias.

Isso indica que produtos de casa e cozinha tendem a ter ticket médio mais alto.

Electronics e Computers & Accessories equilibrados (79 e 66).

Office Products quase não aparece (1 produto).

➔ Faixa de preço "Alto"

Electronics é de longe a categoria mais representativa com 148 produtos — ou seja, produtos eletrônicos tendem a compor boa parte dos itens caros.

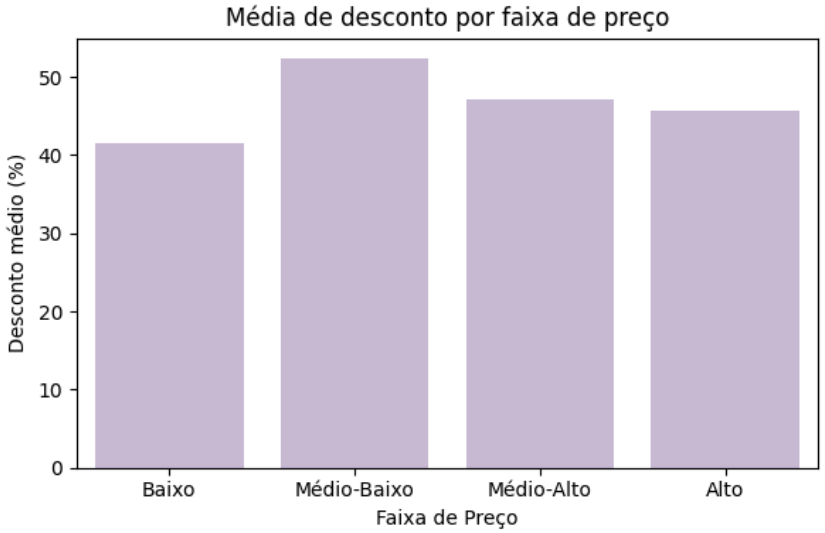
Home & Kitchen aparece forte também (123 produtos), reforçando o posicionamento médio-alto e alto desses itens.

Computers & Accessories cai para 20 produtos, sugerindo que grande parte dessa categoria está concentrada em faixas mais baixas e médias.
Office Products novamente sem relevância.

➔ Conclusões práticas:

Electronics domina o topo da faixa de preço, com maior concentração em Alto.
Home & Kitchen é bem distribuída, mas com tendência a se concentrar nas faixas Médio-Alto e Alto.
Computers & Accessories se concentra fortemente nas faixas Baixo e Médio-Baixo.
Office Products é irrelevante em termos de volume, especialmente nas faixas de preço mais altas.
A categoria Outros é residual em todas as faixas.

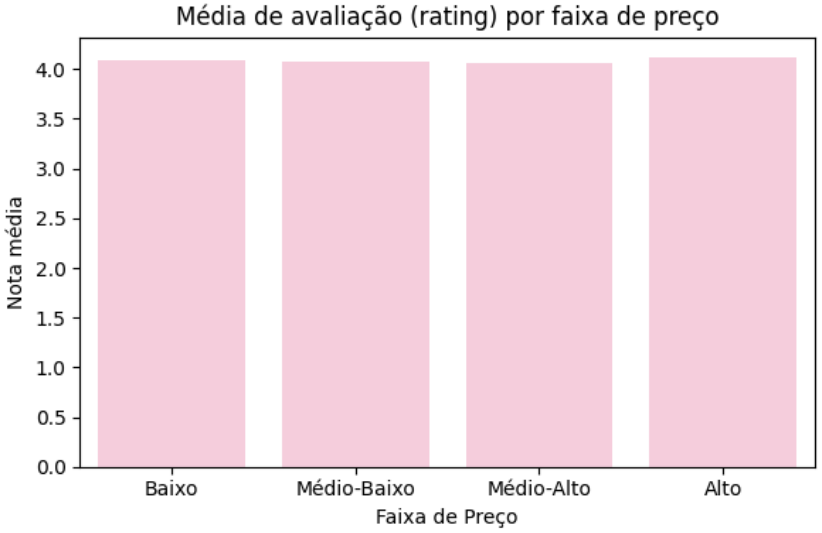
c. Média de Desconto por Faixa de Preço



- Produtos na faixa Médio-Baixo recebem, em média, os maiores percentuais de desconto (52,32%), indicando que essa faixa é potencialmente mais estratégica para ações promocionais agressivas.
- As faixas Baixo, Médio-Alto e Alto têm percentuais médios de desconto relativamente próximos (entre 41% e 47%), com leve destaque para produtos de Médio-Alto (47,17%).
- Curiosamente, produtos Alto não são os mais descontados, o que contraria a hipótese comum de que itens mais caros tendem a ter descontos percentuais maiores.

A política de descontos parece mais concentrada na faixa Médio-Baixo, possivelmente para estimular volume de vendas e girar estoque nesse segmento. Já as faixas Alto e Médio-Alto mantêm descontos relevantes, mas menos agressivos proporcionalmente, o que pode indicar uma estratégia de preservação de margem nesses tickets.

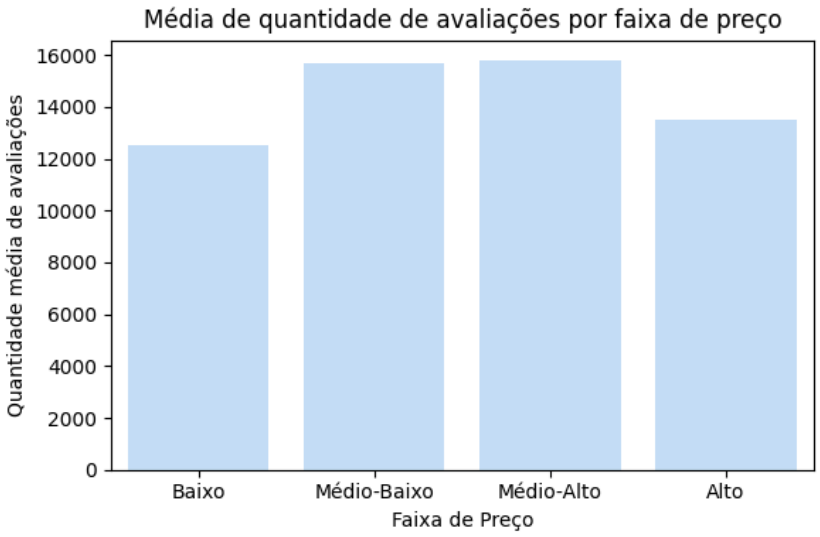
d. Média de Avaliação por Faixa de Preço



- As notas médias são bastante homogêneas entre as faixas de preço, variando de 4,06 a 4,11, o que indica um padrão de avaliação relativamente consistente, independentemente do valor do produto.
- Produtos da faixa Alto possuem a maior nota média (4,11), ainda que a diferença para as demais faixas seja discreta. Isso sugere que produtos mais caros tendem a gerar ligeiramente maior satisfação entre os consumidores.
- As faixas Baixo e Médio-Baixo têm notas praticamente iguais (4,08), reforçando que, no conjunto, produtos mais acessíveis não necessariamente têm avaliações piores.
- Produtos Médio-Alto apresentam a menor média (4,06), mas a diferença não é estatisticamente relevante a princípio, dado o intervalo reduzido.

O preço não exerce uma influência significativa sobre a nota média dos produtos. A leve superioridade na avaliação dos itens de faixa Alto pode estar associada a expectativas mais altas atendidas ou a maior seletividade no consumo desses produtos, mas o padrão geral aponta para boa aceitação em todas as faixas de preço.

e. Média de Quantidade de Avaliações por Faixa de Preço



- As faixas Médio-Baixo e Médio-Alto concentram os produtos com maior média de avaliações, superando 15 mil avaliações por produto.
- A faixa Baixo apresenta a menor média de avaliações (12.498), o que pode indicar menor engajamento ou menor volume de vendas desses produtos, apesar de serem mais acessíveis.
- Curiosamente, produtos de preço Alto têm um volume médio de avaliações (13.476) inferior aos produtos de faixa intermediária. Isso sugere que produtos muito caros, embora bem avaliados (como visto na análise anterior), são avaliados por menos pessoas — possivelmente por serem itens de menor volume de vendas ou consumo mais seletivo.

O volume médio de avaliações tende a ser mais alto entre produtos de preço intermediário (Médio-Baixo e Médio-Alto), indicando que esses produtos podem combinar preço competitivo e boa aceitação de mercado, impulsionando tanto vendas quanto avaliações. Já produtos muito baratos ou muito caros possuem menor engajamento médio em número de avaliações.

4. Validação de Hipóteses

a. Existe correlação entre percentual de desconto e nota média do produto?

- Resultado: Coeficiente de Pearson ≈ –0,17 (correlação negativa fraca).
- Validação: Verdadeira (há correlação), mas no sentido contrário ao que muitas vezes se imagina: maiores descontos associam-se, em média, a notas ligeiramente menores. O efeito é fraco, mas estatisticamente mensurável.

b. Faixas de desconto maiores levam a um maior engajamento (rating_count)?

- Resultado: Os produtos nas faixas de preço intermediárias—que também receberam os maiores descontos médios (~52% na faixa "Médio-Baixo")—foram os que concentraram o maior volume médio de avaliações (> 15 000). Entretanto, as faixas de desconto absolutas mais elevadas (70–100%) não mostraram necessariamente o engajamento mais alto.
- Validação: Falsa no sentido de que descontos estritamente maiores não garantem mais avaliações; o engajamento máximo ocorreu para descontos médios aplicados em segmentos de preço intermediário, e não para as faixas de desconto mais extremas.

c. Qual a relação entre preço real (actual_price) e avaliação média (rating)?

- Resultado: Correlação de Pearson ≈ +0,12 (correlação positiva fraca).
- Validação: Verdadeira, mas muito sutil: existe uma tendência leve de produtos mais caros receberem notas um pouco superiores, ainda que o impacto seja pequeno e não explique a maior parte da variação de rating.

d. Que categorias se destacam em termos de nota média e desconto médio?

- Resultado:
  - Desconto médio mais alto: WearableTechnology (66%), Accessories (62%) e Headphones (60%).
  - Nota média mais alta: GeneralPurposeBatteries (4,35), Accessories (4,34) e CraftMaterials (4,34).
- Validação: Verdadeira: conseguimos identificar claramente quais segmentos lideram em cada métrica, o que atende diretamente à hipótese de descoberta de categorias de destaque.

5. Risco Relativo

Calculamos o Risco Relativo (RR) para medir quão mais (ou menos) provável é um produto com desconto alto ($\geq 50\%$) receber uma avaliação "alta" (nota $\geq 4{,}5$) em comparação a produtos com desconto menor ($< 50\%$).

- Definição de exposição
  - Produtos com desconto_alto = True (desconto $\geq 50\%$) são nosso grupo exposto.

    - Produtos com desconto_alto = False são o grupo controle.
- Definição de desfecho

  “nota_alta = True” quando o rating ≥ 4,5, indicando excelente satisfação.

O resultado indicou que, para produtos com desconto ≥ 50%, a probabilidade de obter uma nota alta (≥ 4,5) é de 7,31%, enquanto para produtos com desconto < 50% essa probabilidade sobe para 8,26%.

O Risco Relativo (RR) calculado foi 0,88, o que significa que: Produtos com descontos altos têm 12% menos chance de receber nota alta em comparação aos produtos com descontos menores. Essa relação reforça o insight de que descontos muito agressivos não elevam a satisfação do cliente e podem até reduzir a percepção de valor ou qualidade.

Links de Interesse:

Apresentação

Github