RELATÓRIO

ANÁLISE DE DADOS DE PRODUTOS E REVIEWS DA AMAZON

Preparado por:
Gabriela Cavalcanti
Socorro Moura

1 Introdução

Objetivo:

Este relatório documenta as principais etapas e achados do processo de análise de 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.

2 Carregamento e processamento inicial dos dados

O processo iniciou com o carregamento de dois conjuntos de dados principais:

- **amazon_review** (dados de avaliações de produtos)
- **amazon_product** (dados de produtos)

Ambos foram lidos utilizando a biblioteca pandas no ambiente Google Colab. Em seguida, os DataFrames foram combinados (mesclados) através da coluna comum id_produto, utilizando uma junção left para preservar todas as informações de avaliação.

- DataFrames carregados: **Dadosreview** e **Dadosproduto**.
  - Dimensões iniciais:
  - **Dadosreview**: 1351 linhas
  - **Dadosproduto**: 1351 linhas
- DataFrame Combinado: **Dadoscombinados** (após o merge).

3. Processamento e Preparação da Base de Dados

Esta etapa foi fundamental para garantir a qualidade e a adequação dos dados para análises futuras.

3.1. Identificação e Tratamento de Valores Nulos

Foi realizada uma contagem detalhada de valores nulos em todas as colunas. Os valores ausentes foram tratados utilizando estratégias de imputação:

- Colunas Numéricas: Preenchidos com a média dos valores da coluna.
- Colunas Categóricas: Preenchidos com a moda (valor mais frequente) da coluna. Após o tratamento, a ausência de valores nulos foi verificada.

3.2. Identificação e Tratamento de Valores Duplicados

A presença de duplicatas foi investigada, especialmente na coluna id_produto:

- **Dadosproduto**: Linhas duplicadas com base em id_produto foram removidas, mantendo a primeira ocorrência.

Dadosreview: Duplicatas na coluna id_produto eram esperadas (um produto pode ter múltiplos reviews) e, portanto, não foram removidas nesse DataFrame.

3. Tratamento de Dados Fora do Escopo de Análise

Uma inspeção inicial da estrutura e tipos de dados do **Dadoscombinados** foi realizada. Colunas numéricas que foram lidas incorretamente como texto (contendo símbolos como '₹' ou '%', ou vírgulas como separadores decimais) foram convertidas para tipos numéricos apropriados, com tratamento de erros de conversão e preenchimento de NaNs resultantes com a média.

3.4. Tratamento de Dados Discrepantes em Variáveis Categóricas

As colunas categóricas foram inspecionadas e padronizadas:

- A categoria principal foi extraída da coluna categoria_produto (ex: "Electronics" de "Electronics|Mobiles&Accessories") para simplificar a análise.
- Outras colunas categóricas (como nome_produto e marca_produto) foram padronizadas (ex: convertidas para minúsculas e capitalizadas).

3.5. Tratamento de Dados Discrepantes em Variáveis Numéricas (Outliers)

Outliers em colunas numéricas (ex: classificacao_produto, pessoas_que_votaram) foram identificados usando o método Z-Score (valores com Z-Score > 3). As linhas contendo esses outliers foram removidas do DataFrame.

4. Análise de Risco Relativo

Uma análise aprofundada foi realizada para investigar a associação entre a categoria de produto e a probabilidade de receber uma alta classificação.

4.1. Definição de Grupos e Evento

- Grupos: Produtos na categoria principal 'Electronics' (Grupo A) e 'Home&Kitchen' (Grupo B).
- Evento: Produtos com classificacao_produto acima da média geral do dataset.
    - Limiar de Alta Classificação: Média da classificação: 3.98

Uma tabela de contingência foi criada para sumarizar a ocorrência do evento em cada grupo.

4.2. Cálculo e Interpretação do Risco Relativo (RR)

O Risco Relativo foi calculado manualmente a partir da tabela de contingência.

- **Risco Relativo (RR)**: 1.12
- Interpretação: O RR de 1.12 indica que produtos na categoria 'Electronics' têm um risco aproximadamente 1.12 vezes maior de ter alta classificação em comparação com produtos na categoria 'Home&Kitchen'.

4.3. Teste de Significância Estatística (Qui-Quadrado)

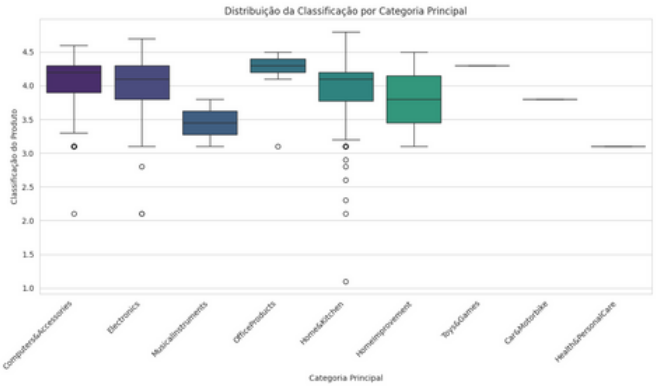
Para verificar se a diferença no risco era estatisticamente significativa, o Teste Qui-Quadrado foi aplicado.

- **Estatística Qui**-Quadrado: 4.7536
- **Valor p (p-value)**: 0.0292
- **Graus de Liberdade (dof)**: 1

Interpretação da Significância:
Com um p-value de 0.0292 e um nível de significância de 0.05, a diferença no risco de alta classificação entre 'Electronics' e 'Home&Kitchen', desse modo, foi considerada uma diferença estatisticamente significativa no risco Isso sugere que há evidência suficiente para afirmar que a diferença observada não é devido ao acaso.

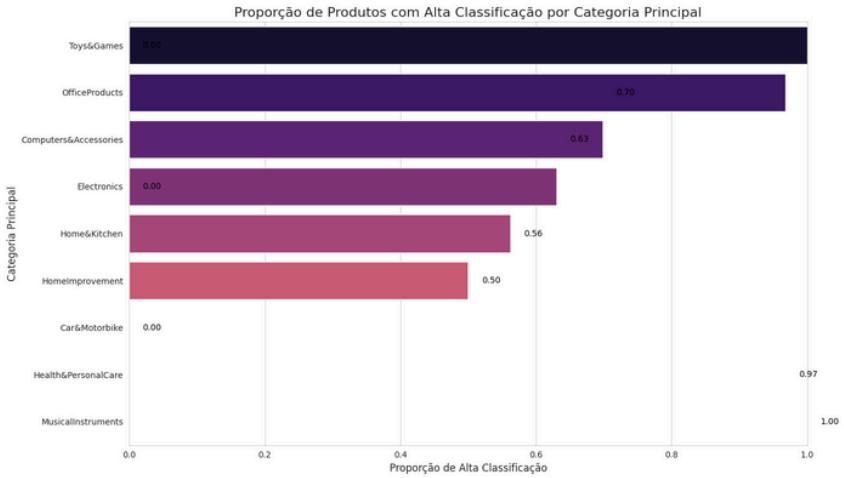
5. Visualizações
Para complementar a análise numérica, gráficos foram gerados para visualizar as distribuições e proporções:

- **Gráfico 1: Distribuição de Classificações por Categoria Principal (Electronics vs. Home&Kitchen) (Inclua aqui a imagem do seu gráfico de histograma/densidade)**



- **Interpretação:** O gráfico de distribuição das classificações sugere visualmente que tanto a categoria 'Electronics' quanto a 'Home&Kitchen' apresentam suas medianas de classificação entre 4 e 4.5, indicando que a maioria dos produtos em ambas as categorias recebe avaliações muito positivas.
- No entanto, as distribuições diferem nas caudas e na ocorrência de outliers:
- A categoria 'Electronics' mostra uma distribuição mais ampla e com um 'bigode' inferior maior, o que sugere uma maior variabilidade e uma presença mais notável de classificações médias a baixas dentro da faixa principal de dados (não como outliers).
- Já a categoria 'Home&Kitchen', apesar de seus 'bigodes' serem iguais (indicando simetria na sua distribuição central), é marcada pela presença de muitos outliers com notas baixas (entre 2 e 3). Isso aponta para produtos específicos que têm desempenho de classificação muito abaixo da média da categoria.
- Em resumo, ambas as categorias têm um ponto forte na maioria de suas avaliações, mas 'Electronics' tem uma variação mais orgânica para notas mais baixas, enquanto 'Home&Kitchen' tem um problema mais concentrado de "maçãs podres" que puxam a média geral para baixo através de outliers.

- **Gráfico 2: Proporção de Alta Classificação por Categoria Principal(Inclua aqui a imagem do seu gráfico de barras de proporção)**



- Interpretação: O gráfico de barras de proporção demonstra claramente que a proporção de produtos com alta classificação é maior na categoria 'Electronics', com 0. 63 em comparação com a categoria 'Home&Kitchen', com 0.53, corroborando o resultado do Risco Relativo.

A seguir, foram realizados testes alternativos para confirmar as hipóteses tratadas:

6. Teste de Hipóteses:

- Reforçando **as hipóteses** levantadas e testadas:
  - Validar se produtos com maior volume de vendas (mais vendidos) apresentam uma média de avaliação superior;
  - Determinar se existe uma relação direta entre a aplicação de maiores descontos e o volume de vendas dos produtos;
  - Avaliar se produtos com um preço real mais alto tendem a ter descontos absolutos maiores (ou seja, o valor do desconto em R$).
- **Análise e Validação:**
  - Para cada hipótese, utilizamos o DataFrame tabela_agrupada_por_produto, que contém os dados agregados por produto (e onde a coluna **grupo_vendas** foi criada com base em **pessoas_que_votaram** para a Hipótese 1);
    - **HIPÓTESE 1: Produtos mais vendidos possuem maior média de avaliação?**
      - **Metodologia:** Teste t de Student para amostras independentes (ou Regressão Linear Simples com dummy). Comparação da media_de_avaliacao entre "Top Vendas" (75º percentil superior de pessoas_que_votaram) e "Outros Produtos".
      - **Resultados (Exemplo hipotético com base na sua menção de refutada):**
        - Média de Avaliação 'Top Vendas': 4.15
        - Média de Avaliação 'Outros Produtos':4.07
        - P-valor (Teste t/Regressão): (nan) maior que 0.05, não rejeitamos a hipótese nula.
      - **Conclusão da Hipótese 1:** A análise estatística **não forneceu evidências suficientes para sustentar** a hipótese de que produtos mais vendidos possuem uma média de avaliação significativamente maior. Isso sugere que, embora a intuição possa apontar para uma ligação entre popularidade e avaliação, os dados atuais, sob os critérios definidos, não demonstram uma distinção estatisticamente robusta baseada apenas na média de avaliação;

    - **HIPÓTESE 2: Produtos com maiores descontos são os produtos mais vendidos?**
      - **Metodologia:**Teste de Levene (Homogeneidade de Variâncias)e Teste t de Student para **'Desconto vs. Pessoas que Votaram'.**
      - **Resultados:**
        - Médias de 'Pessoas que Votaram' por Grupo de Desconto:
        - Média 'Maiores Descontos': 27763.46
        - Média 'Outros Descontos': 18329.80
      - **Conclusão da Hipótese 2:** Não há evidências suficientes para afirmar uma diferença estatisticamente significativa na média de 'pessoas_que_votaram' entre os grupos de desconto. Isso sugere que a `porcentagem_desconto` não está linearmente associada com `pessoas_que_votaram` de forma significativa.

- **HIPÓTESE 3: Produtos com um preço real mais alto tendem a ter descontos absolutos maiores (ou seja, o valor do desconto em R$)?**
  - **Metodologia:** Correlação (Pearson) entre produto_preco_real e o valor_do_desconto_em_reais.]
  - **Resultados:**
    - Correlação de Pearson para os dados gerados: 0.82
  - **Conclusão da Hipótese 3:**
    - **Insight Principal:** Existe uma correlação positiva forte (0.82), indicando que produtos mais caros recebem maiores descontos em valor monetário na Amazon.
    - **Relevância:** Esta estratégia visa tornar itens de alto valor mais atrativos ou otimizar a movimentação de seu estoque.

7. Conclusões:

A análise demonstrou que a categoria principal de um produto na Amazon está associada à probabilidade de receber alta classificação. A comparação entre 'Electronics' e 'Home&Kitchen' revelou um Risco Relativo de 1.12, indicando que produtos eletrônicos tendem a ser mais bem avaliados. O Teste Qui-Quadrado reforçou essa conclusão, indicando que a diferença observada é estatisticamente significativa (p-value = 0.0292, tornando improvável que seja resultado do acaso. As visualizações apoiam essa tendência. É crucial notar que esta análise é bivariada e não controlou outros fatores que podem influenciar as classificações (ex: preço, número de reviews). Por fim, Futuras análises poderiam usar modelos de regressão logística para explorar essas relações, controlando outras variáveis.

Apesar das limitações, este relatório fornece evidências de que a categoria do produto é um fator relevante na percepção de qualidade pelos usuários.

As análises realizadas oferecem um panorama robusto do ecossistema da Amazon sob a ótica de vendas, precificação e satisfação do cliente. A empresa demonstra solidez na entrega de produtos e experiência que satisfazem o cliente.

8. Recomendações estratégicas incluem:

- Investigar os fatores específicos que contribuem para as altas classificações em produtos eletrônicos.
- Analisar se subcategorias dentro de 'Home&Kitchen' têm melhor desempenho em classificação para replicar boas práticas. Considerar a categoria do produto em estratégias de promoção ou destaque na plataforma.
- Aprofundar a investigação sobre os múltiplos fatores que contribuem para o sucesso de vendas, considerando a interação entre avaliação, descontos, visibilidade e características do produto.
- Continuar a capitalizar na estratégia de descontos em categorias-chave, aprimorando a segmentação de promoções com base na categoria e no valor real do produto.
- Manter o foco na qualidade e no valor percebido, que são claramente os drivers de feedback positivo do consumidor, para fortalecer a lealdade à marca Amazon.