Relatório Técnico – Análise de Mercado de Suprimentos para Impressoras HP (Sprint 2)

Lancelot Chagas Rodrigues / 554707
Ana Carolina Martins da Silva / 555762
Kauan Alves Batista / 555082

Link repositório: https://github.com/lancelot-rodrigues/HP-Scraping-Sprint2

---

1. Introdução

Este documento detalha o processo de desenvolvimento e os resultados obtidos na segunda fase do nosso projeto de analise de dados, focado no mercado de suprimentos para impressoras HP. Dando continuidade ao trabalho inicial, esta etapa teve como objetivo principal aprimorar nossa metodologia de coleta, expandir a base de dados com uma nova fonte de e-commerce e aprofundar a análise exploratória para extrair conclusões de negocio mais solidas.

Para alcançar maior organização e manutenibilidade, o projeto foi reestruturado de forma modular. O processo de coleta foi dividido em scripts especificos para cada plataforma, que são orquestrados por um script principal, consolidando os dados para a subsequente etapa de análise.

2. Estrategia de Coleta de Dados

2.1. Estrutura Modular e Fontes de Dados

A arquitetura do nosso sistema de coleta evoluiu para uma abordagem modular, composta por:

- **mercado_scraper.py**: Contém a lógica de scraping especifica para o Mercado Livre.
- **magazine_scraper.py**: Contém a lógica de scraping para a Magazine Luiza.
- **driver_config.py**: Centraliza a configuração do Selenium WebDriver.
- **scraping.py**: Orquestra a execução, chamando os scrapers de cada plataforma e unificando os dados.

Esta modularização nos permitiu isolar a logica de cada site e facilitou a manutenção do código. Como fonte de dados adicional, escolhemos a Magazine Luiza para complementar os dados do Mercado Livre, proporcionando uma visão mais completa ao incluir uma grande varejista com um modelo de negócio distinto.

2.2. Desafios na Coleta e Mitigação

Um desafio persistente durante a coleta foi a detecção e o bloqueio por sistemas anti-bot. Notamos que, mesmo com a implementação de user-agents variados e pausas aleatorias, longas sessões de scraping ou um grande número de requisições sequenciais ocasionalmente resultavam em bloqueios por CAPTCHA.

Para mitigar este problema, a estrutura do nosso script principal (scraping.py) foi projetada para iniciar e encerrar uma nova instancia do navegador para cada busca de links e para cada produto individual. Embora esta abordagem seja menos eficiente em termos de tempo de execução, ela se mostrou mais resiliente ao criar sessões "limpas" para cada tarefa, reduzindo a probabilidade de acumular um historico de comportamento suspeito e ser detectado pelos sistemas de segurança dos sites.

3. Tratamento e Enriquecimento dos Dados

Apos a consolidação dos dados brutos em um unico CSV, realizamos um processo rigoroso de tratamento para garantir a qualidade e a consistencia da nossa base de análise.

3.1. Limpeza e Padronização dos Dados

O tratamento inicial dos dados envolveu as seguintes etapas:

- **Conversão de Preços:** As strings de preço (ex: "R$ 199,90") foram processadas para remover simbolos monetários e espaços. Os separadores de milhar foram removidos e as vírgulas decimais foram convertidas para pontos, resultando em um formato numerico *float* (ex: 199.90).
- **Conversão de Avaliações:** Os campos de nota e numero de avaliações foram convertidos para os tipos numéricos corretos (*float* e *integer*, respectivamente). O numero de avaliações, que por vezes continha pontos como separadores de milhar (ex: "1.234"), foi devidamente convertido para um valor inteiro.
- **Tratamento de Dados Ausentes:** Registros com informações essenciais faltantes, como preço ou título, foram removidos para não comprometer a integridade das analises estatisticas.

3.2. Criação de Colunas Derivadas (Enriquecimento)

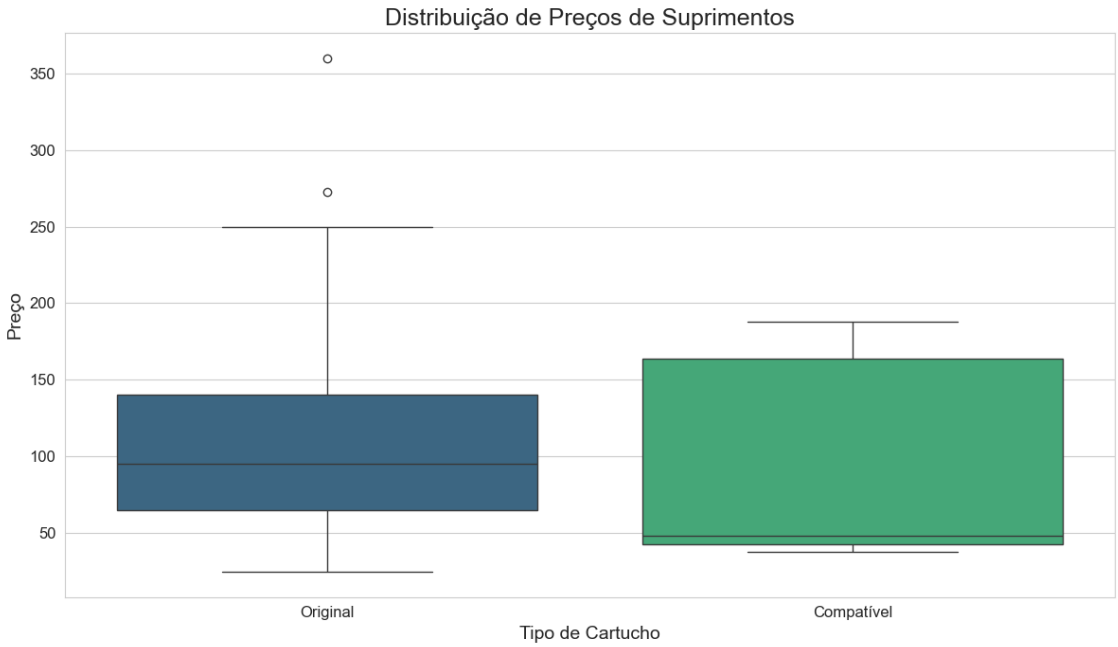
Para possibilitar uma analise mais aprofundada, criamos um conjunto de novas colunas (features) a partir dos dados existentes:

- **categoria_produto:** Classifica cada item como "Suprimento de Impressão", "Notebook" ou "Impressora". Esta coluna foi fundamental para filtrar nossa analise e focar apenas nos cartuchos e toners, evitando a distorção causada por produtos de alto valor.
- **compatibilidade:** Segmenta os produtos em "Original" ou "Compatível", permitindo uma comparação direta entre as duas categorias.
- **capacidade:** Diferencia cartuchos de capacidade "Padrão" e "XL (Alto Rendimento)", crucial para a analise de custo-benefício.
- **modelo_cartucho:** Extrai o numero do modelo (ex: 664, 662) para permitir a analise de popularidade por linha de produto.
- **custo_por_pagina:** Uma metrica calculada (preço / rendimento_paginas) que serve como o principal indicador de valor para o consumidor.

4. Análise Descritiva e Visual

Com os dados tratados e enriquecidos, geramos as seguintes visualizações para extrair insights sobre o mercado.

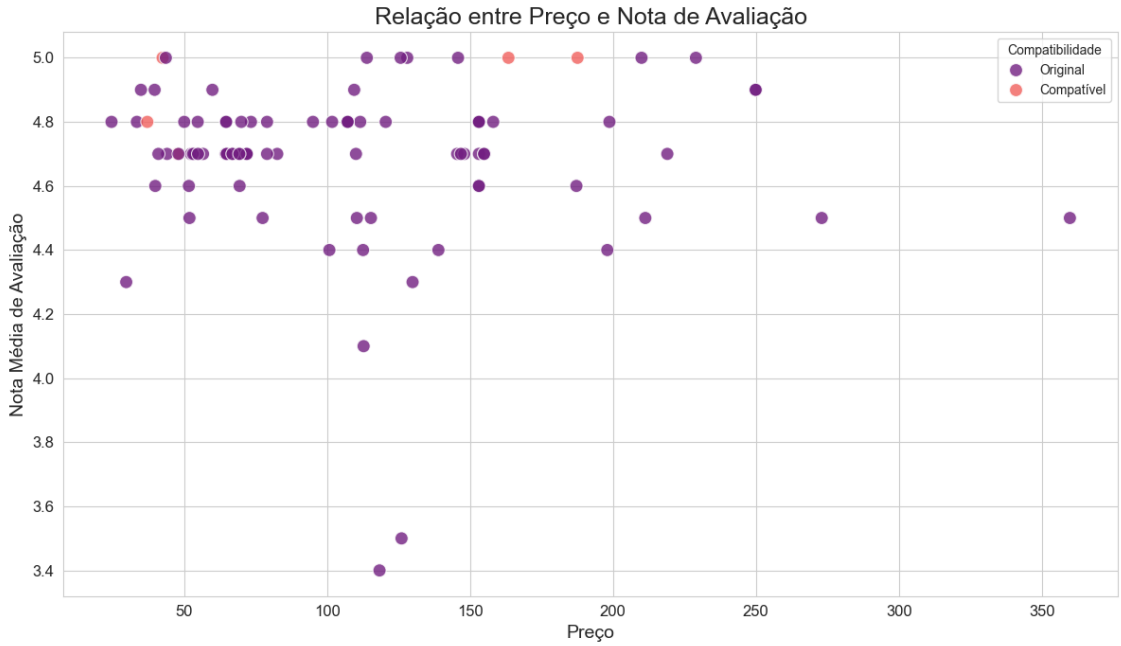
Gráfico 1: Distribuição de Preços de Suprimentos



Análise:
O gráfico de box plot demonstra claramente a diferença de posicionamento de preço. Os cartuchos Originais apresentam uma mediana de preço superior e uma maior dispersão de valores. Em contraste, os Compatíveis se posicionam como uma

alternativa de baixo custo, com preços mais baixos e mais concentrados, validando a estrategia de mercado de cada categoria.

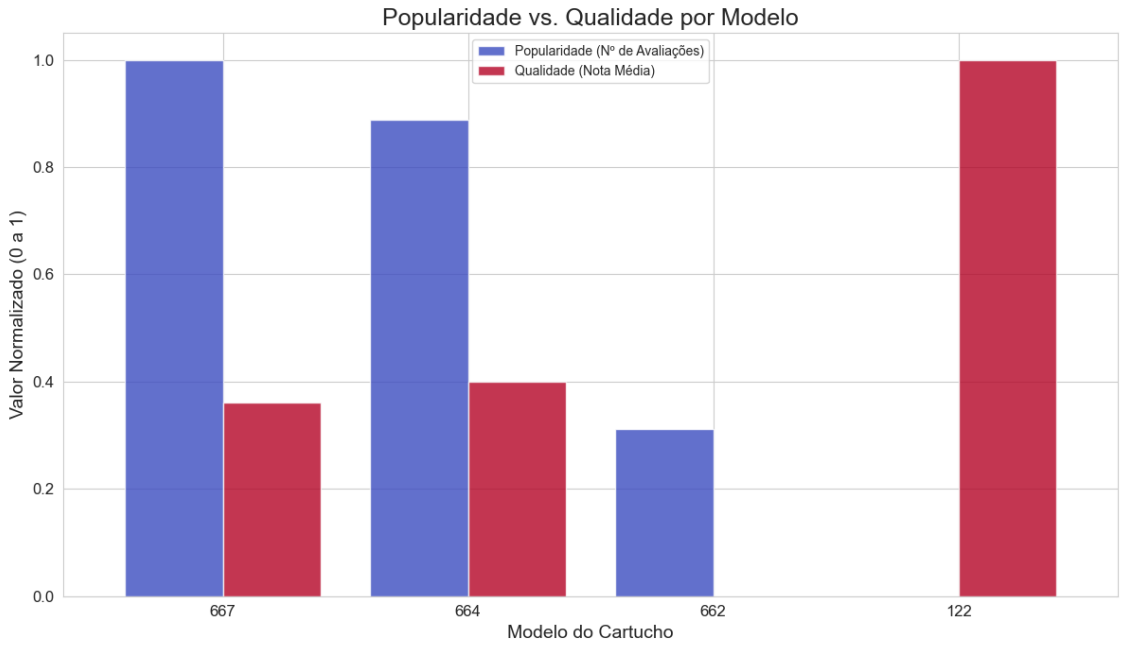
Gráfico 2: Relação entre Preço e Nota de Avaliação



Análise:

O gráfico de dispersão mostra que não ha uma correlação direta entre o preço e a satisfação do cliente. Produtos em diversas faixas de preço, incluindo os mais acessíveis, alcançam notas de avaliação muito altas (entre 4.5 e 5.0). Este resultado sugere que a percepção de qualidade do consumidor não está necessariamente atrelada ao valor pago.

Gráfico 3: Popularidade vs. Qualidade por Modelo

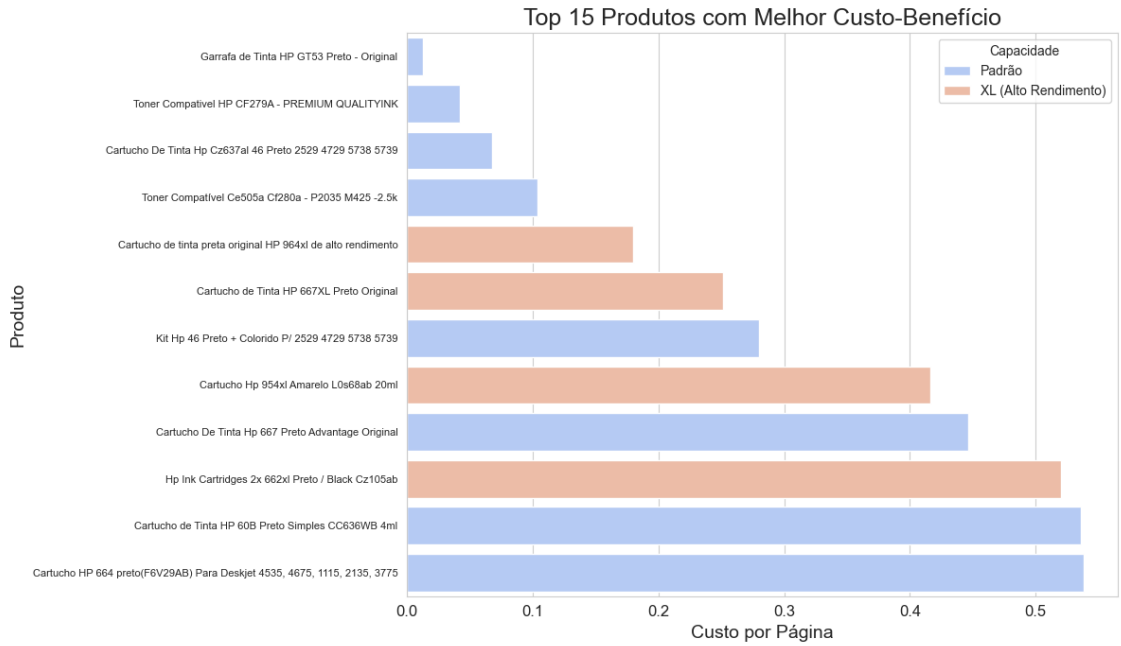


Análise:
Este gráfico compara o volume de avaliações (Popularidade) com a nota media (Qualidade), ambos normalizados em uma escala de 0 a 1 para uma comparação justa.

- O Modelo 667 se destaca por ter a maior qualidade percebida, enquanto o Modelo 664 apresenta a maior popularidade.

Insight: Existe um claro *trade-off* entre o alcance de mercado e a satisfação máxima do cliente. O modelo mais vendido não é necessariamente o mais bem avaliado, o que pode indicar uma oportunidade de marketing para destacar a qualidade superior de modelos de nicho.

Gráfico 4: Análise de Custo-Benefício



Análise:
Este grafico classifica os produtos com base no seu custo por pagina, a métrica mais relevante para o consumidor que busca economia.

- Os cartuchos XL (Alto Rendimento), representados pelas barras laranjas, dominam as posições de melhor custo-benefício.

Conclusão: Fica provado visualmente que, apesar de um custo inicial mais alto, os cartuchos de alta capacidade são a escolha mais economica a longo prazo.

5. Conclusão Final

Esta segunda fase do projeto aprimorou com sucesso nossa metodologia de coleta de dados e aprofundou a analise de mercado. A modularização do código e a decisão de pivotar para uma fonte de dados mais estavel foram cruciais para o sucesso da entrega. As conclusões indicam um mercado complexo, onde o preço não é um indicador direto de qualidade, e o custo-benefício real (custo por página) é um fator chave que pode ser destacado para o consumidor.