# **Relatório Sprint 2 - Análise Exploratória do Dataset HP**

**Integrantes do grupo:**

Danilo Ramalho Silva | RM: 555183

Israel Dalcin Alves Diniz | RM: 554668

João Vitor Pires da Silva | RM: 556213

Matheus Hungaro | RM: 555677

Pablo Menezes Barreto | RM: 556389

Tiago Toshio Kumagai Gibo | 556984

**link do projeto GitHub:** <https://github.com/DaniloRamalhoSilva/HP_Cartucho_Pirata_Detector>

**link do colab:** <https://colab.research.google.com/drive/1wWo_4J3_X4Ympwp5adFQedaf_4WANK46#scrollTo=6kMHE0skJDQJ>

# **1. Limpeza e Preparação**

* O dataset data/dataset\_hp.csv possui 68 registros coletados no Mercado Livre.
* Filtro de busca para a coletada dos dados utilizado foi “**cartucho hp 667 preto**”
* A coluna review\_amount continha valores com textos ("1 avaliação", "1,869" etc.). Foi realizado tratamento para manter apenas números.
* Após a limpeza, as principais colunas numéricas puderam ser analisadas sem erros.

## **2. Visão Geral dos Dados**

Resultado:

* 40 anúncios rotulados como **suspeito**.
* 28 anúncios rotulados como **original**.

### **3. Estatísticas de Preço por Rótulo**

| **label** | **média** | **mínimo** | **máximo** | **mediana** | **contagem** |
| --- | --- | --- | --- | --- | --- |
| original | 317,44 | 64,61 | 979,00 | 260,91 | 28 |
| suspeito | 286,26 | 1,35 | 830,27 | 185,40 | 40 |

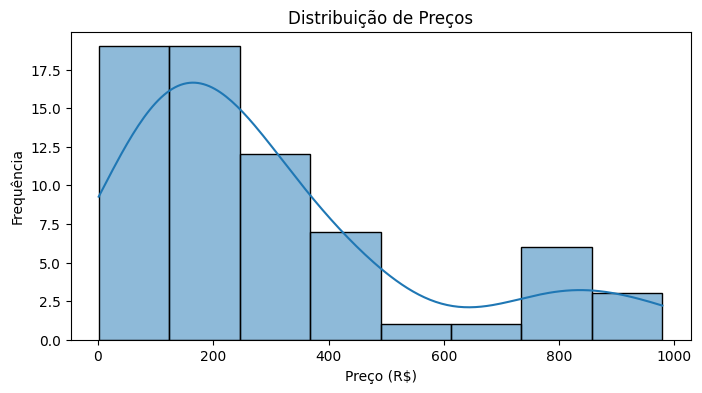
* A média de preço dos cartuchos **originais** é levemente maior que a dos **suspeitos**.

### **4. Avaliações**

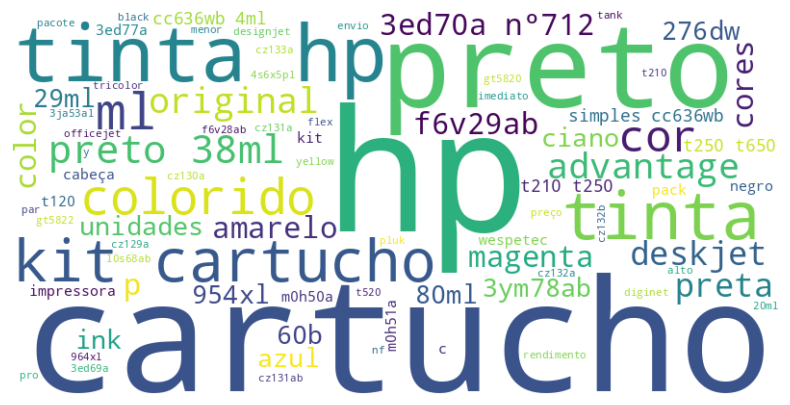
* Nota média dos anúncios: **4.71** estrelas.
* Total de avaliações somadas: **45.404** reviews.
* Os 5 vendedores mais recorrentes são: OBERO INFORMATICA, INK LASER INFO, vanvan, Eshop e Tec Print.

## **5. Visualizações e Análises**

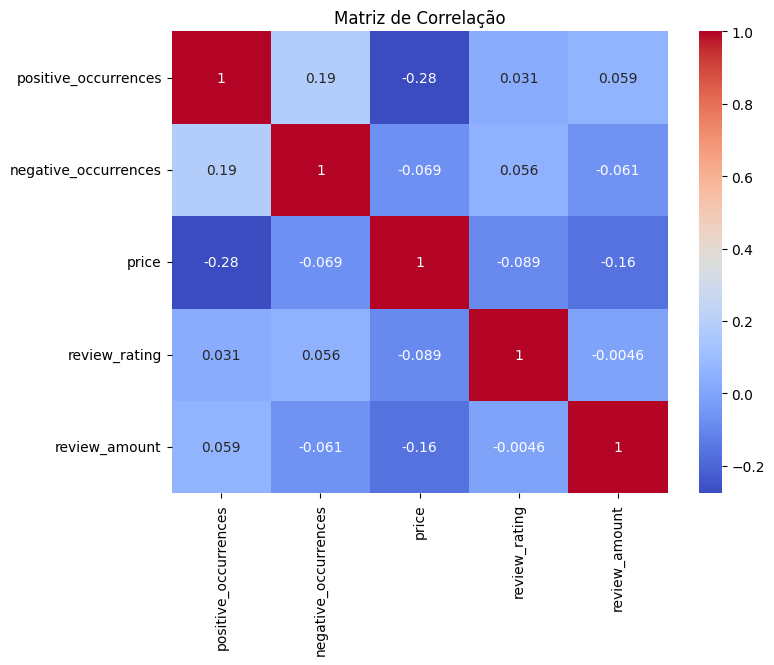
**5.1 Histogramas dos preços.**



**5.2 Wordclouds das 10 palavras mais frequentes nas descrições:**



1. hp 75
2. cartucho 53
3. tinta 39
4. preto 32
5. kit 18
6. 664 14
7. cartuchos 12
8. ml 12
9. 667 11
10. colorido 10

**5.3 Correlações simples entre variáveis numéricas.**

## Não há fortes correlações entre preço e outras variáveis, indicando que outros fatores influenciam a classificação

## **6. Cronograma Macro (CRISP-DM)**

1. **Business Understanding** – Definição do problema: identificar indícios de pirataria em cartuchos HP vendidos online.
2. **Data Understanding** – Coleta dos anúncios no Mercado Livre e inspeção inicial dos atributos.
3. **Data Preparation** – Limpeza do dataset, padronização de preços e campos de avaliações, contagem de ocorrências de palavras-chave.
4. **Modeling** – Utilização de LLM para rotular os dados e criação de features extras (ex.: positive\_occurrences).
5. **Evaluation** – Análise dos rótulos obtidos, validação manual de exemplos e ajustes no prompt.
6. **Deployment** – Exportação de CSV para uso em modelos de machine learning e integração futura em WebApp.