

## **CP 2 - Governança**

## Conteúdo

<b>Documento Técnico .....</b>	<b>3</b>
<b>1. Introdução.....</b>	<b>3</b>
2. Metodologia de Pesquisa e Coleta de Dados .....	3
2.1. Web Scraping .....	3
2.2. Estrutura de Dados e Armazenamento .....	4
3. Fundamentação Teórica.....	4
3.1. Problema de negócio.....	4
3.2. Análise de Dados.....	5
3.3. Gerenciamento de Banco de Dados .....	6
4. Análise de Dados e Resultados .....	7
5. Decisões e Justificativas.....	7
4. Análise de Dados e Resultados .....	7
4.1. Visão Geral dos Dados Coletados .....	7
4.2. Detalhes dos Produtos (products_data).....	8
4.3. Análise de Avaliações (products_review) .....	9
4.4. Próximos Passos para Análise.....	9
5. Decisões e Justificativas.....	9
5.1. Abordagem de Web Scraping .....	9
5.2. Persistência de Dados.....	10
5.3. Modularidade do Código .....	10
5.4. Análise Exploratória de Dados (EDA) .....	10
5.5. Estratégias Anti-Bloqueio .....	11
5.6. Tratamento de Erros.....	11

## Informações do Documento

<b>Projeto:</b>	CP 2 - Governança		
<b>Equipe:</b>	555677 - Matheus Hungaro Fidelis 556389 - Pablo Menezes Barreto 556984 - Tiago Toshio Kumagai Gibo 554668 - Israel Dalcin 555183 - Danilo Ramalho Silva 556213 - João Vitor Pires da Silva	<b>Versão do Documento:</b>	1.0

## Histórico de versão

Versão	Data	Revisado por	Descrição	Nome do Arquivo
1.0	15/06	Tiago Toshio Kumagai Gibo	Criação do documento	CP 2 - Governança
1.1	19/06	João Vitor Pires da Silva	Ajustes	CP 2 - Governança

---

# Documento Técnico

## 1. Introdução

---

Este projeto visa demonstrar a viabilidade e a metodologia para a extração e análise dos anúncios para classificar possíveis vendas de produtos piratas em plataformas de marketplace como o Mercado Livre

## 2. Metodologia de Pesquisa e Coleta de Dados

---

### 2.1. Web Scraping

Para a coleta de dados do Mercado Livre, foi empregada a técnica de web scraping. Dada a natureza dinâmica das páginas do Mercado Livre, que carregam conteúdo via JavaScript, a escolha da ferramenta de scraping foi crucial. Optou-se pelo uso do Selenium, uma ferramenta poderosa para automação de navegadores. O Selenium simula a interação de um usuário real com a página, permitindo a execução de JavaScript e a manipulação do DOM (Document Object Model), o que é essencial para acessar informações que não estariam disponíveis em uma simples requisição HTTP.

Justificativas para o uso do Selenium:

- **Interação com Conteúdo Dinâmico:** O Mercado Livre utiliza JavaScript para carregar grande parte do seu conteúdo, incluindo listas de produtos, detalhes e comentários. O Selenium permite que o script aguarde o carregamento completo desses elementos antes de tentar extrair os dados.
- **Simulação de Comportamento Humano:** Para evitar ser detectado e bloqueado pelos mecanismos anti-bot do site, o Selenium possibilita a simulação de ações humanas, como rolagem da página, cliques em botões (e.g., "mostrar mais comentários") e a introdução de atrasos aleatórios entre as requisições. Isso torna o processo de scraping mais robusto e menos propenso a bloqueios.

Estratégias Anti-Bloqueio Implementadas:

- **Atrasos Aleatórios** (`time.sleep(random.uniform(2.0, 3.0))`): Inserção de pausas aleatórias entre as requisições para imitar o comportamento de um usuário humano e evitar padrões de acesso que poderiam ser identificados como automação.

- Controle de Limites: Definição de limites para o número de páginas, produtos e comentários a serem coletados, controlando a carga sobre o servidor do Mercado Livre e o volume de dados coletados.
- Tratamento de Exceções: Implementação de blocos `try-except` para lidar com erros durante o scraping de produtos ou comentários individuais, garantindo que o processo continue mesmo diante de falhas pontuais.

## 2.2. Estrutura de Dados e Armazenamento

Os dados coletados foram armazenados em um banco de dados SQLite, um sistema de gerenciamento de banco de dados relacional leve, sem servidor e de configuração zero. A escolha do SQLite foi baseada em sua simplicidade, portabilidade e adequação para projetos de pequena a média escala, eliminando a necessidade de um servidor de banco de dados dedicado.

O banco de dados `mercadolivre.db` foi estruturado em três tabelas principais para garantir a organização e a integridade dos dados:

- `products_url`: Armazena as URLs dos produtos que foram identificados para scraping. Inclui campos como `id`, `produto` (nome do produto pesquisado), `url` e `scraped` (indicador se o produto já foi raspado).
- `products_data`: Contém os dados detalhados de cada produto. Campos incluem `id`, `products_url_id` (chave estrangeira para `products_url`), `url`, `title`, `price`, `review_rating`, `review_amount`, `seller` e `description`.
- `products_review`: Armazena as avaliações e comentários dos produtos. Inclui `id`, `products_data_id` (chave estrangeira para `products_data`), `rating`, `review` e `review_date`.

Justificativas para a Estrutura:

- Normalização: A divisão dos dados em tabelas separadas e relacionadas minimiza a redundância e melhora a integridade dos dados, facilitando a manutenção e a consulta.
- Rastreabilidade: A inclusão de campos `data_cadastro` em todas as tabelas permite rastrear a data de inserção dos registros, o que é valioso para auditoria e análise temporal.

## 3. Fundamentação Teórica

---

### 3.1. Problema de negócio

A pirataria nos marketplaces online é um problema crônico no Brasil que gera uma concorrência desleal, prejudica a reputação das marcas e plataformas, expõe consumidores a riscos e causa perdas bilionárias para a economia. A venda de

produtos falsificados, que vão desde eletrônicos e vestuário até cosméticos e brinquedos, cria um ciclo vicioso de prejuízos.

Para os consumidores, o risco vai além do financeiro, envolvendo produtos de baixa qualidade, sem garantia e que podem ser perigosos para a saúde e segurança. Para vendedores legítimos e donos de marcas, o impacto direto se reflete na perda de vendas e na desvalorização da marca.

### O Rastro do Prejuízo: Números do Setor de Eletrônicos

O setor de eletrônicos é um dos alvos preferidos das redes de falsificação, e os números revelam a magnitude do rombo. Segundo dados da Associação Brasileira de Combate à Falsificação (ABCF) referentes a 2024, o mercado ilegal como um todo gerou perdas de **R\$ 471 bilhões** à economia brasileira. Dentro deste universo, o setor de eletrônicos sofre um impacto direto e massivo:

- **Celulares:** O prejuízo com a venda de aparelhos falsificados e contrabandeados atingiu a marca de **R\$9,7 bilhões**.
- **PCs e Softwares:** A pirataria de computadores, componentes e programas de software causou perdas de **R\$8,7 bilhões**.
- **TV por Assinatura (incluindo "TV Box"):** O segmento, impulsionado pela popularização das "TV boxes" piratas que desbloqueiam canais ilegalmente, registrou um prejuízo de **R\$12,1 bilhões**.

## 3.2. Análise de Dados

A análise de dados é o processo de inspecionar, limpar, transformar e modelar dados com o objetivo de descobrir informações úteis, informar conclusões e apoiar a tomada de decisões. É um campo multidisciplinar que envolve estatística, ciência da computação e conhecimento do domínio.

Tipos e Abordagens da Análise de Dados:

1. **Análise Exploratória de Dados (EDA):** É uma abordagem para analisar conjuntos de dados para resumir suas principais características, muitas vezes com métodos visuais. A EDA é usada para ver o que os dados podem revelar além da modelagem formal ou teste de hipóteses. Ajuda a entender a estrutura dos dados, identificar padrões, detectar anomalias e testar suposições.
2. **Análise Descritiva:** Descreve as principais características de um conjunto de dados. Envolve a sumarização de dados usando medidas de tendência central (média, mediana, moda) e dispersão (variância, desvio padrão), bem como distribuições de frequência e visualizações.
3. **Análise Diagnóstica:** Foca em entender por que algo aconteceu. Envolve a exploração de relações de causa e efeito nos dados, muitas vezes usando

técnicas como drill-down, descoberta de dados, mineração de dados e correlações.

4. **Análise Preditiva:** Utiliza dados históricos para fazer previsões sobre eventos futuros. Envolve o uso de modelos estatísticos e algoritmos de aprendizado de máquina para identificar probabilidades e tendências futuras. Exemplos incluem regressão, séries temporais e redes neurais.
5. **Análise Prescritiva:** Vai além da previsão, recomendando ações que podem influenciar os resultados desejados. Combina insights de análises descritivas e preditivas com regras de negócios e otimização para sugerir o melhor curso de ação.

#### Etapas Comuns na Análise de Dados:

1. **Coleta de Dados:** Aquisição de dados de várias fontes, como bancos de dados, APIs, web scraping, etc.
2. **Limpeza de Dados:** Identificação e correção de erros, inconsistências, valores ausentes e duplicatas nos dados.
3. **Transformação de Dados:** Conversão de dados para um formato adequado para análise, incluindo normalização, agregação e criação de novas variáveis.
4. **Modelagem de Dados:** Aplicação de técnicas estatísticas ou de aprendizado de máquina para descobrir padrões e relações nos dados.
5. **Visualização de Dados:** Apresentação dos dados e resultados da análise de forma gráfica para facilitar a compreensão e a comunicação de insights.
6. **Interpretação e Comunicação:** Tradução dos resultados da análise em conclusões acionáveis e comunicação eficaz para as partes interessadas.

#### O Papel da Teoria na Análise de Dados:

A teoria desempenha um papel crucial na análise de dados, fornecendo uma estrutura para a compreensão dos fenômenos e orientando a formulação de hipóteses. Uma teoria pode acelerar o treinamento de um analista, resumindo o que é comumente feito e o que tem sido bem-sucedido. Ela deve reduzir a necessidade de reinventar a roda, fornecendo um ponto de partida para a investigação e ajudando a interpretar os resultados da análise. A análise pode ser vista como a interseção entre teoria e dados, onde a informação é usada para testar previsões teóricas e, por sua vez, os dados podem levar à revisão ou ao desenvolvimento de novas teorias.

### 3.3. Gerenciamento de Banco de Dados

SQLite é um sistema de gerenciamento de banco de dados relacional que é incorporado em um aplicativo final. É uma biblioteca de software que implementa um mecanismo de banco de dados SQL transacional, autônomo, sem servidor e de

configuração zero. É uma escolha popular para aplicações embarcadas e projetos de pequena escala devido à sua simplicidade, portabilidade e ausência de um processo de servidor separado. No contexto deste projeto, o SQLite é ideal para armazenar os dados raspados localmente sem a necessidade de configurar um servidor de banco de dados completo.

## 4. Análise de Dados e Resultados

---

Esta seção apresentará os principais resultados da análise exploratória de dados realizada no notebook `analiseexploratoria.ipynb`. Serão destacadas as descobertas relevantes sobre os produtos, preços, avaliações e vendedores do Mercado Livre, utilizando visualizações e estatísticas descritivas.

## 5. Decisões e Justificativas

---

Esta seção consolida as decisões chave tomadas ao longo do projeto, desde a concepção até a implementação, e as justificativas para cada uma delas. Isso inclui escolhas de ferramentas, metodologias e abordagens para lidar com desafios específicos.

## 4. Análise de Dados e Resultados

---

Esta seção apresentará os principais resultados da análise exploratória de dados realizada no notebook `analiseexploratoria.ipynb`. Serão destacadas as descobertas relevantes sobre os produtos, preços, avaliações e vendedores do Mercado Livre, utilizando visualizações e estatísticas descritivas.

### 4.1. Visão Geral dos Dados Coletados

O notebook `analiseexploratoria.ipynb` inicia com a instalação das bibliotecas necessárias (`db-sqlite3`, `transformers`, `torch`, `tqdm`, `openai`, `pydantic`) e a importação de módulos como `pandas`, `sqlite3`, `re`, `math`, `matplotlib.pyplot`, `seaborn`, `BaseModel`, `OpenAI`, `pipeline` (de `transformers`) e `tqdm`. Embora algumas dessas bibliotecas (como `transformers` e `openai`) sugiram análises mais avançadas, a análise exploratória inicial foca na compreensão da estrutura e conteúdo dos dados.

Os dados são carregados do banco de dados `mercadolivre.db` para DataFrames do Pandas:

```
conn = sqlite3.connect("mercadolivre.db")
products_data = pd.read_sql_query("select * from products_data", conn)
products_review = pd.read_sql_query("select * from products_review", conn)
```

## 4.2. Detalhes dos Produtos (products\_data)

A análise inicial do DataFrame products\_data revela a seguinte estrutura:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                     100 non-null   int64
1   products_url_id       100 non-null   int64
2   url                    100 non-null   object
3   title                  100 non-null   object
4   price                  100 non-null   float64
5   review_rating          100 non-null   object
6   review_amount          100 non-null   object
7   seller                 100 non-null   object
8   description            100 non-null   object
9   brand                  100 non-null   object
10  product_line           100 non-null   object
11  model                  100 non-null   object
12  sales_format           100 non-null   object
13  volume_total           100 non-null   object
14  page_yield             100 non-null   object
15  cartridge_type         100 non-null   object
16  comments_scraped       100 non-null   int64
17  data_cadastro          100 non-null   object
dtypes: float64(1), int64(3), object(14)
memory usage: 14.2+ KB
```

Este DataFrame contém 100 entradas, com 18 colunas. Observa-se que as colunas review\_rating e review\_amount são do tipo object (string), o que indica a necessidade de conversão para tipos numéricos para permitir análises quantitativas. A coluna price já está como float64, o que é ideal para cálculos.

Exemplo das primeiras linhas do products\_data:

	products_url_id	url	title	price	review_rating	review_amount	seller	description
1	1	https://www.mercadolivre.com.br/cartucho-hp-667-colorido	Cartucho Hp 667 Colorido 3yn78ab	71.9	4.7	1,906	vanvan	Caracterist
2	2	https://www.mercadolivre.com.br/cartucho-hp-667-preto	Cartucho HP 667 Preto 2376 2776 6476	60.21	4.7	1,906	OBERO INFORMATICA	cartucho HP
3	3	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp 667 Deskjet Ink Advantage Cor Pr	153.1	4.8	1,960	OBERO INFORMATICA	Itens inclu
4	4	https://www.mercadolivre.com.br/kit-cartucho-hp-664-p	Kit Cartucho Hp 664 Preto E Hp 664 Colorido	146.82	4.7	3,315	Eshop	KIT DE CART
5	5	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp 667 Cor Preto Do 2 Ml	64.49	4.8	3,386	Eshop	Caracterist
6	6	https://produto.mercadolivre.com.br/MLB-2685635387-ki	Kit Cartucho Tinta Original Hp 667 Preto/colorido 237	154.9	4.7	3,759	OBERO INFORMATICA	***REVENDE
7	7	https://produto.mercadolivre.com.br/MLB-3142771512-ki	Kit 2 Cartuchos Tinta Preto Hp 667 Original 1275 2374	179.2	4.8	63	SARAIVA COMERCIO	Kit 2 Carto
8	8	https://produto.mercadolivre.com.br/MLB-1754285285-ki	Kit Cartucho 667 Preto Color Original Hp 2376 2775 27	169.2	4.7	789	Primer Print	KIT COM 2 C
9	9	https://produto.mercadolivre.com.br/MLB-2864354866-2x	2x Cartucho Tinta Original Hp 667xl 2776 2376 3yn81a	419.94	4.6	7	Tec Print	02 cartucho
10	10	https://www.mercadolivre.com.br/cartucho-hp-667-667xl	Cartucho Hp 667 667xl Deskjet 2776 Preto	129.9	4.3	576	TC20240813115050	0 Cartucho
11	11	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp Tricolor 667	68.35	4.7	1,203	Eshop	Cartucho HP
12	12	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp 667 Cor Preto Colorido Do 2 Ml	151.9	4.7	1,906	SCOTCH	Kit Cartuch
13	13	https://produto.mercadolivre.com.br/MLB-3837111410-ki	Kit Cartuchos Tinta Hp 667 Preto + Hp 667 Color Mem	147.99	4.7	573	vanvan	*** REVENDE
14	14	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp 667 Cor Preto Do 2 Ml	73.37	4.8	206	JCMULTVENDAS	- Volume de
15	15	https://www.mercadolivre.com.br/kit-cartucho-hp-667-p	Kit Cartucho Hp 667 Preto E 667 Colorido 2376 2776	152.9	4.7	190	TC20240813115050	KIT CONTEND
16	16	https://produto.mercadolivre.com.br/MLB-1852251485-ki	Kit Cartucho Tinta Original Hp 667xl 2776 Preto E Col	439.9	4.9	29	Tec Print	06 Cartucho
17	17	https://produto.mercadolivre.com.br/MLB-4316298376-ca	Cartucho Hp 667 Preto Original 3yn79ab + 667 Color 3y	177.75	4.6	67	Tec Print	Caracterist
18	18	https://produto.mercadolivre.com.br/MLB-3593038716-02	02 Cartucho Tinta Hp 667 100 rows	279.9	4.7	74	Tec Print	Obtema a q
19	19	https://www.mercadolivre.com.br/cartucho-de-tinta-hp	Cartucho De Tinta Hp 667 667xl 2776 2376 3yn81a	156.0	4.6	506	FINHERITEL KTB	Caracterist



### 4.3. Análise de Avaliações (products\_review)

O DataFrame `products_review` contém as avaliações dos produtos. A análise deste DataFrame é crucial para entender a percepção dos consumidores sobre os produtos. O notebook não detalha a estrutura ou exemplos de `products_review`, mas a presença de colunas como `rating` e `review` sugere a possibilidade de análises de sentimento e de tendências nas avaliações.

### 4.4. Próximos Passos para Análise

Com base na análise exploratória inicial, os próximos passos para uma análise de dados mais aprofundada incluiriam:

- **Limpeza e Transformação de Dados:** Converter as colunas `review_rating` e `review_amount` para tipos numéricos adequados.
- **Análise de Preços:** Investigar a distribuição de preços, identificar outliers e analisar a relação entre preço e avaliação.
- **Análise de Vendedores:** Avaliar o desempenho dos vendedores com base nas avaliações e quantidade de produtos.
- **Análise de Comentários (NLP):** Utilizar as bibliotecas de NLP (Transformers, OpenAI) para realizar análise de sentimento nos comentários, extrair tópicos comuns e identificar pontos fortes e fracos dos produtos.
- **Visualizações Avançadas:** Criar gráficos mais complexos para ilustrar as relações entre as variáveis e os insights descobertos.

## 5. Decisões e Justificativas

---

Esta seção consolida as decisões chave tomadas ao longo do projeto, desde a concepção até a implementação, e as justificativas para cada uma delas. Isso inclui escolhas de ferramentas, metodologias e abordagens para lidar com desafios específicos.

### 5.1. Abordagem de Web Scraping

**Decisão:** Utilização do Selenium WebDriver para a coleta de dados do Mercado Livre.

**Justificativa:** A natureza dinâmica das páginas do Mercado Livre, que dependem fortemente de JavaScript para carregar conteúdo, tornou o Selenium a escolha mais adequada. Ferramentas baseadas apenas em requisições HTTP (como `requests`) não seriam capazes de renderizar o conteúdo dinâmico, resultando em dados incompletos ou ausentes. O Selenium permite a simulação de um navegador real, garantindo que todo o conteúdo visível ao usuário seja acessível para scraping.

## 5.2. Persistência de Dados

Decisão: Armazenamento dos dados coletados em um banco de dados SQLite.

Justificativa: Para um projeto de escopo médio e com foco em prototipagem e análise local, o SQLite oferece uma solução de banco de dados leve, sem a necessidade de um servidor dedicado. Sua facilidade de uso, portabilidade (o banco de dados é um único arquivo) e ausência de configuração complexa agilizaram o desenvolvimento e a integração com o restante do projeto. Além disso, a estrutura relacional do SQLite permitiu a organização eficiente dos dados em tabelas normalizadas, facilitando consultas e análises futuras.

## 5.3. Modularidade do Código

Decisão: Divisão do projeto em módulos Python distintos (`database.py`, `main.py`, `mercadolivre.py`).

Justificativa: Esta decisão de design de software promoveu a organização, a reutilização de código e a facilidade de manutenção. Cada módulo possui uma responsabilidade clara:

- `database.py`: Gerenciamento do banco de dados (criação de tabelas, funções de salvar e consultar).
- `mercadolivre.py`: Funções específicas de scraping para o Mercado Livre (listagem, detalhes do produto, comentários).
- `main.py`: Orquestração do fluxo de scraping e integração dos outros módulos.

Essa modularidade facilita a depuração, permite que diferentes partes do projeto sejam desenvolvidas e testadas independentemente, e torna o código mais compreensível para outros desenvolvedores.

## 5.4. Análise Exploratória de Dados (EDA)

Decisão: Realização de uma Análise Exploratória de Dados (EDA) utilizando Jupyter Notebook e bibliotecas como Pandas, Matplotlib e Seaborn.

Justificativa: A EDA é uma etapa fundamental em qualquer projeto de dados. O Jupyter Notebook, em particular, é uma ferramenta excelente para este fim, pois permite a combinação de código executável, visualizações e texto explicativo em um único ambiente interativo. Isso facilitou a compreensão inicial da estrutura dos dados, a identificação de problemas de qualidade (como tipos de dados incorretos) e a visualização de padrões e tendências. As bibliotecas Pandas, Matplotlib e Seaborn são padrões da indústria para manipulação e visualização de dados em Python, fornecendo as ferramentas necessárias para uma EDA eficaz.

## 5.5. Estratégias Anti-Bloqueio

**Decisão:** Implementação de atrasos aleatórios e simulação de interação humana durante o scraping.

**Justificativa:** Sites como o Mercado Livre possuem mecanismos sofisticados para detectar e bloquear atividades de scraping automatizadas. A inclusão de `time.sleep(random.uniform(min, max))` entre as requisições é uma técnica simples, mas eficaz, para simular um comportamento de navegação mais natural, evitando que o scraper seja identificado por padrões de acesso muito rápidos e consistentes. Além disso, a utilização de `driver.execute_script` para rolar a página e clicar em elementos dinâmicos ajuda a imitar a interação de um usuário real, tornando o scraper mais resiliente a bloqueios.

## 5.6. Tratamento de Erros

**Decisão:** Utilização de blocos `try-except` para lidar com exceções durante o processo de scraping.

**Justificativa:** O web scraping é inerentemente suscetível a erros devido a variações na estrutura das páginas web, problemas de rede ou bloqueios. A implementação de `try-except` garante que o script não pare completamente ao encontrar um erro em um produto ou comentário específico. Em vez disso, ele registra o erro e continua o processamento dos demais itens, tornando o processo de coleta de dados mais robusto e tolerante a falhas parciais.