Universidade Católica de Petrópolis

Disciplina: Mineração de Dados Estruturados
Autor: Marcelo Nicolay Santos
Ano/Semestre: 2025/2

Análise Exploratória do Dataset Online Retail

---

1. Introdução

Este relatório apresenta uma análise exploratória (AED) do dataset **Online Retail**, originado do repositório **UCI Machine Learning**.
O objetivo é investigar **padrões de vendas, comportamento de clientes e características dos produtos**, além de identificar **insights relevantes para estratégias de negócio**.

---

2. Metodologia

A análise foi conduzida em **Python**, utilizando as seguintes bibliotecas:

- **Pandas** e **NumPy**: manipulação e transformação de dados;
- **Matplotlib**, **Mplot3d** e **Seaborn**: visualizações gráficas;
- **MLxtend**: geração de regras de associação via algoritmo Apriori;
- **Scikit-learn**: pré-processamento e clusterização (K-Means).

Etapas da Análise:

1. Carregamento e inspeção inicial dos dados;
2. Estatísticas descritivas e tratamento de valores ausentes;
3. Análise de distribuições (quantidade, preço, países);
4. Visualizações exploratórias para identificar padrões;
5. Pré-processamento e criação de novas variáveis;
6. Mineração de dados com **Apriori** e **Clustering (RFM + K-Means)**.

---

3. Exploração Inicial dos Dados

3.1. Primeiros dados coletados

- **Período analisado:** 01/12/2010 a 09/12/2011;
- **Dimensões:** 541.909 registros e 8 variáveis;
- **Principais variáveis:**
    - *InvoiceNo*: número do pedido;
    - *Quantity*: quantidade de itens;
    - *UnitPrice*: preço unitário;
    - *CustomerID*: identificador do cliente (24,93% ausentes);
    - *Country*: país de origem do pedido.

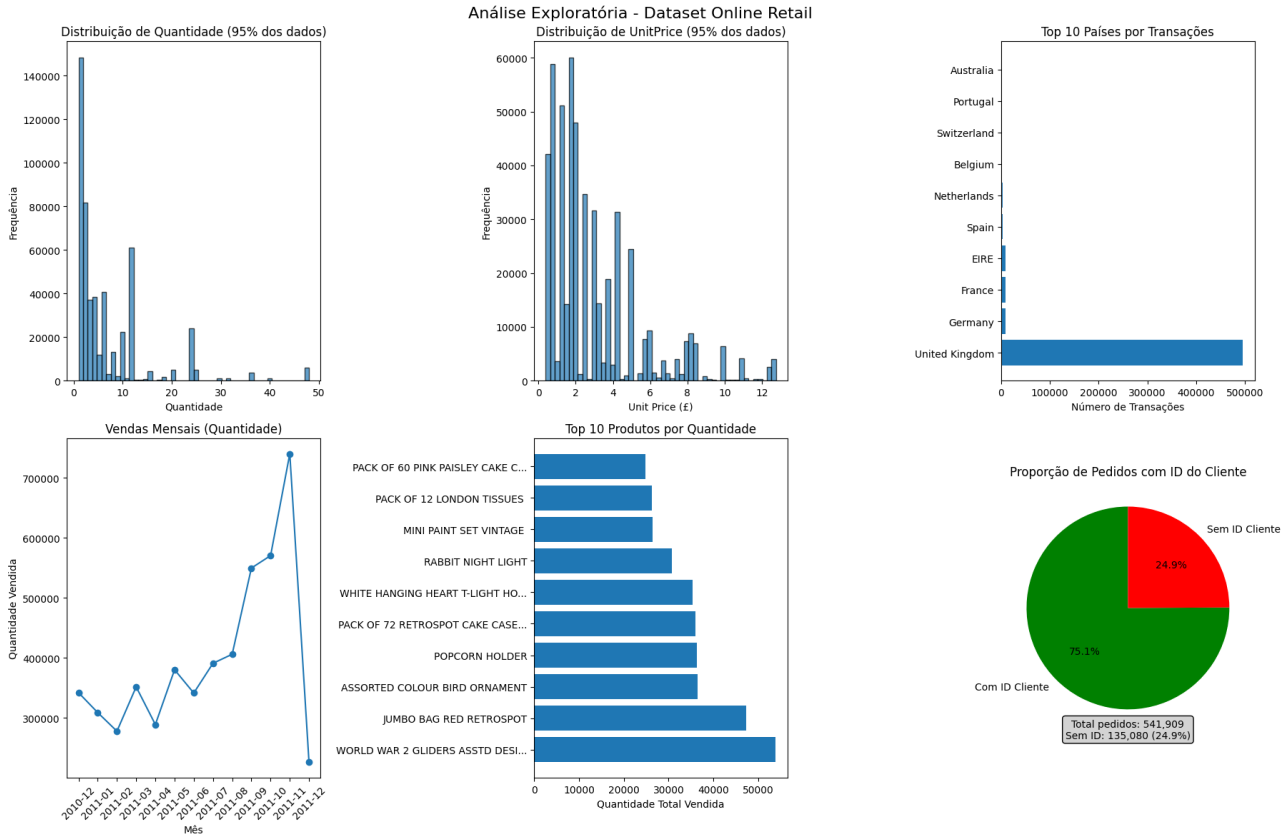
3.2. Estatísticas Descritivas

- **Quantidade:** média de 9,55 unidades por pedido, com valores extremos (mín. –80.995; máx. 80.995), indicando devoluções;
- **Preço unitário:** média de £4,61, com outliers (máx. £38.970);
- **Clientes:** 4.068 únicos, com 24,93% de IDs ausentes.

3.3. Distribuição Geográfica

- **Total de países:** 38;
- **Top 5:** Reino Unido (495.478), Alemanha (9.495), França (8.557), Irlanda (8.196), Espanha (2.533).

3.4. Visualizações Exploratórias



- **Histogramas:** quantidade e preço unitário apresentam distribuição assimétrica, concentrando-se em valores baixos;
- **Vendas mensais:** pico em novembro de 2011, possivelmente devido à *Black Friday*;
- **Produto mais vendido:** *World War 2 Gliders*;
- **Pedidos sem CustomerID:** 24,9% do total, com ticket médio inferior (£11,79 vs. £16,88 com ID).

3.5. Análise de Valores Ausentes

- *CustomerID*: 135.080 ausentes (24,93%);
- *Description*: 1.454 ausentes (0,27%);
- **Impacto:** pedidos sem ID totalizam £199.725,67 em vendas.

3.6. Inferências e Pré-Processamento

- *InvoiceNo* iniciados por **“A”**: ajustes contábeis — removidos;
- *InvoiceNo* iniciados por **“C”**: devoluções — mantidos;
- *Quantity* negativas: devoluções válidas;
- *UnitPrice = 0*: ajustes de estoque — removidos;
- Criação da variável **TotalPrice = Quantity × UnitPrice**.

4. Mineração de Dados

4.1. Regras de Associação (Apriori)

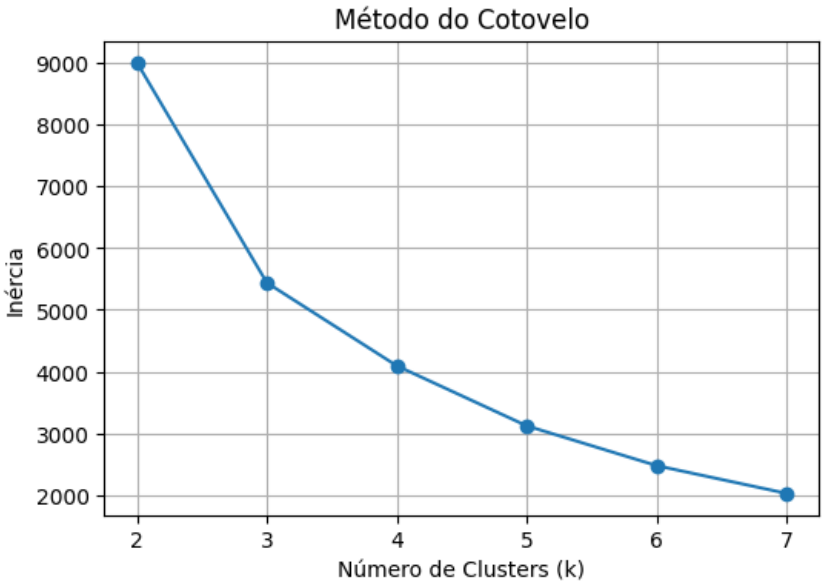
	antecedents	consequents	support	confidence	lift
73	GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	0.020485	0.557190	24.216650
72	PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.020485	0.890339	24.216650
74	PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	0.020485	0.691684	24.188581
71	GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.020485	0.716387	24.188581
5	GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.024270	0.660131	22.289120
4	PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.024270	0.819473	22.289120
70	PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.020485	0.844059	20.723028
75	ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER	0.020485	0.502950	20.723028
7	ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.028595	0.702065	19.095706
6	GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.028595	0.777778	19.095706

Principais insights:

- Produtos da linha *Regency Teacup and Saucer* (cores Green, Pink e Roses) estão fortemente correlacionados;
- *Lift* acima de 20 indica forte dependência — quem compra uma cor tende a adquirir as demais;
- **Estratégias recomendadas:**
    - Criação de *kits promocionais* contendo as três cores;
    - *Cross-selling* automatizado no e-commerce (“Complete sua coleção”);
    - Descontos progressivos e campanhas sazonais.

4.2. Clusterização de Clientes (RFM + K-Means)

Método: análise de *Recency*, *Frequency* e *Monetary*.



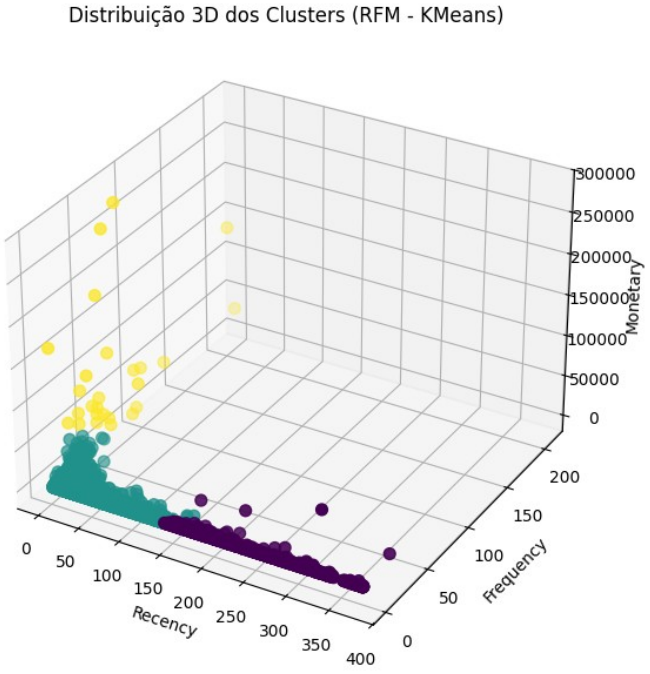
O **método do cotovelo** indicou **k = 3**, resultando em três perfis distintos:

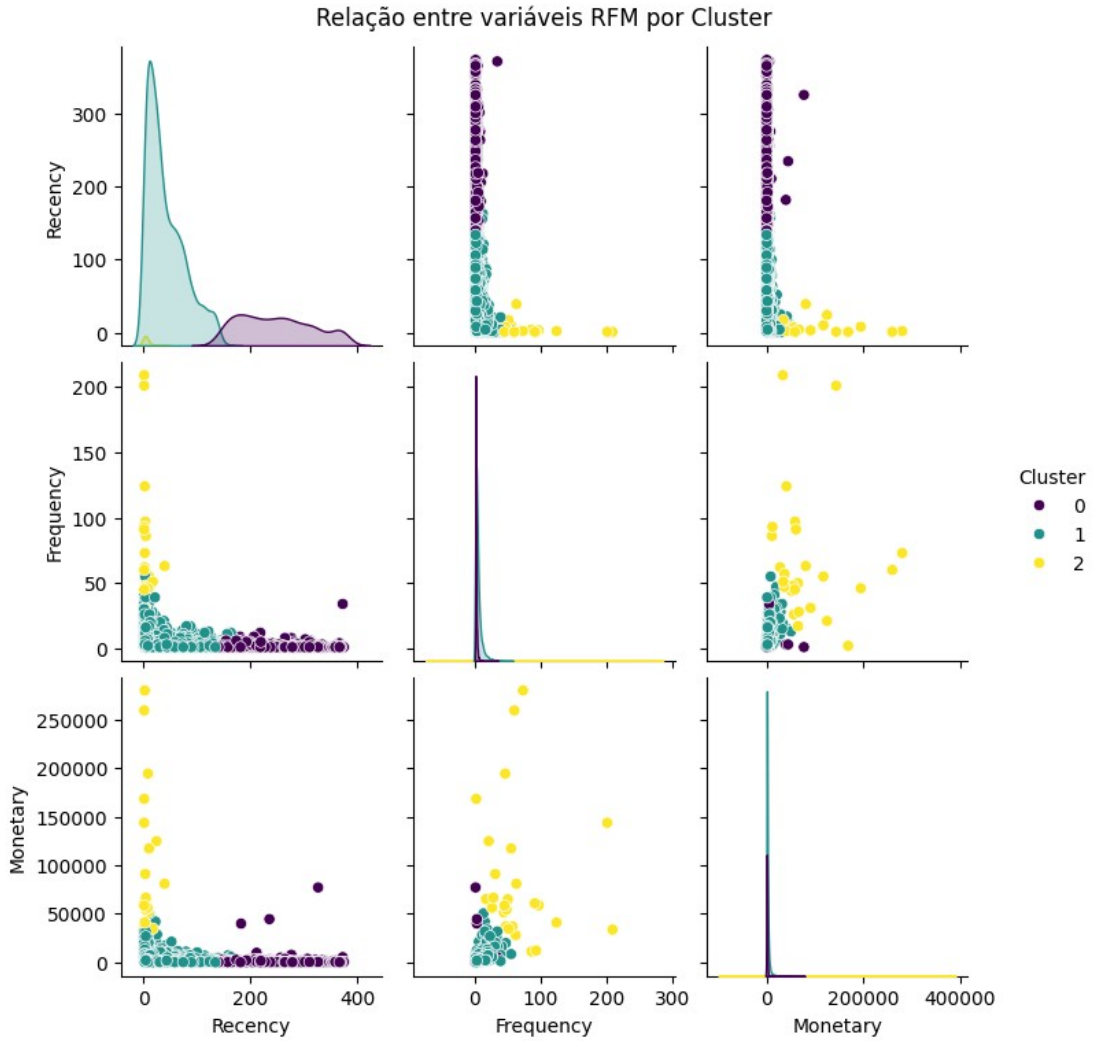
Cluster	Recency	Frequency	Monetary (£)	Qtd. Clientes	Interpretação
0	247.11	1.58	631.42	1082	Clientes inativos ou ocasionais
1	41.45	4.67	1.855,94	3230	Clientes regulares / fiéis
2	6.04	66.42	85.904,35	26	Clientes VIP

Principais Insights:

- **Cluster 2 (VIPs):** clientes de altíssimo valor; requerem tratamento premium;
- **Cluster 1 (Fieis):** principal base ativa — foco em fidelização e aumento do ticket médio;
- **Cluster 0 (Inativos):** potenciais para reativação via cupons e campanhas direcionadas.

Visualizações complementares:





A visualização 3D (Recency × Frequency × Monetary) mostra clara separação entre os grupos, validando a eficácia da segmentação.

---

5. Conclusões

- **Reino Unido** domina 91,4% das transações;
- **Sazonalidade** marcada — pico em novembro;
- **Pedidos sem CustomerID** representam lacuna de informação relevante;
- **Outliers** e devoluções foram tratados adequadamente para integridade da análise.

Após a mineração de dados:

- Estratégias segmentadas podem maximizar resultados:
    - *Cluster 0:* campanhas de reativação;
    - *Cluster 1:* programas de fidelidade e *cross-selling*;
    - *Cluster 2:* atendimento premium e exclusividade.

Relevância prática:

- Segmentação permite campanhas personalizadas e otimização de recursos;
- Análise de associação sustenta ações de *bundle* e recomendação inteligente;
- RFM e Apriori integrados fornecem base sólida para decisões *data-driven*.

---

6. Próximos Passos

1. Implementar kits e promoções baseadas nas regras de associação (*Coleção Regency*);
2. Criar campanhas segmentadas conforme clusters;
3. Desenvolver *dashboard* para monitoramento de comportamento e migração de clientes;
4. Avaliar impacto das estratégias sobre retenção e ticket médio.

A abordagem integrada entre **associação e clusterização** promove um ciclo de:

Entender → Segmentar → Agir → Mensurar

---

7. Referências

- UCI Machine Learning Repository — *Online Retail Dataset*
- Notebook da análise: `AD2.ipynb`

Conclusão Final

A análise exploratória e a mineração de dados aplicadas ao conjunto **Online Retail** revelam padrões consistentes e úteis para ações estratégicas de marketing, fidelização e personalização.
O relatório consolida uma abordagem **orientada a dados**, com potencial direto para **transformação digital e vantagem competitiva sustentável**.