Centro Paula Souza
Faculdade de Tecnologia de Votorantim

Curso de Ciência de Dados para Negócios

Projeto Integrador III - Análise de Dados I

Stefanie Mayumi Inacio Kobayashi

03/2025

Este projeto aborda o problema central enfrentado por muitas lojas virtuais: Como utilizar os dados históricos de transações para identificar padrões de compra significativos, segmentar a base de clientes de forma eficiente e inteligente, e identificar fatores que levam aos cancelamentos, com o objetivo final de aumentar a rentabilidade do negócio? Foram utilizadas as transações registradas no arquivo data.csv, buscando extrair insights valiosos que possam guiar a tomada de decisão estratégica da empresa.

1. Carregamento e Exploração Inicial do Dataset

O conjunto de dados utilizado neste projeto consiste em transações de um e-commerce internacional, fornecido no arquivo data.csv. O carregamento inicial foi realizado utilizando a biblioteca Pandas, especificando a codificação latin-1 devido a possíveis caracteres especiais.

A exploração inicial revelou um dataset com 541909 linhas e 8 colunas: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID e Country.

Realizando análises iniciais foi possível identificar os seguintes pontos:

- A coluna InvoiceDate estava no formato object (string), necessitando conversão para datetime para análises temporais;
- A coluna CustomerID estava como float64 e continha um número significativo de valores ausentes (NaN), representando transações sem identificação do cliente, sendo um desafio direto para a segmentação de clientes;
- A coluna Description também possuía valores ausentes;
- Presença de valores negativos em Quantity (indicando possíveis cancelamentos/devoluções) e UnitPrice igual a zero em algumas transações.

2. Limpeza e Preparação dos Dados

Para preparar os dados para análise e modelagem, algumas etapas de limpeza e transformação foram realizadas, resultando em um novo dataframe chamado df_so:

- Tratamento de Outliers (Quantidade): foi possível identificar a presença de valores extremos na coluna Quantity. Utilizando o método do Intervalo Interquartil (IQR) para definir limites inferior e superior, registros com Quantity fora desses limites foram removidos;
- Remoção de Duplicatas: linhas duplicadas exatas foram removidas do dataframe filtrado;

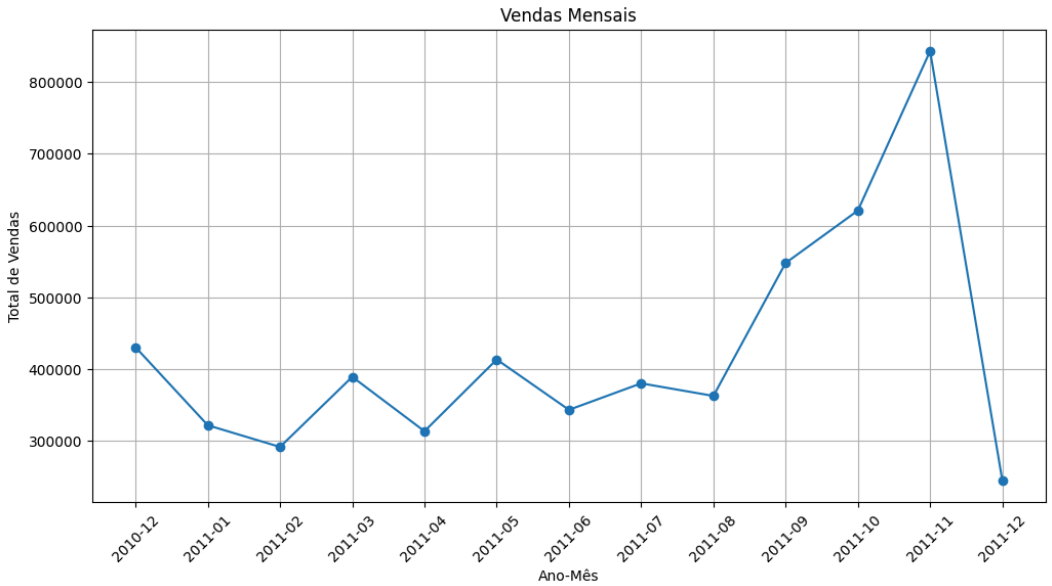
- Conversão de Tipos: a coluna InvoiceDate foi convertida para o formato datetime, e CustomerID foi convertido para o tipo string;
- Criação de novas colunas: foram criadas novas colunas para facilitar a análise: TotalPrice: Calculada multiplicando Quantity por UnitPrice, representando o valor total por item da transação. YearMonth: extraído de InvoiceDate, agrupando as transações por mês e ano para análise de tendências sazonais.

Após essas etapas, o dataset foi consideravelmente reduzido, resultando em 478140 linhas prontas para a análise exploratória.

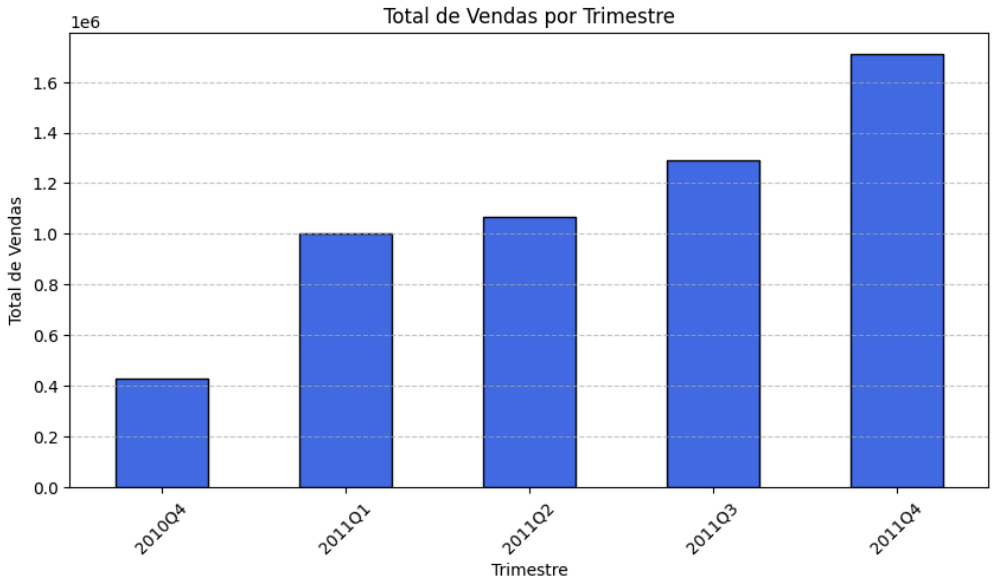
3. Análise Exploratória de Dados (EDA) e Insights

Após a limpeza e preparação, o dataset df_so foi submetido a uma Análise Exploratória de Dados (EDA) para descobrir padrões de compra, entender associações entre produtos e investigar os cancelamentos. Foram utilizadas visualizações e análises estatísticas para extrair insights relevantes.

- Vendas Mensais: um gráfico de linha foi gerado para visualizar as vendas por mês. É possível observar um pico de vendas em novembro de 2011 e um crescimento geral nas vendas ao longo do período.



- Vendas Trimestrais: para uma visão mais agregada, as vendas foram agrupadas por trimestre (Quarter) e visualizadas em um gráfico de barras. O gráfico de barras trimestral confirmou a tendência de crescimento ao longo dos trimestres.



- Produtos Mais Vendidos por Trimestre: foi realizada uma análise para identificar os 3 produtos com maior venda em cada trimestre. Essa análise mostrou que os principais produtos em termos de receita foram DOTCOM POSTAGE e REGENCY CAKESTAND 3 TIER. Isso ajuda a entender quais itens são os motores de receita da empresa e se há sazonalidade na popularidade dos produtos.

	Quarter	Description	TotalPrice
779	2010Q4	DOTCOM POSTAGE	24671.19
1994	2010Q4	REGENCY CAKESTAND 3 TIER	9103.00
1872	2010Q4	POSTAGE	4386.00
3493	2011Q1	DOTCOM POSTAGE	35808.81
4812	2011Q1	REGENCY CAKESTAND 3 TIER	26267.27
4677	2011Q1	POSTAGE	13159.83
6410	2011Q2	DOTCOM POSTAGE	29613.34
7799	2011Q2	REGENCY CAKESTAND 3 TIER	24530.96
7463	2011Q2	PARTY BUNTING	15134.95
9467	2011Q3	DOTCOM POSTAGE	41418.92
10850	2011Q3	REGENCY CAKESTAND 3 TIER	22762.74
10721	2011Q3	POSTAGE	15825.57
12544	2011Q4	DOTCOM POSTAGE	74733.22
13899	2011Q4	REGENCY CAKESTAND 3 TIER	21136.89
13770	2011Q4	POSTAGE	19766.90

- Países com Maior Volume de Vendas por Trimestre: Similarmente, foram identificados os 3 países com maior venda em cada trimestre. Os resultados mostraram que os principais mercados foram United Kingdom e Germany. Isso direciona os esforços de marketing e logística para as regiões mais lucrativas.

	Quarter	Country	TotalPrice
21	2010Q4	United Kingdom	393089.050
9	2010Q4	Germany	10211.350
6	2010Q4	EIRE	7083.210
51	2011Q1	United Kingdom	863427.990
33	2011Q1	Germany	27934.420
32	2011Q1	France	26610.620
82	2011Q2	United Kingdom	925963.941
65	2011Q2	Germany	31789.350
61	2011Q2	EIRE	25578.570
111	2011Q3	United Kingdom	1102066.573
94	2011Q3	Germany	39027.050
90	2011Q3	EIRE	39009.950
141	2011Q4	United Kingdom	1499475.440
123	2011Q4	Germany	46142.520
122	2011Q4	France	40274.790

- Análise de Cesta de Compras (Market Basket Analysis - Foco no Reino Unido): para identificar produtos frequentemente comprados juntos, foi aplicada a análise de regras de associação utilizando o algoritmo Apriori. Esta análise focou especificamente nas transações do Reino Unido e considerou apenas compras efetivadas. Essa análise mostrou que os produtos ROSES REGENCY TEACUP AND SAUCER e GREEN REGENCY TEACUP AND SAUCER são frequentemente comprados juntos de PINK REGENCY TEACUP AND SAUCER, a chance é de quase 70% dos clientes que compram os dois primeiros produtos (ROSES REGENCY TEACUP AND SAUCER e GREEN REGENCY TEACUP AND SAUCER) também comprarem o terceiro (PINK REGENCY TEACUP AND SAUCER).

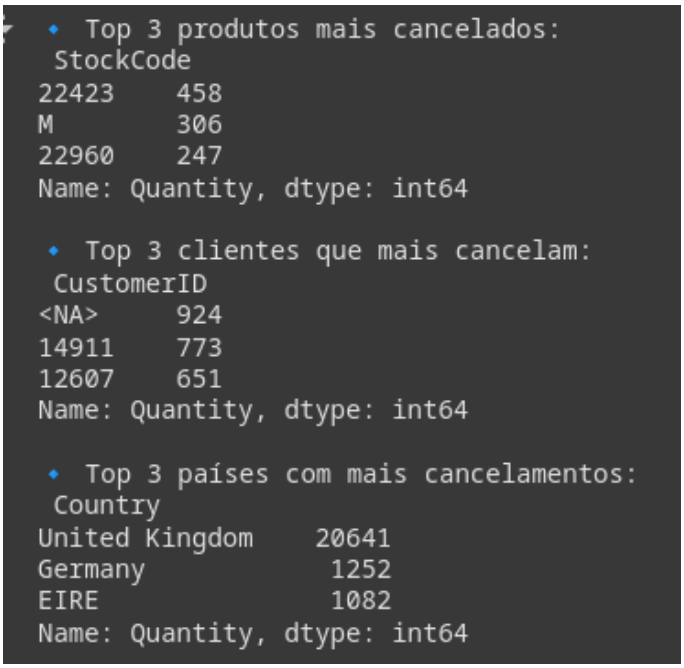
```
       antecedents     consequents  antecedent support  consequent support  \
190  (22699, 22697)         (22698)            0.037317            0.038020
195         (22698)  (22699, 22697)            0.038020            0.037317
191  (22699, 22698)         (22697)            0.028998            0.050615
194         (22697)  (22699, 22698)            0.050615            0.028998
192  (22697, 22698)         (22699)            0.030990            0.051025

      support  confidence       lift  representativity  leverage  conviction  \
190  0.026069    0.698587  18.374241               1.0  0.024650    3.191569
195  0.026069    0.685670  18.374241               1.0  0.024650    3.062653
191  0.026069    0.898990  17.761293               1.0  0.024601    9.398910
194  0.026069    0.515046  17.761293               1.0  0.024601    2.002257
192  0.026069    0.841210  16.486167               1.0  0.024488    5.976282

     zhangs_metric   jaccard  certainty  kulczynski
190       0.982230  0.529132   0.686675    0.692129
195       0.982948  0.529132   0.673486    0.692129
191       0.971881  0.486871   0.893605    0.707018
194       0.994010  0.486871   0.500564    0.707018
192       0.969384  0.465969   0.832672    0.676058
```

Essas regras são valiosas para estratégias de *cross-selling* (oferecer produtos complementares), organização de layout de loja (física ou virtual) e criação de promoções combinadas (bundles), especialmente para o mercado do Reino Unido.

- Análise de Cancelamentos: as transações que representam cancelamentos (identificadas por InvoiceNo começando com 'C' e separadas anteriormente em df_cancelamentos) foram analisadas para identificar padrões. Os 3 produtos com maior volume (em quantidade absoluta) de cancelamentos foram REGENCY CAKESTAND 3 TIER, Manual e JAM MAKING SET WITH JARS. Já os 3 clientes com mais cancelamentos foram <NA> (possivelmente compras sem identificação), ID 14911 e ID 12607. E os 3 países com mais cancelamentos foram United Kingdom, Germany e EIRE.



Identificar os produtos, clientes e países mais associados a cancelamentos permite investigar as causas raiz (ex: problemas de qualidade do produto, descrição inadequada, problemas de logística em certos países, comportamento específico de certos clientes) e direcionar ações para mitigar essas ocorrências.

4. **Segmentação de Clientes e Recomendações Estratégicas**

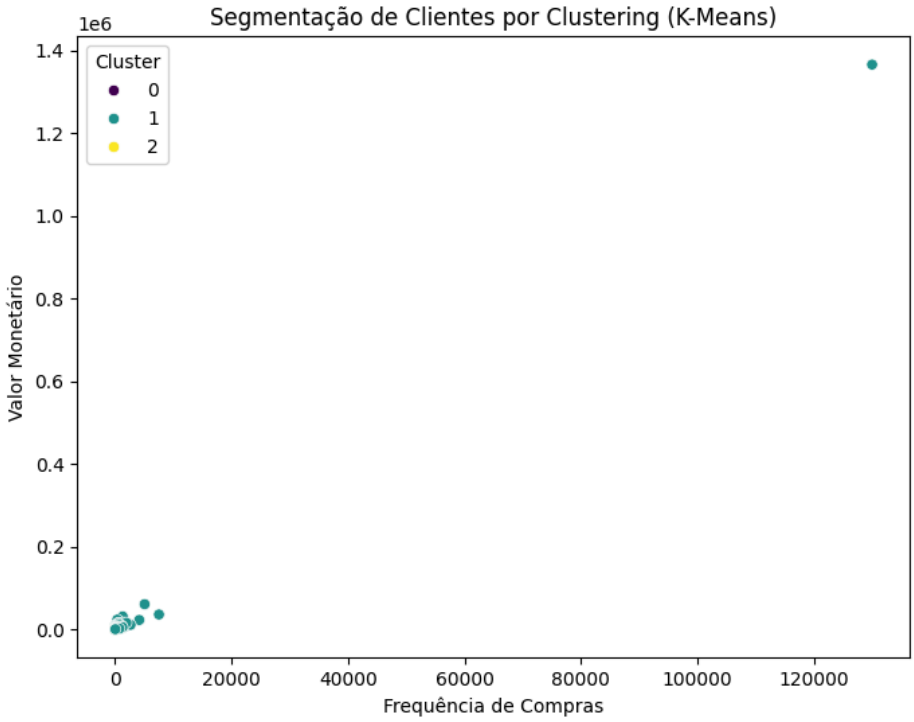
Para direcionar as estratégias de marketing e vendas de forma mais eficaz, foi realizada uma segmentação dos clientes com base em seu comportamento de compra.

A metodologia escolhida combinou a análise RFM (Recência, Frequência, Valor Monetário) com a técnica de agrupamento (clustering) K-Means.

- Métricas RFM: Antes de aplicar o algoritmo de clusterização, foram calculadas as três métricas fundamentais do RFM para cada cliente (CustomerID) identificado no dataset:
    - Recência (Recency): Número de dias desde a última compra do cliente até a data mais recente presente no dataset. Clientes com menor Recência compraram mais recentemente.
    - Frequência (Frequency): Número total de transações (compras) realizadas pelo cliente. Clientes com maior Frequência compram mais vezes.
    - Valor Monetário (Monetary): Soma total do valor gasto pelo cliente em todas as suas compras (TotalAmount). Clientes com maior Valor Monetário gastam mais.

	CustomerID	Frequency	Monetary	Recency
0	12747	83	2100.69	1
1	12748	4143	22912.81	0
2	12749	188	3856.22	3
3	12820	49	824.14	2
4	12821	5	77.12	213

- Clusterização com K-Means: com base na análise do gráfico do Método do Cotovelo, optou-se por segmentar os clientes em 3 clusters. O algoritmo K-Means foi então aplicado aos dados RFM padronizados.



- Análise e Descrição dos Clusters: cada cliente foi atribuído a um dos 3 clusters. Para entender as características de cada segmento, foram calculadas as médias e desvios padrão das métricas RFM originais dentro de cada cluster:
    - Cluster 0: Recency (Média: 294,1, Desvio Padrão: 44,36), Frequency (Média: 21,9, Desvio Padrão: 27,0), Monetary (Média: 288,04, Desvio Padrão: 373,96). Esses clientes possuem uma alta recência, indicando que não fazem compras frequentemente. O gasto médio é baixo e com grande variação. Uma estratégia seria um reaquecimento desses clientes com campanhas de marketing direcionadas e ofertas.
    - Cluster 1: Recency (Média: 32,21, Desvio Padrão: 25,89), Frequency (Média: 154,4, Desvio Padrão: 2566,91), Monetary (Média: 1836,97, Desvio Padrão: 26941,04). Esses clientes são frequentes e gastam muito, mas os dados possuem grande variação, indicando que alguns clientes são mais fiéis que outros, ou que fazem compras esporádicas. Uma estratégia é fidelizar o cliente com bônus ou desconto a cada número de compra realizada.
    - Cluster 2: Recency (Média: 158,73, Desvio Padrão: 37,16), Frequency (Média: 35,31, Desvio Padrão: 44,39), Monetary (Média: 469,14, Desvio Padrão: 493,95). Esses clientes estão menos ativos que os do cluster 1, mas ainda realizam compras com uma frequência moderada, indicando um potencial de crescimento. Uma estratégia é oferecer descontos para aumentar a frequência de compras e também programas de fidelidade.

5. Insights e Recomendações Finais

Este projeto analisou os dados de transações de e-commerce com o objetivo de identificar padrões de compra, segmentar clientes e encontrar oportunidades para aumentar a lucratividade e reduzir cancelamentos. Através de um processo de limpeza de dados, análise exploratória (EDA) e segmentação de clientes (RFM + K-Means).

O dataset inicial apresentava desafios significativos, como um alto volume de CustomerID ausentes, dados duplicados e outliers em Quantity. A etapa de limpeza e preparação foi crucial para garantir a confiabilidade das análises subsequentes, resultando em um dataset mais coeso, embora reduzido em tamanho. A necessidade de tratar dados ausentes, especialmente CustomerID, é vital para análises focadas no cliente.

A análise temporal revelou um crescimento geral nas vendas ao longo do período analisado, e a análise de cesta de compras, focada no Reino Unido, identificou

associações significativas entre produtos, revelando oportunidades claras para estratégias de cross-selling nesse mercado.

Também foi possível identificar os produtos, clientes e países mais frequentemente associados a cancelamentos. Esses padrões sugerem áreas específicas que necessitam de investigação para entender as causas raiz (qualidade, logística, descrição, etc.) e implementar ações corretivas.

A análise RFM combinada com K-Means (k=3) mostrou que a base de clientes não é homogênea. Foi possível identificar segmentos distintos com comportamentos e valores diferentes.

Com base nos insights gerados, as seguintes ações estratégicas são recomendadas para a empresa de e-commerce:

- Otimizar Marketing e Estoque com Base na Sazonalidade: alinhar campanhas de marketing, promoções e gestão de estoque aos padrões sazonais identificados.
- Focar nos Principais Produtos e Mercados: garantir a disponibilidade e visibilidade dos produtos mais vendidos. Concentrar esforços de marketing e otimização logística nos mercados mais lucrativos, com atenção especial ao Reino Unido.
- Implementar Estratégias de Cross-Selling (UK): Utilizar as regras de associação descobertas para criar ofertas combinadas (bundles), sugestões de "compre junto" e campanhas de e-mail marketing direcionadas no mercado do Reino Unido.
- Adotar Marketing Personalizado Baseado em Segmentos: implementar as estratégias de marketing e relacionamento diferenciadas propostas para cada um dos 3 segmentos de clientes identificados.

A análise de dados realizada forneceu uma compreensão do comportamento dos clientes e das dinâmicas de venda do e-commerce. A aplicação dos insights e recomendações, especialmente estratégias personalizadas baseadas na segmentação de clientes e a abordagem proativa para redução de cancelamentos, tem o potencial de aumentar significativamente a retenção de clientes, o ticket médio e, consequentemente, a lucratividade geral do negócio.