Mapeando perfis de consumidores de E-Commerce com RFM e K-Means

Ewerthon J. Kutz, Helen C. M. Senefonte

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina– PR – Brasil

ewerthon.jose.kutz@uel.br, helen@uel.br

Abstract. *Understanding consumers is essential for effective decision-making in an integrated and competitive ecosystem. However, obtaining their characteristics and leveraging them to personalize services and allocate resources efficiently is a non-trivial challenge. This article analyzes data from an e-commerce business with the objective of segmenting its customers using the RFM (Recency, Frequency, Monetary) method combined with the K-Means machine learning algorithm, aiming to interpret these groups for actionable strategies and explore the segmentation process. As a result, four distinct behavioral segments are identified, providing strategic insights into retail consumption in the digital environment.*

Resumo. *Compreender os consumidores é fundamental para a tomada de decisão assertiva em um ecossistema integrado e competitivo. Contudo, obter suas características e trabalhá-las de forma a personalizar serviços e alocar recursos eficientemente é um desafio não trivial. Este artigo analisa dados de um negócio de e-commerce com o objetivo de segmentar seus consumidores a partir do método RFM (Recency, Frequency, Monetary) com o algoritmo K-Means, visando a interpretação desses grupos para ações acionáveis e a exploração do processo de segmentação. Como resultado, quatro segmentos com comportamentos distintos foram identificados, fornecendo insights estratégicos sobre o consumo do varejo no ambiente digital.*

1. Introdução

A utilização de dados gerados por atividades digitais é um catalisador de soluções inovadoras em cidades modernas, capazes de transformar tanto sistemas econômicos quanto a experiência cotidiana dos cidadãos [Ma *et al*. 2024]. O comércio eletrônico (e-commerce) emerge como um componente central desse ecossistema no Brasil, com o setor registrando um crescimento de 351,3% na última década e atingindo R$ 87,4 bilhões em faturamento e 79,7 milhões de consumidores durante a pandemia de Covid-19 em 2020, fato que reforçou sua relevância socioeconômica [Cruz 2021].

Esses consumidores representam a maioria da população economicamente ativa brasileira e compreendê-los por meio de técnicas inteligentes de mineração e análise de dados permite entender a dinâmica de consumo geral e promove maior capacidade de compra do público geral, conveniência de personalização e redução de custos de transações, especialmente em mercados emergentes [Pan et al. 2021].

Analisar as características comportamentais desse mercado também possibilita otimizar as estratégias dos fornecedores, uma vez que esses dados permitem que lojistas se antecipem às tendências de mercado [Cruz 2021]. Além disso, com as estratégias de marketing corretas, o e-commerce representa liberdade e personalização para o consumidor, favorecendo uma experiência de compra assertiva [Cruz 2021].

O desempenho dessa indústria também exerce impacto nos setores de desenvolvimento de software, logística, marketing digital e atendimento ao cliente, inclusive promovendo sustentabilidade com entregas eficientes e sistemas de transporte inteligentes [Gupta *et al*. 2023] e seus dados, integrados a fontes heterogêneas, são fatores relevantes para o desenvolvimento de cidades inteligentes, mas essa integração permanece sendo um desafio a ser superado [Almeida *et al*. 2020].

Agrupar consumidores segundo suas características é um fator determinante de sucesso nesse setor [Amutha e Khan 2023] e dados demográficos ou valor total de gastos são comumente usados para abordar essa atividade [Tang *et al* 2024]. Contudo, esses dados não capturam aspectos comportamentais essenciais para a tomada de decisão estratégica. Para isso, é possível utilizar a metodologia RFM, onde consumidores são agrupados em diferentes quadrantes segundo três dimensões comportamentais, possibilitando aplicar estratégias mais personalizadas e com maiores chances de sucesso [Christy *et al.* 2021].

Este trabalho visa explorar dados de venda de uma empresa de e-commerce e da segmentação de seus clientes utilizando a metodologia RFM e o algoritmo K-Means. Após a coleta, pré-processamento, investigação e segmentação, espera-se que o trabalho possa contribuir com *insights* acerca do setor de e-commerce e dos desafios do processo de segmentação. Além disso, o trabalho visa fornecer análises do comportamento dos consumidores e proporcionar a aplicação de estratégias de decisão orientada por dados.

O trabalho é estruturado da seguinte forma: na Seção 2 são levantados os principais trabalhos relacionados e estado da arte; Na Seção 3, a solução é descrita e os resultados são apresentados na Seção 4. Por fim, a Seção 5 contém as considerações finais.

2. Levantamento Bibliográfico

Esta seção contém a discussão dos trabalhos correlatos ao tema e método proposto e a apresentação do estado da arte do levantamento bibliográfico.

2.1. Trabalhos Correlatos

O estudo de Wei *et al.* aborda os diferentes métodos de extração e segmentação de características comportamentais de consumidores utilizando a metodologia RFM e a evolução dessas técnicas ao longo do tempo. Nesse artigo, o modelo padrão de segmentação é definido como aquele que ranqueia e divide as dimensões RFM dos consumidores e divide-os em quintis, formando "células" que vão de 111 a 555 e indicam a qual grupo um consumidor pertence [Wei *et al.* 2010]. O autor contribui com a pesquisa ao apontar as limitações do modelo RFM e suas variações em termos de interpretabilidade de características dos grupos e sua tendência de negligenciar a representatividade de consumidores com baixas pontuações em alguma de suas dimensões.

Amutha e Khan propõem o uso de modelos de aprendizado de máquina para segmentar clientes a partir de seu histórico de compras como uma abordagem válida e aplicam os algoritmos de aprendizado não supervisionado K-Means, Gaussian Mixture Model e

DBSCAN, avaliando-os a partir do método do cotovelo, que determina o número ideal de centróides (k) para a segmentação a partir da soma do quadrado das distâncias euclidianas intra-segmentos para diferentes valores de k [Amutha e Khan 2023].

O trabalho de Christy *et al.* oferece uma análise abrangente sobre a segmentação de clientes com RFM e detalha como esse método pode ser aplicado para classificar os clientes com base em seu comportamento de compra, combinando-o com os algoritmos de aprendizado de máquina K-Means e Fuzzy C-Means, apresentando uma abordagem para a escolha dos centróides iniciais que resulta em uma redução significativa no número de iterações para alcançar uma clusterização eficaz [Christy *et al.* 2021].

A abordagem Anitha e Patil apresenta uma aplicação inovadora para identificar clientes potenciais no setor de e-commerce ao unir os resultados do modelo RFM com análise exploratória de dados. Além disso, a pesquisa valida os segmentos obtidos com o cálculo do coeficiente de silhueta, avaliando a coesão e separação dos grupos formados. Essa pesquisa aborda a segmentação de forma analítica e fornece um modelo prático que pode ser adotado por empresas do setor [Anitha e Patil 2022]. A análise exploratória dos dados realizada pelos autores traz um nível de interpretação para os segmentos, mas limita-se a visualizar a distribuição das dimensões resultantes dos grupos.

Sobre o enriquecimento do modelo RFM com dados heterogêneos, Ho *et al.* destacam-se ao utilizar também dados demográficos, criando o modelo RFMD, também utilizando o algoritmo K-Means. Esse modelo permite relacionar características comportamentais e demográficas e serem eficientes na abordagem de consumidores [Ho *et al.* 2023].

A utilização das análises de resultados de RFM com K-Means é explorada por Yoshida *et al.*, onde consumidores são agrupados para promover melhoria de vendas em indústrias por meio de recomendações e aumentar eficiência de um portfolio de produtos [Yoshida *et al.* 2014]. Já no trabalho de Lewaa, segmentos de consumidores resultantes do modelo K-Means são classificados em “Potenciais”, “Não podem ser perdidos”, “Em risco” e “Perdidos” [Lewaa 2024]. Esses estudos trazem uma melhoria em interpretabilidade. Contudo, não há delimitação e análise das características dos grupos ou comparação com a interpretação de outros modelos ou com o modelo RFM padrão.

A descrição das características comportamentais de consumidores com RFM não se limita à segmentação. [Tang et al 2024] utiliza o agrupamento dessas variáveis com K-Means em conjunto com um modelo de aprendizado de máquina supervisionado para analisar e predizer a demanda de eletricidade de consumidores, ressaltando a importância desses dados para o planejamento de negócios e cidades como um todo.

Os trabalhos citados avançaram as fronteiras da abordagem RFM ao utilizar aprendizado de máquina e variações em escolha de centróides e adição de dados heterogêneos aos modelos e, em determinados casos, também se utilizou de análise exploratória nas segmentações. Contudo, essas análises restringiram-se aos resultados dos modelos de aprendizado de máquina e limitaram-se a uma classificação categórica dos segmentos.

Este trabalho visa contribuir com a área ao realizar uma análise exploratória abrangente de dados e comparar a interpretabilidade dos resultados do modelo RFM padrão com o modelo RFM com K-Means de forma iterativa - apoiando-se nas conclusões da análise exploratória e dos resultados dos passos anteriores para garantir que todo os perfis de consumidores sejam representados e obtenha-se segmentos interpretáveis e com características bem definidas.

2.2. Estado da Arte

2.2.1. Modelo RFM padrão

A segmentação RFM é utilizada para o ranqueamento e segmentação de consumidores com base no seu comportamento de compras. Esse método é especialmente útil em setores onde há alto número de clientes realizando transações, como no varejo e no e-commerce [Christy *et al.* 2021]. Nesse método, os clientes são agrupados em três dimensões:

1. Recência: refere-se ao número de dias desde que um consumidor e realizou sua última compra. Para o ranqueamento, quanto menor for esse número, maior é a pontuação de recência. A base de consumidores é dividida em quintis, onde os 20% de clientes mais recentes recebem a pontuação máxima de 5, e os demais recebem pontuações decrescentes até 1 [Hughes 1994, Wei *et al.* 2010].
2. Frequência: é definida como o número de compras que um consumidor fez dentro de um período. Também é classificada de 1 a 5 [Hughes 1994, Wei et al. 2010].
3. Valor Monetário: corresponde ao total de dinheiro gasto pelo consumidor em um período. Nesse caso, a pontuação varia de 1 a 5, com critério análogo à dimensão de Recência [Hughes 1994, Wei *et al.* 2010].

Após obter as três dimensões, os dados são ranqueados e os consumidores são classificados de 111 a 555, gerando um total de 125 "células" RFM, onde a célula 555 representa os melhores consumidores e o segmento 111, os menos engajados [Hughes 1994, Wei *et al.* 2010]. Esse tipo de segmentação é apresentado na figura 1 e permite às empresas identificar padrões comportamentais e otimizar suas estratégias de marketing ao direcionar campanhas específicas para cada grupo.

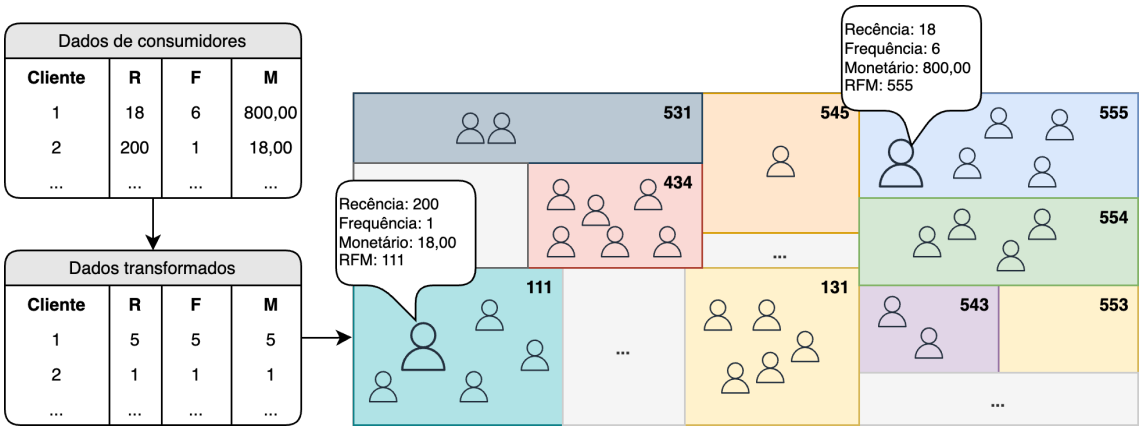

Figura 1. Processo de Segmentação RFM

2.2.2. Segmentação com K-Means

O algoritmo K-Means é um método de agrupamento não-hierárquico de aprendizado de máquina não supervisionado utilizado para segmentação, onde os dados são divididos em um número pré-determinado de clusters (k) [MacQueen 1967, Yoshida *et al.* 2024]. O objetivo principal do K-Means é minimizar a variação dentro de cada cluster e obter grupos compactos. Essa variação intra-cluster é medida pela soma das distâncias quadráticas euclidianas entre as observações e os centróides de seus respectivos clusters [MacQueen 1967, Anitha e Patil 2022].

A distância Euclidiana utilizada no algoritmo, onde x_i representa um ponto específico e c_j representa o centróide do cluster j, é dada pela Equação (1):

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - c_{jm})^2} \quad (1)$$

Para avaliar a qualidade dos agrupamentos, utiliza-se a soma das distâncias quadráticas euclidianas (SDQE). Quanto menor o valor de dessa distância, mais concentrados estão os pontos em torno de seus centróides. Esse método também é conhecido como escore de distorção e, considerando que x_i é um ponto dentro do cluster j, c_j é um centróide e n_j é o número de pontos em j, apresentada na Equação (2):

$$SDQE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_i - c_j||^2 \quad (2)$$

O processo do K-Means envolve associar cada instância ao centróide mais próximo com base na distância euclidiana e recalcular os centróides de forma iterativa com distâncias médias dos agrupamentos. Este processo se repete até que os centróides não mudem significativamente entre as iterações, indicando que o algoritmo atingiu a convergência [Kodinariya e Makwana 2013].

O método do cotovelo é a técnica mais comum para determinar o número ideal de clusters (k) desse algoritmo. Nesse método, inicia-se com $k = 2$ e aumenta-se progressivamente esse número, comparando o SDQE de cada iteração. O "cotovelo" é o ponto onde essa métrica começa a reduzir de forma menos acentuada, indicando um equilíbrio entre a variabilidade explicada e o número de clusters [Yoshida *et al.* 2024, Marutho *et al.* 2018]. Apesar do valor de SDQE não ser interpretável isoladamente, compará-lo entre os valores de k é válido para determinar um número de centróides considerado ótimo [Kodinariya e Makwana 2013, Marutho *et al.* 2018].

2.2.3. Transformação Box-Cox

A transformação Box-Cox busca aproximar a distribuição de um conjunto numérico à uma distribuição gaussiana, normalizando-os. Essa transformação pode melhorar resultados de algoritmos de aprendizado de máquina, especialmente em casos com outliers e alta assimetria [Blum et al 2022]. Essa transformação depende de um parâmetro λ aplicado a uma variável y, apresentado na Equação (3):

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{(\lambda)} - 1}{\lambda}, & se\ \lambda \neq 0 \\ \ln(y_i), & se\ \lambda = 0 \end{cases} \quad (3)$$

Para encontrar o valor ótimo de λ, pode-se utilizar o método de máxima verossimilhança ($L_{\max(\lambda)}$), com $n \in [-2, 2]$ [Li *et al.* 2022]. Seu cálculo é apresentado na Equação (4):

$$L_{\max(\lambda)} = -\frac{n}{2} ln\left[\sum_{i-0}^{n-1} \frac{\left(x_i(\lambda) - \bar{x}(\lambda)\right)^2}{n}\right] + (\lambda - 1)\sum_{i=0}^{n-1} \ln(x_i), onde: \quad (4)$$

$$\bar{x}(\lambda) = \frac{1}{n}\sum_{i=0}^{n-1} x_i\,(\lambda)$$

3. Descrição dos Dados e Metodologia

A segmentação de consumidores foi aplicada em dados de uma empresa de varejo em e-commerce de atuação nacional. Para isso, os dados foram coletados, tratados e analisados de forma exploratória com técnicas estatísticas e visuais.

Após a conclusão da análise exploratória, aplicou-se o modelo RFM tradicional aos dados. Em seguida, utilizou-se o algoritmo K-Means para realizar uma nova segmentação com o método RFM. Posteriormente, os dados foram submetidos a uma transformação com o método Box-Cox e o mesmo algoritmo de clusterização foi reaplicado. O processo completo de segmentação é apresentado na figura 2.

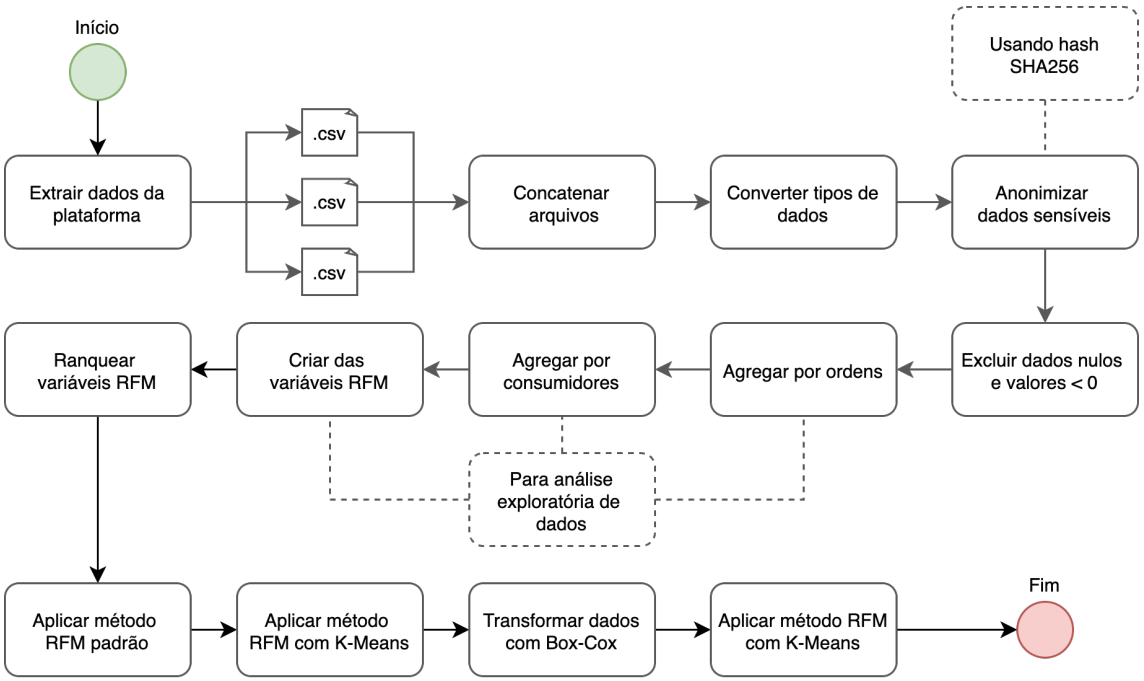

Figura 2. Processo de Análise e Segmentação

Cada método de segmentação foi analisado com o objetivo de entender melhor o mercado de e-commerce e a dinâmica de consumo do varejo online em termos de interpretabilidade das características e separação dos grupos gerados.

3.1. Descrição dos Dados

Os dados foram coletados a partir de extrações na plataforma de e-commerce do negócio em formato CSV (Comma-separated values). O período de corte dos dados é de 02/01/2023 até 11/11/2024 e cada instância do conjunto de dados representa um item de uma ordem de compra. O processo de tratamento dos dados incluiu:

1. Concatenação dos arquivos extraídos.

2. Conversão dos dados para seus tipos corretos.
3. Anonimização de dados sensíveis do consumidor com técnicas de hashing.
4. Exclusão de linhas com IDs nulos e com valores menores que zero.
5. Agregação dos dados por ordens e por consumidor.
6. Criação das variáveis RFM (recência, frequência e valor monetário).

O dataset final para aplicação dos modelos de segmentação contém a identificação anonimizada e única do consumidor e as variáveis RFM.

3.2. Análise Exploratória e Segmentação

Visando investigar as particularidades dos dados, uma análise exploratória foi realizada. Esse processo abrange o cálculo de medidas de localização (média e mediana), variabilidade (desvio padrão) e distribuição (assimetria e curtose) para observar diferentes aspectos das variáveis [Bruce P. e Bruce A. 2019].

Para complementar a análise, as variáveis RFM foram observadas de forma visual com o a utilização de gráficos de densidade. Esse tipo de Gráfico apresenta a distribuição das frequências dos valores em uma linha contínua, permitindo observar as concentrações de densidade e proporção dos dados ao longo de sua distribuição [Bruce P. e Bruce A. 2019]. A análise exploratória possibilita maior entendimento das variáveis em termos estatísticos e comportamentais, sendo valiosa para entender os consumidores e servindo como base para a aplicação das técnicas de segmentação e interpretação de seus resultados.

O processo de segmentação utilizado foi iterativo, onde cada método aplicado beneficiou-se dos resultados anteriores para aproximar-se dos objetivos propostos pelo trabalho. A escolha desse formato deve-se à desvantagem do modelo RFM padrão quando aplicado a dados assimétricos, que pode desconsiderar consumidores com pontuações mais baixas [Wei *et al.* 2010]. Esse modelo também não gera centróides para seus segmentos. Contudo, pode-se calculá-los a partir do ponto médio das posições dos indivíduos de cada grupo, possibilitando obter seu escore de distorção e compará-lo com os demais modelos.

Para capturar padrões de comportamento mais sutis nos segmentos, técnicas de aprendizado de máquina são alternativas válidas [Amutha e Khan 2023] e, devido ao seu desempenho em trabalhos anteriores [Ho *et al.* 2023], o K-Means foi o algoritmo escolhido para esse processo. Contudo, esse algoritmo tem uma particularidade: é necessário determinar um valor para o número de centróides (k). Para isso, o método do cotovelo é uma técnica eficaz que se utiliza do escore de distorção para determinar esse valor [Amutha e Khan 2023, Ho *et al.* 2023].

Mesmo modelos avançados de aprendizado de máquina podem falhar em obter segmentos claros em dados de alta variância e heterogeneidade [Ho *et al.* 2023]. Por isso, o método Box-Cox foi aplicado para estabilizar os dados e aproximá-los de uma distribuição normal. Esse método é conhecido por melhorar resultados desse tipo de modelos [Blum et al 2022].

A cada iteração de segmentação e transformação, o escore de distorção foi comparado com o modelo anterior e os segmentos foram analisados em termos de separação e interpretabilidade com o intuito de determinar se os resultados atingidos eram satisfatórios. Por fim, a classificação de cada segmento foi apresentada e suas características foram discutidas a partir dos padrões qualitativos e quantitativos dos grupos encontrados [Ho *et al.* 2024, Christy *et al.* 2021].

4. Resultados

Nesta seção são apresentados os resultados da análise exploratória dos dados e das segmentações. Cada variável RFM é abordada a partir de suas medidas analíticas e visualizações e os resultados das segmentações são expostos e discutidas em relação aos objetivos de interpretabilidade e separação do trabalho.

4.1. Análise Exploratória de Dados

A análise dos dados agregados por ordens de compra (tabela 1) revela um cenário marcado por heterogeneidade e influência de valores extremos. A superioridade do valor médio em relação à mediana evidencia uma distribuição assimétrica à direita, característica de uma distribuição de cauda longa com outliers. Essa disparidade sugere que, embora a maioria das transações seja de baixo valor, existem ordens de alto montante. O elevado desvio padrão e a curtose acentuada reforçam essa instabilidade, indicando que estratégias simples e baseadas apenas em médias podem ser enganosas.

Tabela 1. Medidas analíticas das ordens de compra

Tipo	Medida	Valor
Localização	Média	R$ 236,14
	Mediana	R$ 179,33
Variabilidade	Desvio padrão	R$ 170,54
Distribuição	Assimetria	4,48
	Curtose	62,75
Absoluta	Contagem de registros	61.614

Na análise por consumidor, a dimensão de Recência, que tem seus valores concentrados acima de 374 dias (média) em um período total de 680 dias, aponta para um fenômeno negativo: a maioria dos consumidores não retorna a fazer transações há mais de um ano. Isso aponta para uma baixa retenção e sugere estagnação no crescimento da base de clientes. A curva relativamente simétrica em torno da média observada na figura 3 indica que a recência pode ser um bom parâmetro para segmentação, embora seu perfil destaque o desafio estrutural de reengajar clientes após o primeiro ano.

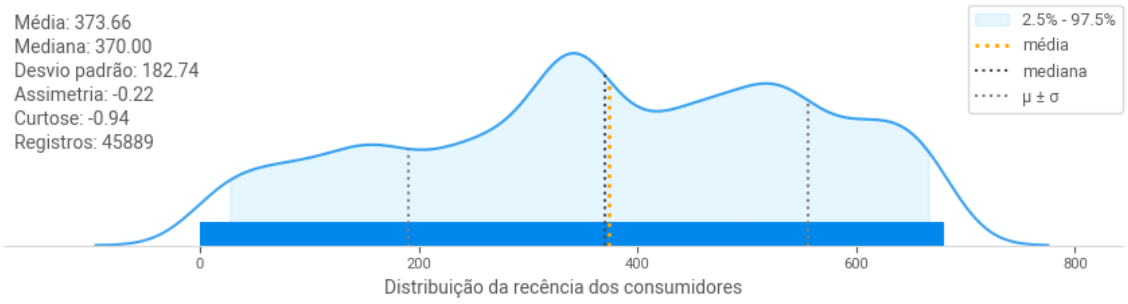

Figura 3. Gráfico da curva de distribuição de recência (com medidas analíticas)

A distribuição de frequência é a mais crítica entre as três dimensões. Com 45.889 consumidores gerando apenas 61.614 ordens, a concentração aguda em torno de uma compra e a curtose elevada revelam uma base pouco fidelizada e engajada. Apesar da existência de outliers (clientes com mais de 20 compras), a cauda longa enviesada à direita no gráfico da figura 4 confirma a predominância de consumidores com transações únicas.

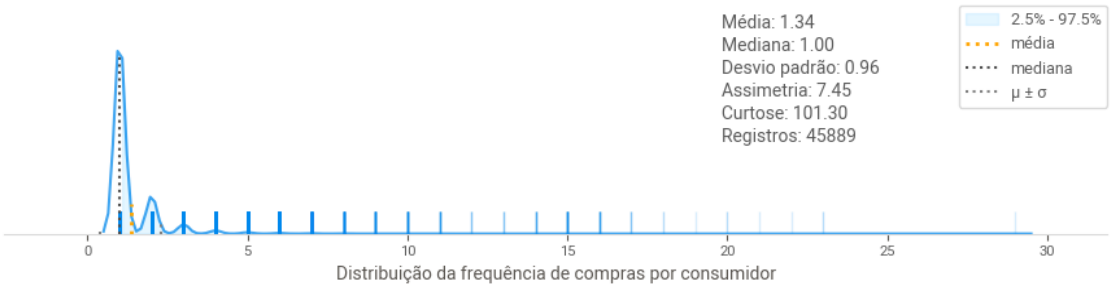

Figura 4. Gráfico da curva de distribuição de frequência (com medidas analíticas)

A dimensão monetária, embora também afetada por outliers, apresenta maior variabilidade comparada à frequência. Na figura 5 é possível visualizar uma curva de densidade de cauda longa e centrada em consumidores com valor monetário na faixa de R$ 200,00 a R$ 300,00. A curtose menor em comparação com a frequência sugere que, há maior diversidade nos padrões de gasto, o que pode viabilizar segmentos mais nítidos.

Figura 5. Gráfico da curva de distribuição de valor monetário (com medidas analíticas)

Panorama da análise exploratória para a segmentação de consumidores:

- A dimensão de recência tem uma concentração mais bem distribuída e, consequentemente, mais fácil de segmentar em conjunto com as demais variáveis. Apesar disso, indica que o crescimento da base de consumidores do negócio, inserido no setor de varejo, vem diminuindo.
- A dimensão de frequência tem a distribuição mais desafiadora para segmentação, sendo um dado de valor numérico inteiro e com uma concentração bastante aguda e afetada por outliers. Além disso, mostra dificuldade na retenção e fidelização dos consumidores: poucos realizaram mais do que duas compras no período.
- A dimensão de valor monetário também tem concentração aguda, mas apresenta maior variabilidade do que a dimensão de frequência. Como duas das três dimensões tem presença de outliers com assimetria e curtose altas, técnicas mais simples de segmentação podem ter dificuldade para separar os dados.

4.2. Segmentação RFM padrão

A aplicação da metodologia RFM tradicional encontrou uma limitação decorrente da natureza assimétrica da dimensão de Frequência, extremamente concentrada em uma única compra. Foi necessário aplicar uma categorização binária (≤2 e >2 compras). Essa solução limitou o modelo a obter 45 células RFM de segmentos de consumidores, indo de 111 até 525 (Figura 6).

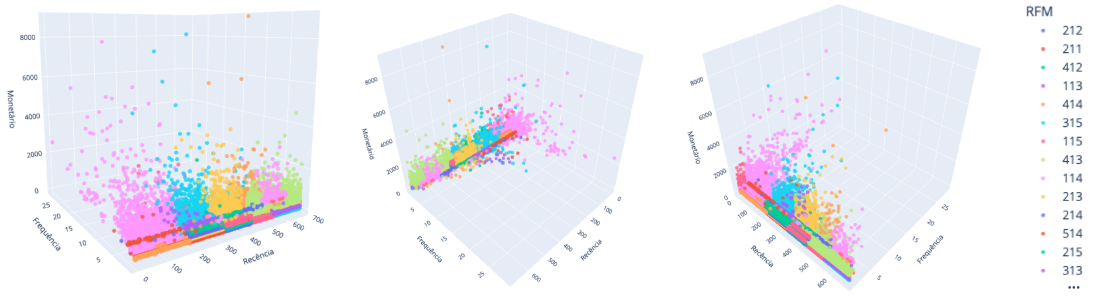

Figura 6. Gráfico de dispersão tridimensional dos segmentos RFM padrão

O alto número de segmentos impossibilita uma visualização e interpretação razoável das características de cada grupo, gerando uma fragmentação excessiva onde grupos podem ter pouquíssimos indivíduos e comprometendo a utilidade prática do modelo (tabela 2).

Tabela 2. Top 5 Segmentos com maiores e menores concentrações

Segmento	Número de indivíduos		Segmento	Número de indivíduos
411	2286		123	3
511	2272		222	2
311	2248	...	523	2
113	2226		321	1
512	2085		522	1

O escore de distorção obtido para esse método resultou em 2.032.706.479 (~2 milhões). Por fim, a rigidez estatística em contextos de assimetria e o desafio de visualizar e interpretar um alto número de segmentos fragmentados colocam o método RFM padrão como um modelo não satisfatório para atingir os objetivos propostos.

4.3. Segmentação RFM com K-Means

A utilização do algoritmo K-Means trouxe avanços para a segmentação. O número de segmentos pelo método do cotovelo foi de $k = 5$ (figura 7), simplificando a separação dos segmentos e obtendo um escore de distorção de 1.469.601.870 (~1,4 milhões).

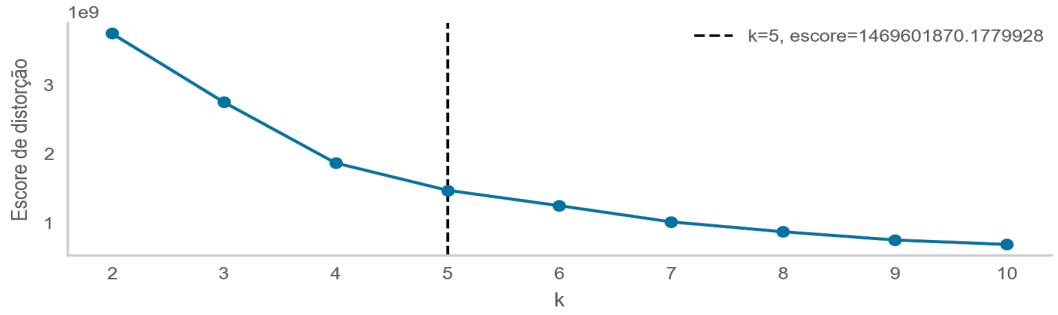

Figura 7. Gráfico do cotovelo para RFM com K-Means

Com o uso de K-Means, é possível observar um agrupamento de consumidores de menor valor monetário por recência de compra (figura 8), indicando que, para valores baixos de compra, há diferença no comportamento de clientes que realizaram compras recentemente e no período anterior a 400 dias. Contudo, a visualização mostra uma limitação crítica: o algoritmo não considerou a dimensão de frequência na segmentação.

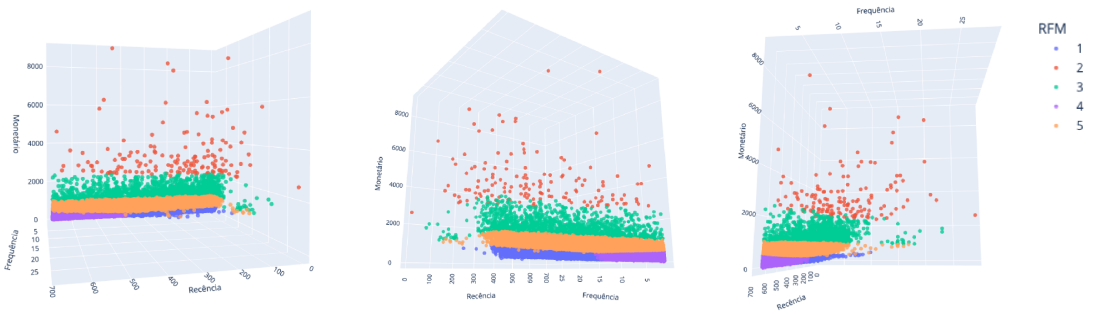

Figura 8. Gráficos de dispersão dos segmentos RFM com K-Means

A possível dificuldade na segmentação da variável de frequência levantada durante a análise exploratória se confirmou com esse modelo. Apesar dos avanços parciais, como a redução de segmentos e o agrupamento de consumidores de baixo valor monetário por recência, esse modelo não considera todas as dimensões RFM. Seu resultado também não é considerado satisfatório para a interpretação das características dos consumidores.

4.4. Segmentação RFM com K-Means e Box-Cox

A transformação Box-Cox trouxe separação para os segmentos. Ao normalizar a distribuição das variáveis, o algoritmo capturou relações entre as variáveis e padrões comportamentais anteriormente desapercebidos, obtendo quatro segmentos ($k = 4$) de clientes com um escore de distorção de 38.784 (~0,3% do valor original). Os segmentos resultantes apresentam separação mais clara e possuem comportamento inteligível, sendo definidos a partir da análise visual da figura 9 e das medidas analíticas da tabela 3 e fundamentados nos trabalhos correlatos [Ho *et al.* 2024, Christy *et al.* 2021].

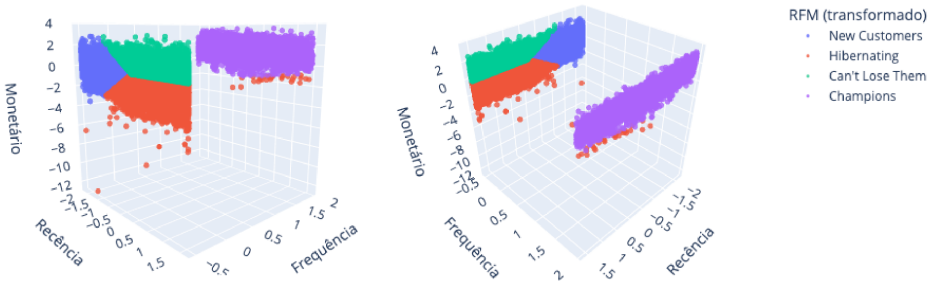

Figura 9. Gráficos de dispersão dos segmentos RFM com K-Means e Box-Cox

Os quatro segmentos resultantes são interpretados da seguinte forma:

- *Hibernating*: inclui 16.226 clientes com recência alta (483 dias), frequência única e valor monetário baixo (R$ 137,00). São consumidores inativos que realizaram apenas uma compra há mais de um ano, com baixo engajamento e valor contribuído, indicando risco elevado de perda definitiva. Porém, sem impacto significativo no faturamento do negócio. É importante abordá-los com cautela e estratégias de criação de valor de marca podem ser utilizadas para reavaliar seu potencial. A prioridade, porém, deve ser a retenção de segmentos mais estratégicos.
- *New Customers*: composto por 10.544 clientes de compras recentes (170 dias), frequência única e valor monetário moderado (R$ 238,00). Este grupo representa

consumidores em fase de teste que, apesar da significância monetária, ainda não consolidaram um hábito de compra. Estratégias de pós-venda (como acompanhamentos personalizados e amostras grátis) são relevantes para converter essa base de forma recorrente.

- *Can't Lose Them*: são 9.286 clientes de alto valor monetário (R$ 378,00) mas recência elevada (493 dias), indicando um perfil outrora valioso que se distanciou da marca. Esse perfil apresenta risco estratégico: perder esses consumidores pode impactar o negócio de forma relevante. Investigar as causas do afastamento (satisfação, concorrência e experiência de compra) e implementar campanhas de retorno é essencial para a sustentabilidade da marca.
- *Champions*: com 9.833 clientes, este grupo destaca-se por valores ótimos em frequência (3 compras) e valor monetário (R$ 638,00) com recência moderada (298 dias), representando os principais geradores de receita e engajamento. Sua fidelidade e valores elevados os tornam aliados estratégicos críticos para o negócio. É interessante valorizá-los com programas de fidelização e acesso antecipado para manter seu interesse a longo prazo.

Tabela 3. Medidas analíticas dos segmentos

Medida	Hibernating	New customers	Can't Lose Them	Champions
Recência (média)	483 dias	170 dias	493 dias	298 dias
Frequência (média)	1 compra	1 compra	1 compra	3 compras
Valor monetário (média)	R$ 137,00	R$ 238,00	R$ 378,00	R$ 638,00
Clientes	16.226	10.544	9286	9833

A comparação dos modelos na tabela 4 reforça a necessidade iterativa do processo: o RFM padrão não obteve resultados satisfatórios e o método com K-Means apresentou desempenho marginalmente superior. O salto qualitativo ocorreu somente com a combinação K-Means + Box-Cox, que gerou uma segmentação mais simples (4 vs. 45 segmentos) e coesa, reduzindo o escore de distorção de 2 milhões para 39 mil. Além disso, esse modelo trouxe intepretabilidade ao capturar as relações comportamentais das variáveis RFM e permitir que os segmentos fossem caracterizados e descritos.

Tabela 4. Comparação entre os métodos de segmentação

Método	Escore de distorção	Número de segmentos
RFM padrão	2 milhões	45
RFM com K-Means	1.4 milhão	5
RFM com K-Means e Box-Cox	39 mil	4

5. Considerações finais

Analisar segmentos de consumidores e seu comportamento traz competitividade e fornece uma base para estratégias orientadas por dados no setor de e-commerce, onde a alocação eficiente de recursos e abordagem ao consumidor exerce papel crítico. Essas estratégias, quando integradas a sistemas urbanos inteligentes, influenciam positivamente cadeias de produção e inovação no varejo e favorecem a adaptabilidade de cidades a demandas dinâmicas de consumo, fortalecendo a sinergia entre competitividade econômica e sustentabilidade.

Modelar e obter valor acionável dessas características é um processo complexo. Neste trabalho, os dados foram extraídos, tratados, analisados e três métodos de segmentação foram utilizados para obter o resultado desejado: RFM padrão, RFM com K-Means e RFM com K-Means e Box-Cox. Cada método utilizou-se dos resultados anteriores e da análise exploratória de dados para obter uma segmentação interpretável e com características distintas, onde cada um dos quatro grupos definidos (*Hibernating*, *New Customers*, Can't *Lose Them* e *Champions*) apresentam comportamentos particulares e podem ser abordados de forma estratégica e com maiores chances de sucesso.

Ao comparar os resultados do primeiro método com o último, nota-se uma redução significante do escore de distorção e do número de segmentos, indo de 45 para 4. Cada um dos 4 segmentos finais obtidos tem uma interpretação definida e apoiada por suas características de recência, frequência e valor monetário e de suas medidas analíticas.

Outro ponto evidenciado é a necessidade de explorar os dados com diligência. Diferentes características nas medidas de localização, variabilidade e distribuição exigem diferentes abordagens na aplicação dos métodos de segmentação. Neste caso, foi necessário utilizar de técnicas de aprendizado de máquina e normalização de dados para chegar em um resultado satisfatório.

Por fim, este estudo explora somente uma das possibilidades de segmentação de consumidores. É possível adicionar dados heterogêneos e tratar variáveis de forma personalizada. Além disso, existem alternativas ao uso de K-Means como o algoritmo de segmentação e diferentes formas de tratar seus dados de entrada e hiperparâmetros. Com isso, espera-se que este artigo traga novas perspectivas estratégicas para o relacionamento com o consumidor no setor de e-commerce e sirva como base para ações e reflexões que impulsionem avanços na área.

6. Referências

Almeida J., Silva J., Cavalcante E. (2020). Extração, Integração e Importação de Dados Heterogêneos em Cidades Inteligentes: Um Mapeamento Sistemático, em Anais do IV Workshop de Computação Urbana (COURB), páginas 57-70.

Anitha P., Patil M. (2022). RFM model for customer purchase behavior using K-Means algorithm, em Journal of King Saud University – Computer and Information Sciences, Vol. 33, No. 5 páginas 1785-1792.

Amutha R., Khan A. (2023). Customer Segmentation using Machine Learning Techniques, em Tujin Jishu/Journal of Propulsion Technology, Vol. 44, No. 3.

Blum L., Elgendi M., Menon C. (2022). Impact of Box-Cox Transformation on Machine-Learning Algorithms, em Frontiers in Artificial Intelligence, Vol. 5, 877569.

Bruce P., Bruce A. (2019). Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais, Alta Books, 1° edição.

Christy A., Umamakeswari A., Priyatharsini L. (2021). RFM ranking – An effective approach to customer segmentation, em Journal of King Saud University – Computer and Information Sciences, Vol. 33, No. 1, páginas 1251-1257.

Cruz, W. (2021). Crescimento do e-commerce no Brasil: desenvolvimento, serviços logísticos e o impulso da pandemia de Covid-19, em GeoTextos, Vol. 17, No. 1, páginas 67-88.

Gupta S., Kushwaga P., Badhera U. (2023). Identification of benefits, challenges, and pathways in E-commerce industries: An integrated two-phase decision-making model, em Sustainable Operations and Computers, Vol. 4, páginas 200-218.

Ho T., Nguyen S., Nguyen H. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry, em Business Systems Research Journal Vol. 14, páginas 26-53.

Hughes A. (1994). Strategic Database Marketing, Probus Press, 1° edição.

Kodinariya M., Makwana P. (2013). Review on Determining Number of Cluster in K-Means Clustering, em International Journal of Advance Research in Computer Science and Management Studies, Vol. 1, páginas 90-95.

Lewaa I. (2024). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis, em Journal of Data Science and Intelligent Systems, Vol. 2, No. 1, páginas 29-36.

Li X., Sun Y., Chen X. (2022). Saline-Sodic Soil EC Retrieval Based on Box-Cox Transformation and Machine Learning, em IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 15, páginas 1692-1700.

Ma X., Li J., Guo Z. (2024). Role of big data and technological advancements in monitoring and development of smart cities, em Heliyon, Vol. 10, No. 15, e34821.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations, em University of California Press, Vol. 1, No. 1, páginas 281-297.

Marutho D., Handaka S., Wijaya E. (2018). The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News, em Anais do Seminário Internacional de Aplicação de Tecnologia para Informação e Comunicação de 2018 (ISemantic), páginas 533-538.

Pan C., Bai X., Li F. (2021). How Business Intelligence Enables E-commerce: Breaking the Traditional E-commerce Mode and Driving the Transformation of Digital Economy, em Anais da II Conferência de E-Commerce e Tecnologia da Internet de 2021 (ECIT), páginas 26-30.

Tang Z., Jiao Y., Yuan M. (2024). RFM user value tags and XGBoost algorithm for analyzing electricity customer demand data, em Systems and Soft Computing, Vol. 6, 200098.

Wei J., Lin S., Wu H. (2024). A review of the application of the RFM model, em African Journal of Business Management, Vol. 4, No. 19, páginas 4199-4206.

Yoshida M., Santos M., Freire F. (2024). Modelo RFM aplicado à melhoria de vendas em indústrias usando clusterização e método AHP-gaussiano, em Anais do XII Simpósio de Engenharia de Produção.