Segmentação de dados de venda de empresa de varejo em e-commerce com RFM e K-Means

Ewerthon J. Kutz

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

ewerthon.jose.kutz@uel.br

1. Descrição do Problema

Uma segmentação eficaz da base de dados de clientes pode ser determinante para o sucesso da estratégia de marketing de um negócio [Amutha e Khan 2023]. Com a complexidade de modelar o comportamento dos clientes em transações e o grande volume de dados transacionais gerado atualmente, segmentar uma base de clientes não é uma tarefa trivial.

É comum que um negócio procure segmentar seus clientes com dados demográficos, como a idade e sexo, ou com base no seu valor monetário, utilizando o valor total de suas compras. [Tang et al 2024]. Contudo, esses métodos não capturam aspectos fundamentais do comportamento do usuário para fornecer uma tomada de decisão assertiva a fundamentada em dados para a estratégia de marketing de um e-commerce.

Neste cenário, surge a metodologia de segmentação Recency, Frequency e Monetary (RFM). Neste modelo, o cliente é segmentado em três dimensões:

- Recency (recência): quanto tempo passou desde sua última compra.
- Frequency (frequência): quantas vezes o cliente já comprou.
- Monetary (valor monetário): quando o cliente gastou no total.

A partir disso, a base de clientes é dividida em diferentes quadrantes, a cada qual pode-se aplicar uma estratégia mais personalizada e com maior chance de sucesso [Christy et al. 2021]. Este projeto propõe uma análise detalhada da segmentação de dados de venda de uma empresa de varejo em e-commerce utilizando a metodologia RFM e o algoritmo K-Means. Para atingir esse objetivo, os dados do negócio serão obtidos, explorados e tratados. Com o processamento concluído, a segmentação será aplicada e apresentada em um relatório.

Espera-se contribuir com insights acerca do setor de e-commerce e dos desafios processuais de uma segmentação de cliente. Além disso, o relatório gerado com o projeto deve fornecer insights sobre o comportamento dos usuários do negócio e proporcionar a aplicação de estratégias de marketing orientada por dados.

2. Levantamento de trabalho correlatos

O estudo de [Christy et al. 2021] oferece uma análise abrangente sobre a segmentação de clientes utilizando a metodologia RFM, destacando sua importância na retenção de clientes e na maximização da receita. A pesquisa detalha como a análise RFM pode ser

aplicada para classificar os clientes em segmentos com base em seu comportamento de compra, utilizando algoritmos como K-Means e Fuzzy C-Means.

Além disso, o estudo propõe uma nova abordagem para a escolha dos centróides iniciais no algoritmo K-Means, o que resulta em uma redução significativa no número de iterações necessárias para alcançar uma clusterização eficaz. Essa inovação é particularmente relevante para empresas que buscam otimizar seus processos de marketing e melhorar a eficiência operacional.

A abordagem de [Anitha e Patil 2022] também apresenta uma aplicação inovadora de inteligência de negócios para identificar clientes potenciais no setor de varejo de e-commerce com o uso do modelo RFM. Além disso, a pesquisa valida os cluster obtidos através do cálculo do coeficiente de silhueta, proporcionando uma avaliação robusta da coesão e separação dos grupos formados. Essa pesquisa não só avança o conhecimento sobre segmentação de clientes, mas também fornece um modelo prático que pode ser adotado por empresas em busca de melhorar suas estratégias de marketing e aumentar a satisfação do cliente.

3. Fundamentação Teórica

4.1. O Modelo RFM

A segmentação RFM é amplamente utilizada para o ranqueamento e segmentação de clientes com base no seu histórico de compras. Esse tipo de análise é especialmente útil em setores onde há um alto número de clientes que realizam transações, como no varejo e no e-commerce [Christy et al. 2021]. Esse método agrupa os clientes em três dimensões:

1. Recência: refere-se ao número de dias desde que um cliente realizou sua última compra. Para o ranqueamento, quanto menor for esse número, maior é a pontuação de recência. Os clientes são divididos em quintis, onde os 20% de clientes mais recentes recebem a pontuação máxima de 5, e os demais recebem pontuações decrescentes até 1 para os clientes menos recentes [Wei et al. 2010]. Esta dimensão é considerada uma das mais importantes para prever a probabilidade de recompra [Wei et al. 2010].
2. Frequência: é definida como o número de compras que o cliente fez dentro de um período específico. Um cliente com uma pontuação de frequência alta, também classificado em quintis de 1 a 5, geralmente apresenta forte lealdade, o que sugere uma alta demanda pelo produto e uma maior probabilidade de repetir compras [Wei et al. 2010]. No entanto, o método pode enfrentar dificuldades ao lidar com clientes em quintis inferiores, que podem ter comportamentos de compra similares entre si [Wei et al. 2010].
3. Valor Monetário: corresponde ao total de dinheiro gasto pelo cliente em um período determinado, o que reflete o quão valioso o cliente é em termos de faturamento para o negócio. Nesse caso, a pontuação varia de 1 a 5, onde a pontuação máxima de 5 é atribuída aos 20% de clientes que mais gastam. A literatura sugere o uso do valor médio de compra para reduzir possíveis correlações entre as dimensões de frequência e valor monetário [Wei et al. 2010].

Após calcular as três dimensões, os dados são ranqueados e os clientes são classificados de 111 a 555, gerando um total de 125 células RFM distintas. Esses segmentos podem

variar consideravelmente em termos de taxas de resposta, sendo o segmento 555 representativo dos melhores clientes e o segmento 111 dos menos engajados [Wei et al. 2010]. Esse tipo de segmentação permite às empresas identificar padrões comportamentais e otimizar suas estratégias de marketing ao direcionar campanhas específicas para cada grupo.

4.1. Algoritmo de Clusterização K-Means

O algoritmo k-means é um método de agrupamento não hierárquico amplamente utilizado para segmentação, onde os dados são divididos em um número pré-determinado de clusters (k) [Yoshida et al. 2024]. O objetivo principal do k-means é minimizar a variação dentro de cada cluster, promovendo uma maior compactação dos grupos. Essa variação intra-cluster é medida pela soma das distâncias quadráticas euclidianas entre as observações e os centróides de seus respectivos clusters [Anitha e Patil 2022].

A distância Euclidiana, usada para calcular a proximidade entre cada ponto e o centróide, é dada pela fórmula:

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

onde x_i representa um ponto específico e c_j o centróide do cluster j.

Para avaliar a qualidade dos agrupamentos, utiliza-se a métrica WSS (within-cluster sum of squares), que mede quão compactos os clusters estão. Quanto menor o valor de WSS, mais concentrados estão os pontos em torno de seus centróides. A fórmula para WSS é:

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_i - c_j||^2$$

onde x_i é um ponto dentro do cluster j, c_j é o centróide, e n_j é o número de pontos em j.

O processo iterativo do k-means envolve associar cada observação ao centróide mais próximo com base na distância euclidiana e recalcular os centróides com as novas médias dos clusters. Este processo se repete até que os centróides não mudem mais significativamente entre as iterações, indicando que o algoritmo atingiu a convergência [Kodinariya e Makwana, 2013].

Para determinar o número ideal de clusters (k), são comumente utilizados dois métodos:

1. Método do Cotovelo: inicializa-se com $k = 2$ e aumenta-se progressivamente, calculando a WSS para cada valor de k. O "cotovelo" do gráfico representa o ponto onde o WSS começa a reduzir de forma menos acentuada, sinalizando o valor ideal de k, em que há um equilíbrio entre a variabilidade explicada e o número de clusters [Yoshida et al. 2024].

2. Método da Silhueta**:** avalia a qualidade do agrupamento de cada observação com base na média das distâncias euclidianas, gerando uma medida de silhueta para

diferentes valores de k. O valor de silhueta mais próximo de 1 indica o agrupamento ideal, pois demonstra que as observações estão bem ajustadas aos seus clusters [Christy et al. 2021].

Estes métodos permitem uma análise detalhada e precisa dos dados, facilitando a escolha do número de clusters mais adequado para o conjunto de dados.

4. Metodologia

O projeto será realizado em formato de estudo de caso e utilizará dados de uma empresa de varejo em e-commerce para a segmentação dos clientes. Para tanto, será realizada uma etapa inicial de coleta e análise exploratória dos dados. Com a análise em mãos, os dados serão tratados e o cálculo das dimensões de RFM será realizado.

A partir disso, a segmentação se dará de duas maneiras: os clientes serão enquadrados nos princípios de quintis da metodologia RFM e segmentados com a aplicação do algoritmo K-Means. Com a segmentação aplicada, os resultados serão analisados com o método do cotovelo e análise visual e interpretados.

5. Resultados

5.1. Coleta de Dados

Os dados foram coletados a partir de uma extração de tabela da plataforma de e-commerce do negócio com as datas de corte de 02/01/2023 até 11/11/2024 em formato CSV (Comma-separated values), contendo quatro colunas, descritas na tabela 1.

Tabela 1. Dataset

Coluna	Descrição
order	ID único da ordem
creation_date	Data de quando a compra foi realizada
total_value	O valor monetário do item comprado na ordem
client_hash	ID único do cliente

Cada linha na tabela representa um item de uma ordem de compra e o único processamento de dados realizado nesta etapa inicial foi o de anonimização do ID único do cliente com *hash*.

5.2. Análise Exploratória de Dados

Inicialmente, os dados foram agregados por ordem de compra e apresentados na tabela 2 em termos de medidas de localização (média e mediana), variabilidade (desvio padrão), distribuição (assimetria e curtose) e contagem absoluta de registros.

Tabela 2. Medidas analíticas das ordens de compra

Tipo	Medida	Valor
Localização	Média	R$ 236,03
	Mediana	R$ 179,28

Variabilidade	Desvio padrão	R$ 170,57
Distribuição	Assimetria	4,48
	Curtose	62,67
Absoluta	Contagem de registros	61.643

O valor médio por ordem de R$ 263,04 tem um alto desvio padrão (R$ 170,57). Essa média é influenciada pela assimetria positiva de distribuição de cauda longa, o que indica a presença de outliers.

Após a observação em relação às ordens, os dados foram agregados por cliente, obtendo-se as dimensões relacionadas à análise RFM: recência, frequência e valor monetário de cada cliente. Cada dimensão foi explorada visualmente em termos de medidas analíticas.

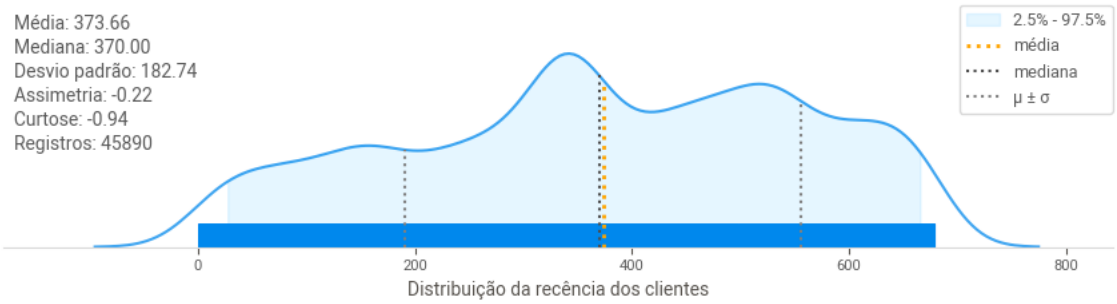

Figura 1. Gráfico da curva de distribuição de recência (com medidas analíticas)

A recência das compras está distribuída de forma ligeiramente uniforme (figura 1), centrada em 370 dias (cerca de um ano). Também se observa maior concentração de valores maiores que 37 que, indicando a existência de maior número de compras no início do período e menor frequência para a base de 45.890 clientes.

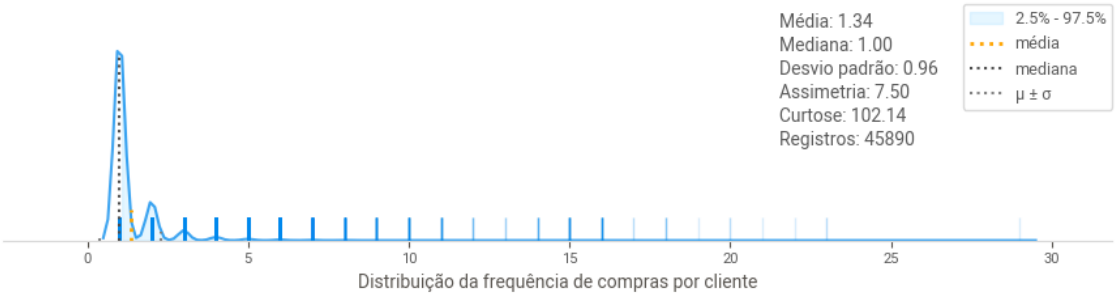

Figura 2. Gráfico da curva de distribuição de frequência (com medidas analíticas)

A curva de distribuição da frequência de compras apresentada na figura 2 reforça a ideia de menor frequência de compras, que tem uma concentração aguda em torno do número 1 e apresenta curtose alta. Além disso, existe uma base de 45.890 clientes para 61.643 ordens de compra - reforçando o comportamento de baixa frequência da base de clientes.

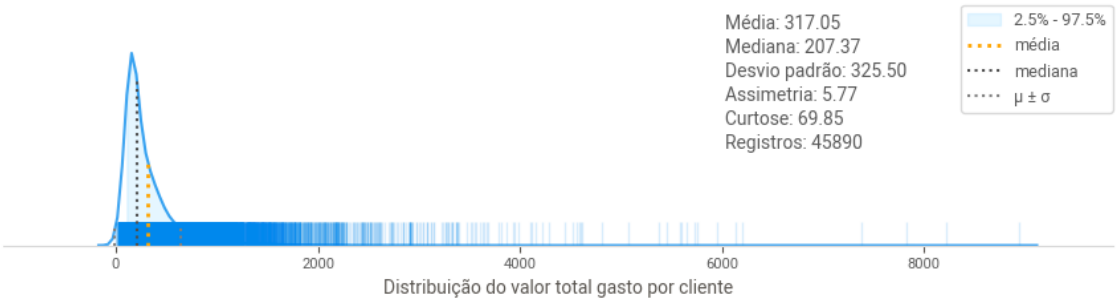

Figura 3. Gráfico da curva de distribuição de valor monetário (com medidas analíticas)

Na figura 3, é possível observar uma curva de distribuição de alta concentração de clientes de baixo valor monetário, fato reforçado pela diferença entre a mediana (R$ 207,37) e a média (R$ 325,50) com alto desvio padrão, no valor de R$ 317,05.

5.3. Segmentação RFM

Com as dimensões criadas, a metodologia de divisão em quintis de [Wei et al. 2010] foi aplicada aos dados, exceto para a dimensão de frequência, que devido à alta concentração apresentada e discutida na figura 3, foi dividida em duas partes: até duas compras e superior a duas compras. Essa combinação gerou 45 segmentos de clientes com pelo menos um individuo em cada cluster (Figura 4).

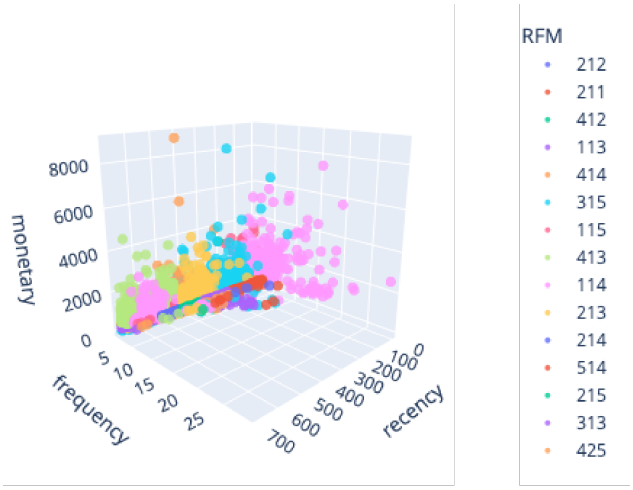

Figura 4. Gráfico tridimensional de dispersão dos segmentos RFM

O alto número de segmentos resultantes dificulta a interpretação e visualização das características de cada grupo. Isso é intensificado com dados de distribuição concentrada, conforme observado na análise exploratória - 15 segmentos apresentaram uma concentração de clientes inferior a 100 indivíduos.

5.4. Segmentação RFM com K-Means

Outra forma de segmentar os grupos nas dimensões RFM é a partir do algoritmo de machine learning K-Means [Yoshida et al. 2024]. Para isso, os dados foram transformados com o método Yeo-Johson de forma a estabilizar sua variância e

normalizá-los para a aplicação das comparações de distância do método do cotovelo, que resultou em um k=5, como observado na figura 5.

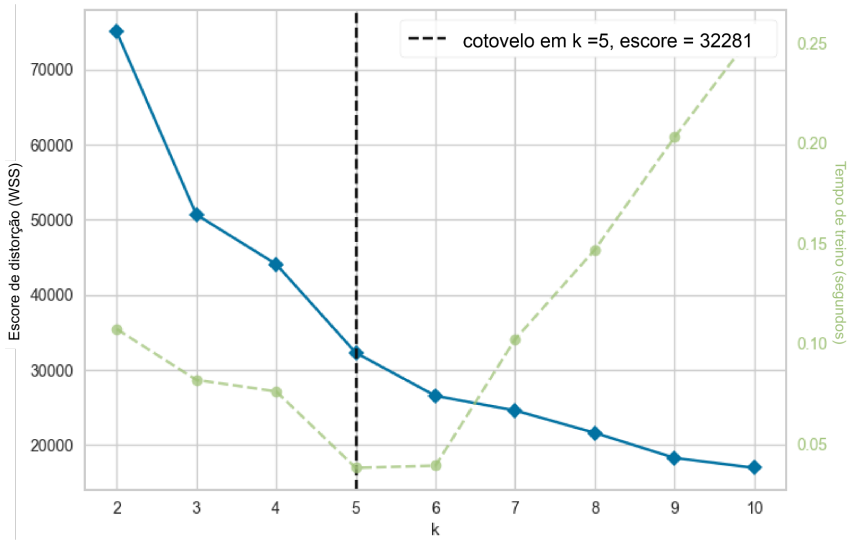

Figura 5. Gráfico do cotovelo para RFM com K-Means

Com 5 segmentos e uma delimitação mais interpretável, os clientes foram divididos conforme suas características: “Hibernating”, “At Risk”, “New Customers”, “Can’t Lose Them” e “Champions” na figura 6.

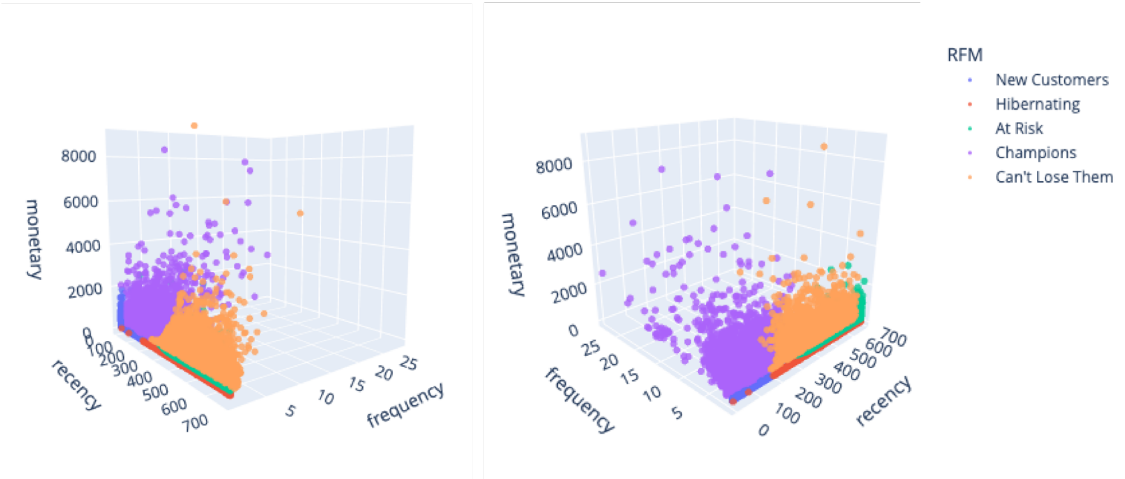

Figura 6. Gráficos tridimensionais de dispersão dos segmentos RFM com K-Means

As características de cada grupo foram determinadas a partir da análise visual dos gráficos, de suas medidas (apresentadas na tabela 3) e das bases de outros trabalhos com RFM [Ho et al. 2024, Christy et al. 2021].

Tabela 3. Dataset

Medida	Hibernating	At Risk	New customers	Can’t Lose Them	Champions
Recência (média)	485 dias	494 dias	174 dias	428 dias	139 dias
Frequência (média)	1 compra	1 compra	1 compra	2 compras	3 compras
Valor monetário (média)	R$ 137,00	R$ 380,00	R$ 236,00	R$ 550,00	R$ 750,00

Clientes	16.004	9137	10859	5454	4400

A interpretação de cada grupo se dá da seguinte forma:

- "Hibernating": clientes que fizeram apenas uma compra há muito tempo e com baixo valor. São compradores ocasionais que não demonstraram interesse ou fidelidade à marca.
- "At Risk": compradores de alto valor que adquiriram um produto, mas nunca voltaram. Representam uma oportunidade perdida, pois possuem grande potencial de gasto.
- "New Customers": novos clientes que fizeram uma única compra recentemente. Ainda estão conhecendo a marca e podem ser fidelizados com a abordagem certa.
- "Can't Lose Them": clientes fiéis no passado, mas que não compram há um bom tempo. Foram compradores recorrentes e de alto valor, mas estão em risco de não retornar.
- "Champions": os clientes mais valiosos e engajados. Compram com frequência, gastam bastante e interagem ativamente com a marca. São o público mais estratégico para o crescimento do negócio.

6. Considerações finais

A segmentação de clientes a partir das características apresentadas nos resultados pode ser um importante diferencial competitivo e fornecer uma base um marketing estratégico e direcionado por dados em uma empresa. Contudo, é importante unir as análises com o ponto de vista do negócio para gerar valor para cada grupo.

Outro ponto evidenciado é a necessidade de explorar os dados com diligência. Diferentes características de localização, variabilidade e distribuição das dimensões de recência, frequência e valor monetário exigem diferentes abordagens na aplicação do método RFM. Nesse caso, foi necessário transformar e normalizar os dados e utilizar uma técnica de machine learning para trazer interpretabilidade para os segmentos.

Por fim, este trabalho explora somente uma das possibilidades de segmentação de clientes. É possível adicionar dimensões de dados demográficos e comportamentais às dimensões características do método RFM e enriquecer o modelo. Além disso, existem alternativas ao uso de K-means como o algoritmo de clusterização e diferentes formas de transformar os dados para uso desses algoritmos. Com isso, espera-se que este estudo traga novas perspectivas estratégicas para o negócio no qual foi aplicado e sirva como base para ações concretas e reflexões que impulsionem avanços na área.

7. Referências

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

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 33, páginas 1251-1257.

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.

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.

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(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. Disponível em: https//www.even3.com.br/anais/12simep/792477-modelo-rfm-aplicado-a-melhoria-de-vendas-em-industrias-usando-clusterizacao-e-metodo-ahp-gaussiano.