57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

ANÁLISE DE SEGMENTAÇÃO DE VAREJO: INTEGRANDO GEOVISUALIZAÇÃO E INSIGHTS ESTRATÉGICOS AO MODELO RFM E K-MÉDIAS

Leonardo Carneiro Andrade
Programa de Pós-graduação em Engenharia de Produção - Universidade Federal Fluminense
Rua Passo da Pátria, 156, Campus Praia Vermelha, Bl. D - sala 309, São Domingos, Niterói-RJ
leoandrade@id.uff.br

Francisco Thomas Barbosa da Silva
Escola Superior de Agricultura "Luiz de Queiroz" (Esalq/USP) - Universidade de São Paulo
Av. Pádua Dias, 11 - Agronomia, Piracicaba - SP, 13418-900
thomasbarbosa@usp.com.br

Diogo Ferreira de Lima Silva, Helder Gomes Costa
Departamento de Engenharia de Produção - Universidade Federal Fluminense
Rua Passo da Pátria, 156, Campus Praia Vermelha, Bl. D - sala 309, São Domingos, Niterói-RJ
diogofls@id.uff.br, heldergc@id.uff.br

RESUMO

Frente à necessidade de otimização no setor varejista, este trabalho propõe o framework RFMKIG (RFM, K-Means, Insights, Geovisualização) para a segmentação estratégica de lojas varejistas. A metodologia combina o modelo RFM e o algoritmo K-Médias, sendo aplicada a dados públicos de 100 lojas norte-americanas, tendo como originalidade a adição de duas camadas: a geovisualização, que revela padrões espaciais, e a análise qualitativa, que traduz os dados em perfis nomeados com insights estratégicos acionáveis. O framework RFMKIG prova-se uma ferramenta robusta para a tomada de decisão baseada em dados, integrando as dimensões quantitativa, espacial e estratégica.

PALAVRAS CHAVE. Ciência de Dados. K-Médias. Segmentação de Clientes.
Tópicos: AS&DS-Análise e Ciência de Dados; EST&AM – PO Analytics em Estatística e Aprendizado de Máquina

ABSTRACT

In response to the need for optimization in the retail sector, this paper proposes the RFMKIG (RFM, K-Means, Insights, Geovisualization) framework for the strategic segmentation of retail stores. The methodology combines the RFM model and the K-Means algorithm, applied to a public dataset of 100 North American stores, with its originality lying in the addition of two layers: geovisualization, which reveals spatial patterns, and qualitative analysis, which translates data into named profiles with actionable strategic insights. The RFMKIG framework proves to be a robust tool for data-driven decision-making, integrating quantitative, spatial, and strategic dimensions.

KEYWORDS. Data Science. K-means. Customer Segmentation.
Paper topics: AS&DS – Data Science and Analysis; EST&AM – Analytics in Statistics and Machine Learning

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

1. Introdução

Em um cenário cada vez mais orientado por dados, empresas que desejam manter-se competitivas precisam compreender com profundidade o comportamento de seus clientes e o desempenho de suas unidades de negócio.

Para Ho et al. [2023], os dados dos clientes são a base para estratégias de negócios bem-sucedidas e a exploração desses dados a fim de identificar informações de relevância sobre os clientes e apoiar a tomada de decisões aumenta o interesse dos negócios.

Nesse contexto, a mineração de dados (Data Mining, ou, abreviando, DM) auxilia na identificação de padrões, tendências e conhecimentos significativos, através de grandes conjuntos de dados [Ernawati et al., 2021], por meio da manipulação e análise desses. Técnicas como a segmentação de clientes destacam-se como fundamentais para o direcionamento eficiente de ações comerciais e operacionais [Nandapala e Jayasena, 2020], e para a definição dos planos de marketing da organização, tendo em vista agrupar os clientes em grupos homogêneos menores (ou relativamente homogêneos), focando em cada grupo individualmente. A adoção desse tipo de técnica auxilia as organizações a identificarem clientes valiosos e a desenvolverem as atividades de marketing [Ernawati et al., 2021].

A segmentação baseada no comportamento de compra, em especial pela ótica dos indicadores de Valor Monetário, Frequência e Recência, também chamada de modelo RFM (Recência, Frequência e Valor Monetário), é amplamente utilizada em iniciativas de marketing [Ernawati et al., 2021]. No entanto, sua aplicação não se limita à análise individual de consumidores: pode também ser estendida à análise comparativa entre filiais ou lojas físicas, permitindo identificar padrões de desempenho e priorizar investimentos.

Ao usar a segmentação de clientes, no caso o RFM, com um algoritmo de agrupamento, é possível obter um resultado bem-sucedido do processo de segmentação [Nandapala e Jayasena, 2020; Ho et al., 2023]. Muitas empresas utilizam-se dessa tática para identificar segmentos ideais de clientes. Muitos estudos propuseram variações ampliadas desse modelo, incorporando novas variáveis ou aplicando várias técnicas analíticas para diferenciar grupos de clientes com mais detalhes e sob vários ângulos [Ho et al., 2023]. Os modelos de aprendizado de máquina têm a capacidade de identificar os padrões dos dados, e almejam encontrar a função ideal ou a mais próxima dessa, que identifique esses padrões [Nandapala e Jayasena, 2020].

Conforme exposto por Zhou et al. [2020], as empresas primeiro devem identificar quais clientes são lucrativos para suas linhas de negócios e, em seguida, qual deve ser a estrutura de segmentação de mercado preferível quando focarem nas preferências desses clientes a fim de retê-los.

A relevância desta pesquisa reside na integração de técnicas de Machine Learning, especificamente o algoritmo K-Médias, com a análise RFM para segmentar 100 lojas de uma rede varejista norte-americana, conforme aplicado em outros trabalhos [Barus et al., 2023; Essayem et al., 2022; Ernawati et al., 2021; Ho et al., 2023; Ling e Weiling, 2025], porém unindo isso à visualização geoespacial das filiais e à íntese do perfil do cluster, por meio da definição de um nome e apontamento de pontos estratégicos para cada cluster. Esta abordagem transcende as aplicações convencionais de marketing ao identificar agrupamentos de lojas com características similares e destacar unidades conforme desempenho, identificando informações relevantes quanto aos perfis de cluster, baseado em dados, fornecendo base estratégica para decisões corporativas e definições de prioridades operacionais.

O trabalho, por meio do método RFMKIG (RFM, K-Means, Insights extraídos e Geovisualização) se atém à detecção de filiais com desempenho atípico, à clusterização de filiais baseada em dados, ao estabelecimento de diretrizes estratégicas práticas para ações comerciais específicas de acordo com cada perfil identificado dos clusters e à geovisualização dessas filiais.

A problemática deste estudo é endereçada particularmente às empresas com redes extensas de lojas físicas que buscam otimizar seu desempenho de forma localizada. No cotidiano varejista, decisões frequentemente baseiam-se em intuições ou dados que não revelam padrões comportamentais significativos, resultando em estratégias generalizadas que desconsideram particularidades. A questão central que norteia esta pesquisa é: como podemos aplicar técnicas

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

avançadas de segmentação para identificar padrões de desempenho entre lojas e traduzir isso em uma identificação do perfil dos clusters, tendo em vista a maximização dos resultados?

2. Fundamentação Teórica

Para Ernawati et al. [2021], clientes são indivíduos que compram produtos ou serviços de uma empresa, as estratégias de marketing são aplicadas para aumentar a satisfação e a fidelidade desses clientes e planos diferentes agradam a grupos de clientes diferentes. Nesse âmbito, Zhou et al. [2020] afirma um conceito base é a segmentação de mercado, tendo em vista a diversidade entre os consumidores e sabendo que normalmente não é factível satisfazer todos os clientes com uma única estratégia.

De acordo com Essayem et al. [2022], a metodologia RFM é descrita conforme a seguir:

- Recência: é definida pelo período (número de dias) entre a transação atual e a anterior, ou seja, quanto menor o período, maior o valor de recência.
- Frequência: é definida pelo número de transações feitas por cada cliente durante um período específico ou estudado. Uma maior frequência tem um impacto positivo.
- Monetário (Valor monetário): é definido pelo valor total ou médio de dinheiro gasto por um cliente durante um período específico. Quanto maior a quantia, maior o valor monetário. Todos os recursos do RFM podem ser extraídos de dados históricos de transações.

Para Ling e Weiling [2025] *Clustering* é o processo de agrupar as observações, de modo que os objetos em um *cluster* são semelhantes entre si e diferentes dos objetos em outros *clusters*. Na segmentação de clientes, o agrupamento de filiais auxilia e propõe um agrupamento dos clientes em clusters com base em suas características, como, por exemplo, comportamentos, necessidades, preferências e fatores psicológicos, aplicando posteriormente estratégias de marketing personalizadas para cada grupo [Ernawati et al., 2021; Ling e Weiling, 2025].

Para Al-Dabbas et al. [2023], um dos métodos de *machine learning* aplicáveis à segmentação da personalidade do cliente é o aprendizado não supervisionado, os quais, treinados com grandes volumes de dados, identificam padrões e similaridades entre os clientes, sem a necessidade de instruções explícitas. Ainda de acordo com o mesmo autor, essa capacidade os torna especialmente eficazes para analisar perfis de personalidade, ao revelar diferenças e tendências sutis que podem passar despercebidas por observadores humanos. O aprendizado de máquina está entre as subáreas da Inteligência Artificial e permite o estudo de grandes quantidades de dados [Nandapala e Jayasena, 2020].

Ho et al. [2023] informa que o *clustering* parcial é um método de agrupamento de dados em um número específico de clusters e que o algoritmo K-Means é um desses métodos.

O K-Means é um algoritmo amplamente utilizado para agrupar pontos de dados semelhantes, operando de forma iterativa ao calcular os centróides dos clusters e realocar os dados ao centro mais próximo. Esse processo se repete até que os resultados entre iterações sucessivas se estabilizem, ou seja, quando os agrupamentos obtidos se tornam semelhantes aos anteriores, garantindo a definição de clusters internamente formados por dados com características semelhantes [Barus et al., 2023; Ling e Weiling, 2025].

Para utilizar Box-Plot, deve-se calcular o limite inferior (L) e o superior (U) conforme a equação a seguir [Wahyuni et al., 2021]:

$$U = Q3 + 1.5(Q3 - Q1)$$

$$L = Q1 - 1.5(Q3 - Q1)$$

Conforme Mufarroha et al. [2022] a definição do valor de k no algoritmo k-médias, que representa a quantidade de clusters, pode ser feita por meio da otimização do método do cotovelo ou do coeficiente de silhueta. De acordo com esse autor, no método *Elbow* (método do cotovelo, inércia) é testada uma faixa de valores de k (por exemplo, de 1 a 12), calculando-se a distorção, ou

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

seja, a soma das distâncias quadradas de cada ponto ao centro do cluster, para cada valor, dessa forma, o ponto ideal de k é identificado visualmente como a inflexão da curva gerada, semelhante a um cotovelo, indicando a melhor configuração do modelo.

Segundo Ling e Weiling [2025], o *Silhouette Score* (pontuação de silhueta) é uma métrica que avalia a qualidade da clusterização com base no grau de separação entre os clusters, variando de -1 a 1. Ele considera a média das distâncias entre um ponto e os demais no mesmo grupo (a), e entre esse ponto e o cluster mais próximo ao qual não pertence (b), sendo calculado por (b - a) / max(a, b). Pontuações mais altas indicam clusters mais bem definidos e uma melhor escolha do valor de K.

A visualização de dados geralmente é usada em conjunto com outras técnicas de data mining para entender melhor os padrões encontrados, levando a uma visão mais aprofundada de suas características, exibindo esses resultados de várias formas, como, por exemplo, visualização duas ou três dimensões, diagramas de dispersão e histogramas [Ernawati et al., 2021].

Conforme metodologia apresentada por Essayem et al. [2022], o processo de limpeza de dados inclui a exclusão de transações com detalhes ausentes, como valor da transação ou quantidade comprada, a exclusão de transações com valor total negativo ou zero (representam produtos devolvidos) e a mesclagem das tabelas de dados para vincular os detalhes do cliente às transações correspondentes.

3. Metodologia

A metodologia desta pesquisa foi estruturada em etapas sequenciais, iniciando-se com a coleta de dados, por meio da extração de informações de vendas de 100 filiais a partir da plataforma Kaggle. Em seguida, procedeu-se ao tratamento dos dados, com eliminação de valores nulos e duplicados, padronização de formatos, remoção de símbolos monetários, criação de novas variáveis como “total” e “total_profit”, transformação de dados de data e hora, e exclusão de registros do ano com dados incompletos. Após isso, realizou-se a agregação dos dados por filial, calculando-se os atributos de Recência, Frequência e Valor Monetário (RFM). A próxima etapa consistiu na identificação e separação de outliers, utilizando análise de quartis e amplitude interquartil para isolar filiais com comportamentos atípicos. Posteriormente, foi aplicada a normalização dos dados não-outliers, com as transformações Box-Cox e MinMaxScaler, a fim de padronizar a escala das variáveis. Em seguida, executou-se a clusterização com K-Médias, utilizando os métodos do cotovelo e da silhueta para definir o número ideal de agrupamentos. Após a clusterização das filiais não-outliers, deu-se início à análise específica das filiais outliers, com a aplicação de uma nova clusterização voltada à detecção de discrepâncias em um ou mais atributos RFM. Foi gerada a visualização gráfica dos agrupamentos, com uso de representações tridimensionais e gráficos do tipo *violin plot*, a fim de facilitar a análise interpretativa dos clusters formados. Em seguida, foram escolhidos nomes para cada cluster, conforme perfil dos dados, elencando pontos estratégicos para a abordagem de cada cluster e, por fim, foi gerada uma geovisualização destes.

4. Estudo de caso: 100 filiais de uma varejista no estado do Texas

O conjunto de dados usado neste relatório foi obtido por meio do website Kaggle, que é uma plataforma que permite aos usuários encontrar e publicar conjuntos de dados. Os conjuntos de dados escolhidos neste estudo atual estão relacionados aos dados de vendas de 100 unidades de uma organização de varejo atuante nos Estados Unidos da América. Este conjunto de dados disponibilizado no Kaggle no formato CSV consiste em 10.051 dados e 11 colunas, que são os seguintes atributos: invoice_id", "Branch", "City", "category", "unit_price", "quantity", "date", "time", "payment_method", "rating" e “profit_margin”. Este estudo utiliza a linguagem de programação Python para o desenvolvimento da análise. A verificação, tratamento e análise dos

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

dados foi realizada através de uma implementação em linguagem de programação Python por meio das seguintes ações:

a) Limpeza, pré-processamento e criação de variáveis:

- Funções “drop” e “drop_duplicates”: empregadas a fim de eliminar as linhas que continham valores “Nan” e valores duplicados;
- Funções “isnull.sum” e “dropna”: verificação e exclusão das linhas nulas;
- Função “replace”: utilizada para retirar o símbolo “$” dos dados da coluna “unit_price”;
- Criação da coluna “total” sendo resultado da mutiplicação dos valores da “unit_price” pelos valores da “quantity”;
- Funções “strip” e “lower”: empregadas a fim de remover os espaços extra nos nomes das colunas e transformar todas as letras do nome destas em letras minúsculas;
- Função “to_datetime” da biblioteca “pandas”: empregada a fim de converter a coluna date (do tipo “str”) para o tipo “datetime64”, que permite operações com datas;
- Funções “day_of_week” e “month”: utilizadas para criação das colunas “day_of_week”, “month” a partir da coluna “date”, representando respectivamente o dia da semana e o mês da respectiva compra, conforme dados da coluna “date”;
- Função “to_datetime”: empregada a fim de transformar os dados da coluna time em dados de hora, variando de 0 à 23.
- Criação da coluna “total_profit”, sendo resultado da multiplicação dos valores das colunas “quantity”, “profit_margin” e “unit_price”;
- Foi verificado se havia algum período com dados faltantes. Com isso, foi identificado a ausência de dados relativos à 9 do total de 12 meses, no ano de 2019, sendo assim, esse ano e seus respectivos dados foram retirados do frame de dados trabalhado.

b) Análise exploratória e transformação dos dados RFM:

- O tratamento de dados resultou no total de entradas de 8969. Além disso, foi criado um DataFrame agregando os dados em três colunas de entradas, sendo elas Valor Monetário, Frequência e Recência, a fim de realizar a análise RFM.
- Foi calculado o total de vendas ao longo dos meses para os quatro anos de dados referidos às 100 filiais, conforme apresentado na Figura 2.

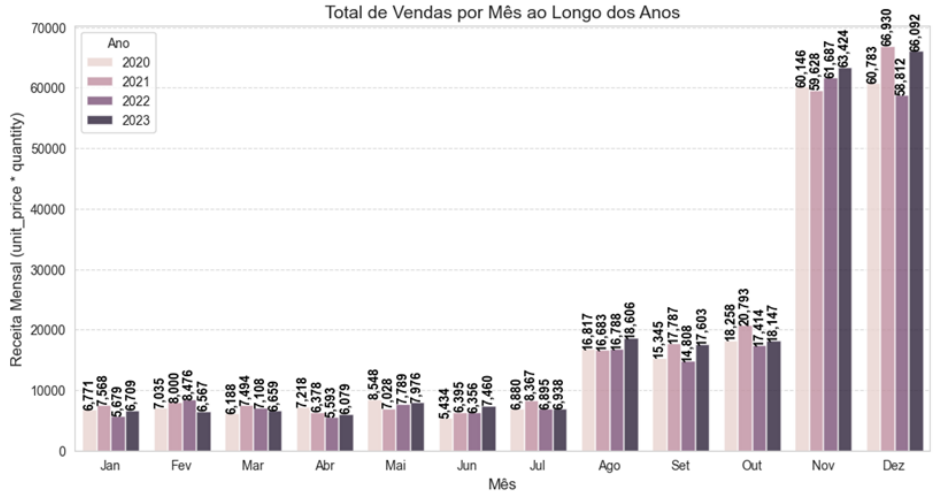


Figura 2: Ilustração de inicialização e interações no k-médias

- Nas Figuras 3 e 4, os dados deste DataFrame agregado de todas as filiais, ou seja, considerando tanto filiais que apresentavam quanto as que não apresentavam valores átipicos para uma das três colunas (“Valor Monetário”, “Frequência” e “Recência”). Os dados foram plotados no histograma e no boxplot, para os valores de cada uma desses três atributos.

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

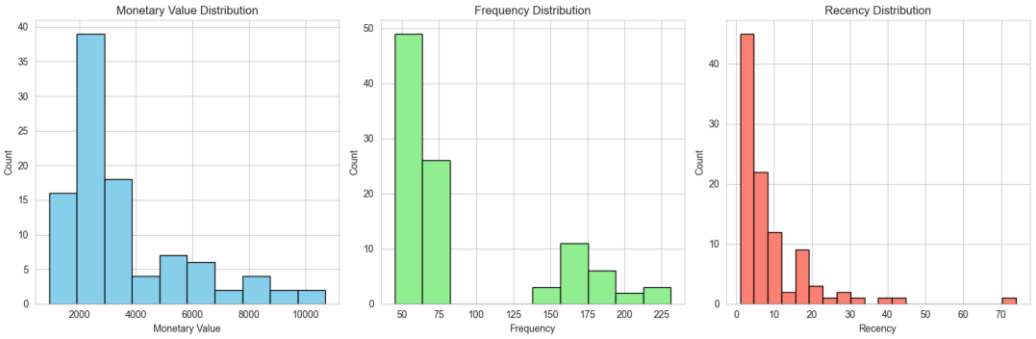


Figura 3: Histograma dos dados RFM das filiais não-*outliers*

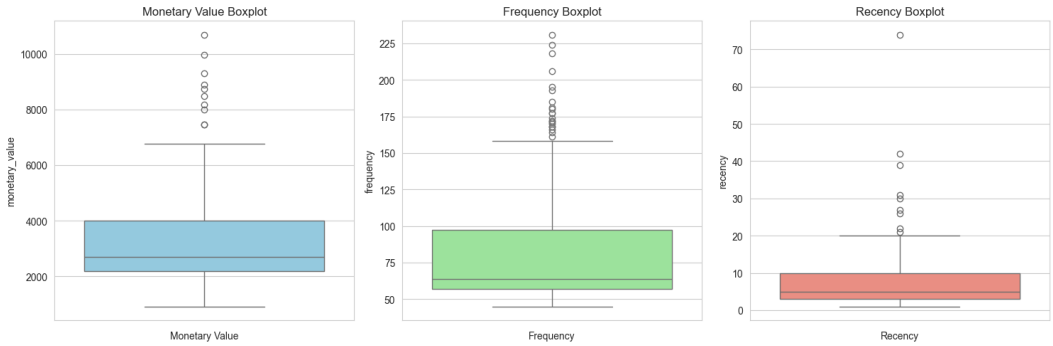


Figura 4: Box-plot dos dados RFM das filiais não-*outliers*

- Foi realizada a separação das filiais com dados *outliers* (atípicos) em um DataFrame específico e em outro DataFrame ficaram armazenados os dados das filiais com dados não-*outliers*, conforme cálculo dos quartis, amplitude interquartil e definição dos bigodes, de modo que 69 das 100 filiais foram identificadas como não-*outliers*.
- Foi aplicada, conforme dados apresentados na Figura 5, a transformação Box-Cox seguida pela MaxMinScaler para colocar variáveis em uma escala comum para algoritmos de aprendizado de máquina, especialmente quando as variáveis apresentam assimetria, de forma que o Box-Cox ajuda a lidar com a forma da distribuição, e o MinMaxScaler garante que todas as variáveis tenham a mesma faixa de valores, o que proporciona inclusive uma melhor visualização do comportamento dos dados, o que pode ser verificado na Figura 6, quando da apresentação dos dados referentes aos atributos RFM em três dimensões.

Dados originais:

	monetary_value	frequency	recency
0	2359.44	64	8
1	2006.28	55	9
3	2726.40	53	3
5	3142.08	66	3
6	2997.60	59	6

Dados transformados com Box-Cox:

	monetary_value	frequency	recency
	13.561166	0.390076	2.585166
	13.112899	0.390072	2.766699
	13.969010	0.390071	1.230260
	14.377023	0.390077	1.230260
	14.240824	0.390074	2.159542

Dados escalados com MinMaxScaler:

	monetary_value	frequency	recency
	0.455857	0.619377	0.627719
	0.373484	0.418900	0.671798
	0.530801	0.356869	0.298727
	0.605776	0.651406	0.298727
	0.580749	0.521468	0.524371

Figura 5: Dados das filiais não-outliers originais, com aplicação do Box-Cox e do MinMaxScaler

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

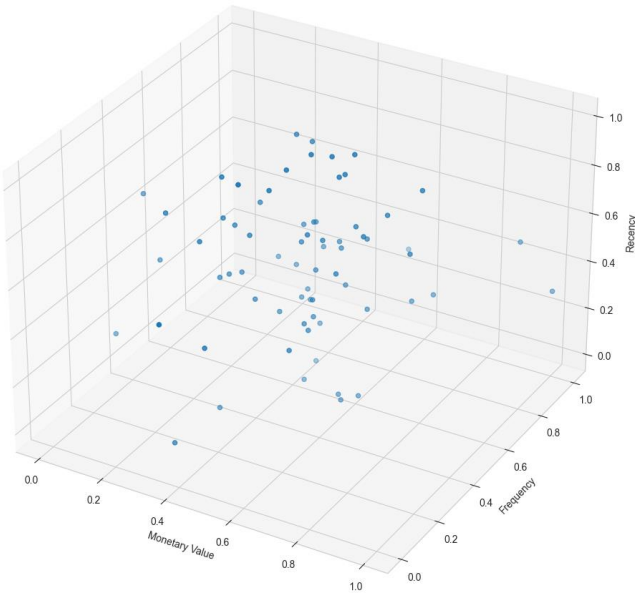


Figura 6: Apresentação em três dimensões dos dados normalizados das filiais não-outliers

c) Escolha do hiperparâmetro "K" e aplicação do algoritmo de aprendizado de máquina:

- A Figura 7 ilustra os resultados do método Elbow aplicado ao DataFrame, exibindo a relação entre o número de *clusters* (K) e a inércia e entre o número de *clusters* (K) e a pontuação de silhueta, calculados a fim de se escolher o valor de K. O gráfico inclui valores para K que variam de 2 a 12, com a Inércia plotada no eixo y no primeiro gráfico e com a Pontuação de Silhueta plotada no eixo y no segundo gráfico. Com base na análise, um valor de K = 9 parece ideal, equilibrando simplicidade e precisão na segmentação dos dados para o agrupamento k-médias no DataFrame.

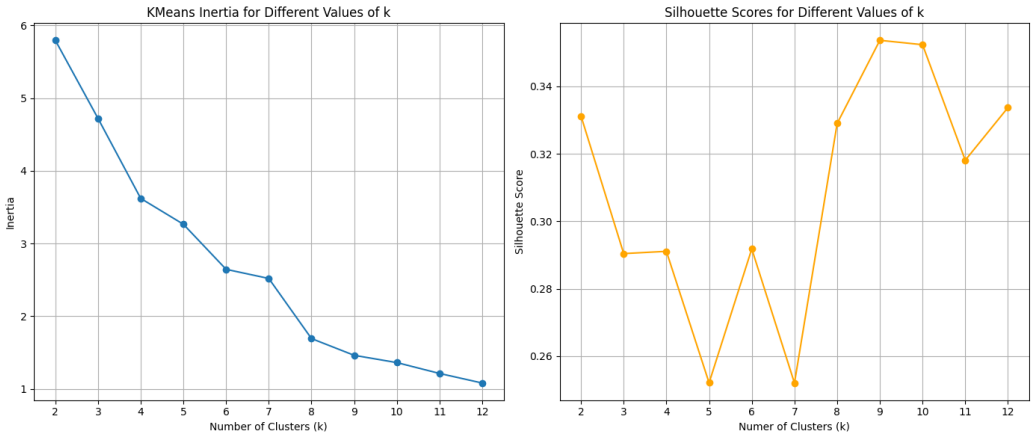


Figura 7: Aplicação do Elbow e do Silhouette Score aos dados das filiais não-outliers

- Após isso, foi realizado a análise e o tratamento dos dados RFM das 31 filiais classificadas como *outliers*, resultando em quatro clusters de filiais: outliers positivos de "Valor Monetário", outliers positivos de "Frequência", outliers negativos de "Recência" e outliers positivos ao mesmo tempo de "Valor Monetário" e de "Frequência".

5. Perfis estratégicos dos clusters e geovisualização

Os dados previamente clusterizados em 13 grupos foram representados por meio de gráficos violin plot, visualizado na Figura 8, na legenda da qual, ao lado da identificação de cada cluster, consta entre parênteses a quantidade de filiais do cluster. Essa visualização permitiu observar a distribuição das métricas de Recência, Frequência e Valor Monetário dentro de cada

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

cluster, oferecendo insights sobre a concentração e variação de desempenho entre as filiais. A partir dessas leituras, foi possível interpretar o perfil dominante de cada grupo e propor estratégias específicas, considerando o posicionamento relativo das lojas nos clusters. Desta forma, a análise gráfica complementa a clusterização numérica ao destacar outliers positivos e permitir decisões mais bem direcionadas com base nas diferenças comportamentais captadas pelo modelo.

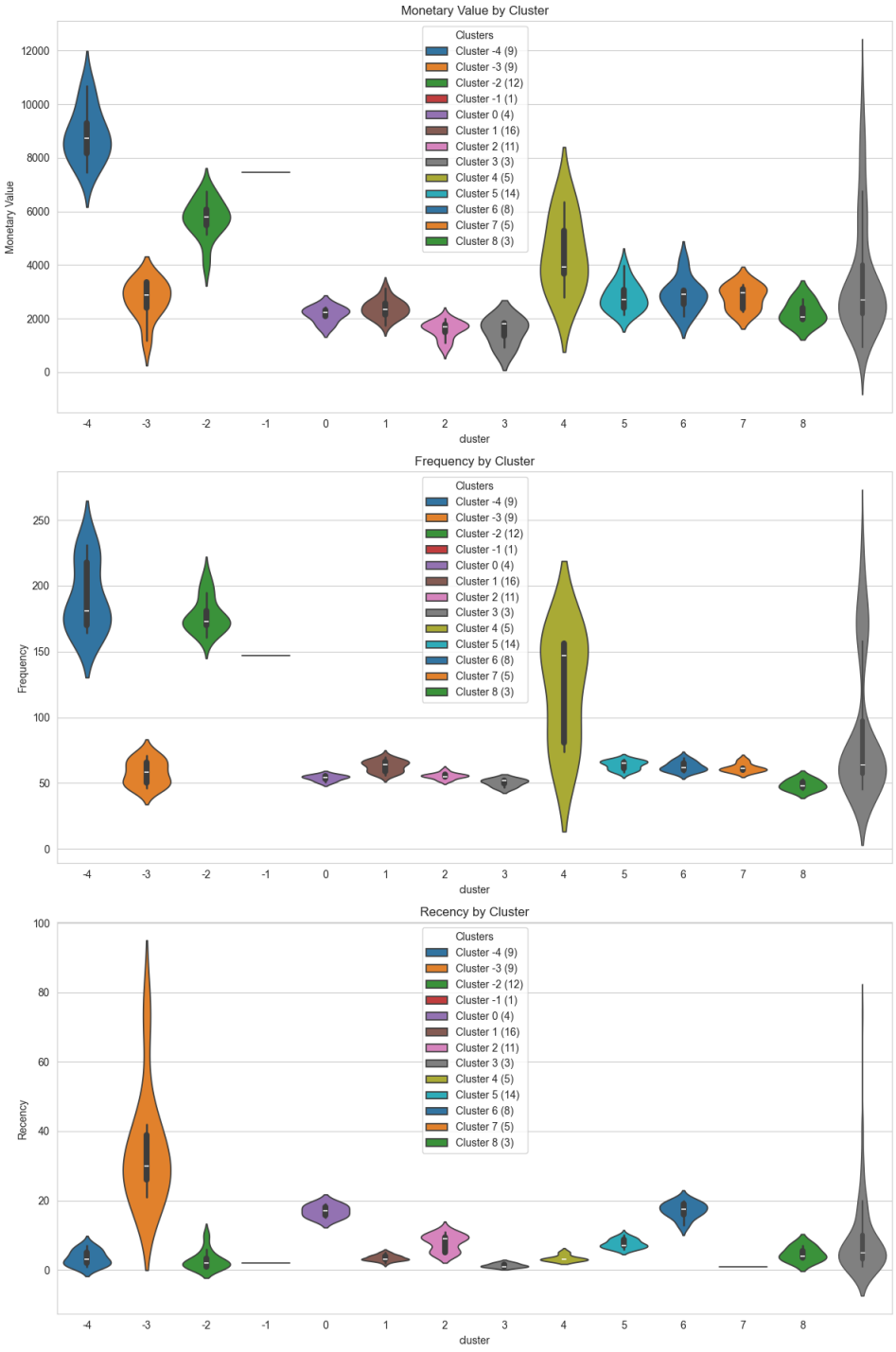


Figura 8: Gráfico violino dos atributos RFM dos 13 clusters

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

De acordo com a aplicação de todos as análises e métodos demonstrados, foram identificados segmentos de clientes distintos, permitindo a elaboração adequada de estratégias a fim de otimizar os resultados das filiais. A segmentação, baseada em modelo RFM, revelou perfis com características e necessidades específicas, demandando abordagens personalizadas, conforme apresentados a seguir:

Cluster	Pontos Estratégicos
Clientes de Altíssimo Valor (Cluster -4: Campeões)	Engajamento máximo e foco na maximização do Lifetime Value (LTV) e retenção.
Clientes Leais em Risco (Cluster -3: Apaixonados)	Risco iminente de churn, exigindo mitigação proativa e reativação preventiva.
Compradores Frequentes (Cluster -2: Leais)	Alto potencial de crescimento, com estratégias para expansão do ticket médio e fidelização.
Unidade Experimental (Cluster -1: Laboratório de Inovação)	Validação estratégica e otimização operacional contínua.
Compradores Moderados (Cluster 0: Promissores)	Potencial latente, com foco no incremento de LTV e frequência de compra.
Compradores Recentes (Cluster 1: Novos Apaixonados)	Alto potencial de conversão, demandando consolidação da primeira experiência.
Compradores Ocasionais (Cluster 2: Paqueradores)	Histórico positivo, com necessidade de aumento de frequência e reativação.
Clientes Sensíveis a Custo-Benefício (Cluster 3: Pechinchadores)	Manutenção da recorrência via estímulos financeiros.
Compradores Esporádicos de Valor (Cluster 4: Don Juans)	Alto ticket com baixa recorrência, requerendo estímulo à frequência.
Clientes com Potencial Latente (Cluster 5: Potenciais Crescentes)	Volume moderado, com foco na conversão em clientes de alto valor.
Clientes em Declínio (Cluster 6: Quase Perdidos)	Risco de churn acentuado, exigindo diagnóstico e reativação.
Novos Compradores de Baixo Valor (Cluster 7: Aprendizes)	Necessidade de educação e incentivo à recorrência.
Compradores Frequentes de Baixo Ticket (Cluster 8: Potenciais Apaixonados)	Estratégias para elevação do valor médio do pedido.

Tabela 1: Nomeação e descrição resumida dos clusters

A abordagem preditiva e adaptativa, com foco em personalização granular, demonstrou ser eficaz. As principais métricas para avaliação do sucesso das estratégias incluem LTV, retenção e ticket médio.

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

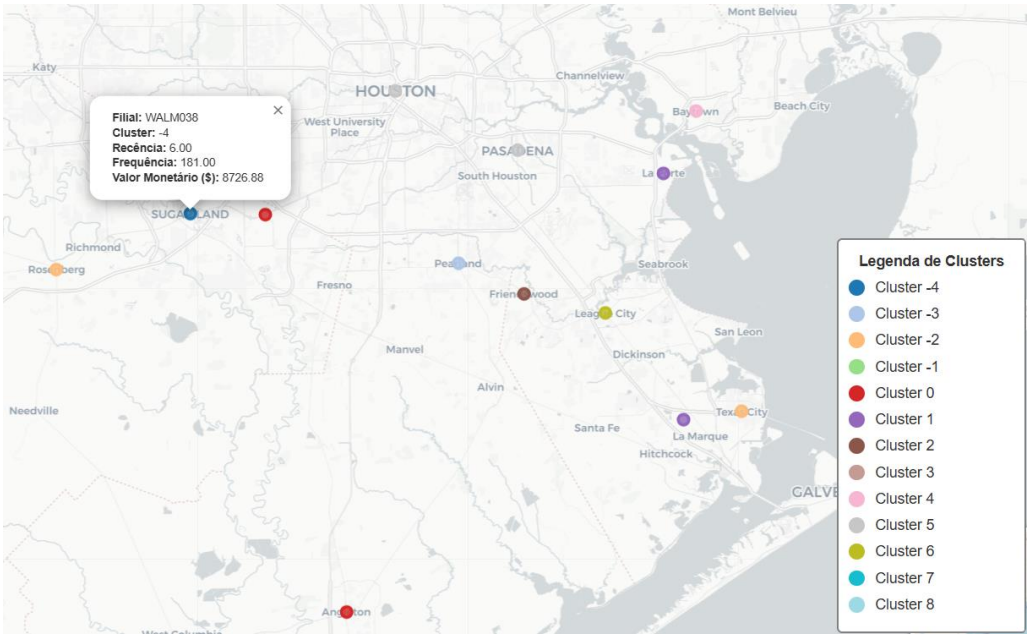


Figura 9: Visualização geoespacial das filiais por clusters

A Figura 9 exibe um recorte da geovisualização interativa, ilustrando a distribuição das lojas por clusters. O exemplo destaca a filial WALM038. Nele, são detalhados o cluster da loja (-4) e seus indicadores de Recência, Frequência e Valor Monetário. É possível notar a variedade de filiais pertencentes à diferentes clusters nas proximidades.

6. Conclusões

Este artigo apresentou a aplicação do método RFMKIG, uma integração entre o modelo RFM, o algoritmo de agrupamento K-Médias (*K-Means*) e estratégias de segmentação, com o objetivo de agrupar 100 lojas de uma rede varejista norte-americana com base no comportamento de compra dos clientes e fornercer um direcionamento prático contundente. Inicialmente, foram calculadas as métricas de Recência, Frequência e Valor Monetário para cada loja, e em seguida, aplicou-se o algoritmo K-Médias para identificar padrões de similaridade entre as unidades. A última etapa consistiu na interpretação qualitativa dos grupos formados e de sua visualização geoespacial, a qual pode servir de insumo para ações comerciais direcionadas.

Os resultados obtidos revelaram clusters coerentes, identificando grupos de lojas com desempenhos similares e destacando unidades com comportamento atípico, tanto positivas quanto negativas. A solução proposta oferece subsídios claros para o planejamento estratégico desse tipo de companhia, alocação de recursos e políticas comerciais personalizadas.

Trabalhos futuros podem aplicar essa abordagem em diferentes setores do varejo ou com outras variáveis comportamentais, além de explorar técnicas alternativas de agrupamento, complementar a análise com fatores psicológicos, com demais formas de identificação do valor de K a ser escolhido, análise das relações inter-filiais e como isso pode afetar os resultados e estratégias de vendas e abastecimento das filiais. Outra possibilidade é a utilização dos resultados de clusterização do presente estudo para definições estratégicas e comerciais específicas e a análise dos resultados gerados, a partir da utilização do modelo, com a criação de um *dashboard* interativo via PowerBI ou Looker, podendo aumentar a robustez da análise e possibilitando a identificação de ainda mais características dos clusters, das filiais separadamente e do comportamento de vendas destas, representando caminhos promissores para aprofundar os resultados obtidos.

Referências

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br

Al-Dabbas, L., Al-Tarawneh, H. e Al-Rawashdeh, T. A. (2023). Customer personality segmentation using K-means clustering. 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 537–543. https://doi.org/10.1109/ICIT58056.2023.10225996

Barus, O. P., Nathasya, C. e Pangaribuan, J. J. (2023). The implementation of RFM analysis to customer profiling using K-means clustering. Mathematical Modelling of Engineering Problems, 10(1), 298–303. https://doi.org/10.18280/mmep.100135

Ernawati, E., Baharin, S. S. K. e Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, 1869(1), 012085. https://doi.org/10.1088/1742-6596/1869/1/012085

Essayem, W., Bachtiar, F. A. e Priharsari, D. (2022). Customer clustering based on RFM features using K-means algorithm. 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Malang, Indonesia, 23–27. https://doi.org/10.1109/CyberneticsCom55287.2022.9865572

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D. S. e Le, T. G. (2023). An extended RFM model for customer behaviour and demographic analysis in retail industry. Business Systems Research Journal, 14(1), 26–53. https://doi.org/10.2478/bsrj-2023-0002

Ling, L. S. e Weiling, C. T. (2025). Enhancing segmentation: A comparative study of clustering methods. IEEE Access, 13, 47418–47439. https://doi.org/10.1109/ACCESS.2025.3550339

Mufarroha, F. A., Suzanti, I. O., Satoto, B. D., Syarief, M., Husni e Yunita, I. (2022). K-means and K-medoids clustering methods for customer segmentation in online retail datasets. IEEE 8th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 223–228. https://doi.org/10.1109/ITIS57155.2022.10010135

Nandapala, E. Y. L. e Jayasena, K. P. N. (2020). The practical approach in customers segmentation by using the K-means algorithm. 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 344–349. https://doi.org/10.1109/ICIIS51140.2020.9342639

Zhou, J., Zhai, L. e Pantelous, A. A. (2021). Market segmentation using high-dimensional sparse consumers data. Expert Systems with Applications, 145, 113136. https://doi.org/10.1016/j.eswa.2019.113136

Wahyuni, E., Surono, S. e Eliyanto, J. (2021). Outlier detection using K-means clustering with Minkowski–Chebyshev distances for inquiry-based learning results in students dataset. 2021 International Conference on Artificial Intelligence and Big Data Analytics, Bandung, Indonesia, 1–5. https://doi.org/10.1109/ICAIBDA53487.2021.9689736

Kaggle. Web page Dataset. https://www.kaggle.com/datasets/najir0123/walmart-10k-sales-datasets . Acessado em: 2025-03-05.

DOI: 10.59254/sbpo-2025-212098
https://proceedings.science/p/212098?lang=pt-br