UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise do comportamento de compra dos clientes no *e-commerce* brasileiro utilizando técnicas de mineração de dados

Gabriella Cristina da Silva Benzi

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Gabriella Cristina da Silva Benzi

Análise do comportamento de compra dos clientes no *e-commerce* brasileiro utilizando técnicas de mineração de dados

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva Neto

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

da Silva Benzi , Gabriella Cristina
da Análise do comportamento de compra dos clientes no e-commerce brasileiro utilizando técnicas de mineração de dados / Gabriella Cristina da Silva Benzi ; orientador Afonso Paiva Neto. -- São Carlos, 2024.
63 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. E-commerce. 2. Mineração de dados. 3. Segmentação de clientes. I. Paiva Neto, Afonso, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

AGRADECIMENTOS

Agradeço ao Professor Dr. Afonso Paiva Neto pela orientação, paciência e apoio no desenvolvimento desse trabalho.

Aos meus amigos, em especial à minha amiga Fernanda que sempre me apoiou e incentivou em diversas etapas importantes da minha vida.

Por fim, aos meus familiares por toda dedicação e suporte que sempre me deram.

RESUMO

S. Benzi, G. **Análise do comportamento de compra dos clientes no *e-commerce* brasileiro utilizando técnicas de mineração de dados**. 2023. 63p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Com o avanço da tecnologia e a expansão do comércio eletrônico, entender o comportamento dos clientes tornou-se fundamental para as empresas que desejam se manter competitivas no mercado. No Brasil, o comércio eletrônico tem crescido significativamente nos últimos anos, impulsionado pela conveniência, variedade de produtos e facilidade de acesso oferecidos aos consumidores. Nesse contexto, este estudo se propõe a analisar o comportamento de compra dos clientes no comércio eletrônico brasileiro, utilizando técnicas de mineração de dados aplicadas a um conjunto de dados reais fornecido pela Olist na plataforma *Kaggle*, com mais de 100 mil pedidos entre os anos de 2016 e 2018. Compreender os padrões e tendências de compra dos clientes pode fornecer *insights* valiosos para estratégias de *marketing*, operações e tomada de decisões no setor de comércio eletrônico. Além disso, este estudo incluirá uma análise de clusterização para segmentar os clientes com base em suas preferências e comportamentos de compra. Através do entendimento dos comportamentos de compras, é possível extrair *insights* valiosos que possam orientar estratégias de negócios, promovendo ações de *marketing* mais assertivas e segmentadas. Ao examinar detalhadamente as datas de compras com as maiores quantidades de pedidos, os tipos de pagamento mais utilizados e o Modelo de Recência, Frequência e Valor (RFV), este estudo busca oferecer uma compreensão mais profunda do comportamento dos clientes no comércio eletrônico brasileiro e destacar áreas de oportunidade para otimização e personalização das estratégias empresariais.

Palavras-chave: Comércio eletrônico. Mineração de dados. Segmentação de clientes. Análise de padrões. Comportamento.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A indústria de comércio eletrônico se expandiu rapidamente, devido ao crescente desenvolvimento da *Internet* o que fez com que os requisitos dos consumidores ficassem mais rígidos em relação às compras *online* (CHIA-JUNG *et al.*, 2020). Desse modo, a atuação neste setor é fundamental para empresas que buscam competitividade no mercado e também ao seu público alvo que almeja por melhores preços, produtos e disponibilidade. Diante dessa premissa, o entendimento do comportamento dos consumidores virtuais é visto como essencial para atender às suas necessidades e às expectativas de consumo, e assim desenvolver estratégias personalizadas.

De acordo com Wang e Huang (2022), mineração de dados pode ser definido como a extração de conhecimento útil de um grande volume de informações a fim de transformá-las em outras formas, como definições, modelos e leis. Assim, a utilização dessa técnica no *e-commerce* contribui para melhoria da experiência do cliente, já que através da identificação de padrões e tendências do comportamento do consumidor, é possível encontrar segmentações e públicos-alvo, prever necessidades futuras, reter e fidelizar clientes.

1.1 Objetivos

Neste projeto de pesquisa propõe-se aplicar técnicas de mineração de dados, para analisar padrões e tendências de comportamento de compra de clientes no comércio eletrônico brasileiro. Os objetivos específicos são:

- Apresentar os diferentes métodos de mineração de dados mais utilizados na análise de comportamento de compras;
- Identificar as principais relações encontradas nas compras dos consumidores, como por exemplo, a frequência de datas de compra, os valores médios gastos, data da última transação, métodos de pagamentos envolvidos;
- Segmentar os clientes de acordo com seu consumo, visando estipular perfis com preferências similares;
- Aplicar a metodologia necessária em um conjunto de dados públicos de comércio eletrônico brasileiro, de modo a avaliar a eficácia dos métodos de mineração de dados utilizados, e fornecer uma análise embasada nos resultados encontrados.

1.2 Estrutura do Trabalho

Este trabalho está estruturado da seguinte forma: no Capítulo 2 são apresentados os trabalhos relacionados ao tema proposto, dos quais foram extraídos as técnicas de mineração utilizadas, e os métodos para comparação de eficácia. No Capítulo 3 exploram-se conceitos que contribuem para uma compreensão mais aprofundada do trabalho e das metodologias empregadas. No capítulo 4, é detalhada a metodologia utilizada, bem como o algoritmo e as características empregados na mesma. No Capítulo 5, aprofundam-se os detalhes do estudo, incluindo a descrição do procedimento, a apresentação dos resultados e a análise correspondente. No capítulo 6, conclui-se o trabalho com um resumo geral e são oferecidas algumas observações sobre possíveis melhorias.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo proporciona uma análise detalhada das técnicas de mineração de dados aplicadas ao estudo do comportamento do consumidor no comércio eletrônico. Aborda-se a jornada do consumidor durante o processo de compra, a importância da mineração de dados nesse contexto e uma comparação aprofundada das técnicas mais relevantes.

As bases do *marketing* moderno estão fundamentadas nas necessidades humanas, com foco especial nas necessidades dos consumidores. Essas necessidades são essenciais para a compreensão do conceito de *marketing*. No cenário competitivo do *marketing*, a capacidade fundamental para a sobrevivência, rentabilidade e crescimento de uma empresa é sua aptidão em identificar e satisfazer as demandas dos consumidores, superando as ações da concorrência (SCHIFFMAN; KANUK, 2000).

Segundo Hawkins e Mothersbaugh (2018), decisões de *marketing* embasadas em premissas e teorias têm maior probabilidade de sucesso do que aquelas baseadas em palpites. Assim, o conhecimento aprofundado sobre o consumidor se torna uma vantagem competitiva, reduzindo consideravelmente o risco de erros e decisões equivocadas.

Kotler e Keller (2018) afirmam que, geralmente, ao efetuarem uma compra, os consumidores costumam passar por cinco etapas: reconhecimento do problema, busca por informações, avaliação de alternativas, tomada de decisão de compra e comportamento pós-compra. Segundo os autores, a primeira fase, reconhecimento do problema ou necessidade, marca o início do processo, sendo desencadeada por estímulos internos ou externos. Em seguida, a busca por informações torna-se uma etapa crucial, na qual os consumidores buscam informações relevantes para embasar suas decisões de compra, influenciados tanto por estímulos externos quanto por estratégias de *marketing*.

A terceira etapa, avaliação de alternativas, evidencia-se como um ponto-chave, onde os consumidores analisam e comparam diferentes marcas e produtos, levando em consideração atributos pertinentes. A tomada de decisão de compra, por sua vez, compreende a escolha da marca, do revendedor, da quantidade, da ocasião e da forma de pagamento, consolidando a intenção de compra do consumidor.

O comportamento pós-compra completa o ciclo, sendo um estágio crítico que abrange a experiência do consumidor após a aquisição do produto. A dissonância cognitiva, eventualmente experimentada, destaca a importância de estratégias de *marketing* contínuas para manter uma percepção positiva do produto.

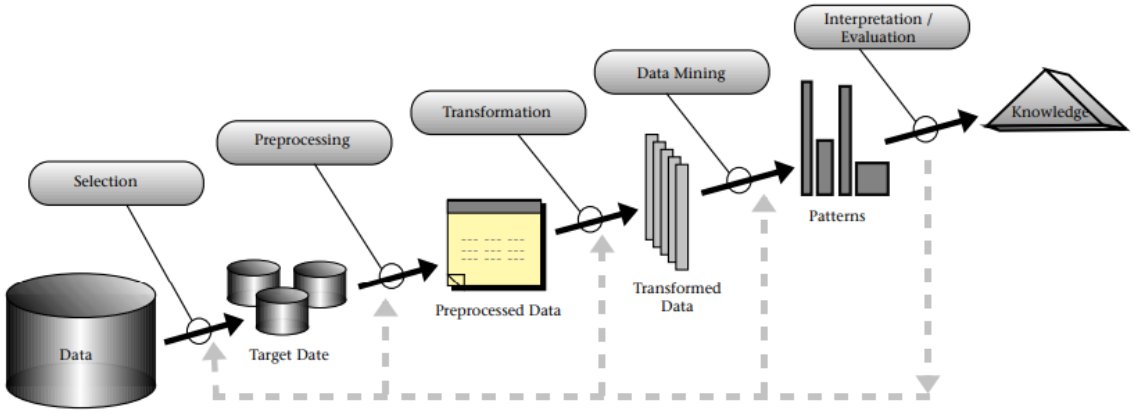
Assim, o profissional de *marketing* desempenha um papel crucial ao longo de todo o processo, não apenas estimulando o reconhecimento da necessidade e facilitando a busca

por informações, mas também monitorando a satisfação pós-compra e garantindo uma comunicação eficaz para promover a fidelidade do cliente e a reputação positiva da marca. Este entendimento profundo do processo de compra é essencial para desenvolver estratégias de *marketing* eficazes e sustentáveis.

No cenário atual, as organizações revelam competência ao coletar grandes volumes de dados. Contudo, muitas ainda enfrentam desafios significativos ao transformar eficientemente essas informações em conhecimento aplicável às suas atividades (CÔRTES; PORCARO; LIFSCHITZ, 2002). Nesse contexto, a Mineração de Dados se destaca como uma ferramenta amplamente reconhecida, capacitando a descoberta de informações e a exposição de estruturas de conhecimento que são cruciais para orientar decisões.

Para atender a esse cenário, surge uma nova área denominada Descoberta de Conhecimento em Bases de Dados (KDD). A mineração de dados, constitui uma das fases dessa área (GOLDSCHMIDT; PASSOS, 2005). A Figura 1 representa uma visão geral das etapas que compõe o processo KDD:

Figura 1 – Etapas do processo de Descoberta de Conhecimento em Bancos de Dados.



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), a mineração de dados constitui uma etapa essencial no ciclo do KDD, na qual é empregada a análise de dados e algoritmos de descoberta dado as restrições de eficiência computacional, visando identificar padrões e modelos específicos nos dados.

A mineração busca por relacionamentos recorrentes em um determinado conjunto de dados (HAN; KAMBER; PEI, 2012). Técnicas como Apriori, um algoritmo para minerar conjuntos de itens frequentes para regras de associação booleanas, classificadores de árvores de descisão e bayesianos, assim como a comparação dos classificadores com base em curvas

de custo-benefício e ROC, são explorados pela mesma.

De acordo com Harrison (1998), não há uma abordagem única que seja capaz de solucionar todos os desafios na mineração de dados. Diferentes técnicas são aplicadas para finalidades distintas, e cada uma delas apresenta seus próprios benefícios e limitações.

As funcionalidades da mineração de dados são aplicadas para definir os tipos de informações a serem extraídas nas atividades de mineração. De modo geral, essas atividades podem ser divididas em duas categorias principais: descritivas, cujo objetivo é caracterizar as propriedades gerais dos dados; e preditivas, que visam fazer inferências a partir dos dados para realizar previsões (CASTRO; FERRARI, 2017).

A análise de *cluster* é uma das ferramentas mais valiosas no contexto da mineração de dados, permitindo a identificação de grupos e a revelação de distribuições e padrões relevantes presentes no conjuntos de dados (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

No contexto do comércio eletrônico, a aquisição de informações dos clientes é amplamente facilitada pela prontidão da dados em grande escala, tornando pertinente a exploração dos serviços de mineração de dados para conferir significado empresarial a essas informações (JIANG; YU, 2008).

Segundo Chia-Jung *et al.* (2020), resultados satisfatórios são atingidos na classificação, predição e extração de regras por meio de árvores de decisão. Tendo em vista prever a recompra do cliente plataforma de *e-commerce*, o autor compara a regressão logística de modelo linear e modelo *XGBoost* baseado em árvore de decisão.

Neto, Ramos e Silva (2019) apresentam o *Random Forest* como um destaque das técnicas de mineração de dados existentes na literatura e propõe uma solução para a concessão de cupons de descontos em comércio eletrônico, utilizando-o como estimador escolhendo o ponto de corte através do método estatístico Kolmogorov Smirnov (KS). O estudo traz como contribuição um direcionamento inicial para construção de soluções de mineração de dados em *web-shop*, diante de desafios como a decisão do método de mineração a ser utilizado, estratégia para tratamento de valores faltantes e escolha do melhor ponto de corte.

Rahardja Untung e Hariguna (2019) propôs uma técnica de mineração de texto usando o algoritmo *K-medoid* para analisar as opiniões dos clientes. A pesquisa realizou análises de sentimento em grandes conjuntos de dados de produtos de dois sites de comércio eletrônico, obtidos de vários clientes online.

Koul e Philip (2021) conclui que a segmentação de clientes é crucial para aprimorar o desempenho do comércio eletrônico, destacando o algoritmo *K-Means* como uma técnica eficaz para essa finalidade.

Jiang e Yu (2008) mostram que a coleta de informações valiosas sobre o comportamento dos compradores faz com que o *marketing* torne-se eficaz para o público-alvo. Através da utilização do algoritmo *K-Means*, com atributos de idade, gênero, tempo, endereço, idioma e tipo de comportamento, segmentou o público em *clusters*.

Namvar, Gholamian e KhakAbi (2010) citam as variáveis populacionais, Recência, Frequência, Monetização (RFM) e Valor Vitalício do Cliente (LTV) como os tipos mais frequentemente empregados no agrupamento de clientes.

Song e Shepperd (2006) identificaram quatro *clusters* de clientes com características distintas no processo de segmentação de clientes com o modelo RFM e o algoritmo *K-means* aplicado aos dados de transações de uma loja de varejo online no Reino Unido. Eles citam a segmentação de clientes como uma ferramenta poderosa para obtenção *insights* sobre o comportamento dos clientes e para impulsionar estratégias de negócios.

Hossain (2017) utiliza os algoritmos *K-means* e *DBSCAN* para agrupar dados, representando abordagens centróides e baseadas em densidade, respectivamente. Ambos os algoritmos mostraram eficácia na segmentação de clientes, porém, o *DBSCAN* destacou-se na identificação de clientes com hábitos de consumo incomuns. Os resultados da implementação do *DBSCAN* foram particularmente significativos, sugerindo sua relevância para uma segmentação precisa dos clientes.

3 CONCEITOS BÁSICOS

Seguindo a definição de Linoff e Berry (2011), a Mineração de Dados visa identificar padrões e regularidades a partir da exploração e análise automatizada ou semi-automatizada de grande bases de dados. Os autores enfatizam que o propósito da mineração de dados é capacitar as organizações a adquirir informações para o desenvolvimento de estratégias aprimoradas em áreas como *marketing*, vendas e suporte, contribuindo, dessa forma, para o progresso de suas atividades.

Para melhor entendimento da análise do comportamento de compra dos clientes no contexto do *e-commerce* brasileiro, que será descrito com maiores detalhes mais a frente, vamos passar por alguns conceitos básicos. Inicialmente, será discutida a definição do modelo RFV. Em seguida, serão abordados os algoritmos pertinentes para este trabalho, como *K-means, K-medoids* e *DBSCAN*, juntamente com a métrica de silhueta para avaliação da qualidade dos *clusters* obtidos.

3.1 Modelo RFV

O conceito de RFV (ou RFM) refere-se a três componentes essenciais: recência, frequência e valor monetário do cliente, no qual recência mede o período decorrido desde a última interação, frequência avalia quão frequentemente um cliente realiza transações e Valor Monetário representa o gasto médio por transação efetuado pelo cliente (PINHO, 2009).

Dado um determinado período de tempo, é possível calcular esses três indicadores para cada cliente e em seguida, agrupar em segmentos de mesmo tamanho com base na sua recência. Em seguida, subdividir os clientes de cada segmento em novos segmentos de tamanho igual com base no número de compras efetuadas e por fim, novamente subdivir em um novo conjunto de segmentos com base no valor monetário das compras (ALENCAR *et al.*, 2006). Esse procedimento tem por fim categorizar os clientes com base em diferentes aspectos do seu comportamento de compra.

Observações recorrentes indicam que clientes que fizeram aquisições recentes de um ou mais produtos têm a tendência de manter compras regulares e de aportar quantias significativas à empresa durante um período específico, o que eleva a probabilidade de realizarem novas compras.

Embora o amplamente utilizado, o RFM possui limitações, como se concentrar quase exclusivamente nos melhores clientes, já que se baseia no comportamento de compra de quando e quanto foi comprado, e também por muitas vezes possuir metade dos clientes representados como os que não compram com frequência, gasta, pouco ou não compraram

recentemente (MIGLAUTSCH, 2002).

3.2 Clustering

De acordo com Anitha e Patil (2022), um *cluster* pode ser definido como um grupo conceitualmente significativo de objetos que compartilham características comuns. As análises de agrupamento, as vezes categorizados como identificação de padrões ou exploração de dados, consistem em um conjunto de métodos empregados para identificar a estrutura subjacente em um conjunto de dados, com a finalidade de dividir observações similares em agrupamentos ou categorias que possuam relevância ou utilidade (STAHL; SALLIS, 2012) .

A técnica de *clustering* demonstrou sua eficácia na realização da segmentação de clientes. Essa abordagem é uma forma de aprendizado não supervisionado, capaz de identificar grupos em conjuntos de dados que não possuem rótulos predefinidos (KANSAL TUSHAR E BAHUGUNA, 2018).

3.2.1 Métodos hierárquicos

Um método de clusterização hierárquica subdivide as observações em uma estrutura de *clusters* hierárquica por meio de uma estrutura de árvore.

Existem duas abordagens para decompor os dados:

- Abordagem aglomerativa: objetos inicialmente distribuídos em diferentes *clusters* e em seguida, de forma iterativa, combinados em grupos maiores até que uma condição de parada seja atendida ou que todos os objetos estejam no mesmo *cluster*.
- Abordagem divisiva: objetos inicialmente agrupados no mesmo cluster, e de maneira iterativa, esse *cluster* é dividido em vários grupos até que uma condição de parada seja alcançada ou todos os objetos estejam distribuídos em diferentes *clusters*.

As iterações são irreversíveis, o que faz com que não seja possível a correção de decisões que se mostraram incorretas ou não ideais, porém traz vantagem significativa em relação a outros métodos devido ao seu custo substancialmente reduzido. Para mitigar esses tipos de decisões, é recomendado realizar um pré-processamento, que possa reduzir o conjunto de dados, diminuindo o número de variáveis consideradas, ou reduzir os erros iniciando com um algoritmo de hierarquia aglomerativa e, posteriormente, aprimorando os resultados com o algoritmo divisivo.

3.2.2 K-Means

O algoritmo *K-Means* é uma técnica amplamente reconhecida em segmentação de clientes, utilizado para agrupar dados de forma não supervisionada com base na

similaridade. Kansal Tushar e Bahuguna (2018) citam o algoritmo como sendo o mais simples de *clustering* baseado em princípio de particionamento.

O funcionamento do *K-Means* envolve o cálculo de distâncias ou similaridades entre as observações no conjunto de dados e a formação de *clusters* com base nessa proximidade. Neste modelo, a distância euclidiana é empregada para quantificar a distância entre os pontos. Além disso, o algoritmo requer a especificação do número de *clusters* na amostra como um parâmetro de entrada. No entanto, é comum desconhecer o número de *clusters* existentes em um conjunto de dados.

O método de cotovelo é utilizado para estimar um valor de K (centróides), no qual avalia-se uma métrica de qualidade para os agrupamentos dos modelo para diferentes valores (KAUFMAN; ROUSSEEUW, 2009, pp. 110). Normalmente a métrica utilizada é a Soma dos Quadrados dos Erros (SSE) dado por:

$$SSE = \sum_{i=1}^{n} \min_{c_j \in C} ||x_i - c_j||^2$$

onde:

- n é o número total de observações no conjunto de dado;
- c_j é o centróide do cluster j;
- C é o conjunto de todos os K centróides;
- $||x_i - c_j||$ representa a distância euclidiana entre a observação x_i e o centróide c_j.

O algoritmo *K-Means* busca encontrar K centróides $c_1, c_2, \ldots, c_K$ de modo que a SSE seja minimizada. Isso é feito por meio de um processo iterativo que envolve a atribuição de cada observação ao *cluster* mais próximo com base na distância euclidiana e a atualização dos centróides dos *clusters*. O algoritmo continua iterando até que não haja mais alterações significativas nas atribuições de observações aos *clusters*.

Em resumo, o *K-Means* possui implementação do algoritmo relativamente simples e escalável para grandes conjuntos de dados. No entanto, possui limitações que devem ser consideradas, como as iterações do algoritmo frequentemente convergirem para um ponto de ótimo local, o que significa que o ponto de ótimo global não é garantido, além da necessidade de determinar previamente o número de *clusters* (K), envolvendo a construção de diferentes modelos, com a seleção do K baseada no desempenho do modelo e por ser sensível a *outliers*, o que pode afetar a qualidade dos agrupamentos.

3.2.3 K-medoids

O *K-medoids* é uma variação do *K-means*, na qual principal diferença está na escolha dos representantes de cada *cluster*. De acordo com Rahardja Untung e Hariguna

(2019), o *K-medoids* seleciona um ponto de dados real do conjunto como o representante do cluster, tornando-o mais robusto a *outliers* e ruído nos dados, uma vez que o ponto medoide é menos sensível a valores extremos do que a média. Os k *clusters* são formados ao atribuir cada ponto de dados restante ao medoide mais próximo, utilizando alguma métrica de distância como critério. Enquanto o *K-Means* utiliza a soma das distâncias euclidianas quadradas entre os pontos de dados e os centróides para definir os *clusters*, o *K-Medoids* escolhe k objetos de dados como medóides, minimizando a soma das dissimilaridades entre os objetos de dados e seus medóides correspondentes, o que reduz ruídos e valores discrepantes. (ARORA; DEEPALI; VARSHNEY, 2016)

No entanto, sua utilização em situações que envolvem grandes volumes de dados e um alto número de *clusters* a serem identificados pode se tornar dispendiosa. Uma abordagem para superar essa restrição é treinar os modelos em uma amostra aleatória dos dados da população, com a ressalva de que o ideal é que essa amostra seja representativa do conjunto original de dados, uma vez que os objetos representativos de cada *cluster* serão selecionados com base nessa amostra.

3.2.4 DBSCAN

O método *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) é um algoritmo de agrupamento fundamentado na densidade, permitindo a formação de *clusters* de qualquer formato e tamanho em conjuntos de dados, inclusive quando há presença de ruído e valores discrepantes (KHAN *et al.*, 2014). Esta metodologia revela-se eficaz na criação de agrupamentos de alta qualidade, destacando-se por suas vantagens em comparação com outros algoritmos convencionais de análise de agrupamento.

A configuração dos *clusters* é influenciada pela densidade de pontos em regiões específicas do conjunto de dados. Nessa abordagem, a principal premissa reside na consideração da vizinhança de um raio específico para cada ponto, exigindo a presença de, um número mínimo de pontos dentro desse raio. Os pontos de *cluster* referem-se àqueles que estão agrupados, exibindo uma densidade mais elevada em comparação aos pontos fora dos *clusters*. Por outro lado, os pontos de ruído são caracterizados por não pertencerem a nenhum dos *clusters* identificado (KOUL; PHILIP, 2021).

Logo, é necessária a definição de dois parâmetros:

1. **Eps:** Raio máximo dentro do qual dois pontos são considerados membros do mesmo *cluster*.

2. **MinPts:** Quantidade mínima de pontos em uma região para a densidade desejada.

3.3 Métrica de Silhueta

O índice de silhueta é uma técnica usada para interpretar e validar a consistência dos grupos obtidos através da clusterização, analisando a distância entre *clusters* resultantes. Este índice fornece uma medida que avalia o quão semelhante um ponto de dados é aos outros pontos dentro do mesmo *cluster* (coesão) e em comparação aos pontos dos *clusters* vizinhos (separação).

O índice de silhueta varia de -1 a 1 e é expresso pela fórmula:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

onde:

- $S(i)$ representa o índice de silhueta para um ponto específico;
- $a(i)$ é a distância média entre o ponto e os outros pontos no mesmo *cluster*;
- $b(i)$ é a distância média entre o ponto e os pontos no cluster vizinho mais próximo.

Uma pontuação de silhueta próxima de 1, indica que o ponto está bem ajustado ao seu próprio *cluster*, enquanto se próxima de 0, sugere que o ponto de dados pode ser igualmente adequado para outro *cluster*, e próxima de -1 indica que o ponto de dados provavelmente está no *cluster* errado (SHAHAPURE; NICHOLAS, 2020).

Tabela 1 – Interpretação de resultados da silhueta média.

Silhueta Média	**Interpretação Separação**
0,71 - 1,00	Excelente
0,51 - 0,70	Razoável
0,26 - 0,50	Fraca
$\leq$ 0,25	Ruim

Fonte: Adaptado de Burney e Tariq (2014).

Em resumo, quanto mais próxima a pontuação de silhueta estiver de 1, melhor será a clusterização do ponto de dados, enquanto valores próximos de 0 ou -1 indicam problemas na clusterização.

4 METODOLOGIA

Este capítulo se dedica a esclarecer os métodos utilizados para o processo de análise do comportamento de compra dos clientes no *e-commerce* brasileiro utilizando técnicas de mineração de dados. Com esse propósito, os experimentos conduzidos têm como objetivo abordar as seguintes questões:

P1: Quais são as variáveis, incluindo métricas de RFV e outros atributos relacionados ao comportamento de compra, que desempenham um papel mais significativo na identificação de segmentos de clientes com diferentes padrões de compra e interação em um ambiente de *e-commerce*?

P2: Entre os métodos de segmentação de clientes, *K-means* e *K-medoids* e *DBSCAN*, qual demonstra ser mais eficaz na segmentação de clientes com base em seus padrões de compra e interação no ambiente de *e-commerce*?

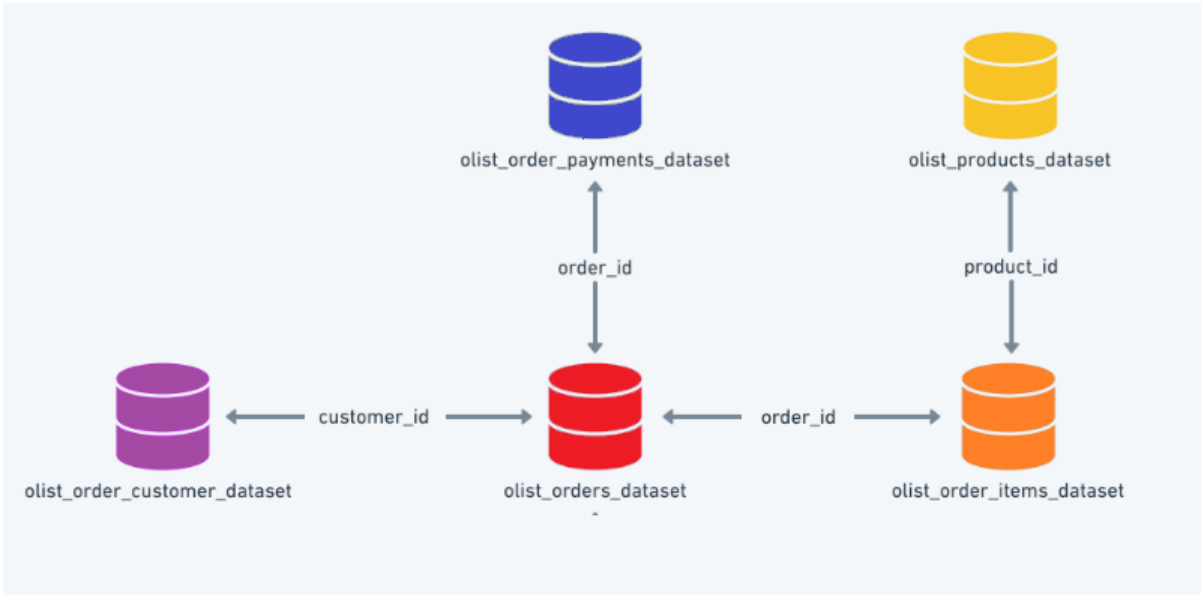
4.1 Base de Dados

Para a realização dos experimentos mencionados, utilizamos a base de dados fornecida pela empresa Olist por meio da plataforma Kaggle. Esses dados fornecem informações relativas a pedidos, produtos, clientes, pagamentos e itens de um comércio eletrônico brasileiro com detalhes especificados em cada uma das tabelas conforme destacado na tabela 2. Além disso, a figura 2 fornece uma representação visual que resume as conexões entre as tabelas utilizadas.

Tabela 2 – Descrição do Conjunto de Dados.

Tabela	**Descrição**
olist_orders_dataset	Informações principais que permitem a identificação de outras informações a partir de cada pedido.
olist_customers_dataset	Informações sobre o cliente, onde cada pedido é atribuído a um ID de cliente exclusivo. O customer_unique_id permite a identificação de recompras pelo cliente.
olist_products_dataset	Informações sobre os produtos vendidos.
olist_order_items_dataset	Informações sobre os itens comprados em cada pedido.
olist_order_payments_dataset	Informações sobre as opções de pagamento de pedidos.

Figura 2 – Conjunto dos Dados.



Fonte: Adaptado de Olist; Sionek, André (2018).

4.2 Variáveis e Métricas Relevantes ao Comportamento de Compras Selecionadas

Nesta seção, descreveremos as variáveis utilizadas na análise, destacando sua importância no entendimento do comportamento de compras dos clientes no contexto do *e-commerce* brasileiro.

Tabela olist_orders_dataset

- Order_id:
  - **Definição:** Identificador único de cada pedido.
  - **Formato:** Texto.
  - **Importância:** Permite identificar pedidos exclusivamente e facilita o acompanhamento do ciclo de vida de um pedido, desde a compra até a entrega.
- Customer_id:
  - **Definição:** Identificador do cliente associado ao pedido.
  - **Formato:** Texto.
  - **Importância:** Associa cada pedido a um cliente específico e facilita a análise do comportamento de compra individual assim como a identificação de padrões de compra de clientes específicos.
- Order_purchase_timestamp:

- **Definição**: Data indicando quando o pedido foi realizado.
- **Formato**: Data e hora.
- **Importância**: Fornece informações sobre a data das transações, possibilita análises sazonais e compreensão de tendências ao longo do tempo.

- Order_status:
  - **Definição:** Indica o estado atual do pedido, como pendente, processando, enviado, entregue, cancelado.
  - **Formato:** Texto.
  - **Importância:** Fornece inforamção sobre o progresso do pedido.

Tabela olist_customers_dataset

- Customer_id:
  - **Definição:** Identificador único de cada cliente para cada pedido.
  - **Formato:** Texto.
  - **Importância:** Permite a identificação única de cada cliente em cada pedido, sendo essencial para associar informações de diferentes tabelas.
- Customer_unique_id:
  - **Definição:** Identificador único e de cada cliente.
  - **Formato:** Texto.
  - **Importância** Fornece uma identificação única do cliente, permitindo a rastreabilidade do histórico de compras de um cliente ao longo do tempo.

Tabela olist_order_payments_dataset

- Order_id:
  - **Definição:** Identificador único de cada pedido.
  - **Formato:** Texto.
  - **Importância:** Serve como chave de ligação entre as tabelas, permitindo a integração de informações relacionadas a pedidos e pagamentos
- Payment_type:
  - **Definição:** Tipo de pagamento em cada pedido.
  - **Formato:** Texto.
  - **Importância:** Permite analisar os métodos de pagamento preferidos pelos clientes.

Tabela olist_order_items_dataset

- Price:
  - **Definição:** Valor monetário pago em cada item do pedido.
  - **Formato:** Valor numérico.
  - **Importância:** Representa o valor associado a cada item por transação, crucial para análises financeiras e para o entendimento da receita gerada por cada pedido.
- Order_id:
  - **Definição:** Identificador único de cada pedido associado à avaliação.
  - **Formato:** Texto.
  - **Importância:** Facilita a associação entre os itens e os pedidos, possibilitando a compreensão do contexto das compras.
- Freight_value:
  - **Definição:** Valor monetário pago no frete do pedido.
  - **Formato:** Valor numérico.
  - **Importância:** O custo do frete pode influenciar diretamente a decisão de compra do consumidor. Frequentemente, os clientes levam em consideração o valor total da compra, incluindo os custos de envio, ao decidir efetuar uma compra.

4.3 Pré Processamento dos Dados

O primeiro passo na construção de modelos de segmentação envolve a preparação dos dados no formato adequado. Isso inclui a carga dos dados, a identificação de valores ausentes e discrepantes e seleção dos campos relevantes para análise. Um *notebook* em *Python* foi utilizado para o processo de pré-processamento, com o objetivo efetuar a leitura dos dados e organizá-los de forma a torná-los adequados para análises subsequentes, através da plataforma online *Google Colab*. As bibliotecas utilizadas, pandas, matplotlib, numpy, sklearn, são públicas, e foram vistas durante o curso.

Tratamento da tabela de Pedidos

Na tabela denominada *olist_orders_dataset*, foram registrados 99.441 pedidos, abrangendo uma variedade de status de compras [3].

A fim de concentrar a análise focar a análise nas transações em que o pedido foi concluído, foram realizadas operações de filtragem nos pedidos que alcançaram os estágios

Tabela 3 – Quantidade de pedidos por status.

Status	**Quantidade**
delivered	96478
shipped	1107
canceled	625
unavailable	609
invoiced	314
processing	301
created	5
approved	2

de *shipped* e *delivered*. Subsequentemente, a variável temporal *order_purchase_timestamp* foi convertida para o formato de data, proporcionando maior eficiência nas análises temporais e na detecção de tendências ao longo do período. Este processo de transformação foi implementado com o propósito de assegurar uma representação temporal adequada associada a cada pedido. O próximo estágio consiste na exploração das características dos métodos de pagamento.

Na tabela *olist_payments_dataset*, foram selecionadas as colunas *order_id*, *payment_type* e *payment_value*. Inicialmente, a variável categórica *payment_type* foi submetida a uma transformação através da técnica de *one-hot encoding*, gerando variáveis *dummies* para cada categoria única de método de pagamento. Posteriormente, estas variáveis *dummies* foram incorporadas ao *DataFrame* de pagamentos original por meio da função "pd.concat"do pacote Pandas. Após essa concatenação, os dados foram agrupados pelo identificador único de cada pedido (*order_id*), consolidando as informações pertinentes a cada transação. A tabela resultante 4, representa uma visão agregada dos dados, onde cada linha corresponde a um pedido único, e as variáveis *dummies* indicam a presença ou ausência de cada método de pagamento.

Tabela 4 – 10 primeiros índices da tabela de pagamentos resultante.

order_id	value	boleto	credit_card	debit_card	not_defined	voucher
00010242fe8c5a6d1ba2dd792cb16214	72.19	0	1	0	0	0
00018f77f2f0320c557190d7a144bdd3	259.83	0	1	0	0	0
000229ec398224ef6ca0657da4fc703e	216.87	0	1	0	0	0
00024acbcdf0a6daa1e931b038114c75	25.78	0	1	0	0	0
00042b26cf59d7ce69dfabb4e55b4fd9	218.04	0	1	0	0	0
00048cc3ae777c65dbb7d2a0634bc1ea	34.59	1	0	0	0	0
00054e8431b9d7675808bcb819fb4a32	31.75	0	1	0	0	0
000576fe39319847cbb9d288c5617fa6	880.75	0	1	0	0	0
0005a1a1728c9d785b8e2b08b904576c	157.60	0	1	0	0	0
0005f50442cb953dcd1d21e1fb923495	65.39	0	1	0	0	0

Na tabela *olist_customers_dataset* foram selecionados os campos *customer_id*, *customer_unique_id*, e *customer_state*, proporcionando uma visão simplificada para as análises subsequentes. A identificação única do cliente, representada por *customer_unique_id*, mostrou que existem 96.096 valores distintos. Esse dado sugere que existem

clientes que realizaram mais de uma compra, indicando uma recorrência e permitindo uma avaliação da frequência de compras individual.

Optou-se por não utilizar as variáveis da tabela *olist_order_items_dataset* previamente selecionados, uma vez que o valor a pagar da varíavel *payment_value* já reflete a o preço dos produtos acrescido do custo de frete.

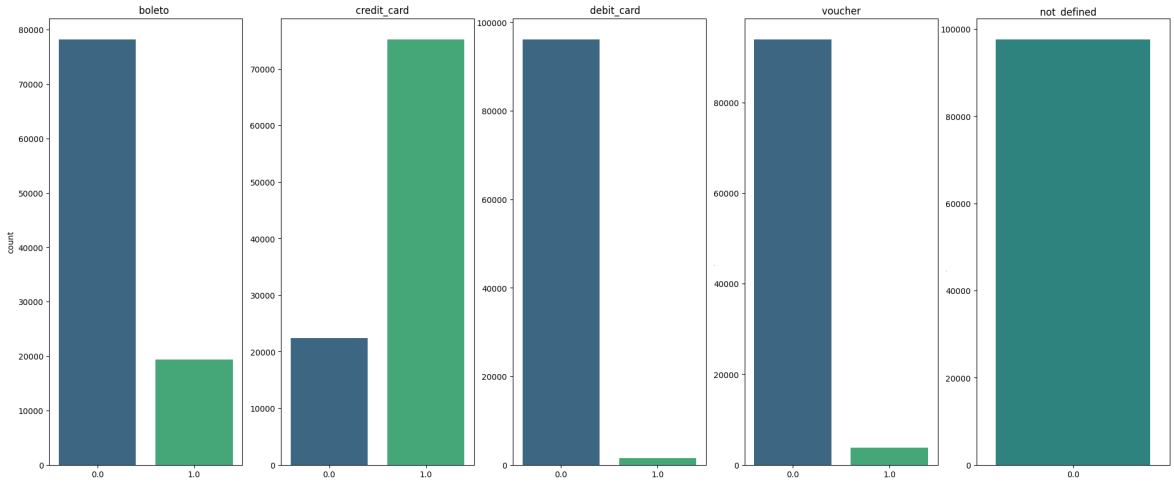
A integração de dados é uma etapa essencial do processo de preparação dos dados, visando unificar informações relevantes de diferentes tabelas. No contexto deste trabalho, a integração envolveu a junção de múltiplos *dataframes*, com o objetivo de gerar a tabela de pedidos final, denominada *df_orders_final*. Inicialmente, os dados dos clientes foram mesclados aos dados dos pedidos, empregando a coluna *customer_id* como chave primária para estabelecer a relação entre eles. Posteriormente, os dados de pagamentos foram incorporados, utilizando a coluna *order_id* para efetuar a junção.

O procedimento, conduzido com métodos de mesclagem apropriados, proporcionou uma visão abrangente dos dados de pedidos, clientes e pagamentos em um único *dataframe* resuntando na tabela *df_orders_final* com 97.585 registros. Cabe destacar que a análise de clusterização é sensível a valores ausentes, e uma abordagem para detecção e correção desses campos é essencial para resultados mais confiáveis e significativos. Assim, a etapa de pré-processamento incluirá a verificação e tratamento adequado dos campos nulos, preparando os dados para a aplicação eficaz das técnicas de clusterização.

Durante a análise de dados, foram identificados valores ausentes em apenas dois registros, afetando atributos de pagamento como *value*, *boleto*, *credit_card*, *debit_card*, *not_defined* e *voucher*. Dada a natureza limitada dessas ocorrências, optou-se pela exclusão desses registros do conjunto de dados.

Adicionalmente, observou-se a ausência de registros com valor 1 na coluna *not_defined*:

Figura 3 – Quantidade de Clientes para Tipos de Pagamento (0 e 1).



Fonte: Autor.

Diante da inexistência de ocorrências únicas nesta categoria e, portanto, sua irrelevância para a análise, optou-se pela exclusão desta coluna do conjunto de dados. Essa medida visa simplificar a estrutura do conjunto de dados, retendo apenas as informações relevantes para a análise de clusterização em questão.

Ao término do processo de tratamento, foi obtida uma tabela de pedidos refinada, contendo 97.583 registros. Esta tabela incorpora as colunas pertinentes para análises subsequentes.

Agregação da Tabela de Clientes

Para obter as variáveis de interesse, recência, frequência e valor, a tabela de pedidos *df_orders_final* será agregada pela coluna *customer_unique_id*. As variáveis de interesse são definidas da seguinte forma:

- **Recência:** A recência será calculada subtraindo a data de cada compra da data máxima presente na tabela. Assim, temos:

$$\text{Recência} = \text{Data Máxima} - \text{Data da Compra}$$

  Posteriormente trazemos a recência mínima pelo customer_unique_id, para garantirmos a compra mais recente do cliente.

- **Frequência:** A frequência será obtida contando o número de ocorrências únicas de customer_unique_id na tabela de pedidos.

- **Valor:** O valor será calculado como a média dos valores dos pedidos para cada customer_unique_id.

O resultado é a tabela *df_rfv*, que contém as variáveis RFV para cada cliente.

Identificação do Uso dos Métodos de Pagamento

Após a agregação da tabela de clientes (*df_rfv)* com as variáveis RFV, procedeu-se à identificação da presença de compras em categorias específicas, representadas por variáveis binárias. As categorias de interesse incluem métodos de pagamento (*boleto*, *voucher*, *credit_card*, *debit_card*).

Cada uma das variáveis binárias passou por uma abordagem de identificação de uso da categoria, com o objetivo de determinar a porcentagem utilizada pelo cliente em suas transações. O método adotado envolveu a extração do valor médio presente na tabela consolidada de pedidos (*df_orders_final*) para cada cliente único (*customer_unique_id*). Assim, o valor médio dessas variáveis binárias reflete a percentagem utilizada pelo cliente do método de pagamento em suas compras.

As variáveis adicionais incorporadas fornecem uma perspectiva importante sobre as preferências de pagamento dos clientes, enriquecendo assim a compreensão do comportamento do consumidor. Como resultado deste processo, obtivemos a base final contendo as variáveis:

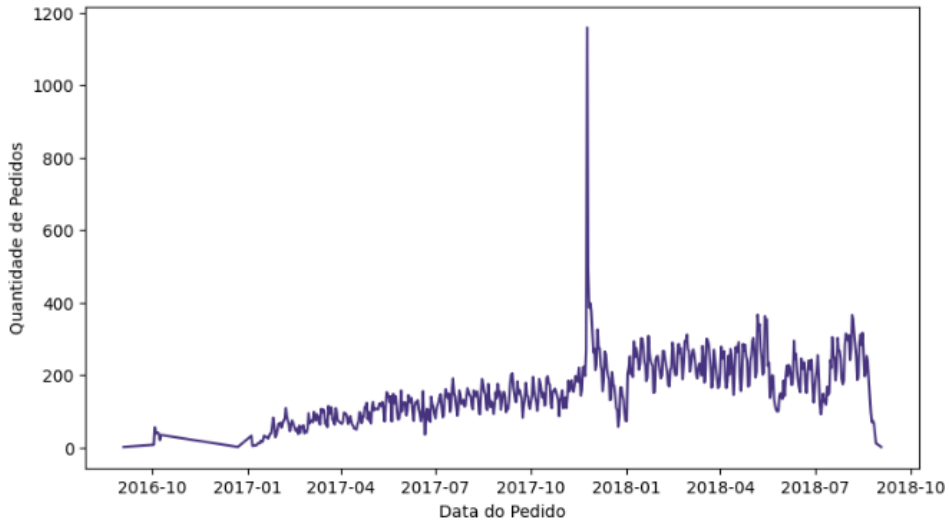
Tabela 5 – Descrição das variáveis.

Coluna	**Tipo**
customer_unique_id	object
valor	float64
recencia	int64
frequencia	int64
boleto	float64
credit_card	float64
debit_card	float64
voucher	float64
valor_2	float64

Análise Exploratória e Descritiva

Abaixo vemos a distribuição dos pedidos por dia dos pedidos entregues e enviados durante os anos 2016 a 2018. Inicia-se em 2016-09-04 e termina no dia 2018-09-03, resultando em 729 dias com uma média de 159,2 pedidos por dia e um pico de pedidos nos dia 2017-11-24, onde foram realizados 1.159 pedidos, possivelmente impulsionada pela Black Friday, data que é tradicionalmente associada a descontos e promoções especiais oferecidos pelos varejistas.

Figura 4 – Pedidos Por Dia.



Fonte: Autor.

A análise da tabela 6 apresenta as datas de compras com as maiores quantidades de pedidos, o que sugere que eventos como a Black Friday, Dias das Mães e dos Pais,

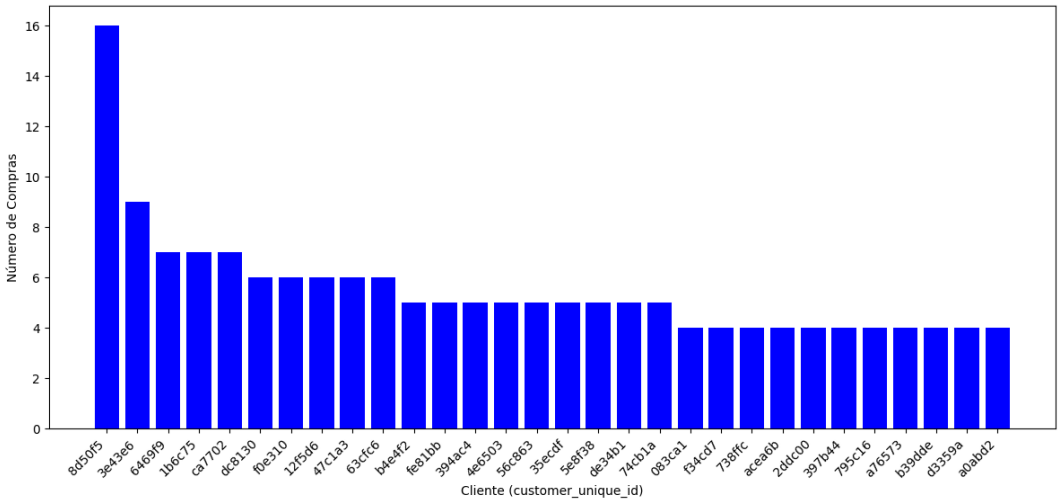
podem exercer uma significativa influência no comportamento de compra dos clientes. Esses períodos de intensa atividade de compras podem refletir campanhas promocionais ou datas comemorativas que estimulam a demanda e impulsionam o volume de vendas, indicando a importância desses eventos como oportunidades estratégicas para os varejistas..

Tabela 6 – Datas de Compra com Maiores quantidades de Pedidos.

Data de Compra	**Total de Pedidos**
2017-11-24	1159
2017-11-25	493
2017-11-27	398
2017-11-26	386
2017-11-28	374
2018-05-07	366
2018-08-06	365
2018-05-14	362
2018-08-07	354

No gráfico dos 30 primeiros clientes com mais compras, é possivel ver que um cliente se destaca com o total de 16 pedidos, enquanto os demais variam na faixa de 9 a 4 pedidos.

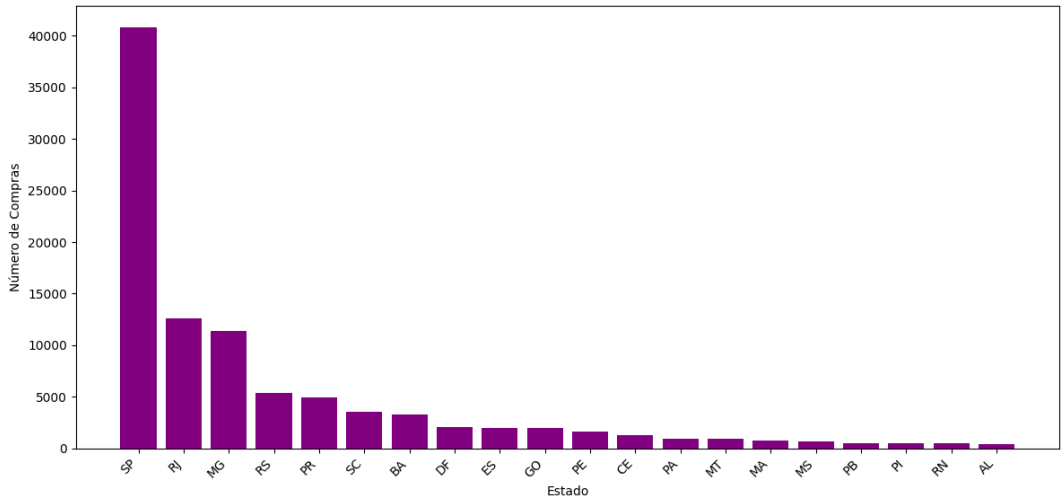
Figura 5 – Top 30 clientes com Maiores Compras.



Fonte: Autor.

No gráfico com os estados com maiores compras, é possível observar que grande quantidade está alocada em SP, superando 40 mil pedidos, seguido por RJ com 12 mil e posteriormente MG 11 mil. Os demais estão abaixo do 6 mil.

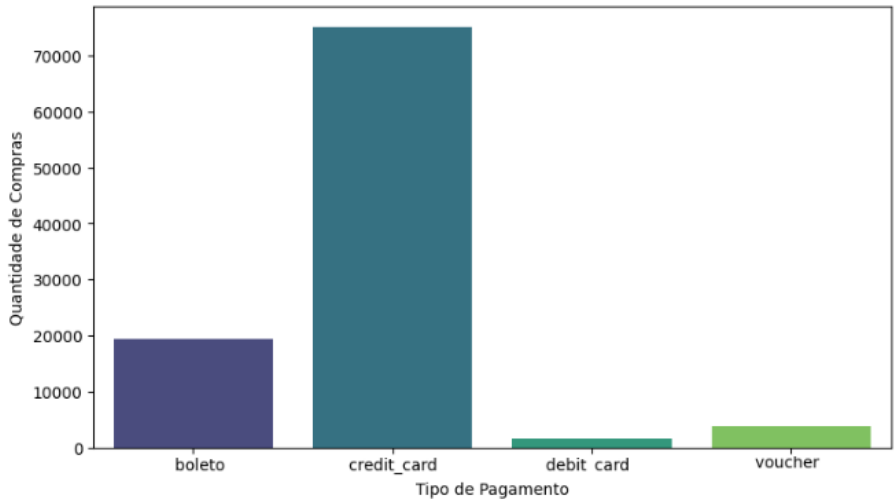
Figura 6 – 20 Estados Com Mais Compras Realizadas.



Fonte: Autor.

A análise da distribuição da quantidade de compras por tipo de pagamento revela padrões distintos. A maior parte das transações concentra-se no pagamento com cartão de crédito, ultrapassando 60 mil compras, seguido do boleto bancário, que totaliza quase 10 mil compras. As demais categorias de pagamento apresentam quantidades consideravelmente menores.

Figura 7 – Quantidade de Compras por Tipo de Pagamento.



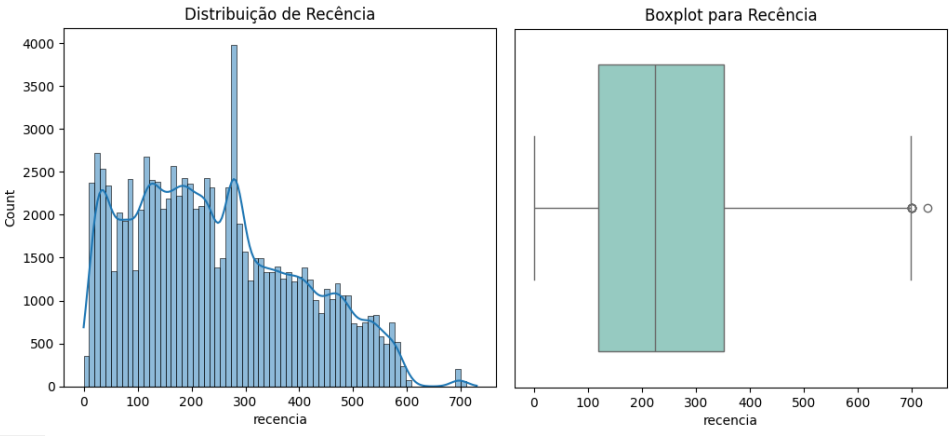
Fonte: Autor.

Esta distribuição destaca a predominância do pagamento com cartão de crédito entre os clientes, evidenciando a importância dessa modalidade de pagamento em relação às demais opções disponíveis.

Prosseguimos nossa análise considerando o Modelo RFV, que contempla três métricas distintas:

Ao examinar a métrica de recência, observa-se uma densidade mais significativa de valores em períodos mais recentes, especialmente abaixo de 1 ano. Essa distribuição sugere uma tendência de compras mais recentes por parte dos clientes, indicando que a base de clientes está atualmente engajada e realizando transações no curto prazo.

Figura 8 – Distribuição e Boxplot da Recência.

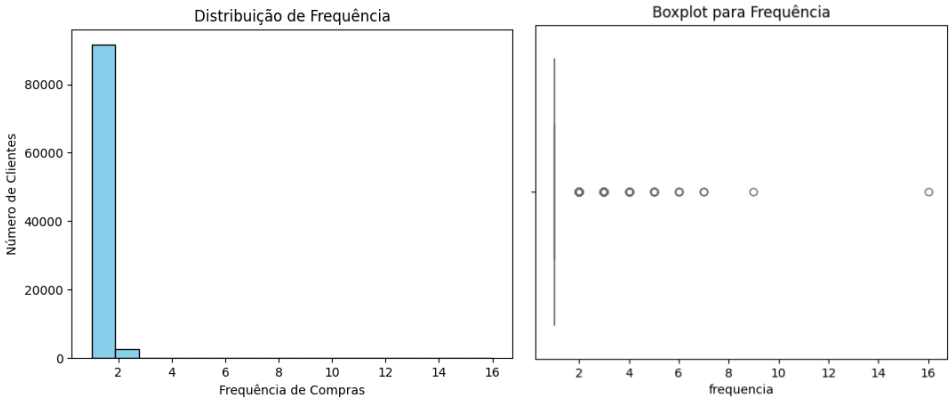


Fonte: Autor.

Além disso, o boxplot da recência também indica que não há muitos outliers, o que sugere uma consistência nas transações dos clientes. A ausência de valores extremos reforça a ideia de uma base de clientes que mantém padrões regulares de compras, sem eventos atípicos que impactem significativamente a métrica de recência.

A métrica de frequência, por sua vez, concentra-se principalmente no início, indicando que a grande maioria dos clientes realizou apenas uma compra. Essa observação sugere uma predominância de compras únicas na base de clientes, indicando que a maioria dos consumidores pode não ter uma frequência de compra recorrente.

Figura 9 – Distribuição e Boxplot da Frequência.

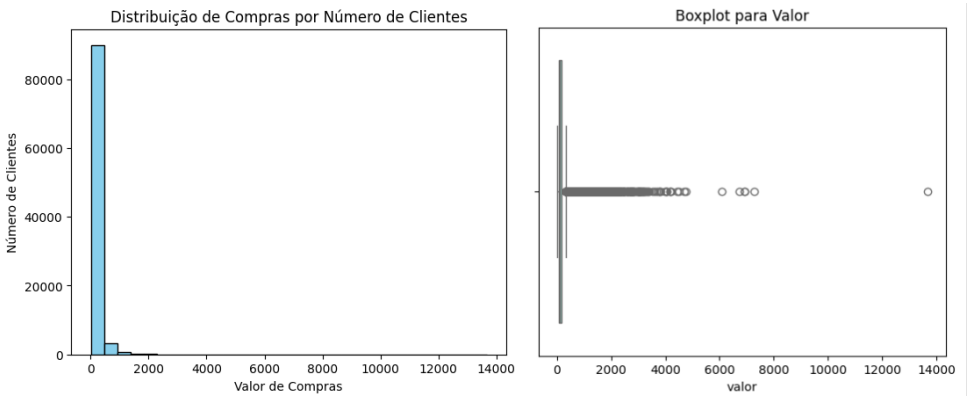


Fonte: Autor.

Entretanto, vale notar que alguns clientes possuem frequências mais altas, com até 16 compras. Essa variação na frequência de compras sugere a presença de um segmento de clientes mais engajadoos.

A métrica de valor concentra-se principalmente nos valores iniciais, especialmente entre 0 e 200. Isso sugere que a maioria dos clientes realiza compras de tickets mais baixos. À medida que o valor das compras aumenta, a quantidade de clientes em cada intervalo diminui, indicando que transações de alto valor são menos comuns.

Figura 10 – Distribuição e Boxplot do Valor.

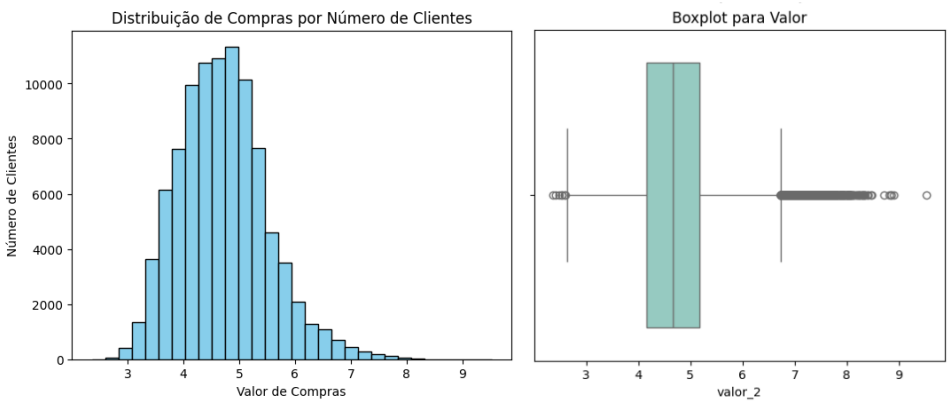


Fonte: Autor.

Além disso, observamos a presença de *outliers* que variam até 14 mil, indicando transações excepcionalmente elevadas. Esses *outliers* podem representar clientes de alto valor ou eventos de compra únicos que impactam significativamente a métrica de valor.

A transformação logarítmica revela uma distribuição mais homogênea dos valores, permitindo uma análise mais equilibrada da métrica de valor. Assim, foi atribuído o logaritmo do valor à variável valor_2. Essa abordagem ajuda a mitigar a influência dos *outliers* e proporciona uma visão mais precisa das tendências de compra.

Figura 11 – Distribuição e Boxplot do Valor Transformção Logarítma.



Fonte: Autor.

Padronização

Os modelos de clusterização comumente utilizam medidas de distância para agrupar dados semelhantes, e a amplitude dos valores em cada característica exerce influência ponderada na formação desses agrupamentos. Portanto é necessário realizar a padronização

dos dados antes de aplicá-los, devido à disparidade nos intervalos de distribuição de valores de cada característica. A padronização dos dados visa equalizar as escalas de todas as características, criando assim uma base homogênea para análise.

A padronização é expressa pela seguinte equação:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

onde:

- X_{std}: é o valor padronizado;
- X: é o valor original;
- μ: é a média da característica;
- σ: é o desvio padrão da característica.

Nesse processo, os valores originais são transformados para uma escala com média zero e desvio padrão unitário, garantindo que as variáveis contínuas contribuam de maneira equitativa para a formação dos clusters.

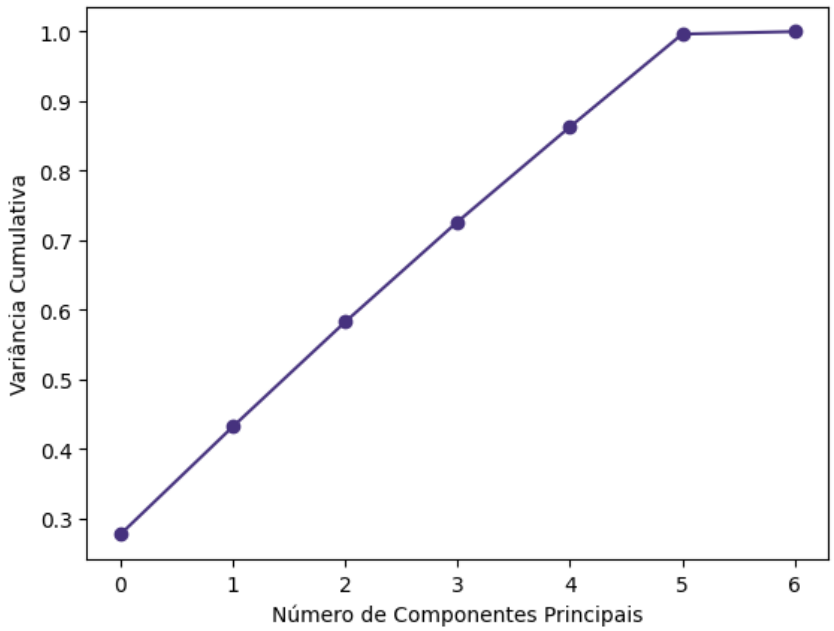
A padronização dos dados foi efetuada empregando o método StandardScaler do módulo *preprocessing* da biblioteca Sklearn, obtendo assim a base padronizada *df_rfv_standard*.

4.4 PCA

Definido e ajustado o vetor de características [***R***, ***F***, ***V***, ***boleto***, ***credit_card***, ***debit_card***, ***voucher***], prosseguimos com a redução da dimensionalidade utilizando a Análise de Componentes Principais (PCA).

A análise de componentes principais (PCA) é uma técnica estatística amplamente empregada para reduzir a dimensionalidade de um conjunto de características (VINODHINI; CHANDRASEKARAN, 2014). Ela é empregada para explorar a estrutura subjacente dos dados e identificar padrões significativos. Isso é particularmente útil quando lidamos com um conjunto de características extenso, visando simplificar a representação sem perder informações importantes.

Figura 12 – Gráfico Variância Cumulativa.



Fonte: Autor.

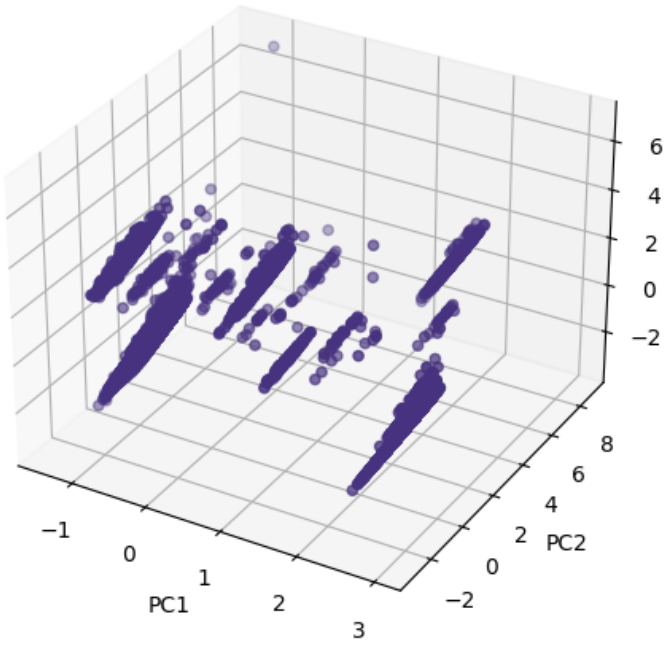
O gráfico de variância cumulativa revelou que, com a inclusão de 4 componentes principais, conseguimos capturar aproximadamente 86% da variância total nos dados. Essa descoberta sugere que a projeção para um espaço de quatro dimensões é capaz de preservar a maioria das informações originais contidas nas variáveis, proporcionando uma representação eficiente e compacta do conjunto de dados.

A escolha de reter 4 componentes principais visa equilibrar a redução de dimensionalidade com a retenção de informação significativa, facilitando interpretações futuras e a aplicação de técnicas subsequentes de análise de clusterização. Prosseguimos então com a aplicação do PCA ao vetor de características mencionado para observar como as variáveis estão inter-relacionadas e como podem ser representadas de maneira mais eficiente em termos de dimensionalidade.

O processo de redução de dimensionalidade foi realizado através do módulo *decomposition* da biblioteca Sklearn. Após a aplicação do PCA, foram obtidos quatro componentes principais, os quais foram armazenados em um *dataframe* denominado *pca_df*, utilizando a biblioteca pandas. Esse DataFrame é composto por quatro colunas: 'PC1', 'PC2', 'PC3' e 'PC4', representando as quatro principais componentes extraídas do conjunto de dados original.

Por fim as componentes foram concatenadas à base original. Essa junção permite uma análise integrada dos dados originais e suas representações de menor dimensionalidade. As exibições gráficas serão representadas pelas 3 primeiras componentes principais que explicam 76% da variância cumulativa. O gráfico tridimensional foi gerado utilizando a biblioteca Matplotlib e importando a classe Axes3D do módulo mpl_toolkits.mplot3d.

Figura 13 – PCA: 3 Componentes Principais.



Fonte: Autor.

O Gráfico do PCA evidenciou algumas separações, indicando a presença de padrões que podem sugerir a existência de *clusters* no conjunto de dados. Essa observação é significativa, pois sugere a possibilidade de agrupamentos ou categorias entre as observações, mesmo que não sejam completamente isoladas.

Esses indícios de separação no Gráfico de PCA fornecem uma base promissora para a aplicação de técnicas subsequentes de clusterização, visando identificar e caracterizar grupos.

4.5 Clusterização

Seguindo para a etapa de clusterização, a seleção do algoritmo é condicionada pela natureza dos dados disponíveis, pelos objetivos e pelas limitações intrínsecas ao estudo. Em muitos casos, existe a possibilidade de optar entre diferentes algoritmos, tornando essencial a aplicação de múltiplos métodos e a análise comparativa dos resultados alcançados.

Além disso, a habilidade de interpretar os agrupamentos produzidos pelo algoritmo é um aspecto indispensável ao definir a abordagem de clusterização, considerando que a análise de agrupamentos é frequentemente empregada por pesquisadores em busca de *insights* e compreensão de padrões nos dados. Três métodos serão comparados: o *K-Means, K-Medoids* e *DBSCAN*.

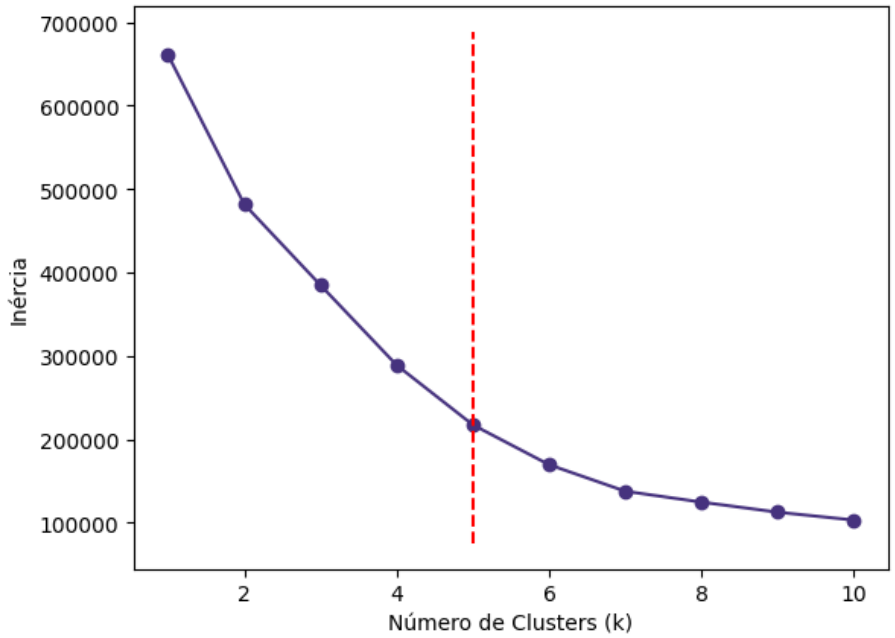
K-Means

O algoritmo *K-Means* foi selecionado devido à sua comprovada eficácia na abordagem de desafios associados à clusterização, tornando-se uma escolha pertinente para a tarefa de segmentação de clientes. Suas vantagens incluem não apenas a simplicidade, mas também a eficiência computacional, destacando-se de maneira significativa ao lidar com extensos conjuntos de dados, uma condição frequentemente encontrada em pesquisas de segmentação. Esta decisão visa contribuir para uma abordagem robusta e eficaz na análise e compreensão dos padrões de comportamento dos clientes por meio da segmentação. Para isso foi empregado o módulo *cluster* da biblioteca Sklearn.

Com o objetivo de evitar escolhas arbitrárias para o parâmetro K e proporcionar uma abordagem mais fundamentada, empregou-se o método do cotovelo para determinar o número ideal de *clusters* (K). A análise foi conduzida variando o número de *clusters* e avaliando a inércia associada a cada configuração, onde a inércia representa a soma das distâncias quadráticas entre os pontos e seus respectivos centros de *cluster*.

O ponto de inflexão identificado no gráfico da inércia em função de K, conhecido como ponto do cotovelo, indicou o valor ótimo de K. Este ponto representa o ponto onde o ganho na redução da inércia diminui significativamente. Assim, com base no método do cotovelo ilustrado e o auxílio da biblioteca *Kneelocator*, definiu-se K = 5, otimizando a eficácia do algoritmo *K-Means* para a nossa análise.

Figura 14 – Método do Cotovelo - K-means.



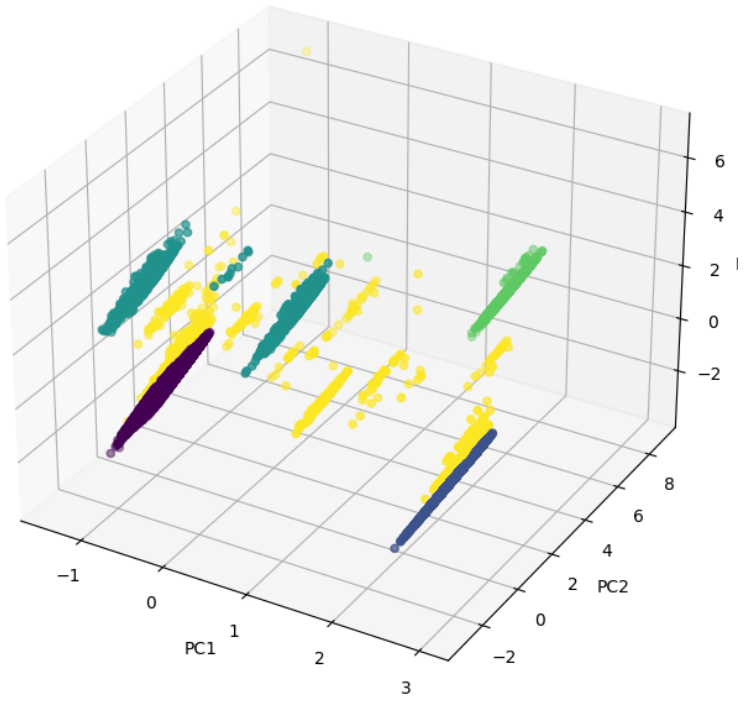
Fonte: Autor.

O algoritmo foi aplicado ao conjunto de dados de clientes, através da criação de uma instância do *KMeans* com 5 *clusters*, onde o modelo é então ajustado aos dados

normalizados pelo método *.fit()*, e os rótulos dos *clusters* são atribuídos aos dados.

Cada *cluster* representa um segmento de clientes com características semelhantes, facilitando a compreensão das preferências e comportamentos de grupos específicos como na figura 15.

Figura 15 – Clusters K-means.



Fonte: Autor.

K-Medoids

A opção pelo algoritmo *K-Medoids* foi fundamentada em sua capacidade de lidar robustamente com valores atípicos. A estratégia de utilizar observações reais como representantes dos *clusters* mostra-se particularmente benéfica em conjuntos de dados nos quais a presença de *outliers* é comum.

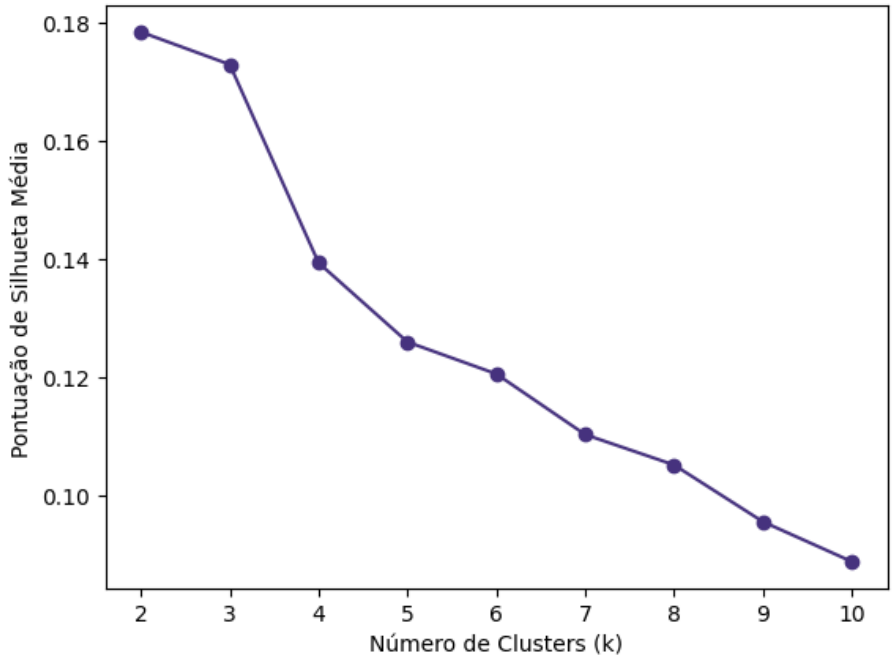
Essa escolha visa fortalecer a eficácia do *K-Medoids* na identificação de padrões mais representativos e resilientes, especialmente em cenários de dados propensos a variações significativas. A abordagem baseada em medóides demonstra-se uma solução robusta e adequada para enfrentar desafios associados à presença de valores extremos. O módulo empregado para o algoritmo foi o *cluster* da biblioteca Sklearn_extra após a instalação do pacote Scikit-learn-extra.

A estratégia adotada para escolha do valor ideal de K no contexto do algoritmo *K-Medoids* foi a análise da silhueta. Contudo, devido à maior exigência computacional associada ao algoritmo e à necessidade de treinamento de múltiplos modelos, optou-se por escolher uma amostra aleatória equivalente a 25% do conjunto dos dados originais

normalizados. Para implementar essa análise foi utilizado o módulo *metrics* do método *Silhouette_score*.

Para cada valor de K, foi treinado um modelo *K-Medoids* com os dados da amostra e previstos nos dados originais normalizados, calculando a pontuação média de silhueta para avaliar a separação dos clusters. As pontuações de silhueta resultantes foram armazenadas em uma lista e plotadas no gráfico da figura 16.

Figura 16 – Análise Silhueta - K-medoids.

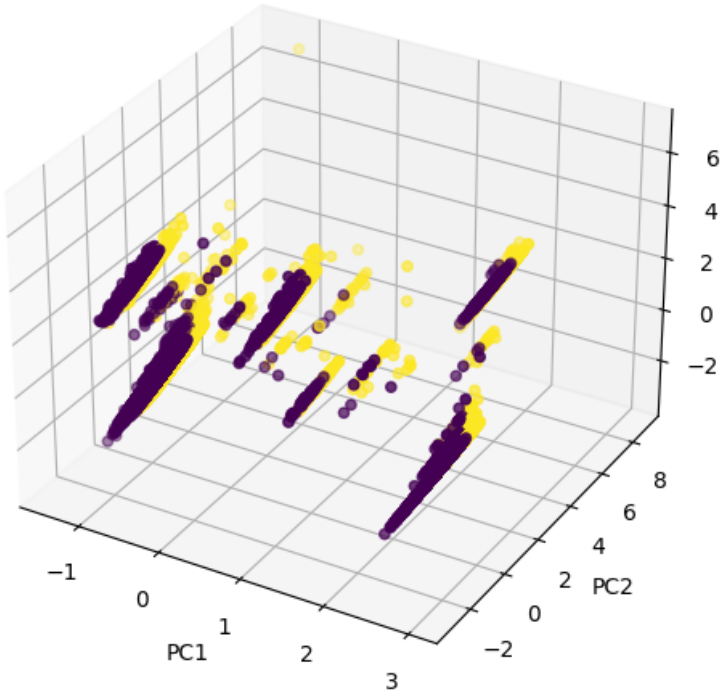


Fonte: Autor.

Após a observação gráfica e visto que as pontuções são uma medida da adequação dos agrupamentos para cada valor de K testado, a decisão tomada foi de configurar o número de *clusters* com o máximo desempenho da silhueta, ou seja, $K = 2$.

Frente à inerente demanda computacional associada ao algoritmo em análise, a decisão foi ajustá-lo à amostra previamente selecionada e utilizar o modelo treinado para predizer os rótulos dos dados.

Figura 17 – Clusters K-medoids.



Fonte: Autor.

DBSCAN

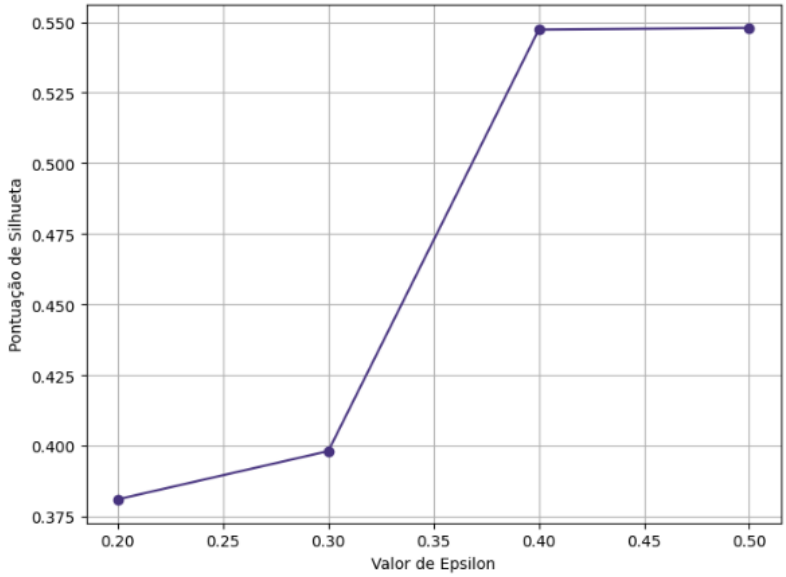
O principal objetivo desta análise de clusterização utilizando o algoritmo *DBSCAN* é identificar estruturas intrínsecas nos dados sem a necessidade de pré-determinar um número fixo de *clusters*. Diferentemente de métodos como o *K-means* e *K-Medoids* vistos anteriormente, o *DBSCAN* é capaz de identificar automaticamente o número de *clusters* com base na densidade dos dados. A escolha do algoritmo visa proporcionar maior flexibilidade na identificação de grupos de diferentes formas e tamanhos, adaptando-se naturalmente à distribuição dos dados sem impor limitações predefinidas.

Ao não fixar um valor específico para o número de *clusters*, a expectativa é capturar de maneira mais precisa a estrutura subjacente dos dados, especialmente em situações em que a distribuição dos *clusters* pode não ser uniforme ou onde a definição de K é desafiadora.

Este enfoque também oferece vantagens na detecção de *outliers* e na adaptação a *clusters* de densidades variáveis, proporcionando uma análise mais robusta e adaptativa à natureza dos dados em questão. O módulo utilizado é o *cluster* da biblioteca Sklearn.

Devido às restrições computacionais, foi necessário limitar os valores de epsilon aplicados ao algoritmo *DBSCAN* e aplicar o modelo ao PCA previamente selecionado com 4 componentes. Visando estabelecer o número ideal de *epsilon* foram testados diferentes valores em uma grade pré-definida [0.2, 0.3, 0.4, 0.5] avaliando o desempenho do modelo para cada combinação.

Figura 18 – Relação Epsilon e Pontuação Silhueta.



Fonte: Autor.

Consequentemente, os testes foram realizados com valores até 0.5, uma vez que valores acima desse limiar não puderam ser computacionalmente explorados. Dentre os valores testados, os melhores desempenhos foram obtidos por *epsilon* igual a 0.4 e 0.5.

Assim, considerando os resultados obtidos, procedeu-se com a avaliação de diferentes valores para o parâmetro *MinPts* em cada caso como descrito na tabela 7. A seleção final dos parâmetros para a clusterização foi baseada no critério do *Silhouette Score* mais elevado, indicando uma melhor estruturação dos *clusters*. Nesse contexto, os valores ótimos escolhidos foram *epsilon* igual a 0.5 e MinPts igual a 15. Essa configuração proporcionou a melhor separação dos dados em *clusters* distintos, conforme avaliado pela métrica do *Silhouette Score*.

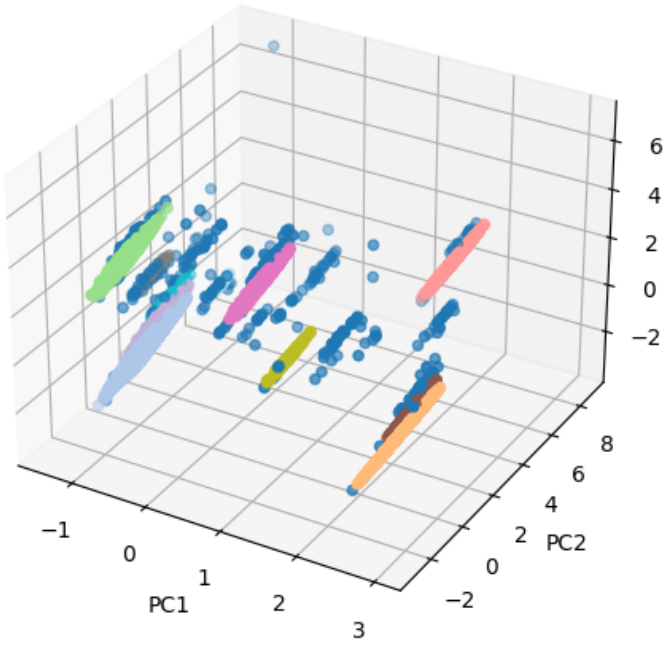
Tabela 7 – Resultados do DBSCAN para diferentes valores de Epsilon e MinPts.

Epsilon	**MinPts**	**Silhouette Score**
0.4	10	0.5489
0.4	15	0.5482
0.4	20	0.5474
0.4	25	0.5470
0.4	30	0.5462
0.5	10	0.5485
0.5	15	0.5491
0.5	20	0.5488
0.5	25	0.5483
0.5	30	0.5481

Portanto, a escolha da clusterização foi feita optando pelo conjunto de parâmetros

que resultou no *Silhouette Score* mais alto, ou seja, $epsilon = 0.5$ e $MinPts = 15$, obtendo a segmentação apresentada na figura 19.

Figura 19 – Clusters DBSCAN.



Fonte: Autor.

4.6 Métrica de Avaliação

A eficaz avaliação dos resultados obtidos por algoritmos de agrupamento desempenha um papel essencial em projetos de mineração de dados, especialmente ao investigar os padrões de compra dos consumidores em ambientes de *e-commerce*. A identificação de *clusters* relevantes e sua interpretação de são fundamentais para extrair *insights* valiosos que possam orientar estratégias de negócios.

Neste contexto, utilizaremos a métrica de silhueta como uma ferramenta robusta para avaliar a qualidade dos *clusters* gerados pelos modelos. A métrica de silhueta oferece uma abordagem sistemática e quantitativa para medir a coesão interna dos *clusters*, fornecendo uma perspectiva clara sobre a eficácia dos algoritmos de agrupamento. Se o valor for alto, o objeto está bem associado ao seu *cluster* e mal associado aos *clusters* vizinhos.

A análise dos resultados dos métodos de clusterização forneceu *insights* sobre a estrutura dos dados. Abaixo estão os índices de silhueta obtidos para cada método:

- **K-Means:** O índice de silhueta de 0.56 indica que o algoritmo foi bem-sucedido na formação de agrupamentos distintos e definidos. A magnitude superior a 0.5 adiciona robustez à interpretação, denotando uma diferenciação razoável entre os *clusters*.

- **K-Medoids:** O coeficiente de silhueta mais baixo, aproximadamente 0.18, sugere uma coesão interna inferior em comparação com os outros métodos, ou seja uma separação ruim. Isso indica possivelmente uma sobreposição entre *clusters* e uma separação menos distintiva.
- **DBSCAN:** O coeficiente de silhueta de 0.55 sugere coesão interna considerável e uma distinção razoável entre os *clusters* formados. Essa pontuação destaca a eficácia do *DBSCAN* na geração de *clusters* definidos e separados, sem a necessidade da escolha de K.

Esses resultados evidenciam as características distintivas de cada método. O *K-Means* demonstra uma diferenciação razoável entre *clusters*, enquanto o *K-Medoids* apresenta uma coesão interna inferior e o *DBSCAN* destaca-se pela distinção entre *clusters* sem a necessidade de definir K.

Em síntese, a interpretação geral aponta que quanto maior o valor de silhueta, melhor é a qualidade dos *clusters*. Porém, a escolha entre os métodos deve considerar não apenas essas pontuações, mas também a natureza dos dados, as limitações computacionais envolvidas e os objetivos específicos da análise, adaptando-se ao contexto particular do estudo. Esses comparativos fornecem *insights* valiosos para a seleção do método mais adequado, considerando as nuances específicas do comportamento de compra dos clientes no ambiente de *e-commerce*.

5 RESULTADOS

Com base nos agrupamentos produzidos por cada um dos modelos aplicados, procedeu-se a uma análise das características de cada grupo com o objetivo de identificar padrões distintos de comportamento dos clientes. Os *clusters* foram gerados através das técnicas vistas anteriormentes.

A seguir, apresentamos as interpretações e *insights* derivados da análise detalhada de cada *cluster*, proporcionando uma compreensão aprofundada dos comportamentos específicos encontrados em cada grupo.

Os valores médios das características do *dataframe* original, conforme apresentados na tabela, constituem uma base fundamental para as análises em curso.

Tabela 8 – Tabela com Médias das Features.

V	R	F	Boleto	Crédito	Débito	Voucher
160.4	242.8	1.0	0.2	0.8	0.02	0.04

5.1 K-Means

Tabela 9 – Tabela com Médias das Features - Kmeans.

Label	Qtd	V	R	F	Boleto	Crédito	Débito	Voucher
0	68446	167.3	241.3	1.0	0.0	1.0	0.0	0.0
1	18258	143.5	252.8	1.0	1.0	0.0	0.0	0.0
2	3476	131.6	263.0	1.0	0.0	0.6	0.0	1.0
3	1445	140.5	172.1	1.0	0.0	0.0	1.0	0.0
4	2772	147.0	224.4	2.1	0.2	0.8	0.0	0.0

Os resultados da análise sugerem que as variáveis de valor e recência podem não contribuir significativamente para a segmentação dos clientes. A falta de distinção clara entre os *clusters* em relação a essas variáveis indica que elas podem não agregar valor à análise de segmentação de clientes.

Quanto à variável de frequência, observamos que os *clusters* 0 e 1 consistem exclusivamente de clientes que realizaram uma compra, enquanto os *clusters* 2 e 3 variam até duas compras e o *cluster* 4 apresenta uma faixa mais ampla, incluindo clientes de 2 até 16 compras, indicando uma frequência de compra mais alta para este grupo específico.

A análise dos valores mínimos e máximos das variáveis de uso de pagamentos revela uma diferenciação clara entre os *clusters*, sugerindo que essas variáveis são capazes de segmentar os clientes de forma significativa. Os *clusters* apresentam padrões distintos de

Tabela 10 – Valores mínimos e máximos da Frequência por Cluster - Kmeans.

Label	Mínimo	Máximo
0	1	1
1	1	1
2	1	2
3	1	2
4	2	16

comportamento de uso de diferentes métodos de pagamento, indicando que essas variáveis desempenham um papel crucial na segmentação dos clientes.

Tabela 11 – Valores mínimos e máximos das colunas de pagamento por Cluster - Kmeans.

Label	Boleto		Crédito		Débito		Voucher	
	min	max	min	max	min	max	min	max
0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
1	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
3	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0
4	0.0	1.0	0.0	1.0	0.0	0.5	0.0	1.0

Assim podemos inferir que:

- **Cluster 0: Baixa Frequência, Pagamento via Cartão de Crédito:**

  Engloba clientes que realizam compras com baixa frequência, de apenas uma compra com a preferência pelo uso do cartão de crédito.

  Esses clientes representam uma oportunidade significativa para as empresas, exigindo estratégias personalizadas que incentivem o uso contínuo do cartão de crédito e promovam o engajamento.

- **Cluster 1: Baixa Frequência e Pagamento Exclusivo via Boleto Bancário:**

  Caracterizado por clientes que realizam compras com baixa frequência, de apenas uma compra com a preferência pelo pagamento exclusivo via boleto bancário. A média de tempo desde a última compra é moderada, sugerindo alguma taxa de atividade.

  A escolha exclusiva pelo pagamento via boleto bancário sugere uma preferência por métodos de pagamento mais diretos ou pode indicar uma falta de acesso a outros métodos de pagamento, como cartões de crédito. Isso pode ser influenciado por fatores socioeconômicos ou preferências pessoais.

- **Cluster 2: Baixa Frequência, Pagamento via Voucher e Cartão de Crédito:**

Caracterizado por clientes que realizam compras com frequência baixa de até 2 compras, com a preferência pelo pagamento via *voucher*, de forma exclusiva ou conjunta ao cartão de crédito.

Estratégias de retenção podem envolver promoções ou benefícios específicos relacionados ao cartão de crédito ou voucher, considerando a preferência destacada.

- **Cluster 3: Baixa Frequência, Pagamento via Cartão de Débito:**

  Clientes que realizam compras com frequência baixa de até 2 compras, com a preferência pelo pagamento pelo cartão de débito.

- **Cluster 4: Alta Frequência, Uso Variado de Pagamentos:**

  Clientes notáveis pela alta frequência de compras, ou seja, 2 ou mais compras. Não há uma preferência por um método específico de pagamento.

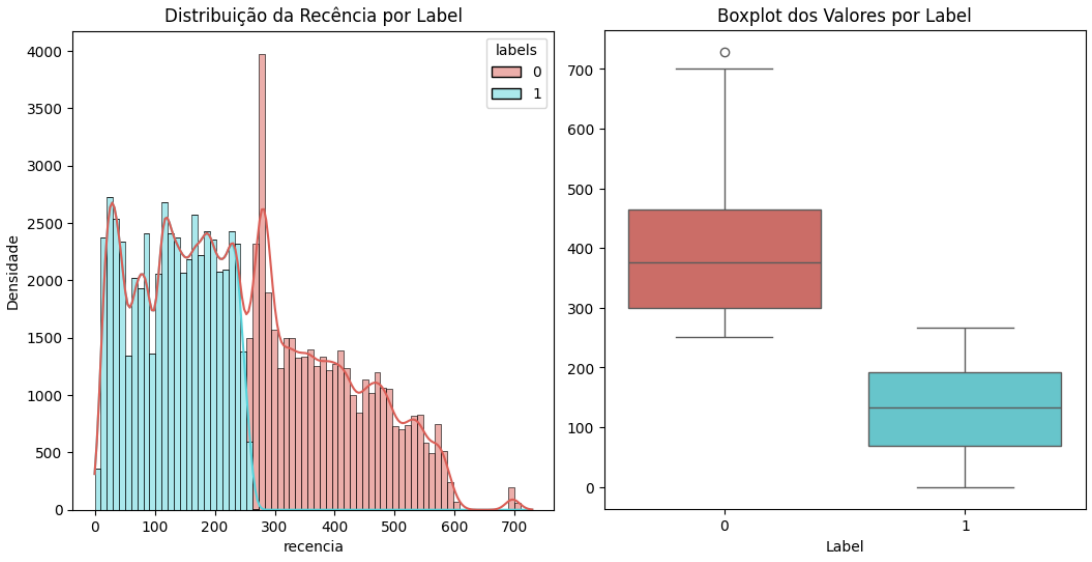
5.2 K-Medoids

Tabela 12 – Tabela de Média das Features - Kmedoids.

Label	Qtd	V	R	F	Boleto	Crédito	Débito	Voucher
0	40737	159.1	389.8	1.0	0.2	0.8	0.0	0.0
1	53660	161.3	131.1	1.0	0.2	0.8	0.0	0.0

A análise das médias na tabela e a visualização dos gráficos relacionados à variável recência sugerem que o *cluster* 0 apresenta uma média de tempo desde a última compra significativamente superior, indicando que os clientes desse *cluster* realizaram compras há mais tempo. Por outro lado, o *cluster* 1 engloba clientes com uma média de tempo desde a última compra menor, sugerindo um maior engajamento desses clientes em atividades de compra mais recentes.

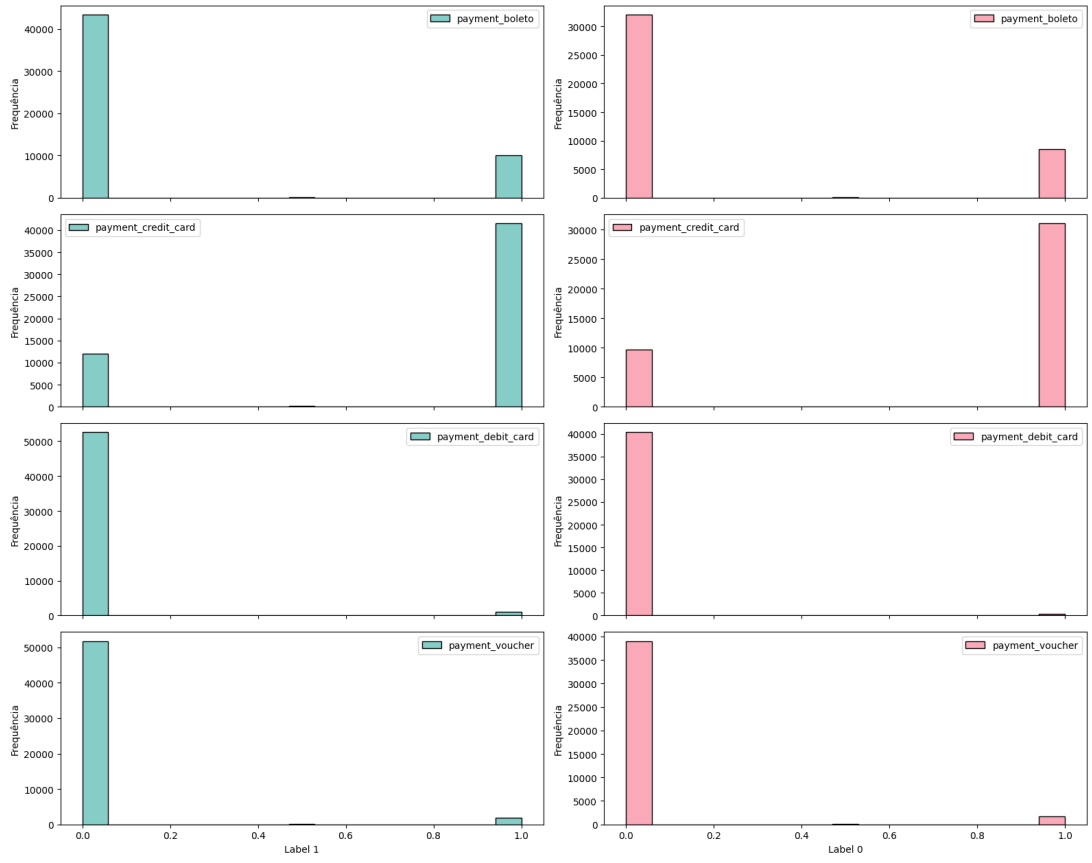
Figura 20 – Distribuição e Boxplot da Recência pelo método K-medoids.



Fonte: Autor.

A análise dos padrões de pagamento entre os *clusters* não sugere a utilização dessas variáveis na segmentação pelo método *K-Medoids*, uma vez que não parece haver um padrão distintivo que diferencie claramente os *clusters*.

Figura 21 – Distribuição dos tipos Pagamentos pelo método K-medoids.



Fonte: Autor.

Tabela 13 – Mínimo e Máximo das Colunas de Pagamento por Label - Kmedoids.

Label	Boleto		Credito		Debito		Voucher	
	min	max	min	max	min	max	min	max
0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
1	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0

As demais variáveis também não demonstraram significância nas separações observadas. Portanto, os grupos obtidos pela clusterização foram:

- **Cluster 1: Baixa Recência**

  Caracterizado por clientes que possuem uma alta média de tempo decorrido desde a última transação, o que sugere uma possível inatividade por parte desses clientes.

  Nesse cenário a necessidade de reativação desses clientes, por meio de estratégias personalizadas de *marketing*, visaria estimular a retomada de atividade e a recorrência de compras.

- **Cluester 0: Alta Recência**

  Abrange clientes que a média de tempo decorrido desde a última compra é baixa, indicando uma atividade contínua por parte desses clientes.

  Nesse cenário a estratégia de *marketing* a ser aplicada pode incluir manutenção desse alto nível de engajamento por meio de estratégias de fidelização e retenção de clientes, para garantia da satusafação com a experiência de compras. Além disso, é necessário monitorar de perto o comportamento desses clientes para identificar possíveis sinais de redução na atividade de compra e agir proativamente para evitar a perda desse segmento valioso de clientes.

Parece que, entre as variáveis disponíveis, o método de segmentação está considerando apenas uma delas de forma significativa para separar os *clusters*, o que resulta em uma simplificação excessiva do problema. Tal abordagem pode não capturar adequadamente a complexidade e os padrões presentes nos dados, levando a uma segmentação menos precisa e uma compreensão limitada da estrutura subjacente.

5.3 DBSCAN

Ao analisar as variáveis de recência e valor, não se observa uma contribuição significativa para a diferenciação dos *clusters* . A ausência de uma diferenciação clara entre os *clusters* com base nessas variáveis sugere que elas não exercem um papel significativo na segmentação dos clientes.

Tabela 14 – Tabela com Médias das Features - DBSCAN.

Label	Qtd	V	R	F	Boleto	Crédito	Débito	Voucher
-1	343	358.5	228.5	2.6	0.3	0.4	0.1	0.4
0	68445	167.1	241.3	1.0	0.0	1.0	0.0	0.0
1	18254	142.1	252.8	1.0	1.0	0.0	0.0	0.0
2	2026	145.5	265.3	1.0	0.0	1.0	0.0	1.0
3	1427	135.0	172.1	1.0	0.0	0.0	1.0	0.0
4	1824	141.9	228.8	2.0	0.0	1.0	0.0	0.0
5	372	152.0	241.2	2.0	1.0	0.0	0.0	0.0
6	1366	101.0	259.1	1.0	0.0	0.0	0.0	1.0
7	54	133.0	160.1	2.0	0.0	1.0	0.0	0.5
8	162	122.4	193.8	2.0	0.5	0.5	0.0	0.0
9	109	145.7	200.8	3.0	0.0	1.0	0.0	0.0
10	15	111.3	99.0	2.0	0.0	0.5	0.0	0.5

Tabela 15 – Valores mínimos e máximos de frequência por Cluster - DBSCAN.

Label	Mín	Máx
-1	1	16
0	1	1
1	1	1
2	1	1
3	1	1
4	2	2
5	2	2
6	1	1
7	2	2
8	2	2
9	3	3
10	2	2

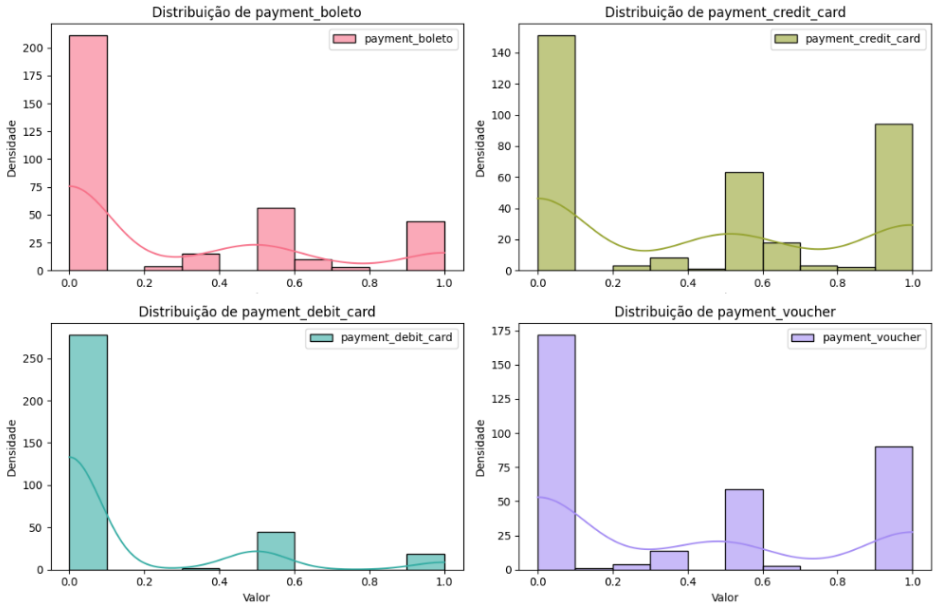
Todos os clientes com o número de compras acima de 3 foram considerados ruídos. O restante dos grupos ficaram divididos em *clusters*: 0,1,2,3 e 6 em clientes com 1 compra, *clusters*: 4,5,7,8 e 10 em clientes com 2 compras e *cluster* 9 em clientes com 3 compras.

Quanto as variáveis dos métodos de pagamento, podemos observar a partir da tabela 16 de mínimos e máximos as seguintes caracterísitcas: clientes que utilizaram exclusivamente algum tipo de pagamento (*clusters*: 0,1,3,4,5,6,9) ou utilizaram 2 métodos de pagamentos, sendo utilizados um método distinto para cada compra (*clusters*: 8 e 10) ou utilizados o método em conjunto em uma das compras (*clusters*: 2,7). Os métodos que não tiveram um padrão claro sobre o uso dos pagamentos entre as compras foram classificadas como ruídos.

Tabela 16 – Mínimo e Máximo para Cada Método de Pagamento por Cluster - DBSCAN.

Label	Boleto		Crédito		Débito		Voucher	
	min	max	min	max	min	max	min	max
-1	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
1	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0
3	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
4	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
7	0.0	0.0	1.0	1.0	0.0	0.0	0.5	0.5
8	0.5	0.5	0.5	0.5	0.0	0.0	0.0	0.0
9	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.5	0.5	0.0	0.0	0.5	0.5

Figura 22 – Distribuição dos Métodos de Pagamentos dos Ruídos - DBSCAN.



Fonte: Autor.

Conseguimos inferir os seguintes grupos:

- **Cluster 0: Clientes com 1 compra com Pagamento Exclusivo com Cartão de Crédito:**

  Demonstram baixa frequência de compras e têm uma preferência marcante pelo pagamento com cartão de crédito.

  Embora os clientes neste *cluster* demonstrem uma preferência clara pelo pagamento com cartão de crédito, a baixa frequência de compras pode indicar um menor

engajamento ou interesse em realizar transações adicionais. Portanto, estratégias de reengajamento e fidelização podem ser necessárias para incentivar compras futuras e aumentar a frequência de transações. Ofertas que oferecem benefícios exclusivos aos clientes que optam pelo pagamento com cartão de crédito podem ser atraentes para esse grupo como, por exemplo, cashback ou pontos de fidelidade.

- **Cluster 1: Clientes Com 1 compra com Pagamento Exclusivo de Boleto:**

  Representa clientes que realizaram apenas uma compra e optaram por pagar exclusivamente com boleto.

  A baixa frequência de compras pode indicar menor engajamento ou interesse do cliente e a preferência exclusiva pelo boleto pode indicar uma preferência por métodos de pagamento tradicionais, descontos associados a essa forma de pagmento ou a falta de acesso a outras opções de pagamento.

- **Cluster 2: Clientes Com 1 compra com Pagamento com Cartão de Crédito em conjunto com Voucher:**

  Realizaram uma única compra, utilizando tanto cartão de crédito quanto *voucher* como métodos de pagamento.

  A combinação de pagamento com cartão de crédito e *voucher* pode indicar uma disposição para aproveitar ofertas ou descontos especiais. Este *cluster* pode representar clientes sensíveis a preços ou que buscam vantagens adicionais ao fazer compras.

- **Cluster 3: Clientes com 1 compra com Pagamento Exclusivo com Cartão de Débito:**

  Este grupo de clientes fez uma única compra e escolheu pagar exclusivamente com cartão de débito.

  O uso exclusivo do cartão de débito pode indicar uma preferência por métodos de pagamento diretos e imediatos, enquanto a baixa frequência de compras sugere uma menor atividade de compra.

- **Cluster 4: Clientes com 2 compras com Pagamento Exclusivo com Cartão de Crédito:**

  Realizaram duas compras, utilizando exclusivamente cartão de crédito como método de pagamento.

  A preferência exclusiva pelo cartão de crédito em duas compras pode indicar confiança financeira ou preferência por conveniência ao fazer compras online.

- **Cluster 5: Clientes com 2 compras com Pagamento Exclusivo de Boleto:**

  Engloba clientes que fizeram duas compras, optando exclusivamente pelo pagamento com boleto.

O uso exclusivo do boleto em duas compras pode indicar uma preferência por métodos de pagamento mais tradicionais, possíveis descontos associados ao método de pagamento ou a falta de acesso a outras opções de pagamento.

- **Cluster 6: Clientes com 1 compra com Pagamento Exclusivo de Voucher:**

  Clientes com apenas uma compra que optaram por pagar exclusivamente com *voucher*.

  O pagamento exclusivo com *voucher* pode indicar uma disposição para utilizar benefícios específicos ou descontos oferecidos pela empresa.

- **Cluster 7: Clientes com 2 compras com Pagamento de Cartão de Crédito em conjunto com Voucher em 1 compra**

  Este grupo de clientes fez duas compras, utilizando cartão de crédito e *voucher* como métodos de pagamento, respectivamente, em uma das compras.

  A combinação de pagamento com cartão de crédito e *voucher* em uma das compras pode indicar uma busca por benefícios adicionais ou descontos especiais em determinadas ocasiões.

- **Cluster 8: Clientes com 2 compras com Pagamento em Boleto ou Cartão de Crédito:**

  Realizaram duas compras, utilizando tanto boleto quanto cartão de crédito como métodos de pagamento.

  Este *cluster* indica uma preferência por métodos de pagamento flexíveis, com clientes alternando entre boleto e cartão de crédito em suas compras.

- **Cluster 9: Clientes com 3 compras com Pagamento em Cartão de Crédito:**

  Representa clientes que fizeram três compras, utilizando exclusivamente cartão de crédito como método de pagamento.

  A frequência de três compras pode indicar uma maior atividade de compra e potencial engajamento do cliente com a marca. O uso exclusivo do cartão de crédito pode sugerir preferências pelos benefícios como flexibilidade financeira, parcelamentos, acúmulo de pontos e milhas.

- **Cluster 10: Clientes Com 2 compras com Pagamento em Voucher ou Cartão de Crédito:**

  Clientes neste *cluster* realizaram duas compras, utilizando *voucher* e cartão de crédito como métodos de pagamento.

  A alternância entre pagamentos com *voucher* e cartão de crédito indica uma possível sensibilidade a descontos ou benefícios especiais, combinada com a conveniência e flexibilidade do cartão de crédito.

Essas análises evidenciam que cada modelo realizará a segmentação considerando características distintas, e a seleção do modelo não deve ser baseada apenas na métrica da silhueta, mas também na finalidade da clusterização. Elas proporcionam uma compreensão detalhada dos comportamentos variados dos clientes em cada cluster, oferecendo *insights* valiosos para estratégias de *marketing*.

O método *K-Means* destacou mais os métodos de pagamentos e frequência enquanto o *K-medoids* baseou-se nas métricas de recência, e o *DBSCAN* nos métodos de pagamentos e frequência.

6 CONCLUSÃO

6.1 Contribuições

Este estudo, centrado na análise do comportamento de compra dos consumidores no cenário do *e-commerce* brasileiro, teve como principal objetivo a criação de segmentos de clientes. Essa categorização foi fundamentada em variáveis como valor das compras, recência, frequência e modalidades de pagamento (boleto, cartão de crédito, débito, *voucher*). O foco central foi aprimorar a eficácia das estratégias de *marketing*, fazendo uso de uma das técnicas de mineração de dados mais reconhecidas pela sua utilidade e eficiência nesse contexto: a clusterização, através dos algoritmos *K-means*, *K-medoids* e *DBSCAN*.

Para colocar esses conceitos em prática, realizamos uma análise utilizando uma base de dados real de um *e-commerce* brasileiro. Após a limpeza, processamento e transformação adequados dos dados, calculamos os atributos, incluindo a porcentagem de utilização dos meios de pagamento pelos clientes. Esses atributos serviram como base para a segmentação dos clientes, resultando na identificação de *insights* relevantes a partir dos dados coletados.

Para comparação, aplicamos os algoritmos *K-means*, *K-medoids* e *DBSCAN* na base de dados com os atributos RFV e métodos de pagamento. Nosso objetivo era obter novos *clusters* para cada cliente, permitindo uma segmentação mais refinada. Após análise, observamos que o algoritmo *K-means* obteve o melhor resultado de *Silhouette Score*, conforme detalhado na seção correspondente. No entanto, notamos que a estrutura de agrupamento produzida pelo *DBSCAN* parecia estar mais alinhada com a natureza das variáveis em questão, sugerindo uma segmentação potencialmente mais significativa com base na intenção desses atributos.

Essa abordagem possibilitará o desenvolvimento de novas estratégias de *marketing* direcionadas, adaptadas às características específicas de cada segmento de clientes, o que potencializará as ações de *marketing*, assim como a eficácia das campanhas.

Os resultados deste estudo fornecem uma base para a aplicação de técnicas similares em diferentes áreas e setores, especialmente no que se refere à segmentação de clientes. Essas descobertas têm o potencial de guiar profissionais que dependem des análises similares, oferecendo *insights* para aprimorar estratégias de *marketing* e auxiliar na tomada de decisões. Ao compartilhar essas experiências e resultados, esperamos contribuir para o avanço no campo da segmentação de clientes e sua aplicação prática.

6.2 Trabalhos Futuros

Neste estudo, enfrentamos algumas limitações que merecem ser destacadas. Primeiramente, a questão da inclusão de variáveis, como sexo, idade, nível socioeconômico, escolaridade e renda, poderia enriquecer significativamente nossa análise. A incorporação dessas variáveis adicionais poderia contribuir para a formação de grupos de clientes mais específicos e refinados, ampliando assim nossa compreensão do comportamento do consumidor no contexto do *e-commerce* brasileiro.

Além disso, deparamo-nos com limitações relacionadas à capacidade computacional, o que nos impediu de explorar valores maiores de *epsilon* para o algoritmo *DBSCAN* e de utilizar uma amostra mais significativa no algoritmo *K-medoids*. Essas restrições afetaram diretamente a qualidade dos *clusters* gerados, podendo ter levado a resultados menos consistentes. A ampliação da capacidade computacional poderia ter permitido uma exploração mais abrangente dos parâmetros dos algoritmos, resultando em *clusters* mais precisos e robustos. Isso, por sua vez, teria potencializado a criação de campanhas de *marketing* mais especializadas, adaptadas às necessidades e preferências específicas de cada segmento de clientes.

Embora este estudo tenha avançado na compreensão e análise dos dados do *e-commerce*, há oportunidades para investigações futuras que podem expandir e aprimorar ainda mais as descobertas e os resultados apresentados. Algumas sugestões para pesquisas futuras incluem a implementação de modelos de classificação para segmentar novas visitas do *e-commerce* nos respectivos segmentos. Isso poderia envolver o uso de algoritmos de aprendizado supervisionado, como árvores de decisão, regressão logística ou máquinas de vetores de suporte (SVM), para prever a categoria ou segmento a que uma visita pertence com base em características relevantes.

REFERÊNCIAS

ALENCAR, A. J. *et al.* Optimized rfv analysis. **Marketing Intelligence Planning**, v. 24, n. 2, p. 106–118, 2006.

ANITHA, P.; PATIL, M. M. Rfm model for customer purchase behavior using k-means algorithm. **Journal of King Saud University-Computer and Information Sciences**, Elsevier, v. 34, n. 5, p. 1785–1792, 2022.

ARORA, P.; DEEPALI; VARSHNEY, S. Analysis of k-means and k-medoids algorithm for big data. **Procedia Computer Science**, v. 78, p. 507–512, 2016. ISSN 1877-0509. 1st International Conference on Information Security Privacy 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050916000971.

BURNEY, S. A.; TARIQ, H. K-means cluster analysis for image segmentation. **International Journal of Computer Applications**, Foundation of Computer Science, v. 96, n. 4, 2014.

CASTRO, L. N. D.; FERRARI, D. G. **Introdução à mineração de dados**. [*S.l.: s.n.*]: Saraiva Educação SA, 2017.

CHIA-JUNG, L. *et al.* Machine learning-based e-commerce platform repurchase customer prediction model. **PLoS ONE**, Public Library of Science, San Francisco, CA, v. 15, n. 12, p. e0243105, 2020. Disponível em: https://doi.org/10.1371/journal.pone.0243105.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados-funcionalidades, técnicas e abordagens**. [*S.l.: s.n.*]: PUC, 2002.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. [*S.l.: s.n.*]: Gulf Professional Publishing, 2005.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2, p. 107–145, 2001. Disponível em: https://doi.org/10.1023/A:1012801612483.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. San Francisco: Morgan Kaufmann, 2012.

HARRISON, T. H. **Intranet data warehouse: ferramentas e técnicas para a utilização do data warehouse na intranet**. [*S.l.: s.n.*]: Berkerley/ABDR, 1998.

HAWKINS, D. I.; MOTHERSBAUGH, D. L. **Comportamento do consumidor: construindo a estratégia de marketing**. [*S.l.: s.n.*]: Elsevier Brasil, 2018.

HOSSAIN, A. S. Customer segmentation using centroid based and density based clustering algorithms. *In*: **2017 3rd International Conference on Electrical Information and Communication Technology (EICT)**. [*S.l.: s.n.*], 2017. p. 1–6.

JIANG, Y.; YU, S. Mining e-commerce data to analyze the target customer behavior. *In*: **First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)**. Adelaide, SA, Australia: [*S.l.: s.n.*], 2008. p. 406–409.

KANSAL TUSHAR E BAHUGUNA, S. e. S. V. e. C. T. Segmentação de clientes usando clustering k-means. *In*: **2018 Conferência Internacional sobre Técnicas Computacionais, Eletrônica e Sistemas Mecânicos (CTEMS)**. [*S.l.: s.n.*], 2018. p. 135–139.

KHAN, K. *et al.* Dbscan: Past, present and future. *In*: **The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)**. [*S.l.: s.n.*], 2014. p. 232–238.

KOTLER, P.; KELLER, K. L. **Administração de Marketing**. 15. ed. São Paulo: Pearson Education do Brasil, 2018. Título original: *Marketing management*. ISBN 978-65-5011-047-5.

KOUL, S.; PHILIP, T. M. Customer segmentation techniques on e-commerce. *In*: **2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)**. [*S.l.: s.n.*], 2021. p. 135–138.

LINOFF, G. S.; BERRY, M. J. **Data mining techniques: for marketing, sales, and customer relationship management**. [*S.l.: s.n.*]: John Wiley & Sons, 2011.

MIGLAUTSCH, J. Application of rfm principles: What to do with 1–1–1 customers? **Journal of Database Marketing & Customer Strategy Management**, v. 9, p. 319–324, 2002.

NAMVAR, M.; GHOLAMIAN, M. R.; KHAKABI, S. Um método de clustering em duas fases para segmentação inteligente de clientes. *In*: **2010 Conferência Internacional sobre Sistemas Inteligentes, Modelagem e Simulação**. [*S.l.: s.n.*], 2010. p. 215–219.

NETO, R. F. D. O.; RAMOS, R. A.; SILVA, C. D. D. Uma solução de mineração de dados para concessão de cupons de descontos em comércio eletrônico: um estudo de caso. v. 11, n. 3, p. 122–132, 2019. Disponível em: https://seer.upf.br/index.php/rbca/article/view/9077.

Olist; Sionek, André. **Brazilian E-Commerce Public Dataset by Olist**. 2018. Disponível em https://www.kaggle.com/dsv/195341.

PINHO, A. G. de. Análise rfv do cliente por algoritmos genéticos na otimização de estratégias de marketing. **Revista Pensamento Contemporâneo em Administração**, Universidade Federal Fluminense, v. 3, n. 2, p. 86–98, 2009.

RAHARDJA UNTUNG E HARIGUNA, T. e. B. W. M. Mineração de opinião sobre dados de comércio eletrônico usando análise de sentimento e cluster k-medoid. *In*: **2019 Décima Segunda Conferência Internacional sobre Computação Ubi-Media (Ubi-Media)**. [*S.l.: s.n.*], 2019. p. 168–170.

SCHIFFMAN, L. G.; KANUK, L. L. **Comportamento do Consumidor**. 6. ed. Rio de Janeiro: Editora LTC, 2000. 476 p. ISBN 8521612206.

SHAHAPURE, K. R.; NICHOLAS, C. Cluster quality analysis using silhouette score. *In*: **2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)**. [*S.l.: s.n.*], 2020. p. 747–748.

SONG, Q.; SHEPPERD, M. Mining web browsing patterns for e-commerce. **Computers in industry**, Elsevier, v. 57, n. 7, p. 622–630, 2006.

STAHL, D.; SALLIS, H. Model-based cluster analysis. **WIREs Computational Statistics**, v. 4, p. 341–358, 2012.

VINODHINI, G.; CHANDRASEKARAN, R. M. Opinion mining using principal component analysis based ensemble model for e-commerce application. **CSI Transactions on ICT**, v. 2, n. 3, p. 169–179, November 2014. ISSN 2277-9086. Disponível em: https://doi.org/10.1007/s40012-014-0055-3.

WANG, X.; HUANG, J. Enterprise decision-making and analysis based on e-commerce data mining. **Wireless Communications and Mobile Computing**, Hindawi, v. 2022, p. Article ID 9493775, 11 pages, 2022.