

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Thiago Fernandes Teodoro Pavani

**Aplicação de Técnicas de Mineração de Dados
para Identificação de Padrões de Vendas em
*Marketplaces***

Uberlândia, Brasil

2026

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Thiago Fernandes Teodoro Pavani

**Aplicação de Técnicas de Mineração de Dados para
Identificação de Padrões de Vendas em *Marketplaces***

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Orientador: Elaine Ribeiro de Faria Paiva

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2026

Thiago Fernandes Teodoro Pavani

Aplicação de Técnicas de Mineração de Dados para Identificação de Padrões de Vendas em *Marketplaces*

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 10 de março de 2026:

**Prof^a. Dr^a. Elaine Ribeiro de Faria
Paiva**
Orientador

Professor

Professor

Uberlândia, Brasil
2026

Resumo

O crescimento dos *marketplaces* digitais tem intensificado a competitividade no comércio eletrônico, tornando cada vez mais relevante o uso de técnicas de Ciência de Dados para apoiar a tomada de decisão dos vendedores. Nesse contexto, este trabalho tem como objetivo extrair conhecimento a partir de dados de *marketplaces*, buscando identificar padrões de vendas e perfis de pedidos por meio da análise de dados. Para isso, foi utilizada uma base de dados pública de *e-commerce* brasileiro, disponibilizada na plataforma *Kaggle*, contendo informações reais sobre pedidos, clientes, pagamentos, produtos, avaliações e aspectos logísticos.

A metodologia adotada envolveu a integração das diferentes tabelas da base de dados, seguida por etapas de pré-processamento, análise exploratória e análise de correlação entre os principais atributos. Como etapa central do estudo, foi aplicada uma técnica de Aprendizado de Máquina não-supervisionado, utilizando o algoritmo de agrupamento *K-Means*, após a normalização dos dados e a definição do número adequado de *clusters* por meio do método do cotovelo.

Os resultados obtidos permitiram identificar diferentes perfis de pedidos no *marketplaces*, evidenciando comportamentos distintos relacionados ao valor das compras, parcelamento, custo de frete e nível de satisfação dos clientes. A análise demonstrou que a aplicação de técnicas de agrupamento possibilita a extração de padrões relevantes que não seriam facilmente observados apenas por meio de análises descritivas.

Conclui-se que o uso de técnicas de Ciência de Dados, em especial o agrupamento de dados, constitui uma abordagem eficaz para a análise de vendas em *marketplaces*, contribuindo para uma melhor compreensão do comportamento dos pedidos e oferecendo subsídios para o apoio à tomada de decisão no contexto do comércio eletrônico.

Palavras-chave: *marketplaces*; extração de conhecimento; agrupamento de dados; análise de vendas.

Lista de ilustrações

Figura 1 – Etapas de Extração de Conhecimento da Base	21
Figura 2 – Volume de vendas de acordo com o mês/ano	31
Figura 3 – Porcentagem de vendas por região	32
Figura 4 – Top 10 estados com maiores pedidos	33
Figura 5 – Meios de pagamento por pedido	34
Figura 6 – Quantidade de parcelas	35
Figura 7 – Matriz de Correlação Entre os Atributos	36
Figura 8 – Número de clusters(K)	38
Figura 9 – Número de clusters(K) - Experimento 2	38

Lista de tabelas

Tabela 1	– Síntese dos trabalhos relacionados	19
Tabela 2	– Descrição das tabelas que compõem a base de dados utilizada	22
Tabela 3	– Principais atributos da base de dados integrada e seus tipos	25
Tabela 4	– Distribuição dos pedidos por <i>cluster</i>	39
Tabela 5	– Média dos atributos por <i>cluster</i>	39
Tabela 6	– Distribuição da quantidade de pedidos por cluster - Experimento 2 . .	40
Tabela 7	– Média dos atributos por <i>cluster</i> - Experimento 2	41

Sumário

1	INTRODUÇÃO	8
1.1	Objetivo	10
1.2	Organização do Trabalho	11
2	REVISÃO BIBLIOGRÁFICA	12
2.1	<i>Marketplaces</i>	12
2.1.1	Exemplos de <i>Marketplaces</i>	12
2.2	Análise de Dados de Marketplaces	13
2.3	Mineração de Dados	14
2.3.1	Descoberta do conhecimento em bases de dados	14
2.3.2	Pré-processamento de dados	14
2.3.3	Transformação e padronização dos dados	15
2.3.4	Agrupamento de dados (<i>Clustering</i>)	15
2.3.5	Algoritmo <i>K-Means</i>	16
2.3.6	Avaliação e Interpretação	16
2.4	Trabalhos Relacionados	17
3	EXTRAÇÃO DE CONHECIMENTO EM MARKETPLACES	20
3.1	Visão Geral Para Extração de Conhecimento em <i>Marketplaces</i>	20
3.2	Base de Dados Utilizada	21
3.3	Pré-processamento	22
3.4	Geração da Base de Dados a partir das sub-bases	23
3.5	Extração de Conhecimento	25
3.5.1	Gráficos Sumários	26
3.5.2	Correlação Entre Atributos	26
3.5.3	Tarefa de Aprendizado de Máquina: Agrupamento de Dados	27
3.6	Considerações Finais	28
4	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	30
4.1	Gráficos Sumários	30
4.2	Análise de correlação entre variáveis	35
4.3	Análise do Agrupamento de Dados	37
4.3.1	Definição do Número de <i>clusters</i>	37
4.3.2	Aplicação do <i>K-Means</i> e Análise dos <i>clusters</i>	38
4.3.2.1	Experimento 1	38
4.3.2.2	Experimento 2	40

4.4	Considerações Finais	41
5	CONCLUSÃO	42
5.1	Principais Contribuições	42
5.2	Trabalhos Futuros	43
	REFERÊNCIAS	44

1 Introdução

O *e-commerce* brasileiro está em uma trajetória de crescimento acelerado, atingindo movimentação de 186 bilhões de reais em 2023, mais do que um dobro em relação a 2019 e um crescimento de 9,5% em relação a 2022, segundo dados do (STATISTA, 2024). O crescimento dos *marketplaces* digitais como Mercado Livre¹, Amazon² e Shopee³ tem transformado o comércio online em um ambiente extremamente competitivo.

Do ponto de vista social e econômico, esses canais digitais passaram a desempenhar papel relevante para pequenos e médios empreendedores que buscam alternativas acessíveis de comercialização. Segundo o (Ebit|Nielsen, 2024), os *marketplaces* responderam por aproximadamente 78% de todo o faturamento do *e-commerce* brasileiro em 2023, o que demonstra sua forte participação no comércio digital. Apenas o Mercado Livre movimentou mais de R\$107 bilhões no Brasil, enquanto a Shopee alcançou R\$60 bilhões em GMV, números que reforçam a expressividade dessas plataformas no cenário nacional (CONVERSION, 2024).

O avanço do *e-commerce* brasileiro também pode ser observado quando analisamos o desempenho individual das principais plataformas do país. O Mercado Livre permanece como o maior *marketplaces* da América Latina, movimentando R\$ 107 bilhões em GMV somente no Brasil em 2023, representando crescimento de 21% em relação ao ano anterior Conversion (2024). A Shopee, por sua vez, consolidou-se rapidamente no mercado nacional, alcançando R\$ 60 bilhões em GMV no mesmo período, impulsionada pela estratégia de frete subsidiado e forte presença no *mobile commerce* (CONVERSION, 2024). Já a Amazon Brasil segue ampliando sua participação com foco em logística, especialmente após a expansão de centros de distribuição e do programa Amazon Prime, que ajudaram a elevar o volume de pedidos e a recorrência de compra dos usuários (E-Commerce Brasil, 2023). Esses números evidenciam que as três plataformas concentram grande parte das vendas online do país, reforçando seu papel como motores do crescimento digital e como ambientes essenciais para pequenos e médios vendedores.

Além do avanço nas vendas, o ambiente dos *marketplaces* se destaca pela intensa geração de dados de pedidos, avaliações, interações dos clientes, logística e estratégias de precificação. Cada transação produz informações valiosas que, se analisadas de forma adequada, auxiliam na tomada de decisões estratégicas, como ajuste de preços, seleção de produtos, previsão de demanda e identificação de comportamentos de compra. O monitoramento desses dados tornou-se indispensável para vendedores que desejam se manter

¹ www.mercadolivre.com.br

² www.amazon.com.br

³ www.shopee.com.br

competitivos, uma vez que permite identificar tendências, antecipar problemas e maximizar o desempenho comercial por meio de ações baseadas em evidências.

A literatura recente busca abordar esses desafios por meio de diferentes abordagens analíticas aplicadas a bases reais de *e-commerce*. Estudos têm explorado análise exploratória de dados, visualização interativa, identificação de padrões de comportamento e técnicas de mineração para agrupamento de consumidores. Trabalhos como os de (MATA, 2021), (ALVARENGA, 2021a), (BENZI, 2023) e (MONTEIRO, 2025) investigam desde o comportamento do consumidor até a segmentação de clientes e a interpretação de padrões de venda. Embora cada estudo contribua de maneira relevante para a compreensão do ambiente digital, eles variam quanto às técnicas aplicadas e revelam oportunidades para aprofundar o uso de métodos mais robustos de extração de conhecimento para auxiliar vendedores.

Como observado nos trabalhos analisados, a utilização de métodos para extração de conhecimento a partir de dados é fundamental para compreender o comportamento dos consumidores e otimizar operações em *marketplaces*. Nesse contexto, destaca-se o processo de *Knowledge Discovery in Databases* (KDD), definido como um conjunto estruturado de etapas que envolve seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados. O KDD permite transformar grandes volumes de dados brutos em informações úteis, sendo essencial para lidar com a complexidade, a diversidade e a escala dos dados gerados pelas plataformas digitais.

Considerando o volume e a complexidade dos dados gerados em *marketplaces*, o uso do processo de KDD mostra-se relevante para transformar essas informações em conhecimento útil e *insights*. A partir disto, torna-se importante desenvolver abordagens que ajudem vendedores a interpretar melhor seus próprios dados e compreender os fatores que influenciam seu desempenho. Uma análise estruturada pode favorecer decisões mais consistentes, contribuindo para estratégias de venda mais elaboradas no contexto competitivo do comércio eletrônico.

O estudo do desempenho de vendedores em *marketplaces* é altamente relevante tanto para a sociedade quanto para a comunidade acadêmica. Com milhares de vendedores oferecendo produtos, destacar-se nesse cenário é um desafio constante. Nesse contexto, a Ciência de Dados surge como uma poderosa aliada para ajudar vendedores a tomarem decisões mais assertivas com base em dados concretos.

O trabalho de Alvarenga (2021a), aplicou análise de dados na base de dados da Olist, com o objetivo de desenvolver painéis interativos que evidenciaram padrões de faturamento, categorias de produtos e avaliações de vendedores. Este trabalho demonstrou como ferramentas simples de Ciência de Dados podem apoiar decisões comerciais. De forma complementar, Leite et al. (LEITE et al., 2024) exploraram técnicas de aprendizado de máquina para entender padrões de consumo em *marketplaces*, revelando que

modelos preditivos podem indicar os clientes com maior probabilidade de compra. Já o estudo de *Santos et al.* (SIMÕES et al., 2019) analisou o desempenho de vendedores que operam em múltiplos *marketplaces* e concluiu que a diversificação de canais gera um aumento significativo nas vendas. Em conjunto, esses trabalhos destacam que tanto análises descritivas quanto abordagens preditivas são úteis para apoiar decisões estratégicas dentro dos *marketplaces*.

Nesse sentido, este trabalho tem como foco a aplicação prática de técnicas de Ciência de Dados para melhorar o desempenho de vendas de um vendedor no ambiente de *marketplaces*. A proposta envolve a análise de dados de vendas, reputação, avaliações, preços e outros atributos que influenciam o comportamento do consumidor na aquisição de tal produto.

1.1 Objetivo

O objetivo geral deste trabalho é aplicar o processo de Descoberta de Conhecimento em Bases de Dados (KDD) a um conjunto de dados de *marketplaces* (⁴ *Brazilian E-Commerce Public Dataset*) com o intuito de identificar padrões de venda que possam apoiar vendedores na compreensão do seu desempenho comercial. A partir da análise dos dados, busca-se gerar *insights* que contribuam para a melhoria do processo de tomada de decisão, permitindo ações mais estratégicas e orientadas. Para isso, será utilizada uma base de dados pública amplamente empregada em pesquisas sobre comércio eletrônico, permitindo a aplicação estruturada de técnicas de análise exploratória e agrupamento de dados.

Objetivos Específicos

Para alcançar o objetivo geral proposto, estabelecem-se os seguintes objetivos específicos:

- Selecionar, extrair e organizar atributos relevantes da base de dados, visando identificar características significativas relacionadas às vendas em *marketplaces*;
- Investigar correlações entre variáveis presentes no conjunto de dados, buscando compreender relações que possam influenciar o desempenho comercial;
- Aplicar técnicas de agrupamento de dados para identificar segmentos e padrões de comportamento entre clientes, produtos ou pedidos;
- Gerar relatórios e visualizações que mostre padrões, oferecendo suporte à interpretação dos resultados e à tomada de decisão por parte dos vendedores.

⁴ <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

1.2 Organização do Trabalho

O trabalho está organizado como descrito a seguir.

O Capítulo 2 contempla a revisão bibliográfica, abordando os principais conceitos que fundamentam o trabalho, incluindo *marketplaces*, análise de dados e o processo de Descoberta de Conhecimento em Bases de Dados (KDD), bem como técnicas de pré-processamento, padronização e agrupamento de dados.

O Capítulo 3 descreve os procedimentos metodológicos adotados, detalhando a base de dados utilizada, as etapas de pré-processamento, a integração das sub-bases e a aplicação das técnicas de análise exploratória, correlação e agrupamento por meio do algoritmo K-Means.

O Capítulo 4 apresenta os experimentos realizados e a análise dos resultados obtidos, incluindo a interpretação dos gráficos sumários, das correlações entre variáveis e dos perfis identificados a partir da técnica de agrupamento.

Por fim, o Capítulo 5 apresenta as conclusões do trabalho, destacando as principais contribuições da pesquisa, suas limitações e sugestões para trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo apresenta os principais conceitos teóricos que sustentam este trabalho. Os três pilares abordados são: **Marketplaces Digitais**, **Análise de Dados de Marketplaces** e **Descoberta do conhecimento em bases de dados**.

2.1 Marketplaces

Os *marketplaces* são plataformas online que conectam múltiplos vendedores a consumidores, funcionando como intermediários que oferecem infraestrutura, audiência e serviços logísticos mediante comissões (LAUDON; LAUDON, 2015).

O modelo ganhou força nos anos 1990 com o eBay e a Amazon e se intensificou no Brasil a partir da década de 2010 com o Mercado Livre, a Americanas S.A. e a Shopee. Atualmente, *marketplaces* respondem por mais de 78% do faturamento total do *e-commerce* brasileiro, refletindo sua centralidade no comércio digital (EBIT, 2024).

Com o crescimento da competitividade nos *marketplaces*, vendedores passaram a utilizar análises baseadas em dados para entender melhor o desempenho de suas lojas e identificar oportunidades de melhoria (LIANG, 2024). Recursos como *dashboards* personalizados, tabelas comparativas de desempenho, acompanhamento de avaliações e análise de reputação tornaram-se ferramentas acessíveis mesmo para pequenos empreendedores.

No contexto do Mercado Livre, por exemplo, muitos vendedores já utilizam informações como taxa de conversão, reputação da conta, tempo de entrega, demanda por localidade e comentários dos clientes para ajustar seus anúncios, descrever melhor seus produtos, melhorar a logística ou rever políticas de atendimento. Essas práticas, baseadas em dados reais, contribuem diretamente para a melhoria das vendas e da competitividade da loja dentro da plataforma.

2.1.1 Exemplos de Marketplaces

Entre os principais *marketplaces* atuantes no Brasil, destacam-se o Mercado Livre, a Shopee e a Amazon, que juntos concentram grande parte do tráfego e das transações do comércio eletrônico nacional. O Mercado Livre, de origem argentina, é hoje o maior *marketplaces* da América Latina, com forte presença no Brasil e investimentos robustos em logística própria e infraestrutura digital. A Shopee, plataforma asiática que se consolidou rapidamente no mercado brasileiro a partir de 2019, atrai um público massivo com estratégias agressivas de frete grátis, cupons e foco em categorias de baixo valor agregado. Já a Amazon, multinacional norte-americana, reforça sua atuação no país com um posici-

onamento voltado à confiabilidade, entrega rápida e integração com o ecossistema Prime. Cada uma dessas plataformas apresenta características únicas que influenciam diretamente o comportamento do consumidor e a forma como os vendedores devem estruturar suas estratégias. Nesse contexto, entender o funcionamento específico de cada *marketplaces* torna-se essencial para aplicar com eficácia técnicas de análise de dados voltadas à melhoria das vendas.

2.2 Análise de Dados de Marketplaces

A análise de dados de *marketplaces* envolve coletar, organizar, interpretar e visualizar dados comerciais para compreender o desempenho de produtos, clientes, campanhas e estoques. Métricas como vendas totais, vendas por geolocalidade, *ticket* médio e taxa de conversão são fundamentais (MARKETPLACER, 2024).

No início, essa análise era restrita a grandes empresas via sistemas de BI. Com ferramentas como *Excel*, *Power BI*, *Tableau* e bibliotecas *Python*, tornou-se possível para negócios menores implementar análises robustas.

Avanços recentes na integração entre sistemas ERP e plataformas de *marketplaces* têm facilitado o acesso a dados relevantes de vendas, avaliações, logística e reputação em tempo real. Essa integração permite que vendedores organizem e analisem informações de forma prática, mesmo com ferramentas simples como planilhas ou *scripts* em *Python* (MARKETPLACER, 2024).

A análise de dados de vendas em *marketplaces* tornou-se uma prática essencial para vendedores que desejam se manter competitivos em um ambiente cada vez mais dinâmico e disputado. Diferente do varejo tradicional, os *marketplaces* oferecem uma grande quantidade de informações sobre o comportamento do consumidor, desempenho dos produtos, sazonalidade, tendências de mercado e posicionamento frente à concorrência (LAUDON; TRAVER, 2020). Ao utilizar esses dados de forma estratégica, o vendedor pode tomar decisões mais assertivas sobre precificação, reposição de estoque, campanhas promocionais e até melhorias na descrição e apresentação dos produtos. Além disso, a análise contínua permite identificar padrões de compra, prever demandas futuras e adaptar-se rapidamente às mudanças do mercado.

A partir dessa estrutura de dados, é possível identificar padrões relevantes de vendas, como os produtos mais vendidos, o impacto de variações de preço, a influência de avaliações negativas no desempenho e as sazonalidades de determinadas categorias. Um exemplo desse tipo de análise foi realizada por Alvarenga (2021a), que utilizou dados reais da Olist para construir *dashboards* interativos e acessíveis, revelando *insights* úteis sobre categorias de produtos, desempenho por período e avaliação de vendedores. Esse tipo de abordagem evidencia como a visualização e análise de dados podem orientar ações

práticas no ambiente competitivo dos *marketplaces*.

2.3 Mineração de Dados

A Mineração de Dados se dá através do processo de descoberta de padrões, relações e conhecimentos relevantes a partir de grandes volumes de dados, por meio da aplicação de técnicas estatísticas, matemáticas e computacionais. Essa área tem se consolidado como um dos principais ramos da Ciência da Computação e da Ciência de Dados, especialmente diante do crescimento da geração e armazenamento de informações digitais. De acordo com [Han, Kamber e Pei \(2011\)](#), a Mineração de Dados envolve a extração automática ou semiautomática de padrões implícitos em bases de dados extensas, contribuindo para a transformação de dados brutos em conhecimento estruturado.

2.3.1 Descoberta do conhecimento em bases de dados

A Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD) é um processo estruturado cujo objetivo consiste na extração de conhecimento válido, compreensível e potencialmente útil a partir de grandes volumes de dados. Diferentemente de análises meramente descritivas, o KDD envolve uma sequência organizada de etapas interdependentes, incluindo seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados ([FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996](#)).

No contexto dos *marketplaces*, o processo de KDD assume papel central, uma vez que essas plataformas geram grandes quantidades de dados provenientes de transações, pagamentos, avaliações e informações logísticas. A aplicação sistemática dessas etapas permite transformar dados brutos em informações estratégicas capazes de apoiar a tomada de decisão e o desenvolvimento de estratégias comerciais mais assertivas.

2.3.2 Pré-processamento de dados

O pré-processamento constitui uma etapa essencial no processo de KDD, pois a qualidade dos resultados obtidos por técnicas de mineração depende diretamente da qualidade dos dados utilizados [Han, Kamber e Pei \(2011\)](#). De acordo com [Han, Kamber e Pei \(2011\)](#), dados reais frequentemente apresentam problemas como valores ausentes, ruídos, inconsistências e formatos heterogêneos, exigindo procedimentos de limpeza e organização antes da aplicação de algoritmos analíticos.

As atividades de pré-processamento podem incluir tratamento de valores ausentes, remoção de registros duplicados, detecção de outliers, padronização de formatos e conversão de tipos de dados. Além disso, a transformação dos dados pode envolver agregações,

criação de variáveis derivadas e reestruturação de atributos, com o objetivo de adequar a base às técnicas que serão aplicadas posteriormente.

2.3.3 Transformação e padronização dos dados

Após o pré-processamento, realiza-se a transformação dos dados, etapa que pode incluir a criação de variáveis derivadas, reorganização de atributos e aplicação de técnicas de reescala. Em algoritmos baseados em medidas de distância, como aqueles utilizados em tarefas de agrupamento, a escala das variáveis exerce influência direta nos resultados (HAN; KAMBER; PEI, 2011).

Quando atributos apresentam magnitudes numéricas distintas, variáveis com valores mais elevados tendem a dominar o cálculo das distâncias, comprometendo a qualidade do agrupamento. Para mitigar esse problema, aplicam-se técnicas de normalização ou padronização (TAN; STEINBACH; KUMAR, 2006).

A normalização do tipo *Min-Max* transforma os valores para um intervalo fixo, geralmente entre 0 e 1, preservando a proporcionalidade relativa entre as observações. Já a padronização, implementada por meio do *StandardScaler*, converte as variáveis para uma escala com média igual a zero e desvio padrão unitário, conforme a Equação 2.1, em que μ representa a média da variável e σ seu desvio padrão (JAMES et al., 2013).

$$z = \frac{x - \mu}{\sigma} \quad (2.1)$$

A padronização é amplamente empregada em métodos baseados em distância, pois reduz o impacto de diferenças de escala e contribui para maior estabilidade no processo de modelagem, especialmente em algoritmos como o *K-Means* (TAN; STEINBACH; KUMAR, 2006).

2.3.4 Agrupamento de dados (*Clustering*)

O agrupamento de dados, também denominado *clustering*, é uma técnica de aprendizado de máquina não supervisionado cujo objetivo é organizar instâncias em grupos de modo que elementos pertencentes ao mesmo grupo apresentem maior similaridade entre si do que em relação aos elementos de outros grupos (HAN; KAMBER; PEI, 2011).

Diferentemente de tarefas supervisionadas, o *clustering* não depende de rótulos previamente definidos. O processo baseia-se em medidas de similaridade ou distância, sendo a distância euclidiana uma das métricas mais empregadas na literatura. Essa técnica é amplamente utilizada em contextos de segmentação de clientes, identificação de perfis de consumo e análise exploratória de dados complexos (TAN; STEINBACH; KUMAR, 2006).

2.3.5 Algoritmo *K-Means*

Dentre os algoritmos de agrupamento mais difundidos na literatura, destaca-se o *K-Means*, formalmente apresentado por MacQueen (MACQUEEN, 1967). O algoritmo particiona um conjunto de dados em K grupos, buscando minimizar a soma dos quadrados das distâncias entre os pontos e o centróide do cluster ao qual pertencem, medida conhecida como *Within-Cluster Sum of Squares* (WCSS) (TAN; STEINBACH; KUMAR, 2006). Essa medida pode ser expressa pela Equação 2.2, onde k representa o número de clusters, C_i corresponde ao conjunto de instâncias pertencentes ao cluster i , x representa uma instância do cluster, e μ_i representa o centróide do cluster i .

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2.2)$$

Essa medida também é conhecida como inércia do modelo, representando o nível de dispersão das instâncias em relação ao centróide de seus respectivos clusters. Valores menores de inércia indicam que os pontos estão mais próximos dos centróides. (TAN; STEINBACH; KUMAR, 2006).

O processo ocorre de forma iterativa, alternando entre a atribuição das instâncias ao centróide mais próximo e a atualização dos centróides até que seja atingido um critério de convergência. Por utilizar distância euclidiana como métrica de similaridade, o *K-Means* é sensível à escala dos dados, reforçando a necessidade da etapa prévia de padronização das variáveis.

A definição do número adequado de clusters pode ser realizada por métodos exploratórios, como o Método do Cotovelo (*Elbow Method*), no qual se analisa o comportamento da redução da inércia para diferentes valores de K , buscando um ponto de equilíbrio entre qualidade do agrupamento e complexidade do modelo (TAN; STEINBACH; KUMAR, 2006).

2.3.6 Avaliação e Interpretação

A etapa de avaliação e interpretação demonstra a fase final do processo de KDD, sendo responsável por verificar a qualidade, consistência e relevância dos padrões identificados durante a mineração de dados. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), essa etapa é essencial para assegurar que os modelos ou estruturas descobertas representem conhecimento válido, útil e compreensível dentro do contexto analisado.

A avaliação envolve a análise do desempenho dos modelos gerados, podendo incluir métricas estatísticas, medidas de erro ou critérios de qualidade específicos da técnica aplicada. Em tarefas de aprendizado de máquina, essa fase também contempla a verificação da capacidade de generalização do modelo, buscando evitar problemas como *overfitting*,

nos quais o modelo se ajusta excessivamente aos dados de treinamento e perde capacidade de adaptação a novos dados (HAN; KAMBER; PEI, 2011). A validação adequada dos resultados é fundamental para garantir que os padrões identificados não sejam fruto de ruído ou de particularidades específicas da base analisada.

Além da avaliação quantitativa, a interpretação desempenha papel igualmente relevante. Nessa etapa, busca-se compreender o significado dos padrões descobertos, estabelecendo conexões lógicas e inferências coerentes com o domínio do problema. De acordo com Tan, Steinbach e Kumar (2006), a interpretação dos resultados permite transformar descobertas técnicas em conhecimento aplicável, facilitando sua comunicação e utilização por diferentes públicos.

A utilização de representações visuais, como gráficos, tabelas e diagramas, contribui para a compreensão dos resultados obtidos, tornando-os mais acessíveis e interpretáveis. Assim, a etapa de avaliação e interpretação consolida o processo de descoberta do conhecimento, assegurando que os achados produzidos pela mineração de dados sejam confiáveis, relevantes e passíveis de aplicação prática.

2.4 Trabalhos Relacionados

Nesta seção, são apresentados estudos que abordam o uso de dados no contexto do comércio eletrônico e dos *marketplaces*, destacando diferentes abordagens adotadas por pesquisadores para analisar comportamento de consumidores, padrões de compra, visualização de informações e técnicas de mineração de dados. A revisão desses trabalhos permite compreender como o tema tem sido explorado academicamente e quais métodos têm sido utilizados para extrair conhecimento a partir de bases reais, oferecendo um panorama geral das contribuições existentes na literatura.

O estudo de Mata (2021) buscou compreender como o *e-commerce* vem crescendo no Brasil e de que forma ele impacta tanto consumidores quanto lojistas. Para isso, o autor combinou uma pesquisa bibliográfica e documental com análises de dados em *Python*, utilizando bibliotecas como *pandas*, *NumPy*, *Matplotlib* e *Seaborn*, além de aplicar um questionário próprio para levantar os hábitos de consumo de usuários. Ele se concentra principalmente em análise exploratória e estatística, sem aplicar técnicas de aprendizado de máquina para predição ou modelagem automatizada, o que destaca a oportunidade de expandir esse tipo de estudo com tarefas mais avançadas, como classificação, *clustering* ou previsão de demanda. Após o tratamento e limpeza dos dados, os resultados evidenciaram informações importantes, como o perfil do consumidor brasileiro, as categorias de produtos mais populares, os principais *marketplaces*, os métodos de pagamento utilizados e a influência de datas promocionais no aumento das vendas. Esses achados oferecem uma visão ampla do mercado nacional e servem de inspiração para o presente projeto, já que

demonstram na prática como o uso estruturado de dados pode orientar estratégias de vendas, especialmente quando aplicados em *marketplaces*.

O trabalho de [Alvarenga \(2021b\)](#), o mais relacionado com este trabalho, explora como a visualização de dados pode se tornar uma ferramenta essencial para a tomada de decisão, especialmente em cenários de *e-commerce*. Usando o banco de dados da *Olist* disponível no *Kaggle*, o autor aplicou técnicas de análise exploratória e visualização interativa com R, criando *dashboards* interativos que ajudam a interpretar os dados. Isso inclui gráficos interativos mostrados por dia, por estado e por cidade, além de análises de método de pagamento, volume de pedidos e padrões geográficos em diferentes escalas territoriais. Essa abordagem interativa torna possível destacar *insights*, como picos de compra na *Black Friday*, preferência por cartão de crédito e distribuição regional dos clientes, ajudando gestores a tomar decisões mais informadas. A metodologia também inspira o projeto atual, ao demonstrar como construir visualizações que permitam a exploração de dados de forma intuitiva, retomando a ideia de trazer clareza e ação a partir de informações complexas. O estudo não aplica técnicas de aprendizado de máquina, mas foca em transformar os dados brutos em painéis visuais que facilitam a interpretação por tomadores de decisão. Esse trabalho contribui para a compreensão de como a visualização pode apoiar a estratégia de negócios em ambientes de *e-commerce*, ressaltando sua relevância frente ao volume e à complexidade dos dados gerados pelos *marketplaces*.

O trabalho de [Benzi \(2023\)](#) busca entender melhor como os consumidores se comportam no *e-commerce* brasileiro, aplicando mineração de dados a um conjunto real da *Olist*, com mais de 100 mil pedidos entre 2016 e 2018. A autora utilizou técnicas como segmentação por *cluster* — incluindo *K-Means*, *K-Medoids* e *DBSCAN* — e cálculos de recência, frequência e valor (modelo RFV/RFM) para agrupar os clientes com base em seus hábitos de compra. Todo o trabalho envolveu a preparação dos dados em *Python* (utilizando plataformas como *Google Colab* e bibliotecas como *pandas*, *sklearn*, *matplotlib* etc.), análise exploratória de padrões sazonais — como picos de compras na *Black Friday* — e uso de PCA para reduzir a complexidade dos dados. O resultado é uma abordagem mais refinada para criar perfis de clientes e direcionar estratégias de marketing.

O trabalho de [Monteiro \(2025\)](#) analisa como o comércio eletrônico se desenvolveu no Brasil e quais foram seus principais impactos ao longo dos últimos anos. O autor mostra que a popularização da Internet e dos dispositivos móveis transformou a forma de comprar e vender, acelerando mudanças no comportamento do consumidor. Um ponto de destaque é a pandemia de COVID-19, que impulsionou de forma significativa as vendas online e a entrada de empresas em *marketplaces*. Apesar do avanço, o estudo também chama a atenção para alguns entraves, como questões logísticas, tributárias, de segurança e fidelização de clientes. Ao mesmo tempo, ressalta o papel do *e-commerce* como uma oportunidade para pequenos negócios se inserirem no mercado e ampliarem seu alcance.

Essa visão ampla do setor ajuda a embasar o meu projeto, pois traz elementos estruturais que precisam ser considerados ao propor estratégias de vendas mais eficazes em *marketplaces*.

Tabela 1 – Síntese dos trabalhos relacionados

Trabalho	Objetivo	Técnicas Utilizadas	Resultados / Relação
Mata (2021)	Analisar o crescimento do <i>e-commerce</i> no Brasil	Análise exploratória; estatística descritiva; interpretação de padrões	Perfil do consumidor; categorias mais vendidas; panorama macroeconômico
Alvarenga (2021)	Evidenciar a importância da visualização de dados	Análise exploratória; visualização interativa; investigação de padrões espaciais e temporais	Identificação de sazonalidade; distribuição regional de clientes; apoio à tomada de decisão
Benzi (2023)	Segmentar consumidores e identificar padrões de compra	Clusterização (<i>K-Means</i> , <i>K-Medoids</i> , <i>DBSCAN</i>); modelo RFM; redução de dimensionalidade (<i>PCA</i>)	Perfis de clientes; <i>insights</i> de segmentação; aplicações práticas em marketing
Monteiro (2025)	Analisar evolução, desafios e oportunidades do <i>e-commerce</i>	Revisão bibliográfica; análise de tendências; estudo descritivo	Desafios logísticos e tributários; impactos da pandemia; oportunidades para PMEs

3 Extração de Conhecimento em Marketplaces

A metodologia deste trabalho foi estruturada de forma prática e alinhada ao objetivo principal: utilizar técnicas de KDD para extrair *insights* relevantes a partir de dados de *marketplaces* e com isso observar padrões. Para isso, empregou-se uma base pública amplamente utilizada em pesquisas sobre comércio eletrônico, criação de códigos desenvolvidos por inteligência artificial, além de ferramentas de análise e visualização de dados. As etapas descritas a seguir apresentam os materiais adotados e os métodos utilizados no desenvolvimento da pesquisa.

3.1 Visão Geral Para Extração de Conhecimento em *Marketplaces*

Este trabalho adota uma abordagem prática para a análise de dados de *marketplaces*, com foco na extração de conhecimento a partir de uma base de dados real de comércio eletrônico. A proposta consiste em analisar informações relacionadas a pedidos, clientes, produtos e logística, buscando compreender o comportamento das vendas e identificar padrões relevantes que possam auxiliar o processo de tomada de decisão de vendedores em plataformas digitais.

A Figura 1 ilustra as etapas do processo de extração de conhecimento adotado neste trabalho, organizado de forma sequencial e alinhado ao paradigma de Descoberta de Conhecimento em Bases de Dados (KDD).

O desenvolvimento inicia-se com a utilização da base de dados da *Olist*, amplamente empregada em estudos sobre *e-commerce*, a qual reúne informações detalhadas sobre transações e pedidos realizados em *marketplaces* brasileiros. A partir dessa base, são realizadas etapas de tratamento e integração dos dados, permitindo a consolidação das diferentes informações em uma única estrutura coerente para análise.

Em seguida, é conduzida uma análise exploratória dos dados, por meio da geração de gráficos e estatísticas descritivas, com o objetivo de obter uma visão inicial sobre a evolução das vendas ao longo do tempo, a distribuição geográfica dos pedidos e outros aspectos relevantes do funcionamento do *marketplaces*. Essa etapa possibilita a identificação de tendências, sazonalidades e concentrações que orientam análises mais aprofundadas.

Com base nos *insights* obtidos na análise exploratória, o trabalho avança para a aplicação de técnicas de Mineração de Dados voltadas à extração de padrões nos dados. Entre essas técnicas, considera-se o uso de métodos de aprendizado, como algoritmos



Figura 1 – Etapas de Extração de Conhecimento da Base

Fonte: Elaborado pelo autor, 2026.

de agrupamento, com o intuito de identificar perfis e comportamentos semelhantes entre vendas, clientes ou produtos. Por fim, os resultados são consolidados na forma de análises e relatórios que contribuem para a compreensão do desempenho em *marketplaces* e apoiam decisões estratégicas de vendedores.

3.2 Base de Dados Utilizada

A base de dados utilizada neste trabalho é o ¹ *Brazilian E-Commerce Public Dataset* fornecida pela *Olist*, disponibilizado publicamente na plataforma *Kaggle*. Esse conjunto de dados reúne informações reais de transações realizadas em *marketplaces* brasileiros, abrangendo o período de 2016 a 2018. A base é bastante utilizada em estudos acadêmicos e projetos de Ciência de Dados por sua riqueza de detalhes e representatividade do cenário nacional de comércio eletrônico.

O conjunto de dados é composto por várias tabelas relacionais, cada uma responsável por armazenar informações específicas do processo de compra e venda. Essas tabelas se relacionam principalmente por meio do identificador único do pedido (*order_id*), permitindo a integração das informações em uma única base consolidada para análise. A Tabela 2 apresenta as sub-bases que compõem o conjunto de dados original, bem como a descrição do conteúdo de cada uma delas, antes do processo de integração. Essas tabelas armazenam informações distintas e complementares sobre pedidos, clientes, produtos, pagamentos, avaliações e localização geográfica, sendo utilizadas como base para a construção da base de dados única empregada nas análises deste trabalho.

¹ <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

Tabela	Descrição
<i>olist_orders_dataset</i>	Contém informações gerais sobre os pedidos, incluindo status da compra e datas relacionadas ao processo de compra, envio e entrega.
<i>olist_order_items_dataset</i>	Armazena os itens associados a cada pedido, incluindo produtos vendidos, vendedores responsáveis, preços e valores de frete.
<i>olist_products_dataset</i>	Reúne características dos produtos comercializados, como categoria, dimensões físicas, peso e quantidade de fotos.
<i>product_category_name_translation</i>	Tabela auxiliar responsável pela tradução dos nomes das categorias de produtos do português para o inglês.
<i>olist_customers_dataset</i>	Contém informações dos clientes, principalmente dados de localização, como cidade e estado.
<i>olist_sellers_dataset</i>	Apresenta dados dos vendedores cadastrados no <i>marketplaces</i> , incluindo informações de localização.
<i>olist_order_payments_dataset</i>	Registra os dados relacionados aos pagamentos, como forma de pagamento, valor pago e número de parcelas por pedido.
<i>olist_order_reviews_dataset</i>	Armazena as avaliações realizadas pelos clientes após a entrega dos pedidos, incluindo notas atribuídas e comentários.
<i>olist_geolocation_dataset</i>	Disponibiliza informações geográficas associadas aos códigos postais brasileiros, como latitude e longitude.

Tabela 2 – Descrição das tabelas que compõem a base de dados utilizada

Para viabilizar as análises propostas, as tabelas foram integradas em uma única base de dados consolidada, utilizando como chave principal o identificador do pedido (*order_id*). Esse processo de integração possibilitou uma visão mais completa de cada transação, evidenciando informações de vendas, clientes, produtos, pagamentos, avaliações e localização.

A utilização dessa base integrada permite a aplicação de técnicas de análise exploratória e de Ciência de Dados, contribuindo para a identificação de padrões de comportamento, tendências de vendas e características relevantes no contexto de *marketplaces*.

3.3 Pré-processamento

A etapa de pré-processamento dos dados teve como objetivo preparar as informações para as análises subsequentes, assegurando qualidade, padronização e consistência nos dados utilizados. Considerando que a base empregada neste trabalho é composta por

dados reais de um ambiente de *marketplaces*, essa fase mostrou-se essencial para minimizar ruídos e evitar impactos negativos nos resultados obtidos.

As atividades de pré-processamento foram realizadas utilizando a linguagem de programação *Python*, em ambiente de *notebook* no *Google Colab*, que oferece suporte adequado para manipulação e análise de grandes volumes de dados de forma interativa. Esse ambiente possibilitou o acompanhamento contínuo das etapas executadas e a validação dos resultados intermediários.

Inicialmente, cada uma das sub-bases foi analisada individualmente, permitindo a compreensão de sua estrutura, dos tipos de atributos disponíveis e da presença de possíveis inconsistências. Nessa etapa, foram realizadas verificações quanto a valores ausentes, duplicidades e formatos inadequados de dados. Para os atributos utilizados nas análises posteriores, especialmente os de natureza numérica, os registros contendo valores ausentes foram removidos, garantindo que apenas observações completas fossem consideradas, uma vez que técnicas estatísticas e algoritmos de aprendizado de máquina exigem dados numéricos consistentes para seu funcionamento adequado.

Os atributos temporais estavam originalmente representados como texto, no formato dia/mês/ano seguido da hora, conforme padrão brasileiro. Dessa forma, foi realizada a conversão dessas colunas para o formato de data e hora, garantindo a correta interpretação das informações temporais pelo ambiente de análise. Essa conversão preservou a informação de data e horário e possibilitou a extração de variáveis derivadas, como mês e ano da compra, utilizadas em análises de evolução e sazonalidade das vendas.

Ao final dessa etapa, as sub-bases encontravam-se organizadas e preparadas para o processo de integração, que será descrito na próxima seção. Esse pré-processamento foi fundamental para viabilizar a geração de uma base de dados unificada e consistente, que servirá de suporte para as análises exploratórias e para a aplicação de técnicas de Ciência de Dados ao longo do desenvolvimento do trabalho.

3.4 Geração da Base de Dados a partir das sub-bases

Após a etapa de pré-processamento e padronização, realizou-se a geração de uma base de dados unificada a partir das sub-bases que compõem o *Brazilian E-Commerce Dataset* da *Olist*. Essa etapa teve como finalidade consolidar as informações dispersas em diferentes tabelas, permitindo uma visão integrada do processo de vendas em um *marketplaces*.

A integração dos dados foi realizada por meio da associação entre identificadores comuns presentes nas sub-bases, principalmente o identificador do pedido (*order_id*), que atua como elemento central de ligação entre informações de compras, pagamentos, itens

adquiridos, avaliações, clientes e vendedores.

Durante o processo de consolidação da base de dados, optou-se por transformar o atributo *order_item_id*, originalmente utilizado para enumerar os itens dentro de cada pedido, em uma variável representativa da quantidade total de itens por pedido. Ou seja, a base integrada foi estruturada no ID de pedido, isto é, com uma única linha por *order_id*. Dessa forma, foi realizada uma agregação dos registros por pedido, contabilizando-se o número de itens associados a cada *order_id*, gerando assim a nova variável *quantidade_itens*. Essa transformação permitiu preservar a informação referente ao volume de produtos adquiridos em cada transação, ao mesmo tempo em que manteve a coerência estrutural da base unificada.

Além disso, foi criada a variável correspondente à região geográfica do cliente (região), derivada a partir do atributo *customer_state*, que indica o estado de localização do cliente. Os estados brasileiros foram agrupados nas cinco regiões do país (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), permitindo representar a localização regional. Essa transformação possibilita incorporar informações geográficas à base integrada, permitindo maiores análises relacionadas à distribuição das vendas e ao impacto da localização no padrão de compra e nos custos logísticos.

Como resultado desse processo, obteve-se uma base de dados única e estruturada, contendo atributos relacionados ao pedido, ao vendedor, ao cliente, às condições de pagamento, às avaliações e à localização geográfica. Essa base integrada possibilita análises mais completas e consistentes, permitindo observar padrões de vendas ao longo do tempo, diferenças regionais, comportamentos de consumo e relações entre variáveis distintas.

A Tabela 3 apresenta os principais atributos selecionados da base de dados integrada, juntamente com seus respectivos tipos de dados, que serão utilizados nas análises exploratórias, na análise de correlação e na aplicação da técnica de agrupamento de dados.

Atributo	Tipo de dado	Descrição
order_id	Nominal	Identificador único do pedido no <i>marketplace</i> .
quantidade_itens	Discreto	Quantidade de itens comprados por pedido.
customer_id	Nominal	Identificador do cliente associado ao pedido.
customer_state	Nominal	Estado de localização do cliente.
order_purchase_timestamp	Temporal (datetime)	Data e horário em que o pedido foi realizado.
order_delivered_customer_date	Temporal (datetime)	Data em que o pedido foi entregue ao cliente.
freight_value	Contínuo	Valor correspondente ao custo do frete do pedido.
seller_id	Nominal	Identificador único do vendedor responsável pelo item ou pedido.
seller_state	Nominal	Unidade federativa (estado) de localização do vendedor.
payment_type	Nominal	Forma de pagamento utilizada na transação.
payment_installments	Discreto	Número de parcelas utilizadas no pagamento do pedido.
payment_value	Contínuo	Valor total pago pelo cliente no pedido.
review_score	Ordinal	Nota atribuída pelo cliente ao pedido, variando de 1 a 5.
regiao	Nominal	Região geográfica do cliente.

Tabela 3 – Principais atributos da base de dados integrada e seus tipos

3.5 Extração de Conhecimento

Nessa fase, são aplicadas técnicas de análise exploratória e estatística, apoiadas por visualizações gráficas e análises de correlação, permitindo uma compreensão mais aprofundada do comportamento das vendas, dos consumidores e dos vendedores presentes na base de dados.

As análises realizadas nesta etapa foram organizadas em dois momentos princi-

país: a geração de gráficos sumários, voltados à exploração inicial dos dados, e a análise de correlação entre atributos, conduzida a partir da base de dados integrada. Essa organização permite avançar gradualmente da observação geral dos dados para a investigação de relações mais específicas entre as variáveis.

3.5.1 Gráficos Sumários

Os gráficos foram elaborados com o objetivo de fornecer uma visão geral da base de dados integrada e apoiar a compreensão inicial do comportamento das vendas em *marketplaces*. As visualizações foram construídas a partir da base única resultante da junção das sub-bases da base de dados, permitindo análises mais consistentes ao considerar, de forma conjunta, informações de pedidos, clientes, produtos, pagamentos e localização.

Inicialmente, foi construído o gráfico de vendas por mês, que apresenta a quantidade total de pedidos ao longo do tempo, considerando a distribuição mensal e anual das vendas. Essa visualização permite observar a evolução temporal do volume de pedidos e identificar possíveis padrões de crescimento ou sazonalidade.

Em seguida, foram gerados gráficos voltados à análise da distribuição geográfica das vendas. O primeiro apresenta os top 10 estados com maior volume de pedidos, evidenciando as unidades federativas com maior concentração de transações. Complementarmente, foi elaborado um gráfico que representa a porcentagem de vendas por região do Brasil, agrupando os pedidos conforme as cinco grandes regiões do país, permitindo analisar a participação relativa de cada região no total de vendas registradas.

Também foi desenvolvido um gráfico referente aos meios de pagamento utilizados nas transações, com o objetivo de identificar as formas de pagamento mais recorrentes entre os pedidos analisados. Por fim, foi construído um gráfico das categorias de produtos mais vendidas, possibilitando a identificação das categorias que concentram maior volume de pedidos dentro do conjunto de dados.

De modo geral, essas visualizações fornecem uma visão inicial estruturada da base integrada, contribuindo para a compreensão dos principais padrões presentes nos dados e orientando as etapas analíticas subsequentes.

3.5.2 Correlação Entre Atributos

A análise de correlação entre atributos foi realizada a partir da base de dados integrada, construída por meio da junção das sub-bases que compõem a base de dados. O objetivo dessa etapa é investigar o grau de relacionamento existente entre variáveis quantitativas, como valor do pedido, número de parcelas, avaliação do cliente, valor do frete, e quantidade de itens do pedido.

Para essa finalidade, foi utilizado o coeficiente de correlação de *Pearson*, escolhido por ser adequado à análise de relações lineares entre variáveis numéricas contínuas. A correlação foi calculada entre todos os pares possíveis dentro do conjunto de variáveis quantitativas selecionadas, resultando em uma matriz de correlação que mostra os níveis de associação entre os atributos analisados.

O coeficiente de Pearson assume valores no intervalo de -1 a 1, em que valores próximos de 1 indicam forte correlação linear positiva, valores próximos de -1 indicam forte correlação linear negativa, e valores próximos de 0 indicam ausência de correlação linear significativa (JAMES et al., 2013). A Equação 3.1 apresenta a fórmula da correlação linear de Pearson, onde, x_i e y_i correspondem aos valores observados das variáveis analisadas, $\bar{x}$ e $\bar{y}$ representam as médias dessas variáveis, e n corresponde ao número total de observações.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.1)$$

Para a interpretação dos coeficientes obtidos, foram consideradas faixas usuais de magnitude da correlação de *Pearson*, nas quais valores próximos de zero indicam associação fraca ou inexistente, valores entre aproximadamente 0,30 e 0,50 indicam associação fraca a moderada, e valores mais elevados indicam relações mais intensas. Essa abordagem permitiu analisar a força das relações observadas sem a adoção de um limiar fixo, priorizando a interpretação contextual dos resultados.

3.5.3 Tarefa de Aprendizado de Máquina: Agrupamento de Dados

Nesta etapa do trabalho, foi aplicada uma tarefa de Aprendizado de Máquina do tipo não-supervisionado, utilizando uma técnica de agrupamento de dados (*clustering*). O objetivo dessa abordagem é identificar padrões e perfis semelhantes entre os pedidos realizados no *marketplaces*, sem a necessidade de rótulos previamente definidos, permitindo uma análise exploratória mais aprofundada da estrutura dos dados.

Para a realização do agrupamento, foram selecionados atributos numéricos representativos do comportamento de compra e da experiência do consumidor. No primeiro experimento, foram considerados os seguintes atributos numéricos, valor total do pedido (*payment_value*), número de parcelas utilizadas no pagamento (*payment_installments*) o custo de frete (*freight_value*) e a avaliação dos clientes (*review_score*), conforme apresentado na Tabela 3.

No segundo experimento, foi incluído também o atributo correspondente à quantidade de itens por pedido (*quantidade_itens*) e à região geográfica do cliente (*regiao*). Para utilizar o atributo de região no algoritmo de agrupamento, a variável foi convertida para formato numérico por meio de codificação categórica através do *One-Hot Encoding*,

resultando em novas variáveis binárias representando cada região (Norte, Nordeste, Centro Oeste, Sudeste e Sul). A inclusão desse atributo teve como objetivo avaliar a influência da localização geográfica na formação dos grupos.

O algoritmo escolhido para a tarefa de agrupamento foi o *K-Means*, um método de agrupamento não supervisionado que particiona os dados em K grupos, minimizando a variância intra-cluster (MACQUEEN, 1967). Para a definição do número adequado de *clusters*, foi empregado o Método do Cotovelo (*Elbow Method*), no qual consiste em calcular a inércia (Equação 2.2) para diferentes valores de K e analisar o comportamento da curva resultante. O ponto a partir do qual a redução da inércia passa a ocorrer de forma menos acentuada é interpretado como um indicativo adequado para a escolha do número de grupos.

Antes da aplicação do algoritmo *K-Means*, os atributos selecionados foram submetidos a um processo de padronização por meio do *StandardScaler* do *Python*, técnica que transforma as variáveis para média zero e desvio padrão unitário. Essa etapa foi necessária devido ao fato de o *K-Means* basear-se em distância Euclidiana, sendo sensível a diferenças de escala entre os atributos. A padronização assegura que todas as variáveis contribuam de forma equilibrada no processo de agrupamento.

A aplicação da técnica de agrupamento possibilita a identificação de perfis distintos de pedidos no *marketplaces*, contribuindo para uma compreensão mais aprofundada da heterogeneidade presente nos dados. Esses perfis servem de base para análises posteriores e para a interpretação dos resultados apresentados no capítulo de experimentos, auxiliando na geração de *insights* relevantes para a tomada de decisão no contexto de vendas em *marketplaces*.

Nos experimentos, foram avaliados valores de k no intervalo de 1 a 10, conforme prática comum na literatura, permitindo observar o comportamento da inércia e identificar o ponto de equilíbrio entre qualidade do agrupamento e complexidade do modelo.

3.6 Considerações Finais

Neste capítulo foram apresentados os procedimentos metodológicos adotados para a condução do trabalho, desde a caracterização da base de dados até a definição das técnicas de análise aplicadas. Inicialmente, descreveu-se o processo de extração de conhecimento estruturado em etapas sequenciais, contemplando coleta, integração, pré-processamento e análise dos dados. Em seguida, detalhou-se a construção da base integrada a partir das sub-bases originais, bem como os cuidados adotados na preparação dos dados para garantir consistência e adequação às análises subsequentes.

Também foram apresentadas as técnicas estatísticas e de aprendizado de máquina

empregadas, incluindo a análise de correlação entre variáveis quantitativas e a aplicação da técnica de agrupamento de dados.

Com a definição dos procedimentos metodológicos, estabelece-se a base necessária para a apresentação e discussão dos resultados no capítulo seguinte, no qual são analisados os padrões identificados e os perfis obtidos a partir da aplicação das técnicas descritas.

4 Experimentos e Análise dos Resultados

Este capítulo apresenta os resultados obtidos a partir da aplicação dos procedimentos metodológicos descritos anteriormente. Inicialmente, na Seção 4.1, são apresentados os gráficos sumários gerados a partir da base integrada, com o objetivo de caracterizar o comportamento geral das vendas e identificar padrões descritivos relevantes. Em seguida, na Seção 4.2, é realizada a análise de correlação entre os principais atributos quantitativos, buscando compreender o grau de associação linear existente entre as variáveis selecionadas. Por fim, na Seção 4.3, são discutidos os resultados da técnica de agrupamento de dados, incluindo a definição do número de *clusters* e a interpretação dos perfis identificados a partir da aplicação do algoritmo.

4.1 Gráficos Sumários

Esta seção apresenta os gráficos sumários elaborados a partir da base de dados integrada, com o objetivo de caracterizar o comportamento geral das vendas no *marketplace* analisado. Ao todo, cinco gráficos principais foram construídos: o gráfico da quantidade de vendas por mês, que evidencia a evolução temporal dos pedidos; o gráfico da porcentagem de vendas por região do Brasil, que apresenta a participação relativa das cinco grandes regiões; o gráfico do top 10 estados com maior volume de pedidos, que destaca a distribuição geográfica das vendas; o gráfico dos meios de pagamento mais utilizados, que identifica as formas de pagamento predominantes; e o gráfico da quantidade de parcelas utilizadas nas compras, que permite observar o padrão de parcelamento adotado pelos consumidores. Essas visualizações oferecem uma visão inicial estruturada da base e servem como ponto de partida para as análises subsequentes.

A Figura 2 apresenta o gráfico de vendas por mês permite observar o comportamento temporal dos pedidos registrados no *marketplaces* entre 2017 e 2018. A série temporal apresenta um crescimento gradual ao longo do período analisado, indicando uma expansão consistente do volume de compras na plataforma. Esse comportamento está alinhado ao movimento geral do *e-commerce* brasileiro, que passou por um processo de aceleração no período.

Também é possível identificar algumas variações sazonais, com picos de vendas em determinados meses — especialmente no final de 2017 — que podem estar relacionados a datas promocionais amplamente conhecidas no varejo digital, como *Black Friday* e campanhas sazonais de fim de ano. Da mesma forma, observa-se a ocorrência de períodos de menor movimento, refletindo oscilações naturais da demanda ao longo do tempo.

Esse panorama é importante pois oferece uma visão macro do comportamento das vendas, permitindo identificar tendências e eventuais sazonalidades que influenciam o desempenho dos vendedores. A partir desse entendimento, torna-se possível direcionar análises mais específicas, aprofundando a investigação sobre fatores que impactam o volume de pedidos e a dinâmica comercial dentro dos *marketplaces*.

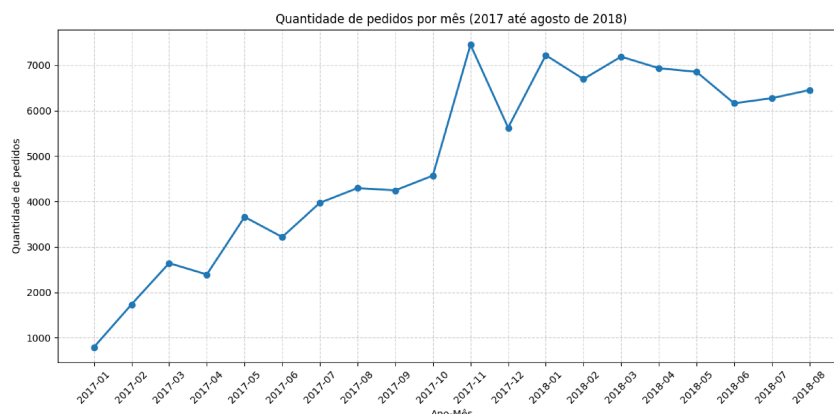


Figura 2 – Volume de vendas de acordo com o mês/ano

Fonte: Elaborado pelo autor, 2026.

A Figura 3 mostra a análise da distribuição percentual dos pedidos por região do Brasil que demonstra uma concentração extremamente elevada no Sudeste, responsável por aproximadamente 68,6% de todas as transações registradas na base de dados. Esse resultado reforça a importância econômica da região no comércio eletrônico nacional, influenciada por fatores como maior densidade populacional, infraestrutura logística avançada, maior disponibilidade de centros de distribuição e maior poder aquisitivo médio da população.

Em seguida, a região Sul aparece com 14,2% dos pedidos, seguida pelo Nordeste com 9,4% e pelo Centro-Oeste com 5,8%. Esses valores indicam uma participação intermediária dessas regiões no volume total de pedidos analisados, evidenciando a presença de atividade relevante no comércio eletrônico, embora em menor escala quando comparadas à região Sudeste.

Por fim, a região Norte registra apenas 1,9%, indicando uma menor participação no volume de pedidos analisado. Esse resultado pode estar relacionado a fatores logísticos e geográficos que influenciam o desenvolvimento do comércio eletrônico, como maiores distâncias entre centros urbanos, limitações na infraestrutura de transporte e maior complexidade na distribuição de mercadorias nessa região, o que pode impactar a frequência de transações realizadas (E-Commerce Brasil, 2021).

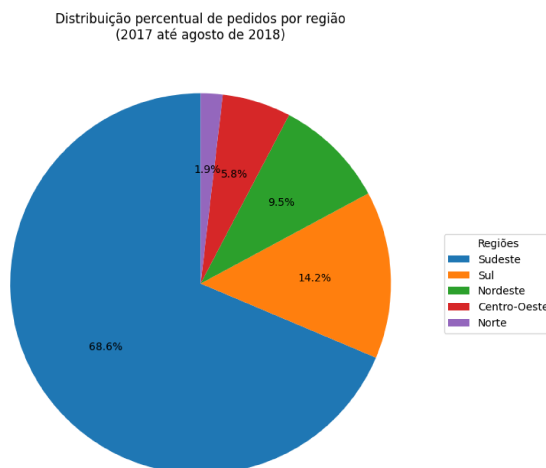


Figura 3 – Porcentagem de vendas por região

Fonte: Elaborado pelo autor, 2026.

A Figura 4 demonstra a análise do gráfico que apresenta os dez estados com maior volume de pedidos. Ela revela uma forte concentração das vendas na região Sudeste, com destaque absoluto para o estado de São Paulo, no qual aparece com um número significativamente superior aos demais, evidenciando-se como o principal polo consumidor do *marketplaces* e concentrando, sozinho, um volume de pedidos maior do que a soma de diversos outros estados. Em seguida, Rio de Janeiro e Minas Gerais reforçam o peso econômico da região Sudeste, que se consolida como o mercado mais relevante para estratégias de vendas, marketing e logística.

A região Sul também demonstra participação consistente, com Paraná, Rio Grande do Sul e Santa Catarina figurando entre os dez primeiros, embora com números bem inferiores aos líderes, indicando um mercado importante, porém não tão expressivo quanto o do Sudeste. Já as regiões Nordeste e Centro-Oeste aparecem de forma mais modesta no ranking, representadas principalmente pela Bahia, Distrito Federal e Goiás, com volumes menores e distribuídos de maneira menos homogênea.

Esses resultados reforçam a disparidade regional no comércio eletrônico brasileiro, indicando que fatores como infraestrutura logística, densidade populacional e poder aquisitivo influenciam diretamente o desempenho das vendas. Para vendedores que atuam em *marketplaces*, essa distribuição oferece *insights* estratégicos sobre onde concentrar esforços de divulgação, otimizar estoques e desenvolver ações específicas para regiões com menor participação, mas com potencial de crescimento.

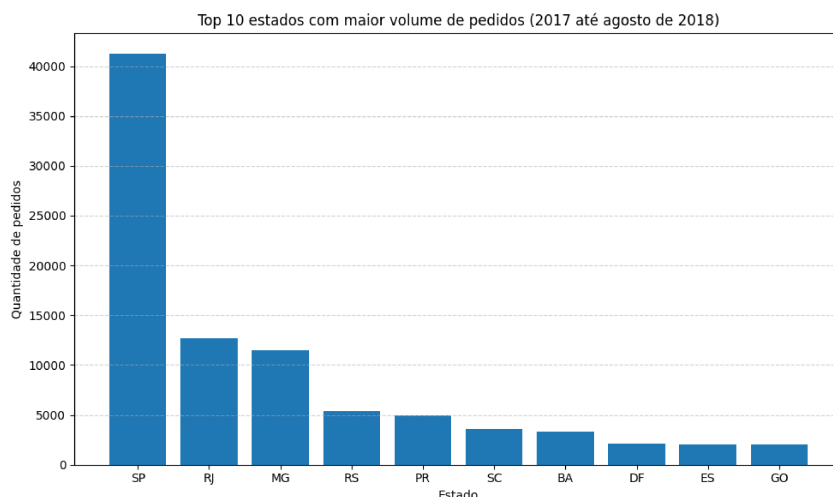


Figura 4 – Top 10 estados com maiores pedidos

Fonte: Elaborado pelo autor, 2026.

A análise dos meios de pagamento utilizados nas transações permite compreender as preferências dos consumidores e o comportamento de compra no contexto dos *marketplaces*. Observa-se que o cartão de crédito é amplamente predominante, concentrando a maior parte dos pedidos realizados. Esse resultado indica a preferência dos consumidores por modalidades que oferecem maior flexibilidade, como o parcelamento das compras, além da praticidade no processo de pagamento. Em seguida, o boleto bancário aparece como a segunda opção mais utilizada, demonstrando que uma parcela significativa dos consumidores ainda opta por pagamentos à vista, possivelmente motivada pela ausência de juros ou por questões de controle financeiro.

Outros meios de pagamento, como voucher e cartão de débito, apresentam participação menor no volume total de pedidos. O uso de voucher está associado, em geral, a créditos promocionais, programas de fidelização ou compensações concedidas pelo *marketplaces*, funcionando como uma estratégia para incentivar novas compras e aumentar a retenção de clientes. Apesar de sua menor representatividade, esse método revela a importância de ações promocionais no ecossistema do comércio eletrônico.

De forma geral, os resultados reforçam a relevância de compreender a diversidade de meios de pagamento disponíveis, uma vez que eles influenciam diretamente a experiência do consumidor e podem impactar as taxas de conversão e vendas dos vendedores nos *marketplaces*, no qual pode ser observado na Figura 5.

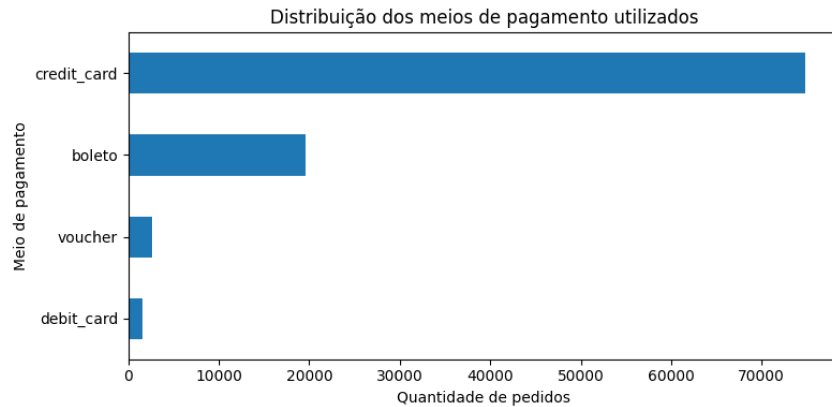


Figura 5 – Meios de pagamento por pedido

Fonte: Elaborado pelo autor, 2026.

A Figura 6 apresenta a distribuição da quantidade de parcelas utilizadas nas transações. Por meio desse gráfico, é possível observar quais faixas de parcelamento são mais utilizadas pelos consumidores, permitindo identificar padrões no modo como os pagamentos são realizados no comércio eletrônico. Os dados evidenciam a frequência de pedidos conforme o número de parcelas escolhido no momento da compra.

Observa-se uma forte concentração de pedidos realizados com uma única parcela, indicando que uma parcela significativa dos consumidores opta pelo pagamento à vista, mesmo quando utiliza o cartão de crédito. Esse comportamento pode estar associado a compras de menor valor ou à preferência por evitar compromissos financeiros prolongados.

Ao mesmo tempo, nota-se uma presença relevante de compras parceladas em duas ou mais parcelas, esse padrão sugere que o parcelamento é um recurso importante para viabilizar aquisições de maior valor, tornando os produtos mais acessíveis ao consumidor final.

Esses resultados indicam que oferecer opções flexíveis de parcelamento é uma estratégia relevante para vendedores em *marketplaces*, pois pode impactar diretamente a decisão de compra, aumentar o *ticket* médio e ampliar o alcance de diferentes perfis de consumidores.

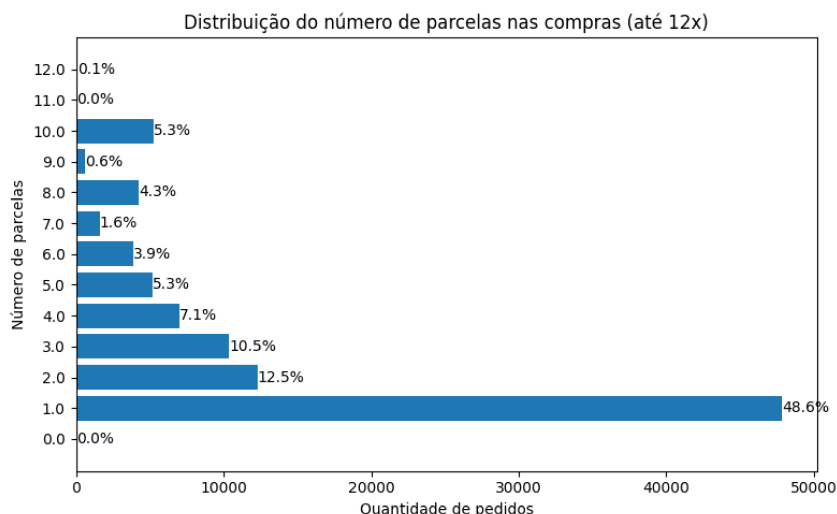


Figura 6 – Quantidade de parcelas

Fonte: Elaborado pelo autor, 2026.

4.2 Análise de correlação entre variáveis

Com o objetivo de compreender as relações entre os principais atributos analisados, foi realizada uma análise de correlação entre as variáveis valor do pedido, número de parcelas, valor do frete, avaliação do cliente e quantidade de itens por pedido, utilizando o coeficiente de correlação de Pearson. Essa análise também teve como finalidade verificar a existência de correlações fortes entre os atributos, uma vez que variáveis altamente correlacionadas poderiam introduzir redundância de informação e influenciar o processo de agrupamento posteriormente realizado. A Figura 7 apresenta a matriz de correlação obtida.

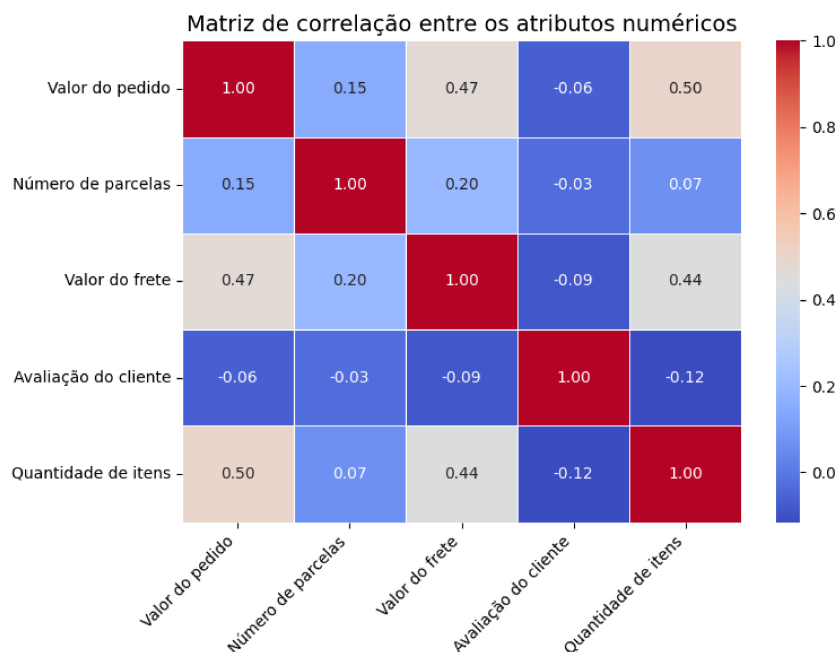


Figura 7 – Matriz de Correlação Entre os Atributos

Fonte: Elaborado pelo autor, 2026.

Observa-se que o valor do pedido apresenta correlação moderada com a quantidade de itens, indicando que pedidos com maior número de itens tendem a possuir maior valor total. Também foi verificada uma correlação moderada entre o valor do pedido e o valor do frete, bem como entre a quantidade de itens e o valor do frete, sugerindo que pedidos maiores geralmente estão associados a custos logísticos mais elevados.

O número de parcelas apresentou correlação fraca com as demais variáveis, indicando que o parcelamento não está fortemente relacionado ao valor da compra ou à quantidade de produtos. Já a avaliação do cliente apresentou correlação muito baixa com todos os atributos analisados, indicando que a nota atribuída não depende diretamente dessas variáveis, podendo estar associada a outros fatores.

Os atributos relacionados à região geográfica não foram incluídos nesta análise, pois exigiriam um estudo mais aprofundado para a definição da medida de correlação mais adequada ao tipo de variável. Ainda assim, essa informação foi utilizada posteriormente na etapa de agrupamento de dados, permitindo avaliar sua influência na formação dos grupos.

De modo geral, os resultados indicam que as variáveis apresentam baixa ou moderada correlação entre si, o que é adequado para a aplicação da técnica de agrupamento, pois permite que diferentes características contribuam para a identificação de perfis distintos de pedidos.

4.3 Análise do Agrupamento de Dados

Com o objetivo de aprofundar a análise dos padrões presentes na base de dados, nesta seção são apresentados os resultados da aplicação da técnica de agrupamento de dados (*clustering*). Antes da aplicação de agrupamento, é importante destacar que a base analisada é composta por 97.916 instâncias, correspondentes aos pedidos realizados após o pré-processamento dos dados. Foram realizados dois experimentos, sem que em cada um deles, um conjunto diferente de atributos foi usado. No experimento 1, foram usados quatro atributos numéricos: valor total do pedido, número de parcelas, valor do frete e avaliação do cliente. Já no experimento 2, foram usados dez atributos, sendo cinco numéricos: valor total do pedido, número de parcelas, valor do frete, avaliação do cliente, quantidade de itens, e cinco binários que representam as regiões: norte, nordeste, centro-oeste, sudeste e sul. Esses atributos serviram como base para a identificação de padrões e perfis semelhantes entre as transações. A partir dos atributos selecionados, o agrupamento permite identificar perfis distintos de pedidos no *marketplaces*, evidenciando comportamentos semelhantes entre grupos de transações.

Inicialmente, é descrito o processo de definição do número adequado de *clusters* para cada experimento, seguido pela aplicação do algoritmo de agrupamento e pela análise dos perfis identificados. Essa abordagem possibilita uma compreensão mais detalhada da heterogeneidade dos pedidos e fornece informações para a interpretação dos resultados no contexto de vendas em *marketplaces*.

4.3.1 Definição do Número de *clusters*

A Figura 8 apresenta o gráfico referente ao experimento 1, resultante da aplicação do método do cotovelo, no qual é possível observar uma redução acentuada da inércia nos primeiros valores de k , seguida por uma diminuição mais gradual a partir de determinado ponto.

Observa-se que a partir de quatro *clusters* a curva apresenta uma mudança de inclinação, indicando que a redução da inércia torna-se menos expressiva. Dessa forma, o valor $k = 4$ foi selecionado para a aplicação do algoritmo *K-Means*, por representar um equilíbrio entre a qualidade do agrupamento e a complexidade do modelo, além de permitir uma interpretação adequada dos perfis identificados nos dados analisados.

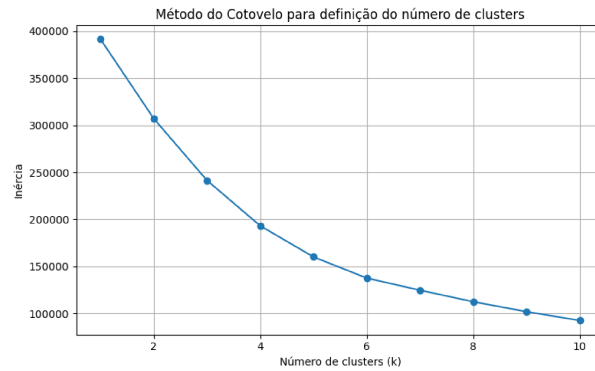


Figura 8 – Número de clusters(K)

Fonte: Elaborado pelo autor, 2026.

Já para o experimento 2, a Figura 9 apresenta um comportamento parecido com o experimento 1, ou seja, a partir de quatro *clusters* a curva apresenta uma mudança de inclinação. Com isso o valor $k = 4$ foi selecionado para o experimento 2.

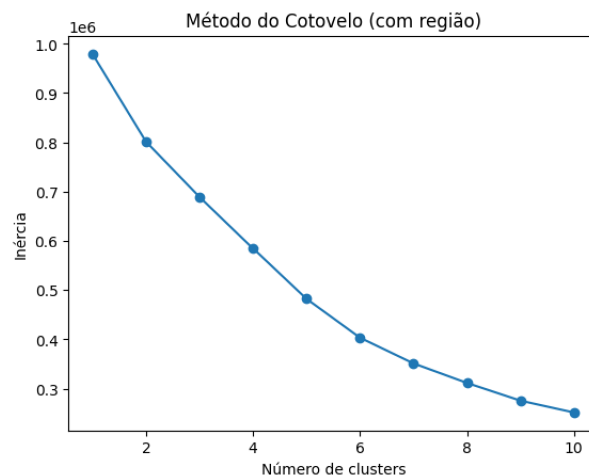


Figura 9 – Número de clusters(K) - Experimento 2

Fonte: Elaborado pelo autor, 2026.

4.3.2 Aplicação do *K-Means* e Análise dos *clusters*

Após a definição do número de *clusters*, o algoritmo *K-Means* foi aplicado aos dois experimentos, utilizando $k = 4$. Como resultado, cada pedido foi atribuído a um dos quatro *clusters* identificados, permitindo a segmentação dos pedidos com base em características semelhantes.

4.3.2.1 Experimento 1

Na Tabela 4, a distribuição dos pedidos entre os quatro grupos formados mostrou que os *clusters* não apresentam tamanhos homogêneos. Observou-se que o *Cluster* 1

concentrou a maior parte dos pedidos, representando aproximadamente 63,10% da base analisada. Em seguida, o *Cluster 2* correspondeu a 19,46%, o *Cluster 0* a 15,84% e, por fim, o *Cluster 3* apresentou a menor representatividade, com 1,60% dos pedidos. Essa diferença de proporção é comum em técnicas de agrupamento aplicadas a dados reais, uma vez que determinados perfis de comportamento tendem a ser mais frequentes que outros.

Cluster	Quantidade de pedidos	Percentual (%)
0	15.514	15,84
1	61.783	63,10
2	19.051	19,46
3	1.568	1,60

Tabela 4 – Distribuição dos pedidos por *cluster*

Para facilitar a análise, foi construída a Tabela 5, na qual apresenta a média dos atributos em cada *cluster*, permitindo identificar as principais características de cada grupo e compreender os perfis distintos de pedidos presentes na base analisada.

Cluster	Valor médio (R\$)	Parcelas	Avaliação	Frete (R\$)
0	317,12	7,94	4,36	27,93
1	133,96	1,81	4,75	19,00
2	174,34	2,31	3,84	22,02
3	2.269,61	5,16	3,58	131,65

Tabela 5 – Média dos atributos por *cluster*

A aplicação do algoritmo *K-Means* resultou na identificação de quatro *clusters* com características distintas. Cada *cluster* representa um perfil específico de comportamento de compra, considerando aspectos financeiros, operacionais e de satisfação do consumidor.

A análise das médias dos atributos em cada *cluster* permitiu identificar perfis distintos de pedidos no *marketplace*. O *Cluster 1*, que concentra a maior parcela dos registros, caracteriza-se por pedidos de menor valor médio, baixo número de parcelas e elevada avaliação média dos clientes. Esse grupo representa o perfil predominante da base, sugerindo transações de menor complexidade e alto nível de satisfação.

O *Cluster 0* apresenta pedidos com valor médio superior ao do grupo dominante, além de um número significativamente maior de parcelas. Esse comportamento indica a presença de consumidores que realizam compras de maior valor e optam por parcelamento como estratégia de pagamento. A avaliação média desse grupo permanece elevada, sugerindo que o aumento do valor do pedido não compromete necessariamente a satisfação do cliente.

O *Cluster 2* distingue-se por apresentar avaliação média consideravelmente inferior aos demais grupos, embora o valor médio do pedido e o número de parcelas permaneçam

em níveis intermediários. Esse perfil pode indicar ocorrências de problemas relacionados à experiência de compra, como atrasos na entrega, divergências no produto ou outras questões operacionais, configurando um grupo de atenção estratégica para análise mais aprofundada.

Por fim, o *Cluster 3* representa um grupo bastante específico e minoritário, composto por pedidos de altíssimo valor médio e frete significativamente superior aos demais *clusters*. Ainda que corresponda a uma pequena fração da base, esse grupo possui grande relevância financeira, pois envolve transações de elevado *ticket* médio. A identificação desse perfil pode auxiliar na formulação de estratégias direcionadas a consumidores de maior poder aquisitivo ou a produtos de maior valor agregado.

De maneira geral, os resultados evidenciam que a aplicação do *K-Means* foi capaz de identificar padrões heterogêneos no comportamento dos pedidos do *marketplace*, revelando perfis distintos que não seriam facilmente perceptíveis por meio de análises exclusivamente descritivas. A segmentação obtida fornece subsídios relevantes para a compreensão da dinâmica das vendas e pode contribuir para a formulação de estratégias mais direcionadas no contexto do comércio eletrônico.

4.3.2.2 Experimento 2

A Tabela 6 apresenta a quantidade de pedidos em cada *cluster*, bem como sua participação percentual em relação ao total de pedidos. Observa-se que o *Cluster 1* concentra a maior parte dos pedidos, representando aproximadamente 68,55% da base, indicando que esse perfil corresponde ao comportamento predominante dos consumidores. Em seguida, o *Cluster 3* representa 16,12% dos pedidos, enquanto o *Cluster 0* corresponde a 9,50% e o *Cluster 2* apresenta a menor participação, com 5,83%.

Cluster	Quantidade de pedidos	Porcentagem (%)
0	9301	9,50
1	67120	68,55
2	5712	5,83
3	15783	16,12

Tabela 6 – Distribuição da quantidade de pedidos por cluster - Experimento 2

A Tabela 7 apresenta os valores médios dos atributos analisados para cada cluster, considerando o valor total do pedido, o número de parcelas, o valor do frete, a avaliação do cliente, a quantidade de itens e a região geográfica predominante em cada grupo.

Cluster	Valor médio	Parcelas	Frete	Aval.	Qtd. itens	Região pre-dominante
0	314,04	3,50	37,38	3,90	1,14	Nordeste
1	185,34	2,83	19,66	4,13	1,14	Sudeste
2	230,11	2,94	26,49	4,08	1,15	Centro-Oeste
3	215,69	3,01	26,24	4,13	1,14	Sul

Tabela 7 – Média dos atributos por *cluster* - Experimento 2

O *Cluster* 0, predominante na região Nordeste, apresentou o maior valor médio de pedido (R\$ 314,04) e também o maior valor médio de frete (R\$ 37,38), além de uma avaliação média de 3,90. Esse resultado indica que pedidos dessa região tendem a ter maior custo de entrega, o que pode estar relacionado à distância em relação aos principais centros logísticos.

O *Cluster* 1, predominante na região Sudeste, apresentou o menor valor médio de pedido (R\$ 185,34) e o menor valor médio de frete (R\$ 19,66), com avaliação média de 4,13. Como esse cluster também representa a maior parte dos pedidos, ele pode ser considerado o perfil mais comum da base, caracterizado por pedidos de menor valor e menor custo de envio.

O *Cluster* 2, predominante na região Centro-Oeste, apresentou valores intermediários tanto para o valor do pedido (R\$ 230,11) quanto para o frete (R\$ 26,49), além de avaliação média de 4,08. Esse grupo representa um perfil menos frequente, mas com características próximas à média geral observada na base.

Já o *Cluster* 3, predominante na região Sul, apresentou valor médio de pedido de R\$ 215,69 e valor médio de frete de R\$ 26,24, com avaliação média de 4,13. Esse cluster apresenta valores próximos da média geral, diferenciando-se principalmente pela região predominante.

De forma geral, os resultados mostram que o algoritmo foi capaz de identificar grupos com características diferentes, especialmente em relação ao valor do frete e à região dos clientes, indicando que a localização geográfica influencia diretamente o comportamento das vendas no marketplace.

4.4 Considerações Finais

As análises apresentadas neste capítulo permitiram identificar padrões importantes na base de dados, evidenciando diferenças no comportamento das vendas ao longo do tempo, na distribuição geográfica e nas características dos pedidos. Além disso, a aplicação do algoritmo K-Means possibilitou a identificação de diferentes perfis de pedidos, demonstrando que a base apresenta comportamentos distintos entre os grupos formados.

5 Conclusão

Este trabalho teve como objetivo utilizar técnicas de Ciência de Dados para extrair conhecimento a partir de dados de *marketplaces*, buscando compreender padrões de vendas e identificar perfis de pedidos que possam auxiliar a tomada de decisão no contexto do comércio eletrônico. Para isso, foi utilizada uma base de dados pública amplamente empregada em estudos sobre *e-commerce*, contendo informações reais de pedidos realizados em *marketplaces* brasileiro.

Inicialmente, foram realizadas análises exploratórias que permitiram compreender o comportamento geral das vendas ao longo do tempo, a distribuição geográfica dos pedidos, os meios de pagamento mais utilizados e as categorias de produtos com maior volume de vendas. Essas análises forneceram uma visão inicial do funcionamento do *marketplaces* e evidenciaram características importantes do cenário analisado.

Em seguida, foi conduzida uma análise de correlação entre os principais atributos da base de dados, possibilitando identificar relações lineares entre variáveis financeiras, logísticas e de avaliação dos clientes. Os resultados indicaram correlações fracas a moderadas entre os atributos analisados, sugerindo que o comportamento do consumidor em *marketplaces* é influenciado por múltiplos fatores e não apenas por aspectos monetários.

Como etapa central do trabalho, foi aplicada uma técnica de Aprendizado de Máquina do tipo não supervisionado, por meio do algoritmo de agrupamento *K-Means*. A partir da normalização dos dados e da definição adequada do número de *clusters*, foi possível identificar diferentes perfis de pedidos no *marketplaces*, evidenciando a heterogeneidade presente nas transações e confirmando a adequação da técnica de agrupamento para a extração de padrões a partir dos dados analisados.

5.1 Principais Contribuições

As principais contribuições deste trabalho estão relacionadas à aplicação prática de técnicas de Ciência de Dados em um contexto real de *marketplaces*. Destaca-se a construção de uma base de dados integrada a partir de diferentes tabelas, permitindo uma visão consolidada do processo de venda e do comportamento dos pedidos.

Outra contribuição relevante é a utilização de técnicas de análise exploratória e de correlação para compreender as relações entre os principais atributos do *e-commerce*, fornecendo subsídios para análises mais avançadas. Além disso, a aplicação do algoritmo de agrupamento possibilitou a identificação de perfis distintos de pedidos, evidenciando padrões de comportamento que não seriam facilmente percebidos por meio de análises

descritivas tradicionais.

Por fim, o trabalho demonstra como técnicas de Aprendizado de Máquina podem ser utilizadas como ferramentas de apoio à tomada de decisão em *marketplaces*, oferecendo *insights* que podem auxiliar vendedores a compreender melhor seus dados e a desenvolver estratégias comerciais mais direcionadas.

5.2 Trabalhos Futuros

Como trabalhos futuros, sugere-se a ampliação do conjunto de atributos analisados, incluindo informações adicionais relacionadas a prazos de entrega, categorias específicas de produtos ou desempenho individual de vendedores. A incorporação de dados mais recentes ou específicos de um único *marketplaces* também pode enriquecer as análises realizadas.

Além disso, podem ser exploradas outras técnicas de Aprendizado de Máquina, como mineração de regras de associação para identificação de padrões de compra entre produtos ou modelos preditivos para estimativa de vendas e demanda. A comparação entre diferentes algoritmos de agrupamento também representa uma possibilidade de extensão deste estudo, contribuindo para análises ainda mais aprofundadas no contexto do comércio eletrônico.

Referências

ALVARENGA, M. V. d. **Visualização de dados e sua importância para a tomada de decisão: uma aplicação usando o conjunto de dados de e-commerce da Olist**. 2021. Trabalho de Conclusão de Curso – Universidade Federal de Ouro Preto (UFOP). Disponível em: <<https://www.monografias.ufop.br/handle/35400000/2994>>. Acesso em: 29 jul. 2025. Citado 2 vezes nas páginas 9 e 13.

_____. Monografia (Bacharelado em Estatística), **Visualização de Dados e sua Importância para a Tomada de Decisão: Uma Aplicação Usando o Conjunto de Dados de E-Commerce da Olist**. Ouro Preto, MG: [s.n.], 2021. 41 p. Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP. Disponível em: <https://monografias.ufop.br/bitstream/35400000/2994/6/MONOGRAFIA_Visualiza%C3%A7%C3%A3oDadosImport%C3%A2ncia.pdf>. Citado na página 18.

BENZI, G. C. d. S. Monografia (MBA em Ciência de Dados), **Análise do comportamento de compra dos clientes no e-commerce brasileiro utilizando técnicas de mineração de dados**. São Carlos, SP: [s.n.], 2023. 63 p. MBA em Ciência de Dados, CEMEAI, ICMC-USP. Disponível em: <https://sites.icmc.usp.br/apneto/MBA/TCC_MBA_2023_Gabriella.pdf>. Citado 2 vezes nas páginas 9 e 18.

CONVERSION. **GMV do e-commerce brasileiro em 2023**. 2024. <<https://conversion.com.br/blog/marketplaces/>>. Acesso em: 20 jul. 2025. Citado na página 8.

E-Commerce Brasil. **Os desafios da logística no e-commerce brasileiro**. 2021. Acesso em: 24 fev. 2026. Disponível em: <<https://www.ecommercebrasil.com.br/artigos/desafios-na-logistica-para-atender-clientes-do-norte-e-nordeste>>. Citado na página 31.

_____. **Amazon investe em logística e impulsiona expansão no e-commerce brasileiro**. 2023. Notícia sobre expansão logística da Amazon no Brasil. Disponível em: <<https://www.ecommercebrasil.com.br/>>. Citado na página 8.

EBIT, N. **Webshoppers 49 – Relatório e-commerce Brasil 2023**. 2024. Relatório oficial da NIQ Ebit. Faturamento do e-commerce brasileiro atingiu R\$254,4bi em 2023, crescimento de 0,7% em relação a 2022. Citado na página 12.

Ebit|Nielsen. **Webshoppers 51ª edição – Panorama do E-commerce no Brasil**. 2024. <<https://www.ebit.com.br/webshoppers>>. Acesso em: jul. 2025. Citado na página 8.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996. Citado 2 vezes nas páginas 14 e 16.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. Burlington: Morgan Kaufmann, 2011. Citado 3 vezes nas páginas 14, 15 e 17.

- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**. New York: Springer, 2013. Citado 2 vezes nas páginas 15 e 27.
- LAUDON, K. C.; LAUDON, J. P. **Sistemas de Informação Gerenciais**. [S.l.]: Pearson Brasil, 2015. Citado na página 12.
- LAUDON, K. C.; TRAVER, C. **E-commerce: Business, Technology, Society**. 16. ed. [S.l.]: Pearson, 2020. Citado na página 13.
- LEITE, G. A. d. A.; PINOCHET, L. H. C.; PARDIM, V. I.; ONUSIC, L. M. Uso de machine learning para entendimento das relações de consumo no modelo de negócio (saas enabled marketplace) do olist. **Revista Tecnologia e Sociedade**, 2024. Citado na página 9.
- LIANG, D. Should competing sellers adopt the data analytics service offered by the online marketplace? **Electronic Commerce Research and Applications**, 2024. Available via ScienceDirect. Citado na página 12.
- MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley, CA: University of California Press, 1967. p. 281–297. Citado 2 vezes nas páginas 16 e 28.
- MARKETPLACER. **Introduction to Marketplace Analytics**. 2024. Available at: Marketplacer Support Center. Discusses key marketplace metrics such as sales, conversion rate, and product performance. Citado na página 13.
- MATA, K. B. d. C. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Computação), **E-commerce: análise de dados sobre o comércio eletrônico no Brasil**. Goiânia, GO: [s.n.], 2021. 53 p. Repositório Institucional PUC Goiás. Disponível em: <<https://repositorio.pucgoias.edu.br/jspui/bitstream/123456789/1761/1/E-commerce%20-%20An%C3%A1lise%20de%20Dados%20sobre%20o%20Com%C3%A9rcio%20Eletr%C3%B4nico%20no%20Brasil.pdf>>. Citado 2 vezes nas páginas 9 e 17.
- MONTEIRO, B. F. Trabalho de Graduação (Tecnologia em Gestão Comercial), **A evolução do comércio eletrônico no Brasil: impactos, desafios e oportunidades**. Marília, SP: [s.n.], 2025. 33 p. Trabalho de Graduação — FATEC Marília, 02 de junho de 2025. Disponível em: <https://ric.cps.sp.gov.br/bitstream/123456789/34578/1/gestao_comercial_2025_1_bruno_fabiano_monteiro_a_evolucao_do_comercio_eletronico_no_brasil.pdf>. Citado 2 vezes nas páginas 9 e 18.
- SIMÕES, N.; SENRA, K. B.; JACOMINO, G. P.; BRITTO, L. R. G. D. Predicting online e-marketplace sales performances: A big data approach. **International Journal of Quality & Reliability Management**, 2019. Citado na página 10.
- STATISTA. **E-commerce in Brazil - Statistics & Facts**. 2024. <<https://www.statista.com/topics/4697/e-commerce-in-brazil/>>. Acesso em: jul. 2025. Citado na página 8.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston: Pearson, 2006. Citado 3 vezes nas páginas 15, 16 e 17.