UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS TRINDADE
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS

Luis Eduardo Fortunato

Segmentação de clientes de um *e-commerce* brasileiro utilizando RFV e métodos de clusterização particionais

Florianópolis
2022

Luis Eduardo Fortunato

Segmentação de clientes de um *e-commerce* brasileiro utilizando RFV e métodos de clusterização particionais

Trabalho de Conclusão de Curso submetida ao Departamento de Engenharia de Produção e Sistemas da Universidade Federal de Santa Catarina para a obtenção do título de Grau de Engenheiro Mecânico com habilitação em Engenharia de Produção.
Orientador: Prof. Mauricio Uriona Maldonado, Dr.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Fortunato, Luís Eduardo
Segmentação de clientes de um e-commerce brasileiro utilizando RFV e métodos de clusterização particionais / Luís Eduardo Fortunato ; orientador, Mauricio Uriona Maldonado , 2022.
73 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2022.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. segmentação de clientes. 3. clusterização. I. , Mauricio Uriona Maldonado. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

Segmentação de clientes de um *e-commerce* brasileiro utilizando RFV e métodos de clusterização particionais

O presente trabalho em nível de Trabalho de Conclusão de Curso foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Lynceo Falavigna Braghirolli, Dr.
Universidade Federal de Santa Catarina

Prof. Maurício Uriona, Dr.
Universidade Federal de Santa Catarina

Profª. Mônica Maria Mendes Luna, Dra.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Grau de Engenheiro Mecânico com habilitação em Engenharia de Produção.

Documento assinado digitalmente
Rogerio Feroldi Miorando
Data: 27/07/2022 22:54:56-0300
CPF: 598.652.290-72
Verifique as assinaturas em https://v.ufsc.br

Prof. Rogério Feroldi Miorando, Dr.
Sub-coordenador do Programa

Documento assinado digitalmente
Mauricio Uriona Maldonado
Data: 27/07/2022 09:12:51-0300
CPF: 010.972.939-04
Verifique as assinaturas em https://v.ufsc.br

Prof. Mauricio Uriona Maldonado, Dr.
Orientador

Documento assinado digitalmente
Lynceo Falavigna Braghirolli
Data: 27/07/2022 10:03:32-0300
CPF: 806.687.200-34
Verifique as assinaturas em https://v.ufsc.br

Prof. Lynceo Falavigna Braghirolli, Dr.
Avaliador

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 27/07/2022 15:27:07-0300
CPF: 314.104.993-91
Verifique as assinaturas em https://v.ufsc.br

Profª. Mônica Maria Mendes Luna, Dra.
Avaliadora

Florianópolis, 12 de julho de 2022.

RESUMO

Este trabalho desenvolve uma estratégia de segmentação baseada no comportamento de compra de clientes de um e-commerce do Brasil. Através da análise de dados, este trabalho busca fornecer embasamento ao direcionamento de esforços comerciais específicos para determinados segmentos de cliente. A abordagem utilizada faz uso do conceito de frequência,recência e valor, comumente chamado de RFV, aliada a uma posterior clusterização dos dados. Após a clusterização, os agrupamentos gerados serão interpretados quanto às suas características de recência, frequência e valor para profundo entendimento acerca dos atributos de cada *cluster*. O fluxo de trabalho é baseado no conceito de epiciclo de análise de dados. Assim, a análise contempla as etapas de definição da pergunta de pesquisa, análise exploratória dos dados, construção de modelos formais, interpretação dos resultados e comunicação dos resultados. A partir da tabela de pedidos por cliente, gerou-se as métricas RFV para cada cliente único. As métricas foram então submetidas a um processo de clusterização que utilizou os dois principais métodos particionais: k-means e k-medoids. Os métodos particionais foram escolhidos devido à escalabilidade dos algoritmos aliada à interpretabilidade do resultado. Através do método do cotovelo, definiu-se cinco *clusters* como parâmetros aceitáveis para os modelos. Em ambos os métodos foi possível identificar grupos equivalentes de clientes: clientes ativos de baixo valor e baixa frequência, clientes inativos de baixo valor e baixa frequência, clientes recorrentes, clientes assíduos e clientes de alto valor.

Palavras-chave: segmentação de clientes. clusterização de dados. RFV. K-Means. K-medoids

ABSTRACT

This study develops a segmentation strategy based on the purchase behavior of customers in an e-commerce in Brazil. Through data analysis, this work seeks to provide a basis for targeting specific commercial efforts for specific customer segments. The approach used makes use of the concept of frequency, recency and monetary, commonly called RFM, combined with a subsequent data clustering. The generated clusters are interpreted in terms of their recency, frequency and monetary characteristics for a deep understanding of the attributes of each cluster. The workflow is based on the data analysis epicycle concept. Thus, the analysis includes the steps of defining the research question, exploratory data analysis, construction of formal models, interpretation of results and communication of results. The RFM metrics were generated for each unique customer. The metrics were then submitted to a clustering process that used the two main partitional methods: k-means and k-medoids. Partial methods were chosen due to the scalability of the algorithms combined with the interpretability of the result. Through the elbow method, five clusters were defined as acceptable parameters for the models. In both methods, it was possible to identify equivalent groups of customers: active low-value and low-frequency customers, inactive, low-value and low-frequency customers, recurring customers, regular customers and high-value customers.

Keywords: Customer segmentation. clustering. RFM. K-Means. K-medoids

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Por décadas as organizações focaram seus esforços na construção de marcas e produtos em detrimento do foco no consumidor. Contudo, quando as empresas falham em atender as expectativas do cliente é comum observar queda nos indicadores de retenção e fidelidade, o que tem impacto direto na receita do negócio. Em contrapartida, exceder as expectativas pode gerar custos operacionais maiores que os necessários e comprometer o retorno de investir na experiência do consumidor. A geração de valor é máxima quando a empresa entrega a experiência que está alinhada com a expectativa do cliente. Portanto, essa relação deve ser vista como uma decisão de investimento (MOSADDEGH *et al.*, 2021; STAHL; MATZLER; HINTERHUBER, 2003).

Dado que cada cliente possui diferentes necessidades e as empresas enfrentam restrições financeiras e operacionais, é inviável que cada consumidor seja tratado de maneira individualizada. Para contornar este problema, divide-se os clientes em grupos, chamados de segmentos de mercado e, então, direciona-se os esforços de *marketing* para os segmentos mais atraentes. Nesse caso, atratividade significa rentabilidade e sustentabilidade. A tarefa de segmentação separa os consumidores em grupos que são internamente homogêneos e heterogêneos em relação aos membros externos (BRITO *et al.*, 2015). O objetivo é garantir que cada cliente tenha a melhor experiência possível em cada interação com a marca, um esforço que se traduz na fidelidade do cliente e melhores resultados de negócios como aumento de receita e retenção de consumidores (HBR, 2021).

A prática de segmentação de clientes tornou-se recorrente nas organizações devido ao advento das novas tecnologias da era do *big data* e da computação na nuvem, que gerou uma grande redução nos custos de armazenamento de dados, e possibilitou o crescimento da aplicação de ferramentas e métodos capazes de processar grandes volumes de informação. Com isso, algoritmos e modelos estatísticos tornaram-se muito mais acessíveis e, desde então, a inteligência de dados virou um grande diferencial competitivo nas organizações modernas. No entanto, com a inundação de dados de fontes internas e externas, a pergunta é onde concentrar esforços para gerar valor ao negócio. Em estudo realizado pela McKinsey Global Institute (2011, p. 5) foram identificadas 5 amplos caminhos de gerar impacto a partir dos dados: (i) criação de transparência, (ii) permitir a experimentação para descobrir necessidades, expor a variabilidade e melhorar o desempenho, (iii) segmentar populações para personalizar ações, (iv) substituir e/ou apoiar a tomada de decisão humana com automação e algoritmos e (v) inovar modelos de negócios, produtos e serviços. De acordo com o estudo, essas iniciativas possuem potencial transformacional e tem implicações em como as organizações são desenhadas, organizadas e geridas (MANYIKA *et al.*, 2011).

Para desenvolver uma estratégia de negócio direcionada pelos dados, a maioria

das organizações planeja aumentar drasticamente o uso de recursos tecnológicos, desde a adoção de ferramentas de *business intelligence* (BI) até a aplicação de técnicas de *Machine Learning* e inteligência artificial. Uma pesquisa da Harvard Business Review Analytic Services com 185 executivos globais em uma ampla gama de setores revelou que a grande maioria (73%) acredita que seus recursos de dados são chave para a criação de valor de negócio, incluindo 33% que dão a classificação mais alta possível: 10 de 10 em termos de importância. O uso da inteligência dos dados permite que as organizações conheçam fundamentalmente mais sobre seus negócios e transformem esse conhecimento em uma melhor tomada de decisão (HBR, 2019). A administração eficaz dos dados nunca foi tão imperativo para as empresas. No entanto, de acordo com uma pesquisa mundial da Gartner (2008), 91% das 196 organizações entrevistadas confirmaram que ainda não atingiram níveis desejáveis de maturidade analítica (GARTNER, 2008). Neste sentido, com a crescente relevância dos dados para para a tomada de decisão empresarial, fica evidente o tamanho da oportunidade de desenvolvimento de trabalhos nesta área.

1.1 DESCRIÇÃO DO PROBLEMA

A segmentação de mercado tem sido considerada um dos conceitos mais fundamentais do *marketing* moderno (WIND, 1978). Este conceito introduz a ideia de que a demanda de determinado mercado é heterogênea e, portanto, as empresas podem aumentar sua lucratividade ao calibrar estratégias comerciais que considerem seu mercado consumidor como grupos cujos elementos possuem características similares. Apesar do conceito ter sido introduzido há mais de 6 décadas, nos últimos anos as estratégias de segmentação ganharam destaque com o advento de ferramentas que permitem o armazenamento e processamento de grandes quantidades de dados. A utilização dessas ferramentas possibilita não só a descoberta e compreensão do comportamento de consumo de cada grupo, mas também fornece meios para que as empresas desenvolvam estratégias que impactam cada segmento.

Este trabalho desenvolve uma estratégia de segmentação baseada no comportamento de compra de clientes de *e-commerce* do Brasil. A abordagem utilizada faz uso do conceito de frequência, recência e valor, comumente chamado de RFV, aliada a uma posterior clusterização dos dados. O RFV é bastante popular na pesquisa de *marketing* pois é econômico na aquisição de importantes informações acerca do comportamento do cliente, é muito valioso para prever a resposta aos estímulos de *marketing* e, por isso, pode aumentar os lucros de uma empresa em curto prazo, possui alto poder de sumarização pois resume longos conjuntos de dados em apenas 3 atributos e, por último, o método facilmente identifica consumidores valiosos para as empresas (JO-TING; SHIH-YEN; HSIN-HUNG, 2010). Apesar de fornecer intuições importantes, apenas o cálculo das métricas RFV não permite a identificação de seg-

mentos relevantes. Este problema é contornado através da utilização de métodos de aprendizado de máquina não supervisionados. Especificamente, este estudo utiliza os métodos de clusterização particionais *k-means* e *k-medoids* para realizar a tarefa de agrupamento dos dados.

O conjunto de dados utilizados nesta pesquisa foi disponibilizado pela Olist através do Kaggle, uma plataforma de aprendizado e prática de *Data Science*. Os arquivos disponibilizados contém informações reais de 100 mil pedidos realizados entre 2016 a 2018 em vários *marketplaces* do Brasil. De maneira geral, é possível encontrar informações de pedidos, pagamentos, clientes e produtos (OLIST; SIONEK, 2018).

A Olist é uma empresa brasileira fundada em 2015 que atua no segmento de *e-commerce*. Dentre as soluções ofertadas está a Olist Store, de acordo com o próprio *site* da empresa o Olist Store é uma solução de vendas que une tecnologia de ponta com inteligência de mercado para aumentar o faturamento de lojistas nos *marketplaces*. Em resumo, os clientes da Olist cadastram produtos na plataforma que encarrega-se de anunciá-los nos mais diversos *marketplaces*. Além disso, a empresa fornece soluções de gestão de vendas para que seja possível o gerenciamento das operações sem a necessidade de acessar cada um dos *marketplaces* individualmente.

Diante da evolução do comércio eletrônico e a consequente entrada massiva de empresas no setor, mais do que nunca é necessário oferecer serviços que atendam às expectativas do cliente. Neste ambiente de alta concorrência, cada vez mais as empresas buscam atingir públicos mais específicos como forma de criar diferenciais competitivos. Para os consumidores impactados por comunicações mais personalizadas, o valor percebido da empresa aumenta, o que reforça a preferência por determinado produto ou serviço. Com isso, gradativamente as empresas optam por desenhar soluções direcionadas a segmentos de clientes em detrimento de soluções padronizadas (ZENONE, 2007). Isto permite não apenas adquirir novos clientes, mas reter os clientes mais valiosos. Como aponta Kotler (1980), atrair clientes é uma tarefa importante, mas manter os clientes é uma tarefa ainda mais importante tendo em vista que perder um cliente significa perder todo fluxo de compras que aquela pessoa faria. No entanto, o desenvolvimento de estratégias focalizadas demanda o entendimento profundo do perfil do cliente. Dito isso, o produto deste trabalho busca fornecer intuições que auxiliem a resolução de potenciais problemas enfrentados pela Olist como baixa fidelidade de consumidores, perda de clientes de alto valor e altos níveis de insatisfação por exemplo. Baseado nesses fatores define-se a pergunta da análise como: Quais os melhores clientes da base de dados da empresa Olist baseado nas métricas de recência, frequência e valor?.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Identificar os principais segmentos de clientes de um empresa de *e-commerce* baseado em métricas de recência, frequência e valor.

1.2.2 Objetivos Específicos

Os objetivos específicos almejam:

a) Construir o modelo de clusterização baseado em métricas de recência, frequência e valor;

b) Selecionar a quantidade de *clusters* através do método do cotovelo.

c) Identificar as características comuns entre os membros de cada *cluster* formado;

d) Identificar as características que difere cada *cluster* dos demais.

1.3 JUSTIFICATIVA

Através da análise de dados, este trabalho busca fornecer embasamento para a tomada de decisão comercial de direcionamento de esforços para grupos de clientes mais atrativos. Apesar da clareza quanto aos objetivos, é necessário traduzir o problema de negócio em uma pergunta de análise. Através do estabelecimento da pergunta será possível nortear o desenvolvimento da pesquisa para que os objetivos estabelecidos sejam alcançados. Para isto, determina-se que tipo de pergunta será realizada para que se tenha uma boa compreensão de quais tipos de conclusões podem ser obtidas. Este trabalho busca identificar as características presentes nos distintos segmentos de clientes de uma base de dados. Dessa forma, a pergunta deverá buscar por padrões, tendências ou relações entre variáveis, portanto, caracteriza-se como uma pesquisa exploratória. Outro aspecto a considerar durante o desenvolvimento da questão é a operacionalização da pergunta em uma análise de dados. Adiantar como seriam as eventuais conclusões pode evitar que tempo seja desperdiçado em um processo cujo resultado não é interpretável. Neste trabalho, as conclusões devem apontar não somente quais são os principais clientes da empresa, mas também quais as características que tornam esses clientes importantes. Este aspecto é imprescindível pois diante da entrada de novos clientes é necessário dispor de um procedimento ou conjunto de regras que permita classificar também esses novos entrantes.

Sabe-se que comumente a identificação de segmentos de clientes é realizada através de *clusterização*. Esse tipo de abordagem busca identificar grupos de objetos com características semelhantes. A importância do cliente é traduzida em métricas

através da abordagem de RFV, este procedimento cria *features* baseadas nas interações comerciais do cliente com a empresa. Os clientes são pontuados de acordo com a recência, frequência e valor das suas compras. Naturalmente, os clientes com interações mais recentes, mais frequentes e de maior valor são apontados como clientes mais valiosos.

O presente estudo é composto por seis capítulos que estruturam o desenvolvimento de uma análise de dados aplicada à segmentação de clientes de um *e-commerce* brasileiro. O capítulo inicial tem como intuito fornecer o contexto da análise, descrever o problema de pesquisa e justificar a relevância do tópico abordado. No capítulo dois é exposto a fundamentação teórica necessária para compreensão dos desenvolvimentos desta pesquisa. Inicialmente, é introduzido o conceito de segmentação de mercado onde é apresentada a ideia de heterogeneidade dos mercado, noção esta que embasa toda a prática da segmentação de clientes. Em seguida, aborda-se os métodos utilizados durante o processo de análise. No terceiro capítulo é apresentado os procedimentos metodológicos adotados. Nesta seção define-se o enquadramento metodológico, delimita-se os materiais e métodos e por fim, os procedimento metodológicos são detalhados. Os capítulo quatro e cinco relatam o processo de análise. No capítulo quatro é realizada uma análise exploratória dos dados, nesta seção busca-se entender se há algum problema com o conjunto de dados utilizado e se a pergunta proposta pode ser respondida com os recursos disponíveis. No quinto capítulo, a modelagem das métricas RFV é realizada e a clusterização dos dados é aplicada. Por fim, no sexto capítulo as conclusões são apresentadas.

2 REFERENCIAL TEÓRICO

2.1 SEGMENTAÇÃO DE MERCADO

O conceito de segmentação de mercado foi inicialmente formalizado por Smith (SMITH, 1956). O autor afirma que a segmentação é representa um ajuste racional e mais preciso do produto e do esforço de *marketing* aos requisitos do consumidor ou usuário. Através da segmentação de mercado, as empresas dividem mercados grandes e heterogêneos em um número de mercados menores e homogêneos que podem mais facilmente serem atingidos. Kotler (1980, p. 213) define que os compradores em qualquer mercado diferem em seus desejos, recursos, localizações, atitudes de compra e práticas de compra. As empresas capitalizam a heterogeneidade das necessidades identificando grupos de clientes que possuem conjuntos de necessidades relativamente homogêneos. Apesar de extensões consideráveis, exploração de variáveis e uma abundância de pesquisas sobre o tema, a essência da teoria da segmentação por quase 50 anos tem sido a noção de heterogeneidade das necessidades do cliente (BLOCKER; FLINT, 2007). De acordo com Veloso (2008, p.37) o conceito de segmentação pode ser resumido em três princípios básicos:

a) É um processo de agregação de consumidores baseado em determinadas características.

b) O agrupamento de consumidores permite identificar grupos que possuem necessidades e desejos similares.

c) Por possuírem características similares, os grupos responderão de maneira similar aos estímulos de marketing.

Os resultados da segmentação podem ser utilizados para embasar um amplo conjunto de decisões de negócios, Tynan e Drayton (1978, p. 346, apud Lunn) apontam que a segmentação de mercado é normalmente empregada para solucionar os seguintes problemas:

a) Definir um mercado: A segmentação auxiliará na compreensão do ambiente a partir da perspectiva do consumidor. Por exemplo, produtos que são vistos como concorrentes pela perspectiva do fabricante podem não ser considerados desta forma do ponto de vista do cliente.

b) Justificar princípios de ação para marcas e produtos: Direciona estratégias para aumentar a retenção de clientes, converter compradores de outras marcas ou atrair um novo grupo de compradores para os produtos e/ou serviços ofertados.

c) Posicionar as marcas e os produtos: Diante dos inúmeros segmentos de mercado existentes, as empresas precisam direcionar seus esforços para os grupos que oferecem maior potencial de retorno.

d) Identificar lacunas no mercado: A segmentação de mercado permite também identificar grupos de consumidores cujas necessidades não estão sendo atendidas. Essas necessidades podem ser satisfeitas a partir do lançamento de novos produtos ou pela modificação de um produto ou serviço existente.

O modelo de segmentação requer a seleção de uma base de segmentação (variável dependente) e variáveis descritoras (variáveis independentes). Além disso, as variáveis podem ser divididas em dois grupos: características gerais do cliente (dados demográficos e características socioeconômicas, por exemplo) e características situacionais (padrão de compra e utilização do produto, por exemplo) (WIND, 1978). Para Kotler (1980, p.213), as principais bases de segmentação são:

a) Geográfica: Divide o mercado em diferentes unidades geográficas, como nações, regiões, estados, condados, cidades ou até mesmo bairros

b) Demográfica: Divide o mercado em segmentos com base em variáveis como idade, estágio do ciclo de vida, gênero, renda, ocupação, escolaridade, religião, etnia e geração.

c) Psicográfica: Divide os compradores em diferentes segmentos com base no estilo de vida ou características de personalidade.

d) Comportamental: Divide os compradores em segmentos com base em seus conhecimentos, atitudes, usos ou respostas a um produto. Muitos profissionais de marketing acreditam que as variáveis de comportamento são o melhor ponto de partida para construir segmentos de mercado.

Contudo, Kotler (1980, p.217) adiciona que raramente os profissionais de *marketing* limitam suas análises a apenas uma base de segmentação. Em vez disso, é comum a utilização de algumas bases de segmentação em um esforço para identificar grupos-alvo menores e mais bem definidos. Quanto aos procedimentos de análise, Wind (1978, p.321) identifica quatro abordagens de modelagem:

a) Segmentação a *priori*: Nos modelos de segmentação *a priori* a escolha da variável dependente (base da segmentação) é realizada antes da coleta e análise dos dados, como por exemplo a marca preferida do cliente. Os clientes são então classificados em segmentos de marcas favoritas e são examinados em relação a suas diferenças em outras características, como dados demográficos, faixa etária e escolaridade por exemplo (variáveis independentes) (GREEN, 1977).

b) Segmentação baseada em clusterização (*post hoc*): Na segmentação baseada em clusterização, a variável dependente é selecionada após a coleta e análise dos dados. Frequentemente, o procedimento de agrupamento é precedido por uma análise fatorial para reduzir o conjunto original de variáveis independentes. As variáveis são agrupadas de acordo com sua correlação

entre si e a quantidade de variância que podem explicar na variável dependente (TYNAN; DRAYTON, 1987).

c) Segmentação flexível: A segmentação flexível difere das abordagens *a priori* e baseadas em *clusters*, pois baseia-se na integração dos resultados de uma análise conjunta e uma simulação computacional do comportamento de escolha do consumidor. Como vantagem, oferece a flexibilidade de "construir" segmentos com base na resposta dos consumidores a ofertas de produtos alternativos (sob várias condições competitivas e condições ambientais). Além disso, a partir da seleção de um segmento, a administração tem informações fáceis (modo interativo) sobre o tamanho estimado do segmento e suas características discriminatórias (WIND, 1978). Tynan e Drayton (1987, p. 305) acrescentam que essa abordagem permite que a administração desenvolva e examine um grande número de segmentos alternativos, cada um composto por um grupo de consumidores que apresentam uma resposta semelhante a novos produtos "teste".

d) Segmentação componencial: Ao contrário das abordagens *a priori* e *post hoc*, na segmentação componencial as pessoas que apresentam perfis de resposta semelhantes não são agrupadas. Pelo contrário, suas respostas individuais são ainda decompostas em contribuições de componentes relacionadas à pessoa (demografia, estilos de vida e assim por diante) e contribuições relacionadas à oferta (como níveis de atributo do produto ou serviço) (GREEN, 1977). Ainda, de acordo com Tynan e Drayton (1987, p. 306), o modelo de segmentação componencial oferece uma nova conceituação para segmentação de mercado porque oferece tanto uma análise do segmento de mercado para uma oferta de produto específica, quanto uma avaliação da oferta ou posicionamento de produto mais desejável.

Claramente, é possível segmentar um mercado de diversas maneiras, no entanto algumas formas não fornecem nenhuma informação útil. Por isto, é necessário garantir que os agrupamentos gerados possuam significado mercadológico. Neste contexto, Kotler (1980) pontua que segmentos úteis devem possuir as seguintes características:

a) Mensurabilidade: Um segmento deve ser fácil de medir para determinar seu tamanho, poder de compra e perfil.

b) Acessibilidade: Os segmentos podem ser efetivamente impactados e servidos Substancialidade: Os segmentos devem ser grandes e lucrativos o suficiente para serem servidos

c) Diferenciabilidade: Os segmentos são conceitualmente distinguíveis e respondem de forma diferente a diferentes elementos e programas do *mix* de

marketing. Acionabilidade: A organização deve possuir recursos o suficiente para ser capaz de impactar o segmento.

A segmentação de mercado é um instrumento poderoso para direcionar decisões de negócio. Os resultados da análise permitem que as empresas melhorem o posicionamento de marcas e produtos perante o consumidor. Como exposto neste capítulo, existem diversas abordagens de segmentação de mercado, na prática as empresas testam diferentes métodos em busca das melhores oportunidades de mercado.

2.1.1 RFV

Para atingir melhores níveis de retenção de clientes, uma empresa precisa alocar eficientemente seus recursos de marketing. Nesse sentido, o método de RFV foi introduzido por Hughes (1994) como ferramenta para avaliar o valor de determinado consumidor para uma empresa. Este modelo é usado em dados transacionais com base em um período de análise e tem como objetivo identificar os clientes mais importantes de uma base de dados através dos seguintes atributos:

a) Recência: Tempo decorrido desde a última interação

b) Frequência: Número de interações em um determinado intervalo de tempo

c) Valor: Valor monetário envolvido nas transações no período selecionado

Os clientes que compram recentemente, com frequência e gastam grandes quantias de dinheiro são mais propensos a responder às campanhas de marketing (MONALISA; NADYA; NOVITA, 2019). Apesar de simples, um estudo realizado por Verhoef (2003) com 228 empresas apontou que aproximadamente 40% delas utiliza o método como forma de segmentar sua base de clientes. O processo para quantificar o comportamento do cliente por meio do modelo RFV é realizado da seguinte forma:

a) Para cada atributo de recência, frequência e valor, divida a lista de clientes em cinco percentis

b) Nos percentis superiores, é atribuída nota 1. Ao segmento adjacente ao superior, é atribuída nota 2 e assim sucessivamente.

c) A aplicação do método resultará em 125 agrupamentos distintos de clientes (5 possíveis notas para cada um dos 3 atributos).

Desde que os dados estejam disponíveis, a aplicação do método é bastante simples. Por meio dos atributos criados pelo modelo, os profissionais de *marketing* poderão direcionar de maneira mais eficaz as estratégias de retenção de cliente, o que leva ao aumento da taxa de resposta dos clientes (KAHAN, 1998). Apesar das vantagens do modelo levantadas anteriormente, há também algumas desvantagens. Visto que o RFV visa identificar clientes valiosos nas empresas, o foco está nos melhores clientes. Desta forma, fornece uma pontuação pouco significativa em recência, frequência e valor, quando a maioria dos clientes não compra com frequência, gastou pouco e não

comprou recentemente. Além disso, o modelo irá gerar 125 grupos distintos de clientes, o que inviabiliza a análise aprofundada de cada segmento. (WEI; LIN; WU, 2010).

2.2 CLUSTERIZAÇÃO

Historicamente, a noção de encontrar padrões úteis em dados recebeu vários nomes. O termo *Data Mining* (Mineração de dados) tem sido usado principalmente por estatísticos, analistas de dados e comunidades de sistemas de informação de gestão (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Para Han & Kamber (2011, p.8), *Data Mining* é definido como

> O processo de descoberta de padrões e conhecimentos interessantes de grandes quantidades de dados. As fontes de dados podem incluir bancos de dados, *data warehouses*, a internet, outros repositórios de informações ou dados que são transmitidos dinamicamente para um sistema.

Já para Nisbet, Elder & Miner (2009, p.17), *Data Mining* pode ser entendido como:

> O uso de algoritmos de aprendizado de máquina para encontrar padrões tênues de relacionamento entre elementos de dados em grandes conjuntos de dados, ruidosos e confusos, o que pode levar a ações em ordem de obter algum benefício de alguma forma (diagnóstico, lucro, detecção, etc.)

Ainda, de acordo com Fayyad, Piatetsky-Shapiro & Smyth (1996, p.83) podemos considerar um padrão como conhecimento se ultrapassar algum limiar de interesse, o que de forma alguma é uma tentativa de definir o conhecimento na visão filosófica ou mesmo popular. A maioria das técnicas de *Data Mining* desenvolveram-se em um campo conhecido hoje como Aprendizado de Máquina (*Machine Learning*) (WITTEN; FRANK, 2002). O Aprendizado de Máquina pode ser classificado em (i) aprendizado supervisionado, (ii) aprendizado não supervisionado e (iii) aprendizado por reforço. Os métodos de aprendizagem supervisionada tem como objetivo aprender uma boa aproximação do verdadeiro mapeamento do vetor de entrada para o vetor de saídas, usando informações contidas em um conjunto de dados de exemplos. Os métodos contidos nesse tipo de aprendizagem, inferem a função a partir de dados rotulados (BERTOLINI *et al.*, 2021). Para “dados rotulados” entende-se um conjunto de dados em que cada exemplo é marcado com a resposta que o algoritmo deve apresentar por conta própria. Ainda, os métodos de aprendizagem supervisionada podem ser classificados em (i) métodos de classificação e (ii) métodos de regressão. Quando a variável resposta é contínua, trata-se de um problema de regressão, e quando a variável resposta é categórica trata-se de um problema de classificação (IZBICKI; SANTOS, 2018). Contudo, há também os problemas em que a variável resposta não está rotulada. Este tipo de problema é chamado de aprendizagem não supervisionada. Um algoritmo de

aprendizagem não supervisionada funciona por conta própria para descobrir padrões ou *clusters* ocultos (agrupamento de dados semelhantes). Os algoritmos de aprendizagem não supervisionada podem ainda ser classificados em: estimativa de densidade e redução de dimensionalidade e agrupamento (clusterização) (MOUNT; ZUMEL, 2019). A estimativa de densidade é um amplo conjunto de técnicas que podem ser usadas para descobrir propriedades úteis (por exemplo, assimetria ou multimodalidade) ou mesmo para gerar uma estimativa de uma função de densidade de probabilidade subjacente não observável, de um conjunto de dados de dados observados (BERTOLINI *et al.*, 2021). A redução de dimensão, por sua vez, é frequentemente usada em *clusterização*, classificação e muitos outros aplicativos de aprendizado de máquina e mineração de dados. Geralmente retém as dimensões (atributos) mais importantes, remove as dimensões ruidosas (atributos irrelevantes) e reduz o custo computacional (DING *et al.*, 2002).

Clusterização refere-se a um conjunto muito amplo de técnicas para encontrar subgrupos, ou *clusters*, em um conjunto de dados. Quando agrupamos as observações de um conjunto de dados, procuramos dividi-las em grupos distintos, de modo que as observações dentro de cada grupo sejam bastante semelhantes entre si, enquanto as observações em grupos diferentes seja bastante diferentes umas das outras. O problema da clusterização é bastante conhecido na literatura de análise de dados por sua vasta aplicação em questões como segmentação de clientes, classificação e análise de tendências. Este é um problema não supervisionado porque estamos tentando descobrir a estrutura - neste caso *clusters* distintos - com base em um conjunto de dados (GARETH *et al.*, 2013). A análise de *cluster* é amplamente utilizada em aplicações como inteligência de negócios, reconhecimento de padrões de imagem, pesquisa na Web, biologia e segurança. Em *business intelligence*, a clusterização pode ser usada para organizar um grande número de clientes em grupos, onde os clientes dentro de um grupo compartilham fortes características semelhantes. Isso facilita o desenvolvimento de estratégias de negócios para melhorar o atendimento ao cliente (HAN; KAMBER, 2006)). De acordo com Xu & Wunsch (2005) o procedimento padrão para análises de clusterização segue as seguintes etapas:

a) Seleção das *features* do modelo: nesta etapa são escolhidas quais features serão consideradas no modelo.

b) Construção do modelo de *cluster*: Os padrões são agrupados de acordo com as features escolhidas.

c) Validação do modelo: Padrões e critérios de avaliação eficazes devem fornecer um grau de confiança nos resultados alcançados.

d) Interpretação dos resultados: O objetivo final do clustering é fornecer *insights* significativos sobre os dados originais.

Ainda, de acordo com os autores a análise de *clusters* não é um processo linear, mas um processo iterativo que não possui critérios universais para orientar a seleção de *features* e quantidade de *clusters*. Além disso, salientam que uma análise robusta passa pela escolha de um algoritmo adequado. Este dever atende aos seguintes requisitos:

a) Escalabilidade:

b) Capacidade de lidar com variáveis de diferentes tipos de variáveis:

c) Capacidade de descobrir *clusters* de formato arbitrário

d) Independência quanto à necessidade de *input* do usuário para determinação de parâmetros

e) Capacidade de lidar com dados ruidosos

f) Possibilidade de receber atualizações incrementais de dados

g) Capacidade de lidar com conjunto de dados de muitas *features*

h) Capacidade de lidar com restrições

i) Interpretabilidade e usabilidade

No entanto, não existe um algoritmo único que possa satisfazer plenamente todos os requisitos colocados. É importante entender as características de cada algoritmo para que o algoritmo adequado possa ser selecionado para o problema em questão (ZAIANE *et al.*, 2002).

2.2.1 MÉTODOS DE CLUSTERIZAÇÃO

De acordo com Han & Kamber (2011, p.443) métodos de clusterização podem ser categorizados como métodos de particionamento, métodos hierárquicos, métodos baseados em densidade e métodos baseados em grades. No entanto, os autores reforçam que essas categorias podem se sobrepor, de modo que um método pode ter características de várias categorias.

Os métodos de particionamento minimizam um determinado critério de agrupamento iterativamente até que uma partição localmente ótima seja encontrada. Em um algoritmo iterativo, como *K-means*, a convergência é local e a solução globalmente ótima não pode ser garantida. No entanto, como o número de observações é sempre finito e, portanto, também o número de partições distintas é finito, o mínimo global pode ser obtido através de métodos de força bruta. Assim, estes métodos são inviáveis para grandes conjuntos de dados uma vez que o número de partições possíveis a serem analisadas é muito alto. Dessa forma, certas heurísticas são utilizadas na forma de otimização iterativa (BLASHFIELD; ALDENDERFER, 1978).

Já os métodos hierárquicos classificam o conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre cada observação. Geralmente, os

resultados são apresentados como uma árvore binária ou dendograma. Existem dois tipos de métodos hierárquicos: aglomerativos e divisivos. Inicialmente nos métodos aglomerativos cada observação é considerada um único *cluster* e a cada iteração dois *clusters* são unidos. Nos métodos divisivos inicialmente tem-se um único *cluster* e a cada iteração um *cluster* é dividido (KAUFMAN; ROUSSEEUW, 2009). Cada técnica constrói sua hierarquia na direção oposta, possivelmente resultados bastante distintos serão alcançados.

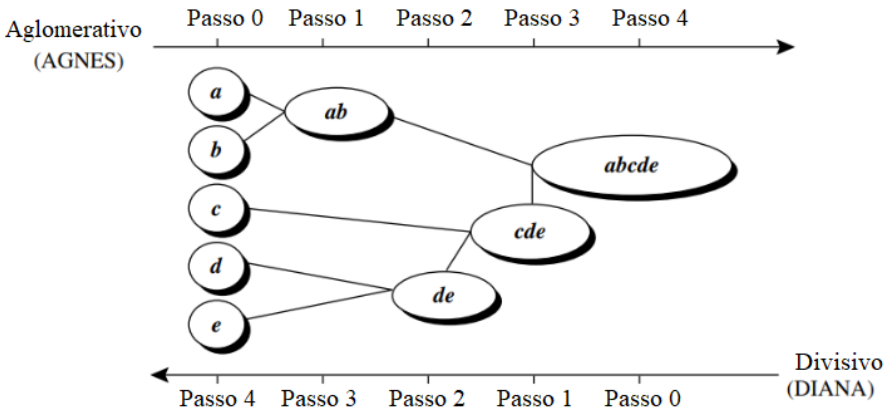
Tanto os métodos de particionamento quanto os métodos hierárquicos são incapazes de lidar com *clusters* de formatos arbitrários. Em contrapartida, os métodos baseados em densidade identificam os *clusters* como áreas de alta densidade de observações separados por áreas de baixa densidade. Dessa forma, os *clusters* podem assumir formatos arbitrários. Além disso, esses métodos não exigem o número de grupos como parâmetros de entrada, nem fazem suposições sobre a distribuição dos dados ou a variação dentro dos *clusters* (KRIEGEL *et al.*, 2011)., 2011). Desta forma, métodos baseados em densidade podem encontrar *clusters* de formato arbitrário e diferentes densidades.

Por fim, os métodos baseados em grades discretizam o espaço em um número finito de elementos que formam uma estrutura de grade na qual todas as operações de agrupamento são realizadas. Essa abordagem é capaz de encontrar *clusters* de formato arbitrário, possui baixa sensibilidade à *outliers* e é eficiente e escalável (XU; TIAN, 2015). O tempo de processamento é rápido pois não depende do número de observações, mas sim do número de células em cada dimensão do espaço (HAN; KAMBER, 2006).

2.2.1.1 Clusterização hierárquica

Um método de *clusterização* hierárquica funciona agrupando observações em uma hierarquia ou “árvore” de *clusters*. A representação de objetos de dados na forma de uma hierarquia é mais informativa do que o simples conjunto de *clusters* retornados pelos métodos particionais.As abordagens de clusterização hierárquica podem ser categorizadas em aglomerativas ou divisivas. A Figura 1 é um exemplo da aplicação de ambas abordagens em um conjunto de dados $\{a, b, c, d, e\}$. Comumente, a abordagem aglomerativa é chamada de AGNES (Agglomerative Nesting) e a abordagem divisiva DIANA (Divisive Analysis)

Figura 1 – Exemplo de aplicação AGNES/DIANA



Fonte: (HAN; KAMBER, 2006)

Inicialmente, o método aglomerativo coloca cada objeto em um *cluster* próprio. Os *clusters* são então unidos passo a passo de acordo com algum critério. Por exemplo, os *clusters* C_1 e C_2 podem ser mesclados se um objeto em C_1 e um objeto em C_2 formam a distância mínima entre quaisquer dois objetos de diferentes *clusters* (HAN; KAMBER, 2006). Para tanto, é necessário definir uma medida da distância entre os *clusters*. Abaixo são apresentadas quatro medidas de distância amplamente utilizadas. Onde, $p - p_i$ denota a distância entre dois pontos, m_i é a média do cluster C_i e n_i é o número de observações em C_i (Figura 2).

Figura 2 – Medidas de distância

Distância Mínima	$dist_{min}(C_i, C_j) = \min_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{\lvert \boldsymbol{p} - \boldsymbol{p'} \rvert\}$
Distância Máxima	$dist_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{\lvert \boldsymbol{p} - \boldsymbol{p'} \rvert\}$
Distância das médias	$dist_{mean}(C_i, C_j) = \lvert \boldsymbol{m_i} - \boldsymbol{m_j} \rvert$
Distância média	$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \lvert \boldsymbol{p} - \boldsymbol{p'} \rvert$

Fonte:(HAN; KAMBER, 2006)

Os algoritmos que utilizam a distância mínima para medir a distância entre *clusters* são normalmente chamados de algoritmo do "vizinho mais próximo"(*nearest-*

neighbor). Além disso, se o algoritmo for encerrado quando a distância entre os *clusters* mais próximos exceder um limite definido pelo usuário, será chamado de algoritmo de ligação única (*single linkage*). Quando o algoritmo utiliza a distância máxima, é comumente chamado de "algoritmo de clusterização do vizinho mais distante"(*farthest-neighbor clustering algorithm*). Desta vez, se o processo for encerrado quando a distância máxima entre os *clusters* mais próximos exceder um limite definido pelo usuário, será chamado de algoritmo de ligação completa. Os métodos de distância mínima e distância máxima representam extremos na forma de medir a distância entre *clusters*, ambos são sensitivos à *outliers* e dados ruidosos. A utilização de métricas intermediárias é uma abordagem alternativa para contornar esse problema (HAN; KAMBER, 2006, pp. 462)

De maneira geral, as principais vantagens da utilização de um algoritmo hierárquico é sua facilidade de interpretação dos resultados, comumente apresentada em forma de dendograma, e sua facilidade de implementação. Contudo, os métodos hierárquicos podem encontrar dificuldades em relação à seleção de pontos de junção ou divisão. Uma vez que um *cluster* de observações é mesclado ou dividido, o processo na próxima etapa irá operar nos *clusters* recém-gerados e a operação anterior não será desfeita. Além disso, a maioria dos algoritmos hierárquicos não escalam para grandes conjuntos de dados, a porção de tempo necessária para rodar um modelo cresce de forma quadrática em relação ao número de observações.

2.2.1.2 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) é um algoritmo de clusterização que utiliza a noção de densidade para encontrar agrupamentos nos dados. Esse entendimento permite a descoberta de *clusters* de formatos arbitrários, uma limitação dos métodos particionais e hierárquicos. A principal razão pela qual reconhecemos os *clusters* é que dentro de cada *cluster* temos uma densidade de pontos que é consideravelmente maior do que fora do *cluster*. Além disso, a densidade nas áreas de ruído é menor do que a densidade em qualquer um dos *clusters*. A ideia chave é que para cada ponto de um *cluster*, a vizinhança de um determinado raio deve conter pelo menos um número mínimo de pontos, ou seja, a densidade na vizinhança deve exceder algum limite (ESTER *et al.*, 1996). O método possui dois parâmetros, o raio da vizinhança ϵ e o número mínimo de pontos na vizinhança para definir um *cluster* $MinPts$. De acordo com Han & Kamber (2011, p.472) a partir desses parâmetros, o algoritmo classifica os pontos em:

a) *core point* p: O número de observações na vizinhança de raio ϵ do ponto p é maior ou igual MinPts

b) *border point*: O número de observações na vizinhança de raio ϵ do ponto q é menor que $MinPts$ e a vizinhança contém pelo menos um *core point*.

c) Ruído: Corresponde a uma observação que não é enquadrada como *border point* ou *core point*

Além disso, a definição do *cluster* é fundamentada nos conceito de

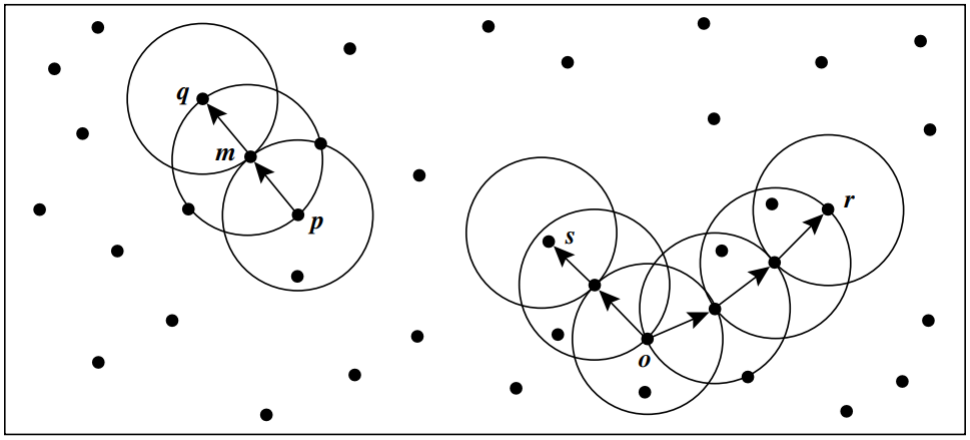
a) Alcance direto por densidade: Um objeto$p$ é diretamente alcançável por densidade do objeto q, com respeito à ϵ e a $MinPts$, se p está na ϵ vizinhança de q onde q é um ponto central.

b) Alcance por densidade: Um objeto p é alcançável por densidade do objeto q com respeito à e $MinPts$ em um conjunto D, se existe uma cadeia de objetos $p_1, ..., p_n$, tais que p_1= q e p_n=p e $p_i + 1$ é alcançável por densidade diretamente de p_i com respeito a e $MinPts$, para $1 \leq i \leq n$, p_i em D.

c) Conexão por densidade: Um objeto p é conectado por densidade ao objeto q com respeito à e $MinPts$ em um conjunto de objetos D, se existe um objeto o em D tal que ambos p e q são alcançáveis por densidade do objeto o.

A Figura 3 apresenta um exemplo dos conceitos de conectividade e alcance. O objeto q é indiretamente alcançável pela densidade de p porque q é diretamente alcançável pela densidade de m e m é diretamente alcançável pela densidade de p. No entanto,p não é alcançável em densidade a partir de q porque q não é um objeto central. Da mesma forma, r e s são alcançáveis pela densidade de o e o é alcançável por densidade a partir de r. Assim, o, r e s são todos conectados por densidade.

Figura 3 – Conectividade e alcance



Fonte(HAN; KAMBER, 2006)

Baseado nesses conceitos, define-se um *cluster* C com respeito a ϵ e $MinPts$ como um subconjunto não vazio de D que satifaz (HAN; KAMBER, 2006, pp. 472):

a) Para quaisquer dois objetos o_1 e $o_2 \in C$, o_1 e o_2 são conectados por densidade.

b) Não existe um objeto $o \in C$ e outro objeto $o' \in (D\text{-}C$), tais que o e o' são conectados por densidade

O DBSCAN tem muitos méritos em relação a outros métodos de *clusterização* como os métodos particionais ou hierárquicos. Neste algoritmo não é necessário especificar o número de *clusters a priori* como no algoritmo *k-means*. Além disso, o DBSCAN possibilita a descoberta de *clusters* de formatos arbitrários e lida de maneira robusta com *outliers*. Em contrapartida, os parâmetros ϵ e $MinPts$ devem ser especificados pelo usuário. Tais parâmetros são normalmente definidos de maneira empírica e possuem grande impacto no resultado entregue pelo algoritmo.

2.2.1.3 K-Means

De acordo com Han & Kamber (2011, p.451) a versão mais simples e fundamental da análise de *clusters* é o particionamento, que organiza os objetos de um conjunto em vários grupos ou *clusters* exclusivos. Para manter a especificação do problema concisa, podemos assumir que o número de *clusters* é dado como conhecimento prévio. Este parâmetro é o ponto de partida para métodos de particionamento.

Dado um conjunto de dados $D = x_{i=i}^{n}$, no algoritmo *k-means* os dados são particionados em um determinado número de *clusters*. Para cada *cluster* C_i existe um ponto z_i que o representa e é denominado centróide. O centróide é definido como a média dos elementos do *cluster*, onde n_i é o número de elementos em cada *cluster*. Necessariamente, cada observação pertence a pelo menos um *cluster* e as observações não pertencem a mais de um *cluster* simultaneamente. Para avaliar a qualidade de cada *cluster* é normalmente utilizado a soma dos erros quadrados dado por: $SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in c_i} (D_{x_j}, z_i)$. Este algoritmo começa por distribuir aleatoriamente os pontos do conjunto do conjunto D em k *clusters* e calcula os centróides através da média dos pontos do *cluster* C_i. Na atribuição de *clusters*, cada ponto $x_j \in D$ é associado ao *cluster* que tem o centróide z_i mais próximo do ponto, isto é, x_j é atribuído ao *cluster* C_j quando $j^* = argmin||x_j - z_i||$. Em cada iteração as observações são atribuídas movendo-as para o *cluster* mais próximo. Novos centróides são recalculados. O processo continua até que todas as observações estejam situadas no *cluster* mais próximo.

De forma geral, a implementação do algoritmo *k-means* é relativamente simples e escala para grandes conjuntos de dados. Especificamente em segmentação de clientes, a literatura científica é bastante rica em aplicações deste algoritmo para identificação de grupos de interesse. (ANITHA; PATIL, 2019; BALAKRISHNAN *et al.*, 1996; CHRISTY *et al.*, 2021; KHAJVAND *et al.*, 2011; RAHIM *et al.*, 2021; SHIN; SOHN, 2004). No entanto, normalmente as iterações são finalizadas em um ponto de ótimo local e o ponto ótimo global não é atingido. Outra desvantagem dessa abordagem, é a necessidade de determinar *a priori* o número de *clusters*. Comumente, diferentes modelos são construídos e o número de *clusters* k é escolhido a partir do modelo com melhores resultados.

2.2.1.4 K-Medoids

A estrutura simples e flexível do *K-means* torna possível a construção de algoritmos modificados sobre ele. Algumas das variações propostas estão baseadas em: (i) escolher diferentes critérios para representação dos *clusters* (*K-medoids*, *K-medians*, *K-modes*), (ii) escolher melhores estimativas para os centroides iniciais (*Intelligent K-means*, *Genetic K-means*) e (iii) aplicar algum tipo de técnica para transformação das *features* (*Weighted K-means*, *Kernel Kimeans*) (AGGARWAL *et al.*, 1999).

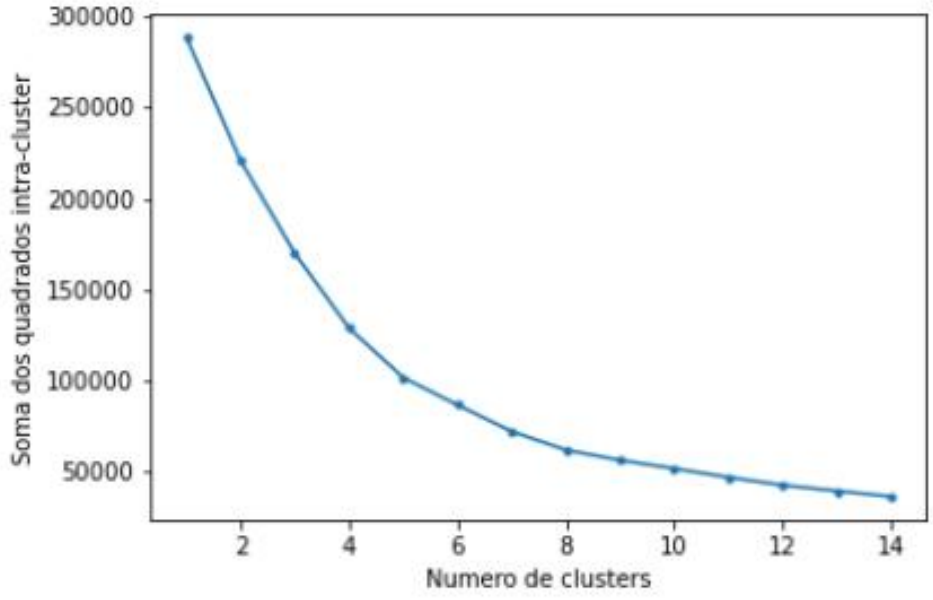
De acordo com Han & Kamber (2011, p.454) o algoritmo *k-means* é sensível a *outliers* porque tais objetos estão longe da maioria dos dados e, portanto, quando atribuídos a um cluster, podem distorcer drasticamente o valor médio do *cluster*. Nesse sentido, o algoritmo *k-medoids* é uma alternativa mais robusta à presença de *outliers*, onde em vez de tomar o valor médio dos objetos de um *cluster* como ponto de referência, escolhe-se objetos reais para representar os *clusters*. Cada objeto restante é atribuído ao *cluster* do qual o objeto representativo é o mais semelhante. Após encontrar um conjunto de *k* objetos representativos, os *k clusters* são construídos pela atribuição de cada objeto do conjunto de dados ao objeto representativo mais próximo (KAUFMAN; ROUSSEEUW, 2009, pp. 68). O particionamento é então realizado minimizando a soma das diferenças entre cada objeto e seu correspondente objeto representativo. Este critério é chamado de erro absoluto definido por: $E = \sum_{i=1}^{k} \sum_{p \in c_i} dist(p, o_i)$. Ainda de acordo com Han & Kamber (2011, p.454) apesar de ser mais robusto que o *k-means* na presença de ruído, a complexidade de cada iteração no algoritmo *k-medoids* torna custoso a adoção deste método em situações onde o conjunto de dados é muito grande e o número de *clusters* procurados é alto. Para contornar este problema, os autores sugerem a possibilidade de treinar os modelos em uma amostra dos dados. Em vez de utilizar todo conjunto de dados, retira-se uma amostra aleatória da população para construir o modelo. Idealmente, a amostra deve representar o conjunto original de dados originais de forma fidedigna uma vez que os objetos representativos de cada *cluster* (medoides) serão escolhidos a partir da amostra.

2.2.1.5 Método do cotovelo

Determinar a quantidade de grupo é uma questão central na análise de *clusters*, não apenas porque alguns algoritmos como *k-means* e *k-medoids* necessitam desse parâmetro de entrada, mas também porque a escolha do número apropriado de *clusters* controla a granularidade da análise (HAN; KAMBER, 2006, pp. 486). Essa escolha representa um *trade-off* entre compressibilidade e acurácia: quanto menor o número de *clusters* maior a sumarização dos dados e menor o nível de informação entre pela análise. Por outro lado, um grande número de *clusters* maximiza a definição de cada grupo enquanto minimiza a sumarização dos dados. Nesse sentido, o método

do cotovelo fornece uma heurística para determinar um valor apropriado de *clusters* através da redução da soma da variancia dentro intra-*cluster*. Este procedimento é baseado na ideia de que aumentar o número de *clusters* permite capturar grupos com objetos mais semelhantes entre si mas o efeito marginal de reduzir a soma das variâncias dentro do *cluster* perde intensidade a medida que mais *clusters* são formados. Assim, a determinação do parâmetro apropriado é realizada através da aplicação de diversos modelos com diferentes valores de *k*. Em cada rodada, calcula-se a distância média quadrada de todos os pontos dentro de um *cluster* para o centróide do *cluster*. Inicialmente, deve-se encontrar a distância euclideana entre todos os pontos de um *cluster* e seu centroide, divide-se o total pelo número de pontos do *cluster* e então calcula-se a média entre todos os *cluster*. Comumente, este procedimento é disponibilizado por padrão nas bibliotecas de implementação de algoritmos de aprendizado de máquina. Os resultados gerados são utilizados para construção de um gráfico onde é possível observar o comportamento da métrica de variância intra-*cluster* em relação ao número de *clusters*. O nome método do cotovelo deriva do formato da curva gerada, normalmente o valor apropriado de *clusters* é aquele representado pelo ponto de inflexão da curva. A Figura 4 apresenta um exemplo da curva gerada pelo método. Como pode ser observado, entre 4 e 6 *clusters* é possível notar uma queda na diminuição marginal da métrica de dispersão e, portanto, esses pontos representam as melhores escolhas de acordo com o procedimento. Contudo, a análise de *clusters* possui caráter exploratório, a escolha final deve estar pautada sobretudo na interpretabilidade dos resultados gerados.

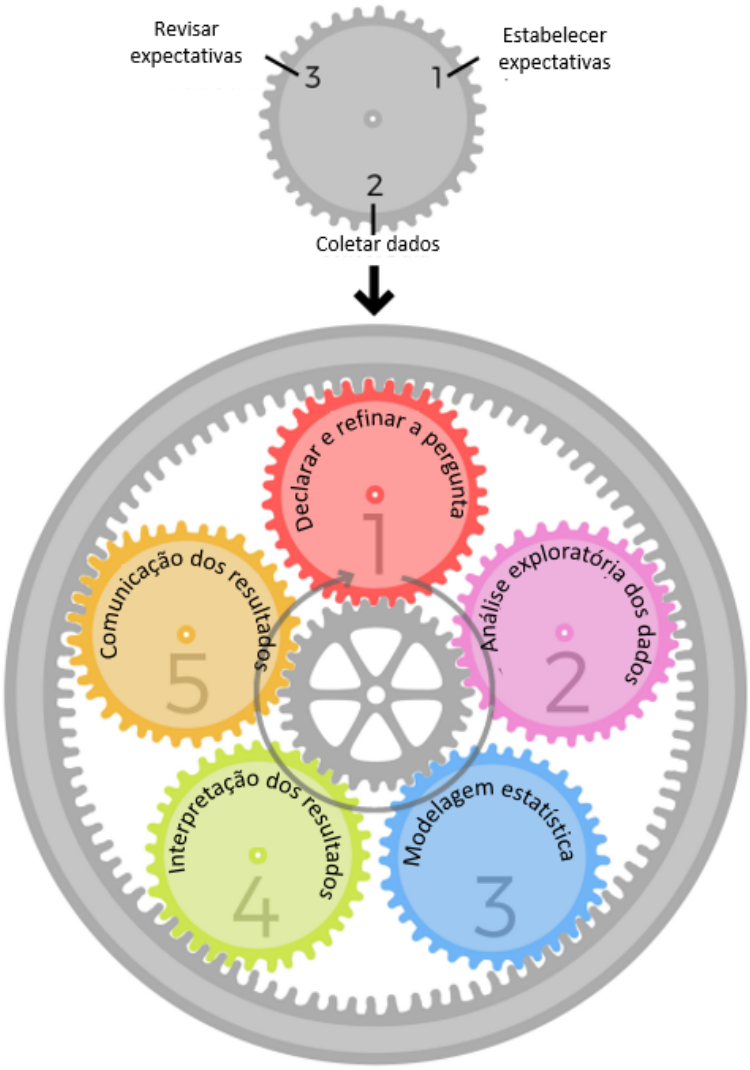
Figura 4 – Exemplo de curva gerada pelo método do cotovelo



2.3 EPICICLO DA ANÁLISE DE DADOS

Este trabalho baseia-se no epiciclo de análise de dados proposto por Peng e Matsui (2015). Para os autores, o processo de análise de dados é análogo ao processo de composição de uma música. Por mais que o compositor tenha conhecimento da teoria musical, isto não garante a qualidade da composição. Em algum momento do processo, é necessário uma faísca de criatividade . Caso contrário, computadores também seriam capazes de escrever boas músicas. No ensaio "*Computer Programming as an Art*" de 1974, Donald Knuth escreve que "a ciência é um conhecimento que entendemos tão bem que podemos ensiná-lo a um computador. Todo resto é arte.". Assim como na música, o processo de análise de dados também é uma arte. O conhecimento da teoria e das ferramentas não garante a qualidade da análise. O analista de dados necessita de uma faísca de criatividade para construir algo que de fato responda uma pergunta relevante. No livro "A arte da ciência de dados", Peng e Matsui destacam que para o leigo a análise de dados pode parecer um processo linear, passo a passo, que no final gera os resultados desejados. No entanto, o processo é altamente iterativo e não linear. A cada etapa novos conhecimentos são adquiridos que informam se é necessário refinar ou refazer a etapa anterior ou se é possível seguir para a próxima etapa. O que configura o que os autores chamam de epiciclo. Um epiciclo é um pequeno círculo cujo centro se move ao redor da circunferência de um círculo maior (Figura 5).

Figura 5 – Epiciclo da análise de dados



Fonte: (PENG; MATSUI, 2015)

2.3.1 Declarar e refinar a pergunta

As perguntas podem ser descritivas, exploratórias, inferenciais, preditivas, causais ou mecanicistas. Abaixo é apresentado um breve resumo sobre cada categoria.

a) Descritivas: As perguntas descritivas buscam resumir uma característica de um conjunto de dados

b) Exploratórias: Os dados devem ser analisados para checar se existem padrões, tendências ou relações entre as variáveis.

c) Inferenciais: Fazem proposições sobre o conjunto de dados

d) Preditivas: Busca identificar resultados futuros prováveis baseado em uma amostra de dados

e) Causais: Procura mudança de um fator mudará outro fator, em média, em uma população

f) Mecanicistas: As questões mecanicistas buscam explicar como a relação entre os fatores se estabelece

Além disso, os autores apontam cinco boas práticas que uma boa pergunta deve contemplar. Inicialmente, a pergunta deve ser de interesse de uma determinada audiência envolvida no contexto da análise, isto é, a pergunta deve apresentar relevância suficiente para que valha a pena ser respondida. Também, é necessário verificar se a pergunta estabelecida já não foi respondida. Neste momento, pesquisas e discussões com especialistas são necessárias para aprofundamento da questão estudada. Mesmo que não haja resposta específica para a questão de interesse da análise, o embasamento teórico adquirido ajuda a refinar a pergunta proposta. A terceira boa prática para construção de uma pergunta é garantir a plausibilidade da questão. Isto é, a pergunta deve ser coerente com o contexto da análise. Caso a pergunta não seja plausível, os resultados serão provavelmente de difícil interpretação e baixa confiabilidade. A pergunta deve também ser passível de ser respondida dada as restrições da análise. Por exemplo, é bastante plausível que existam defeitos no funcionamento de certas células do cérebro que causam autismo, mas não é possível realizar biópsias cerebrais para coletar células vivas para estudar, o que seria necessário para responder a essa pergunta (PENG; MATSUI, 2015). Por último, é importante também especificidade na construção da pergunta. Tornar a questão mais específica refinará sua pergunta e informará diretamente quais etapas devem ser seguidas quando começar a analisar os dados.

2.3.2 Análise exploratória dos dados

A análise exploratória de dados tem como objetivo aprofundar os conhecimentos do pesquisador acerca dos dados estudados. Neste momento, busca-se entender se há algum problema com o conjunto de dados utilizado e se a pergunta proposta pode ser respondida com os recursos disponíveis. Tipicamente, esta etapa inclui examinar a estrutura dos componentes do conjunto de dados, as distribuições das variáveis e as relações entre dois ou mais fatores. Os gráficos são os artifícios mais poderosos na análise exploratória, pois permite sumarizar grandes volumes de dados de uma forma intuitiva e de fácil absorção para o leitor. É possível também utilizar esta etapa para realizar um rascunho do que virá a ser a resposta final da pergunta proposta. Ao final da análise exploratória, o pesquisador deve ser capaz de responder às seguintes perguntas:

a) Você possui os dados corretos? Uma das conclusões possíveis para a análise exploratória é que os dados utilizados não serão capazes de responder a pergunta colocada.

b) Você precisa de outros dados? Embora os dados disponíveis pareçam apropriados para responder à pergunta do problema, talvez seja necessário novas *features* ou maior volume de dados para garantir a solidez do resultado.

c) Você fez a pergunta correta? Através da análise dos dados o pesquisador pode descobrir que a pergunta inicial não estava tão clara, ou que talvez os dados disponíveis sejam capazes de responder perguntas mais relevantes dado o contexto do problema.

2.3.3 Construção da modelagem estatística

A modelagem estatística é uma representação simplificada da realidade que tem como objetivo fornecer um sumário quantitativo dos dados e impor uma estrutura específica à população de onde os dados foram amostrados. Em essência, um modelo estatístico fornece uma descrição de como o mundo funciona e como os dados foram gerados. Em segmentação de clientes, a modelagem estatística é utilizada para identificar indivíduos com características similares. Como cada consumidor possui diferentes necessidades, é inviável que as empresas tratem cada um dos indivíduos de maneira personalizada. Neste tipo de problema, é comum a adoção de modelos de clusterização para identificação dos padrões nos dados. Em suma, os modelos são usados para fornecer redução de dados e fornecer algumas intuições sobre a população estudada. É importante salientar que a modelagem estatística é um processo altamente iterativo que deve seguir as três etapas que dão a dinâmica do epiciclo de dados: estabelecimento das expectativas, coleta de informações e revisão das expectativas. No entanto, é necessário estabelecer o momento de parar as iterações. Peng e Matsui recomendam algumas reflexões para auxiliar essa decisão:

a) Você já esgotou os dados? A análise iterativa dos dados levantará dúvidas que não poderão ser respondidas com os dados em mãos. Talvez seja necessário coletar mais dados para concluir se o padrão encontrado de fato se sustenta ou é apenas acaso estatístico. Outra situação que levanta a necessidade de coletar mais dados é quando a análise foi concluída e resultados satisfatórios foram alcançados, neste momento é muito importante replicar as análises em outro conjunto de dados.

b) Você tem evidência suficiente para responder a pergunta? O objetivo da análise de dados é sustentar um processo decisório, é importante identificar o momento em que os resultados da análise são suficientemente satisfatórios para embasar uma tomada de decisão.

c) Os resultados fazem sentido? É necessário ser altamente crítico quanto aos resultados alcançados. É comum chegar a resultados que não correspondem às expectativas, neste momento é importante certificar que todas as etapas

foram feitas corretamente e que não há problemas com o conjunto de dados. Após garantir que não há erros na análise, o modelo deve ser testado em outro conjunto de dados e os mesmos resultados devem ser atingidos. Uma nova descoberta só é realizada com muito escrutínio.

d) Você ainda tem tempo? Este critério é bastante subjetivo mas representa um grande fator na decisão de avançar na análise. Toda análise possui uma restrição de recursos disponíveis para alocar no processo. É importante reconhecer o momento de parar a iteração da análise de dados devido aos limites de tempo e orçamento.

2.3.4 Interpretação dos resultados

Apesar do procedimento metodológico definir uma etapa exclusiva para interpretação dos resultados, os autores reiteram que a interpretação acontece de maneira contínua durante a análise. Há quatro princípios básicos que devem ser contemplados durante a interpretação dos resultados. Inicialmente, deve-se revisitar a pergunta original para não perder de foco a questão central da análise. É comum que os pesquisadores façam descobertas acidentais durante o processo de exploração ou modelagem estatística que desvia o rumo da análise. Neste momento, é importante garantir que os resultados encontrados de fato podem ser utilizados para embasar uma resposta. O segundo princípio estabelece que deve-se começar um único modelo e focar na direção, magnitude e grau de incerteza do resultado. Com base nessas informações é possível construir uma intuição sobre a natureza dos resultados. Em seguida, os autores recomendam avaliar os resultados encontrados em um contexto de informação externa. As informações externas são conhecimentos gerais que o pesquisador ou membros do time possuem acerca do assunto estudado. Como todo processo de análise de dados é realizado para responder uma pergunta, por último deve-se apontar as implicações dos resultados alcançados. Muitas vezes as implicações não são óbvias e será necessário bastante esforço para atravessar essa etapa.

2.3.5 Comunicação dos resultados

A comunicação é tanto uma ferramenta quanto o produto final da análise de dados. Não há sentido em realizar uma análise se os resultados não forem compartilhados com uma audiência interessada. Também, através da resposta da audiência é possível receber *feedbacks* que indicarão os próximos passos do processo. Em suma, o principal propósito da comunicação é coletar dados. Para um bom planejamento da rotina de comunicação é importante conhecer e selecionar o público apropriado para o tipo de *feedback* que se busca, ser direto e conciso mas fornecer informação suficiente

para o entendimento da audiência, utilizar comunicação clara e simples que evita o uso de jargões e adotar uma postura aberta e colaborativa.

3 METODOLOGIA

3.1 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos adotados no trabalho baseiam-se no epiciclo de análise de dados de Peng & Matsui (2015). Para os autores, a análise de dados não é um fluxo linear com começo, meio e fim mas sim, um processo iterativo composto por 5 etapas: declarar e refinar a pergunta, análise exploratória, modelagem formal, interpretação dos resultados e por fim, comunicação dos resultado. Em cada uma das etapas citadas, é crítico que o pesquisador continuamente defina expectativas, colete dados e revise expectativas.

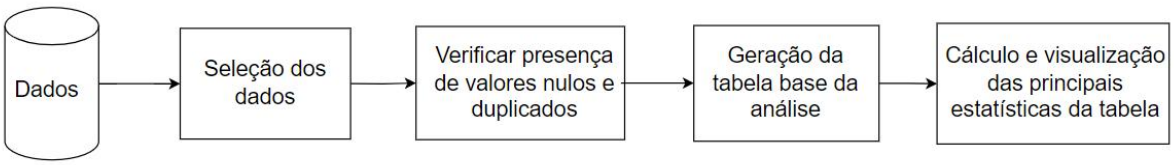
3.1.1 Declarar e refinar a pergunta

Nesta etapa apresenta-se o contexto do problema de pesquisa, justifica-se a relevância do assunto e define-se uma pergunta para nortear a análise de dados. Neste trabalho, o contexto do problema é apresentado na seção introdutória onde é exposto a motivação por trás deste desenvolvimento. Em seguida, na seção 1.1 realiza-se a descrição do problema, a apresentação da empresa cujos dados são objeto de estudo, o valor gerado pela análise e por fim, a pergunta de pesquisa.

3.1.2 Análise Exploratória

Durante a análise exploratória almeja-se detectar eventuais problemas com o conjunto de dados e desenvolver expectativas quanto ao resultado da análise. Esta etapa é realizada durante o capítulo 4 deste trabalho. De forma resumida, os procedimentos adotados na análise exploratória são apresentados na Figura 6:

Figura 6 – Procedimentos da análise exploratória

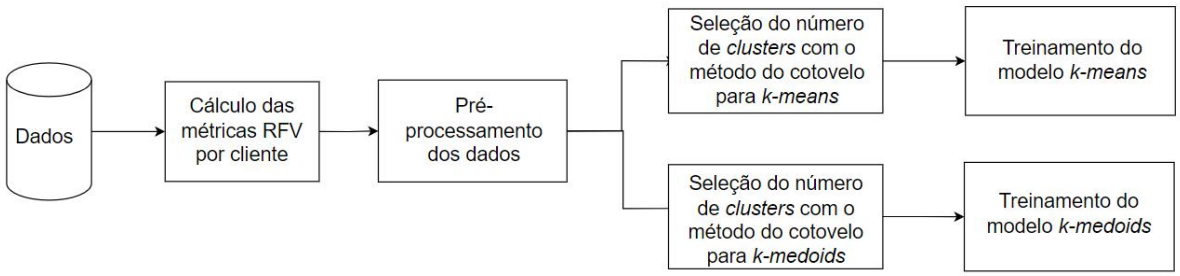


3.1.3 Modelagem de dados

Na seção referente à modelagem de dados calcula-se as métricas RFV por cliente, seleciona-se uma quantidade razoável de *clusters* e treina-se os modelos. A modelagem RFV é descrita na seção 5.0.1. Com as métricas RFV calculadas, realiza-se o pré-processamento dos dados descrito na seção 5.0.2. Por fim, na seção 5.0.3 são apresentados os resultados obtidos com os modelos de clusterização. Na Figura

7 são apresentados os procedimentos adotados nesta etapa do epiciclo de análise de dados.

Figura 7 – Procedimentos da modelagem de dados



3.1.4 Interpretação e comunicação dos resultados

Neste trabalho o processo de interpretação e comunicação ocorre majoritariamente durante cada tópico do processo de análise. A interpretação dos resultados obtidos pela clusterização dos dados é realizada durante as seções 5.0.4 e 5.0.5. Nessas seções, as métricas RFV referentes à cada *cluster* são analisadas e os resultados são comparados entre os modelos. Quanto à comunicação dos resultados, entende-se que o próprio trabalho derivado deste processo de análise de dados materializa-se como a etapa de comunicação.

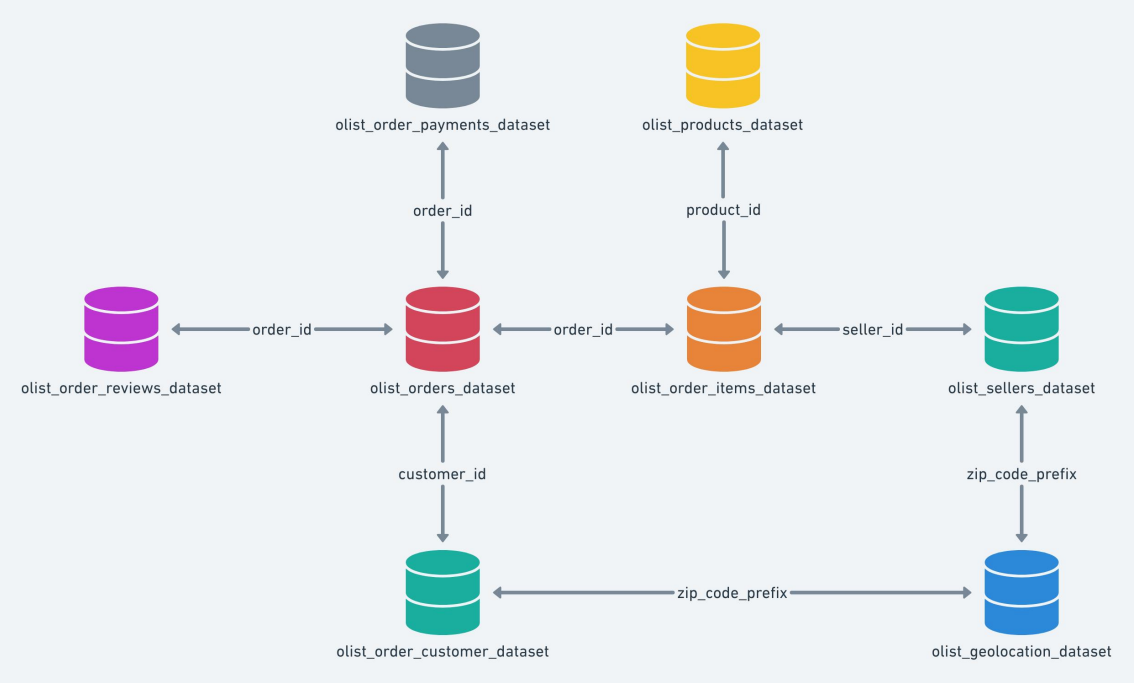
3.2 MATERIAIS

A amostra de dados utilizadas neste trabalho foi disponibilizada pela empresa Olist através da plataforma de aprendizado em *Data Science* chamada Kaggle. Os dados contemplam informações de pedidos, produtos e clientes. As informações contidas em cada uma das tabelas é mostrado no Quadro 1. Na Figura 8 é apresentado um diagrama que resume a forma como as tabelas se relacionam.

Quadro 1 – Resumo do conteúdo de cada tabela do conjunto de dados

Tabela	Descrição
olist_customers_dataset	Este conjunto de dados contém informações sobre o cliente. Cada pedido é atribuído a um ID de cliente exclusivo. Isso significa que o mesmo cliente receberá IDs diferentes para pedidos diferentes. O objetivo de ter um customer_unique_id no conjunto de dados é permitir que você identifique os clientes que fizeram recompras na loja.
olist_geolocation_dataset	Este conjunto de dados possui informações de CEPs brasileiros
olist_order_items_dataset	Este conjunto de dados inclui dados sobre os itens comprados em cada pedido
olist_order_payments_dataset	Este conjunto de dados inclui dados sobre as opções de pagamento de pedidos
olist_order_reviews_dataset	Este conjunto de dados inclui dados sobre as avaliações feitas pelos clientes
olist_orders_dataset	Este é o conjunto de dados principal. A partir de cada pedido é possível encontrar todas as outras informações.
olist_products_dataset	Este conjunto de dados inclui dados sobre os produtos vendidos pela Olist
olist_sellers_dataset	Esse conjunto de dados inclui dados sobre os vendedores que atenderam aos pedidos feitos na Olist
product_category_name_translation	Traduz o productcategoryname para o inglês.

Figura 8 – Esquema de dados



Fonte: (OLIST; SIONEK, 2018)

3.2.1 Ferramentas

Este estudo utilizará a linguagem de programação Python como ferramenta de manipulação de dados. A grande vantagem em utilizar Python deriva principalmente do grande e ativo ecossistema de pacotes e bibliotecas para análise de dados. Neste trabalho, os pacotes utilizados serão:

a) NumPy: é o pacote fundamental para computação científica em Python. É uma biblioteca que fornece uma variedade de rotinas para operações rápidas em matrizes, incluindo matemática, lógica, manipulação de formas, classificação, seleção, E/S , transformadas discretas de *Fourier*, álgebra linear básica, operações estatísticas básicas, simulação aleatória e muito mais (NUMPY, 2022).

b) Pandas: O pandas fornece estruturas e funções de dados ricas projetadas para tornar o trabalho com dados estruturados rápido, fácil e expressivo (MCKINNEY, 2012).

c) Matplotlib: é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python (MATPLOTLIB, 2022).

d) Scikitlearn: é uma biblioteca de aprendizado de máquina de código aberto que oferece suporte ao aprendizado supervisionado e não supervisionado. Ele também fornece várias ferramentas para ajuste de modelos,

pré-processamento de dados, seleção de modelos, avaliação de modelos e muitas outras utilidades (SCIKITLEARN, 2022)

4 ANÁLISE EXPLORATÓRIA

De acordo com Peng & Metsui (2015, p.31), a análise exploratória tem 3 principais objetivos: determinar se existe algum problema com o conjunto de dados do estudo, determinar se a resposta pode ser obtida com os dados disponíveis e desenvolver um rascunho da solução final. Dito isto, inicialmente procedeu-se com a leitura dos dados e descoberta da estrutura da tabela. São três principais tabelas de análise neste trabalho: *olist_orders_dataset*, *olist_payments_dataset* e *olist_customers_dataset* (Quadros 2, 3 e 4). A primeira contempla os registros de pedidos por cliente e a respectiva situação do pedido, a segunda tabela, por sua vez, contém as informações de pagamento dos pedidos como forma de pagamento, número de parcelas e valor total do pedido, a terceira tabela contém os registros de identificação de cada cliente.

Quadro 2 – Colunas da tabela *olist_orders_dataset*

Coluna	Descrição
order_id	Identificador único do pedido
customer_id	Identificador único do cliente.
order_status	Referência ao status do pedido (entregue, enviado, etc).
order_purchase_timestamp	Data e hora de realização do pedido
order_approved_at	Data e hora de aprovação do pagamento.
order_delivered_carrier_date	Hora e data da entrega do pedido ao parceiro logístico.
order_delivered_customer_date	Data que o pedido foi entregue ao cliente.
order_estimated_delivery_date	Data prevista para entrega do pedido.

Fonte: (OLIST; SIONEK, 2018)

Quadro 3 – Colunas da tabela *olist_order_payments_dataset*

Coluna	Descrição
order_id	Identificador único do pedido
payment_sequential	Caso o cliente utilizar mais de um método de pagamento, uma sequência será criada para acomodar todos os pagamentos.
payment_type	Método de pagamento escolhido
payment_installments	Número de parcelas do pedido.
payment_value	Valor do pedido.

Fonte: (OLIST; SIONEK, 2018)

Quadro 4 – Colunas da tabela *olist_customers_dataset*

Coluna	Descrição
customer_id	Chave para o conjunto de dados de pedidos. Cada pedido tem um *customer_id* exclusivo.
customer_unique_id	Identificador único do cliente.
customer_zip_code_prefix	Código postal do cliente.
customer_city	Cidade de origem do cliente.
customer_state	Estado de origem do cliente.

Fonte: (OLIST; SIONEK, 2018)

Como este trabalho é focado na avaliação da base de clientes, nem todos os campos são relevantes para análise. Desta forma, as colunas foram classificadas em 3 grupos quanto ao seu valor analítico (Quadros 5, 6 e 7).

a) Alta relevância: O valor dessas colunas é essencial para a análise.

b) Baixa relevância: A ausência dessas colunas não compromete a realização da análise. No entanto, sua inclusão pode oferecer intuições quanto à resposta buscada.

c) Indiferente: Não tem relevância dado o contexto do problema.

Quadro 5 – Tabela *olist_orders_dataset*

Coluna	Descrição
order_id	Alta relevância
customer_id	Alta relevância
order_status	Baixa relevância
order_purchase_timestamp	Alta relevância
order_approved_at	Indiferente
order_delivered_carrier_date	Indiferente
order_delivered_customer_date	Indiferente
order_estimated_delivery_date	Indiferente

Quadro 6 – Tabela *olist_order_payments_dataset*

Coluna	Descrição
order_id	Alta relevância
payment_sequential	Indiferente
payment_type	Baixa relevância
payment_installments	Baixa relevância
payment_value	Alta relevância

Quadro 7 – Colunas da tabela *olist_customers_dataset*

Coluna	Descrição
customer_id	Alta relevância
customer_unique_id	Alta relevância
customer_zip_code_prefix	Baixa relevância
customer_city	Baixa relevância
customer_state	Baixa relevância

O prosseguimento desta análise irá considerar apenas as colunas que possuem alta relevância para análise. As colunas categorizadas como de alta relevância são aquelas imprescindíveis para a geração das métricas RFV e, portanto, a ausência dessas colunas inviabilizaria o estudo. Já as colunas de baixa relevância são aquelas cujos campos podem complementar a análise. No entanto, como este trabalho é focado em segmentação de clientes baseado em métricas RFV, as colunas de baixa relevância analítica são descartadas. Contudo, a inclusão desses atributos em análises futuras tem o potencial de enriquecer os resultados obtidos. Por fim, as colunas categorizadas como indiferentes não possuem valor analítico para esta análise. Desta forma, 5 colunas foram eliminadas da primeira tabela, 3 colunas eliminadas da segunda e 2 colunas eliminadas terceira tabela. Dentre as 7 colunas restantes no total, uma coluna é numérica (*payment_value*) e as demais categóricas. Como um dos intuitos da análise exploratória é verificar a existência de algum problema no conjunto de dados, inicialmente foi realizada uma contagem de linhas nulas na amostra para atestar a qualidade do dado (Tabelas 1,2 e 3).

Tabela 1 – Valores nulos na tabela *olist_orders_dataset*

Coluna	Descrição
order_id	0
customer_id	0
order_purchase_timestamp	0

Tabela 2 – Valores nulos na tabela *olist_order_payments_dataset*

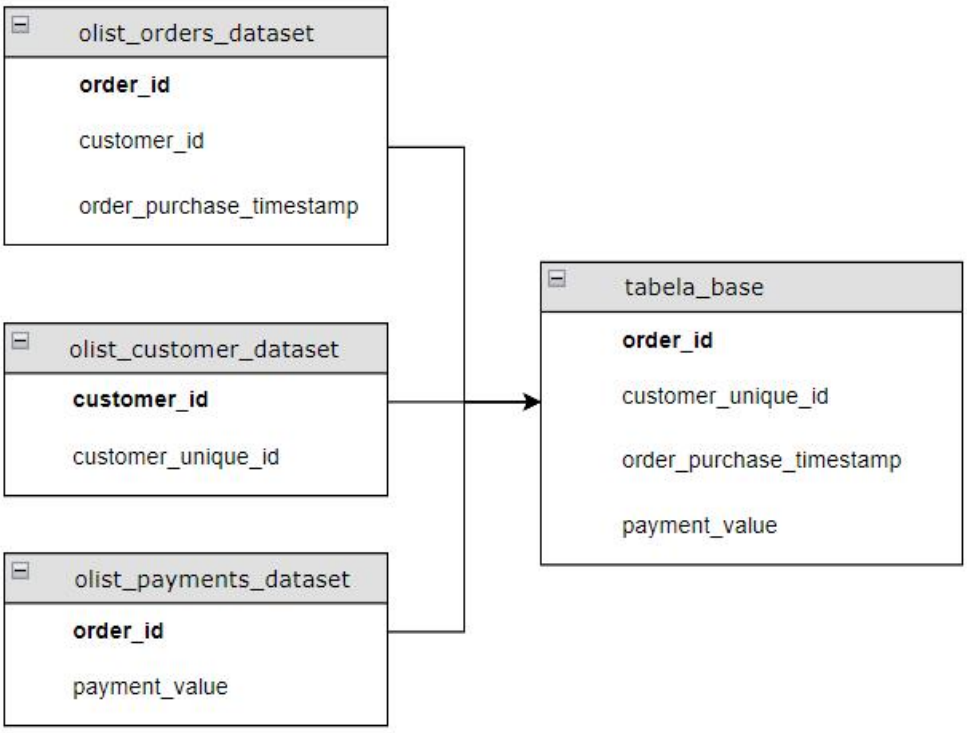
Coluna	Descrição
order_id	0
payment_value	0

Tabela 3 – Valores nulos na tabela *olist_customers_dataset*

Coluna	Descrição
customer_id	0
customer_unique_id	0

Em seguida, realizou-se um teste quanto à duplicidade de dados em cada tabela. No entanto, não foi constatado a presença de linhas repetidos, o que confirma a integridade das tabelas analisadas. É importante ressaltar que as tabelas são retiradas de um banco de dados relacional e, portanto, a informação está estruturada de forma a eliminar redundâncias. Assim, as informações necessárias para construção das métricas RFV, estão dispersas em diferentes tabelas. Após atestar a qualidade do dado, foi realizado a concatenação das tabelas através das suas respectivas chaves. Este procedimento, ilustrado na Figura 9, gerou a tabela base de análise. A transformação ilustrada permite a identificação dos clientes e o valor transacionado nos pedidos, requisitos básicos para construção do modelo RFV.

Figura 9 – Fluxo de transformação das tabelas



A coluna *customer_unique_id*, que representa a identificação de cada cliente, apresenta 96096 valores distintos em um total de 103887 pedidos, o que gera aproximadamente 1,08 pedido por cliente. A análise revelou que aproximadamente 94% dos clientes realizaram apenas 1 pedido no período. Na Figura 10 é possível visualizar a distribuição dos dados. A partir da análise, notou-se também que o cliente com mais pedidos somou 33 ordens de compra no período. A Figura 11 apresenta a contagem de pedidos dos vinte clientes com maior frequência de pedidos na base de dados.

Figura 10 – Histograma da contagem de pedidos por cliente

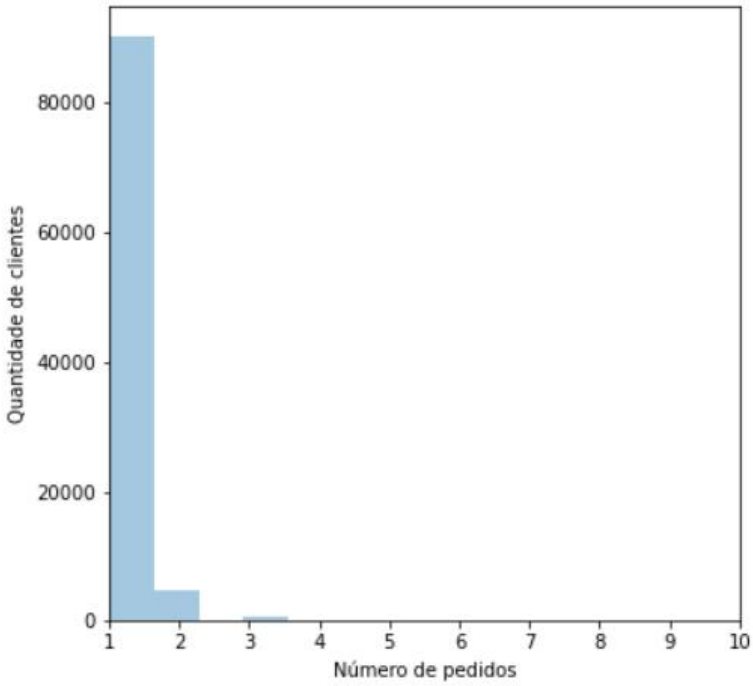
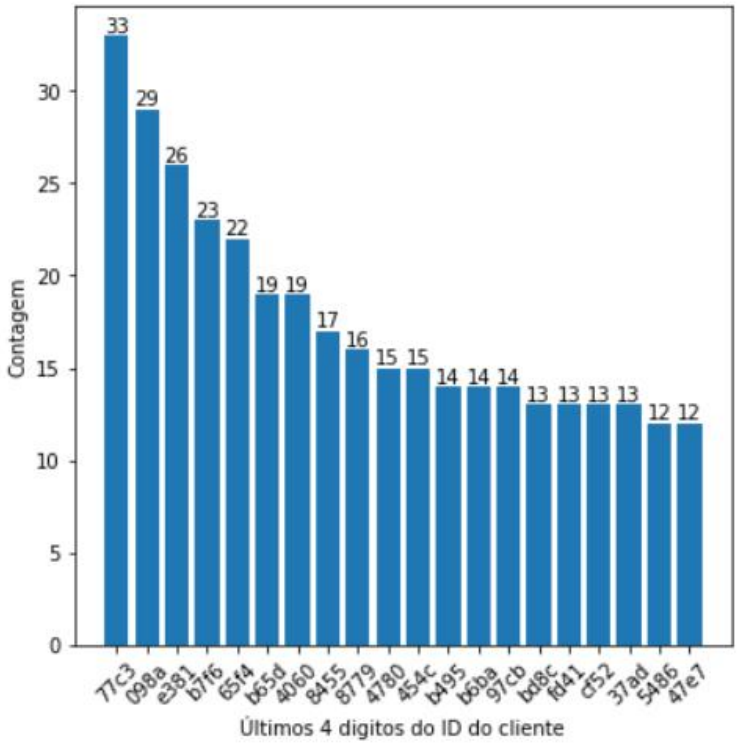


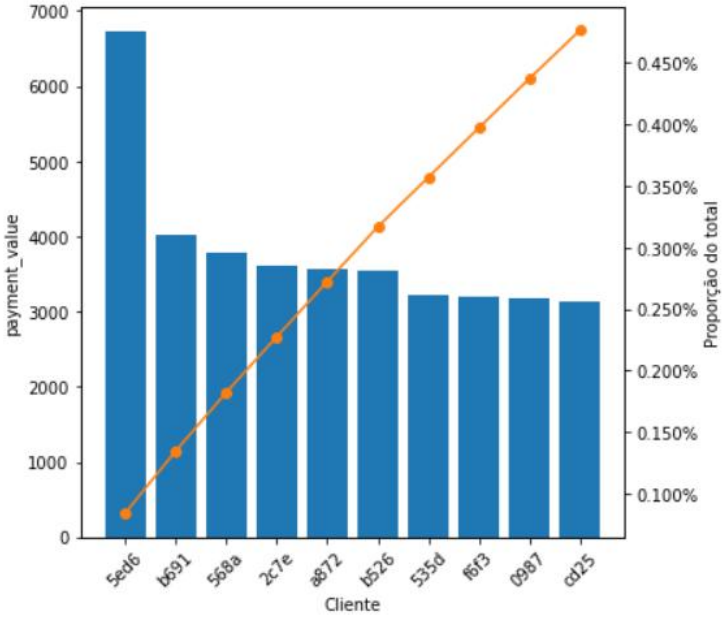
Figura 11 – Principais clientes da base identificados pelos últimos quatro dígitos do código de identificação



Com o objetivo de aprofundar a compreensão quanto ao comportamento de compra dos clientes, aplicou-se o princípio da Análise de Pareto para identificação

dos clientes com maior participação no valor total transacionado no período. Para tanto, somou-se o valor dos pedidos por cliente e ordenou-se de forma descendente. A Figura 12 apresenta a relação dos principais 10 clientes e a participação total que estes representam na base de dados. Esta análise não identificou uma concentração muito relevante em favor de um cliente específico. Nota-se que os principais 20 clientes por valor de pedido representam apenas 0,7% aproximadamente do total transacionado.

Figura 12 – Principais clientes pela soma dos pedidos efetuados



Quanto à dimensão temporal, o principal campo de análise é *order_purchase_timestamp*. Esta coluna registra o momento em que o pedido foi efetivado. Neste conjunto de dados, o primeiro pedido foi registrado em 04/09/2016 e o último em 17/10/2018, o que resulta em 772 dias e uma média de 134,6 pedidos por dia aproximadamente. No entanto, como pode ser visto na Figura 13, a contagem diária de pedidos pode apresentar grandes variações como mostrado pelos picos periódicos no gráfico abaixo. Além disso, é possível observar um pico de pedidos entre 2017 e 2018. A Tabela 4 mostra os dez dias com mais pedidos no período. É possível observar que os cinco maiores pedidos ocorrem entre 24 e 28 de novembro de 2017. Uma causa provável para esse grande aumento no número de pedidos é o acontecimento da *black-friday*, celebrada dia 25 de novembro. Devido aos descontos oferecidos nos mais diversos produtos, as empresas de *e-commerce* enfrentam intenso tráfego de clientes em suas plataformas nessa época do ano.

Figura 13 – Pedidos por dia

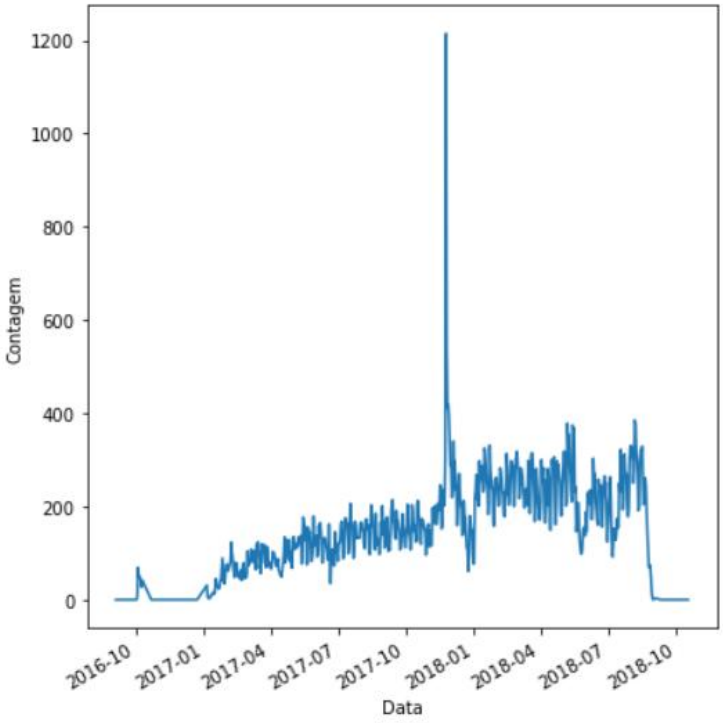


Tabela 4 – Dez principais picos de pedidos

Data	Numero de pedidos
2017-11-24	1214
2017-11-25	538
2017-11-27	420
2017-11-26	411
2017-11-28	393
2018-08-06	386
2018-05-07	379
2018-08-07	379
2018-05-14	375
2018-05-16	369

Nesta análise exploratória foi possível aprofundar a compreensão dos dados que serão utilizados durante a modelagem estatística. Inicialmente, procedeu-se com a leitura das tabelas e o entendimento do conteúdo das colunas. Notou-se que nem todas as colunas possuem valor analítico dado o escopo dessa análise, por este motivo apenas os campos de interesse para construção das métricas RFV foram mantidos. A partir das tabelas individuais, foi realizado um procedimento de transformação de

dados que concatenou em uma mesma tabela as informações de pedidos, clientes e pagamentos. De forma geral, percebeu-se que a base de clientes da empresa é bastante diversificada e as vendas não estão concentradas em alguns poucos clientes. A grande maioria dos clientes realizou apenas uma compra com valor inferior à 250 reais no período de análise.

5 MODELAGEM ESTATÍSTICA

5.0.1 Modelagem RFV

Para construção dos modelos de clusterização calculou-se inicialmente as métricas e pontuações RFV. Tradicionalmente, a partir das métricas, atribui-se pontuações de acordo com o quartil que contempla o respectivo valor da métrica. No entanto, neste trabalho os segmentos de clientes serão identificados a partir da modelagem estatística de modelos de aprendizado não supervisionados. Isto possibilitará a classificação do cliente de acordo com sua importância para o negócio de maneira mais robusta que a convencional. O cálculo de cada uma das métricas é resumido abaixo:

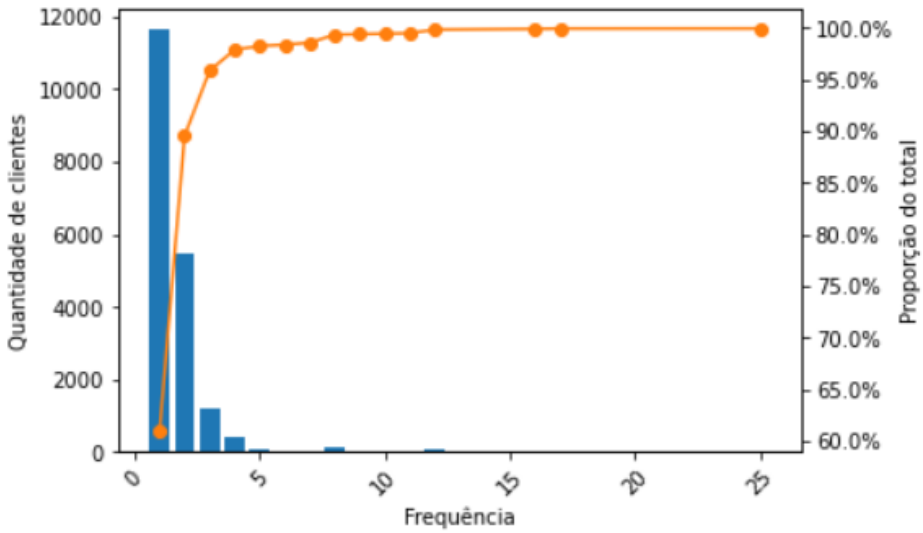
a) Recência: Número de dias desde a última compra em relação à última data do conjunto de dados.

b) Frequência: Quantidade total de pedidos realizados no período.

c) Valor: Soma do subtotal de pedidos.

Como já apontado durante a análise exploratória, a grande maioria dos clientes realizou apenas uma compra, o que justifica a concentração da distribuição da variável de frequência em valores entre 0 e 5 (Figura 10). Constatou-se que aproximadamente 99% da base de cliente está contemplada neste intervalo (Figura 14).

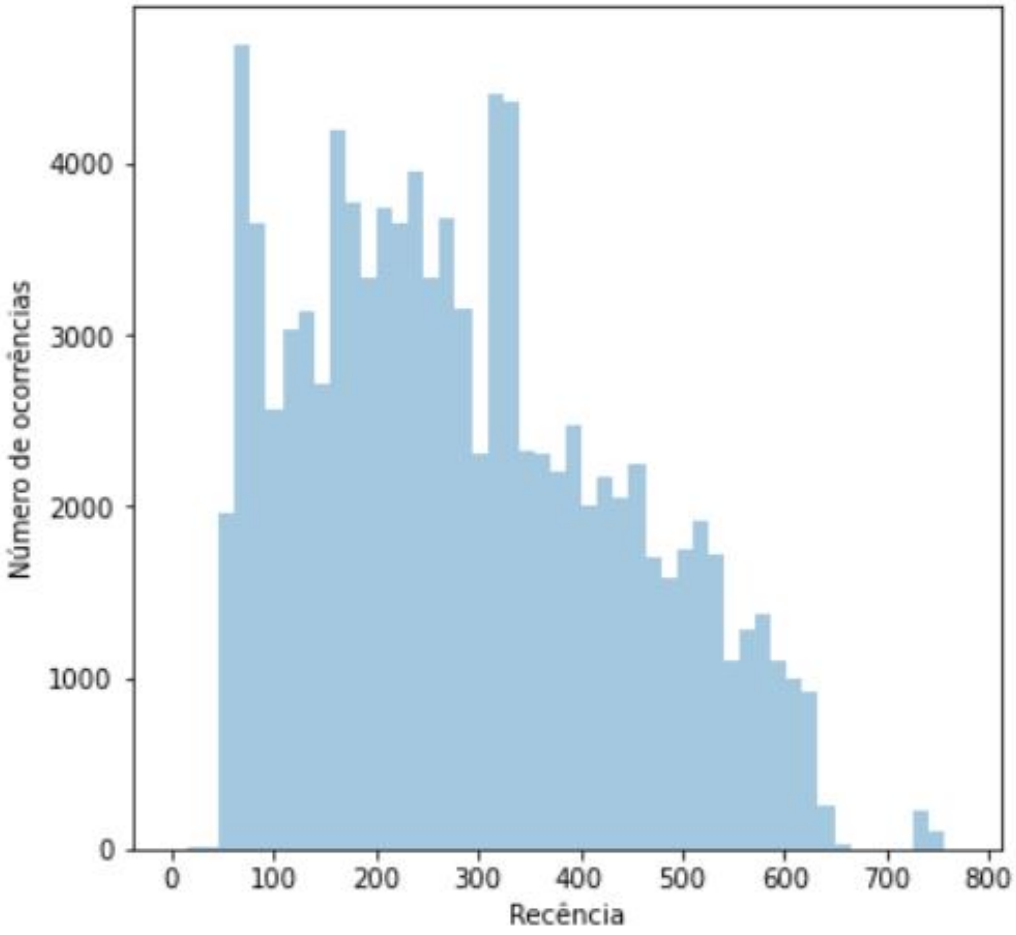
Figura 14 – Distribuição da frequência dos pedidos



Quanto à métrica de recência, nota-se na Figura 15 que a distribuição é visualmente mais dispersa em torno da média em relação à frequência. Além disso, é possível perceber a maior densidade de valores em períodos mais recentes. Assim como visto no capítulo anterior, há uma clara tendência de aumento na quantidade de pedidos no período analisado, o que explica o porquê da maioria dos pedidos terem sido realizados há menos tempo. De acordo com Wei & Lin (2010, p.4200) dentre as

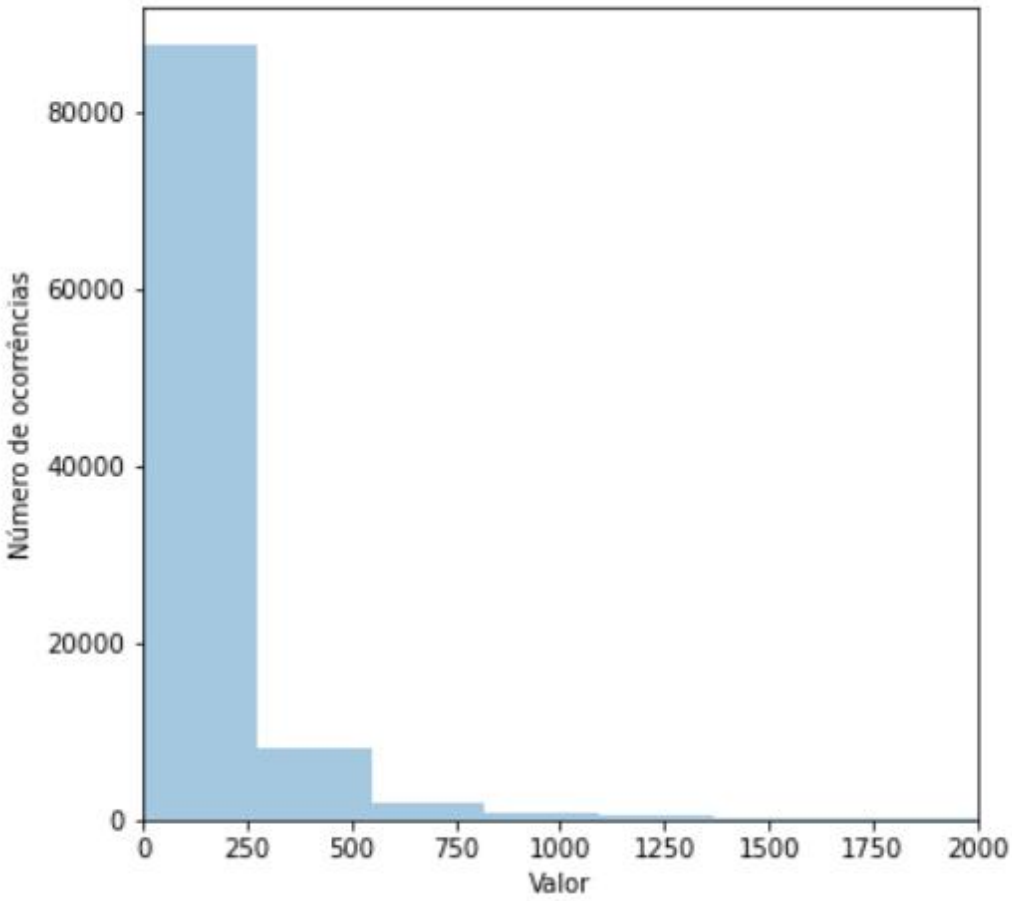
três métricas do RFV, a recência é muitas vezes considerada a métrica mais importante uma vez que o comportamento passado é o melhor preditor do comportamento futuro.

Figura 15 – Distribuição da métrica recência



Quanto à métrica de valor, é possível notar que a maioria dos valores está concentrada no intervalo entre 0 e 250 (Figura 16). A medida que o valor aumenta, menos clientes são encontrados em cada intervalo. Tradicionalmente, a métrica de valor é calculada como a soma dos valores de todas transações realizadas pelo cliente. Há, no entanto, abordagens alternativas que sugerem a utilização da média dos valores das transações para evitar colinearidade entre frequência e valor (MARCUS, 1998).

Figura 16 – Distribuição da métrica valor



5.0.2 Pré processamento dos dados

Como o intervalo de valores de cada uma das distribuições apresentadas acima é bastante distinto, é necessário padronizar os dados antes da aplicação dos modelos de clusterização. Dado que os modelos de clusterização normalmente baseiam-se em medidas de distância para definição dos grupos, o intervalo de valores em cada *feature* atuará como um peso ao determinar como agrupar os dados, o que normalmente é indesejado. A vantagem de padronizar os dados é que todas as *features* passam a contar com a mesma escala. De acordo com Kaufman & Rousseuw (2009, p.9), não são em todas as situações que é benéfico padronizar os dados. Ao padronizar, tenta-se dar a todas as variáveis um peso igual. No entanto, pode ser que algumas variáveis sejam intrinsecamente mais importantes do que outras em uma determinada aplicação e, portanto, durante a padronização essa informação é perdida. O procedimento de padronização é especialmente importante para modelos de clusterização porque muitos destes estão baseados em métricas distância, que são diretamente afetadas pela escada dos dados. Como abordagem inicial neste trabalho, será performado a padronização dada por $z = \frac{(x-u)}{s}$ onde x corresponde à observação, u é a média da

amostra e s é o desvio padrão da amostra (KAUFMAN; ROUSSEEUW, 2009).

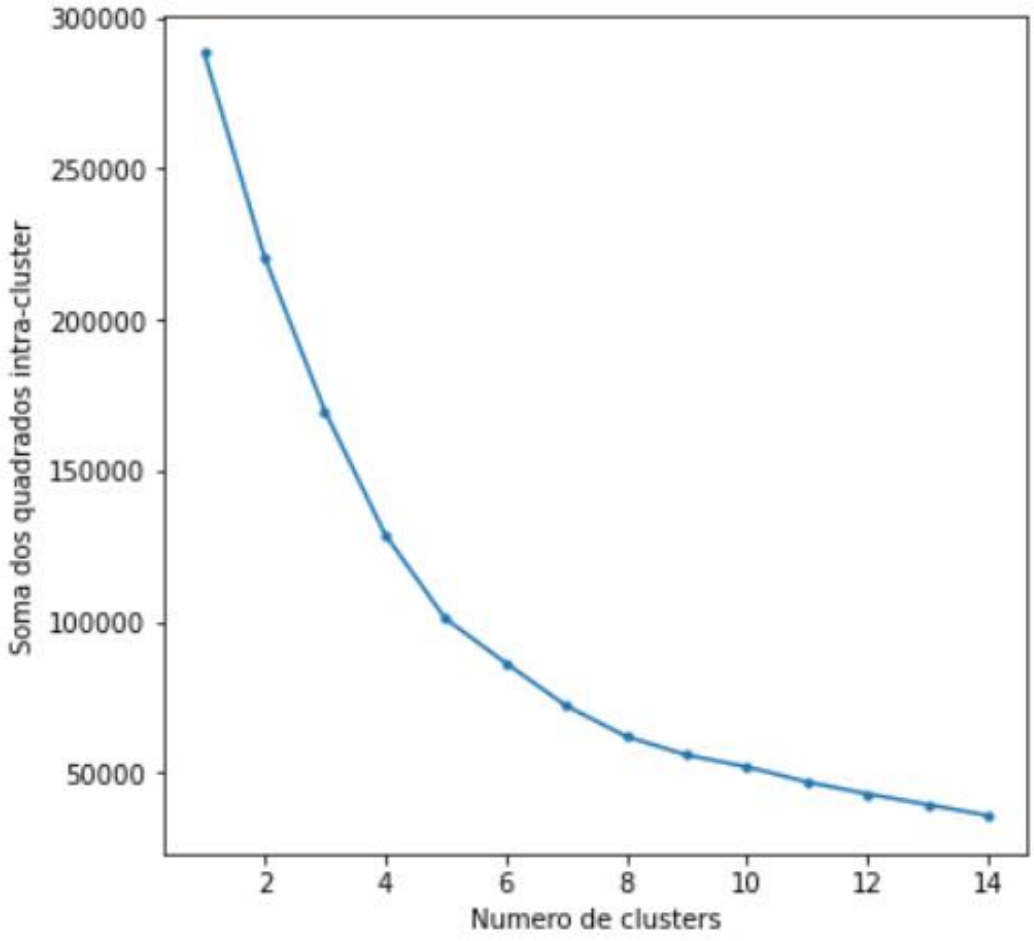
5.0.3 Clusterização

A escolha do algoritmo de clusterização depende do tipo de dado disponível, do objetivo e restrições da pesquisa. Na maioria dos casos diversos algoritmos podem ser utilizados. Nesses casos é importante utilizar mais de um método e comparar os resultados obtidos. A interpretabilidade do resultado também é um fator importante na escolha do método pois a análise de *clusters* é normalmente empregada como ferramenta de exploração dos dados.

5.0.3.1 *K-means*

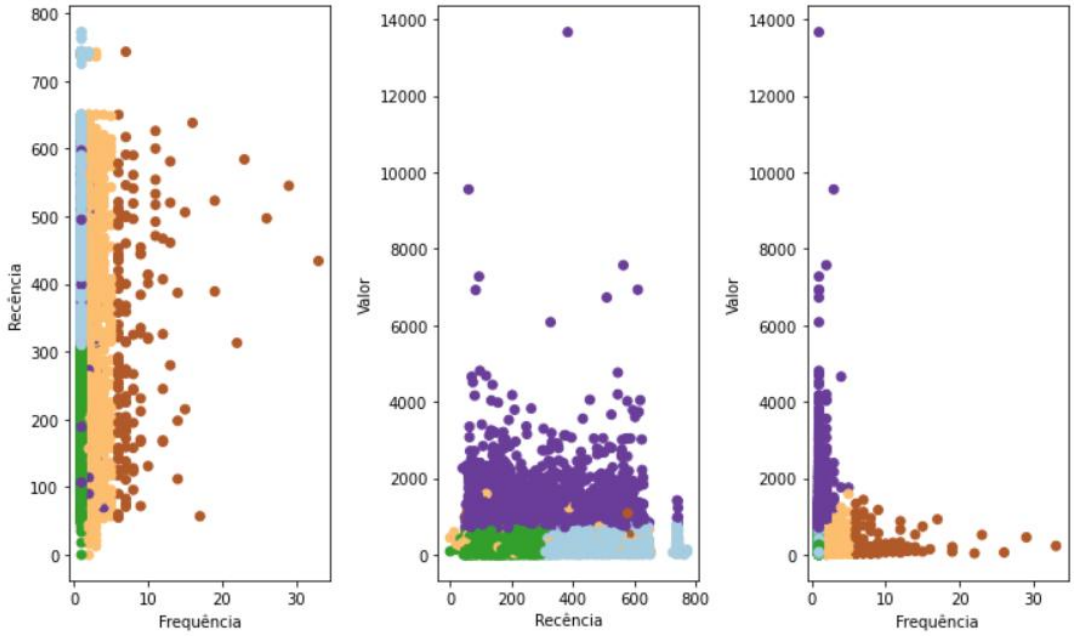
O primeiro modelo utilizado para segmentação da base de cliente é o *k-means*, bastante explorado na literatura neste tipo de problema (LI *et al.*, 2021; EZENKWU; OZUOMBA; KALU, s.d.; KANSAL *et al.*, 2018). Esta modelagem utiliza a distância euclideana para mensurar a distância entre os pontos. Além disso, o *k-means* exige como parâmetro de entrada a definição do número de *clusters* na amostra. No entanto, normalmente não se conhece o número de *clusters* presentes em um conjunto de dados. Para estimar um valor de *k*, aplica-se o modelo para diferentes valores e avalia-se uma métrica de qualidade para os agrupamentos (KAUFMAN; ROUSSEEUW, 2009, pp. 110). Comumente, à medida que aumenta-se a quantidade de *clusters*, as diferenças entre cada *cluster* diminui, enquanto a diferenças entre os pontos dentro de um mesmo *cluster* aumenta. Desta forma, o objetivo é encontrar o ponto de equilíbrio entre a homogeneidade dentro de um mesmo *cluster* e a heterogeneidade entre *clusters* diferentes. Na Figura 17 é possível notar que o ponto de inflexão ocorre para valores de *k* próximos à 5.

Figura 17 – Método do cotovelo



Após a determinação do valor de *k*, treina-se novamente um modelo de clusterização para o número identificado de *clusters*, neste caso 5 *clusters*. A Figura 18 apresenta os gráficos das métricas RFV onde cada ponto é identificado pelo grupo ao qual estão contidos. Cada gráfico apresenta uma perspectiva bidimensional das três *features* deste conjunto de dados.

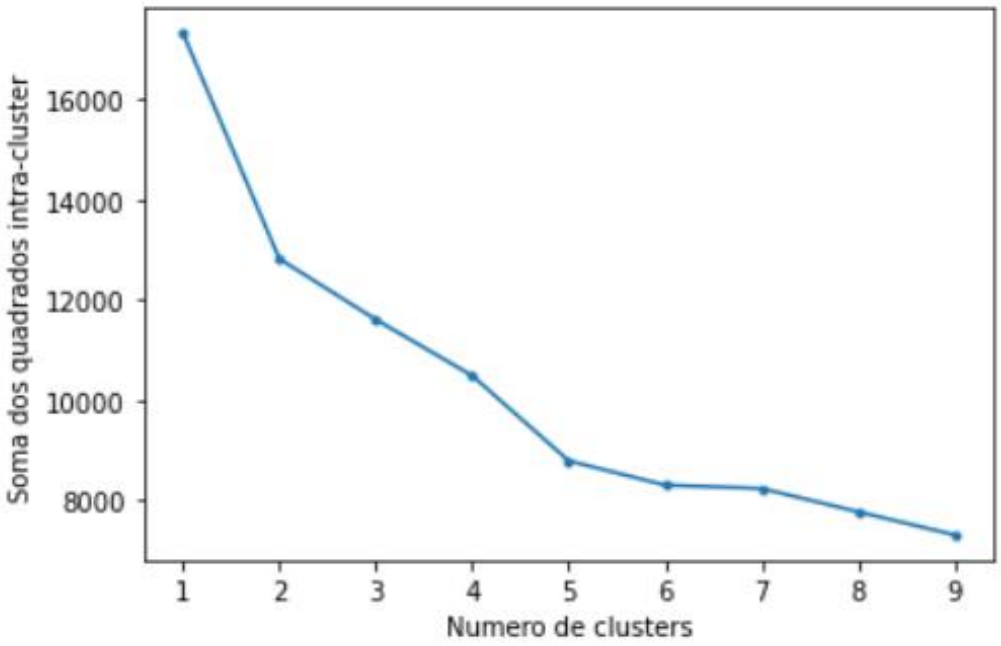
Figura 18 – Agrupamentos *k-means*



5.0.3.2 *K-medoids*

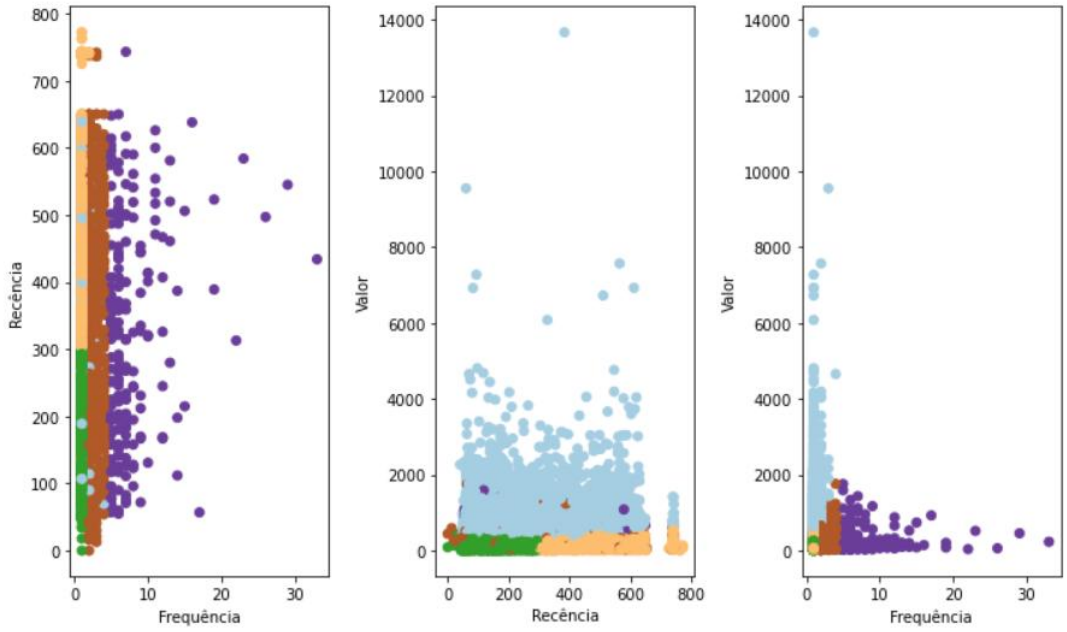
De acordo com Han & Kamber (2006, p.454) o algoritmo *k-means* é sensível a *outliers* porque tais observações estão distantes da maioria dos dados e, portanto, quando atribuídos a um *cluster*, podem distorcer drasticamente o valor médio do *cluster*. Isso afeta diretamente a atribuição de outros objetos aos *clusters*. Neste sentido, o algoritmo *K-medoids* torna-se uma opção viável e mais robusta, no entanto, é uma abordagem computacionalmente mais custosa. Para contornar este problema, o algoritmo é treinado em uma amostra aleatória dos dados. Da mesma forma como foi feito para o algoritmo *k-means*, o método do cotovelo é novamente aplicado para determinar o número de *clusters*. Como naturalmente este algoritmo já é computacionalmente mais custoso e, além disso, o método do cotovelo requer que diversos modelos sejam treinados, utilizou-se uma amostra aleatória correspondente à 20% do total de linhas. A definição deste valor buscou maximizar o número de linhas utilizadas e, ainda assim, obter um tempo de execução viável para o andamento da pesquisa. A Figura 19 apresenta a curva de número de *clusters* pela soma dos quadrados intra-*clusters*. É possível notar que a curva não possui o mesmo comportamento da curva obtida no método *k-means*. No entanto, é possível observar que novamente 5 *clusters* aparenta ser um número razoável para utilização no modelo.

Figura 19 – Método do cotovelo



Após a determinação do número de *clusters*, treinou-se um modelo novamente a partir de uma amostra de 20% dos dados. O modelo foi então utilizado para prever em qual *cluster* cada observação do conjunto completo dos dados está contemplada. A Figura 20 apresenta os resultados obtidos, cada um dos gráficos mostra uma perspectiva bidimensional do conjunto de dados onde cada observação é identificada através da cor pelo *cluster* que está contida.

Figura 20 – Agrupamentos *k-medoids*



5.0.4 Avaliação dos modelos

5.0.4.1 *K-means*

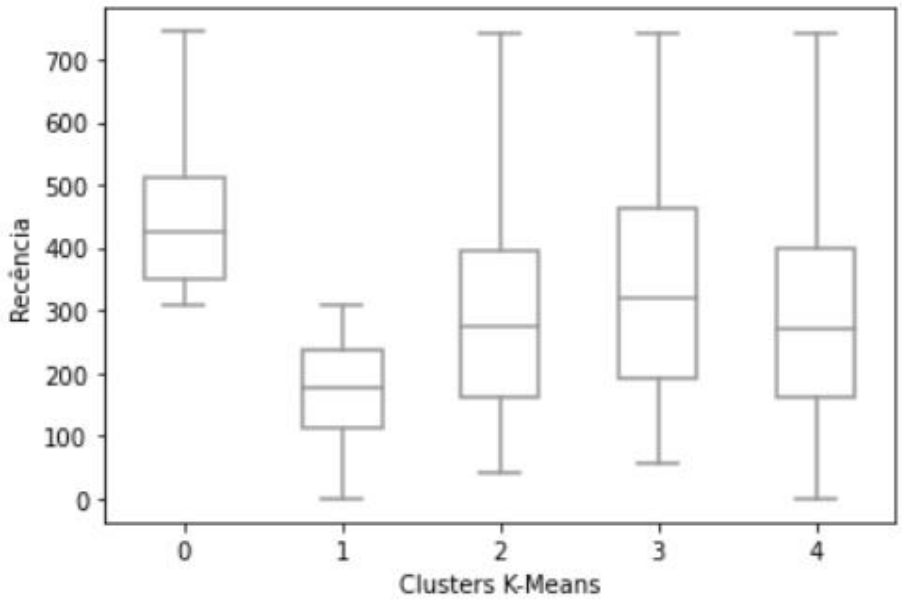
A partir dos *clusters* gerados por cada um dos modelos aplicados, realizou-se uma análise das características de cada grupo com intuito de identificar suas particularidades. Inicialmente, o modelo analisado foi o *k-means*. A Tabela 5 apresenta a média das métricas de recência, frequência e valor para cada um dos *clusters*. É importante ressaltar que os modelos foram aplicados em dados padronizados, no entanto as métricas expostas abaixo foram calculadas a partir dos dados originais. Portanto, a escala dos valores das métricas apresentadas abaixo não influenciou a clusterização dos dados

Tabela 5 – Média das métricas RFV para o modelo *k-means*

Cluster	Média de recência	Média de frequência	Média de valor	Número de clientes
0	437.0	1.0	135.0	37570
1	176.7	1.0	135.3	50511
2	286.7	1.1	1195.7	2492
3	330.2	8.8	259.0	143
4	287.0	2.2	201.7	5380

Nota-se que os grupos 0 e 1 representam os extremos entre os valores da média de recência. Em contrapartida, as médias de frequência e valor desses grupos não só possuem valores aproximadamente iguais como também representam os respectivos limites inferiores da média de cada métrica. Além disso, os grupos citados representam aproximadamente 91,6% do total de clientes. Como mostrado na Figura 21, é justamente a métrica de recência a variável capaz de diferenciar os indivíduos nos grupos analisados.

Figura 21 – *boxplot* da métrica de recência para o modelo *k-means*



Como visto durante a análise exploratória, a grande maioria dos clientes possui apenas uma compra com valor inferior a 150 reais. Neste sentido, é possível identificar os clientes dos *clusters* 0 e 1 como clientes ocasionais de baixo valor inativos e clientes ocasionais de baixo valor ativos respectivamente. Quanto aos *clusters* 2,3 e 4, visualiza-se no gráfico acima que a métrica de recência não é capaz de diferenciar os indivíduos de um grupo em relação aos indivíduos dos outros grupos. Esta diferenciação é realizada a partir da análise das demais métricas. Ainda de acordo com a Tabela 5, percebe-se que quanto à métrica de frequência, o *cluster* 3 destaca-se devido à alta média de compras realizadas por cliente. Nesse sentido, a Tabela 6 apresenta as frequências máximas, mínimas e as médias de cada *cluster*. Como pode ser visto, a mínima frequência do grupo 3 ainda é maior que a máxima frequência dos outros grupos. Portanto, este grupo contempla os clientes mais fiéis da empresa.

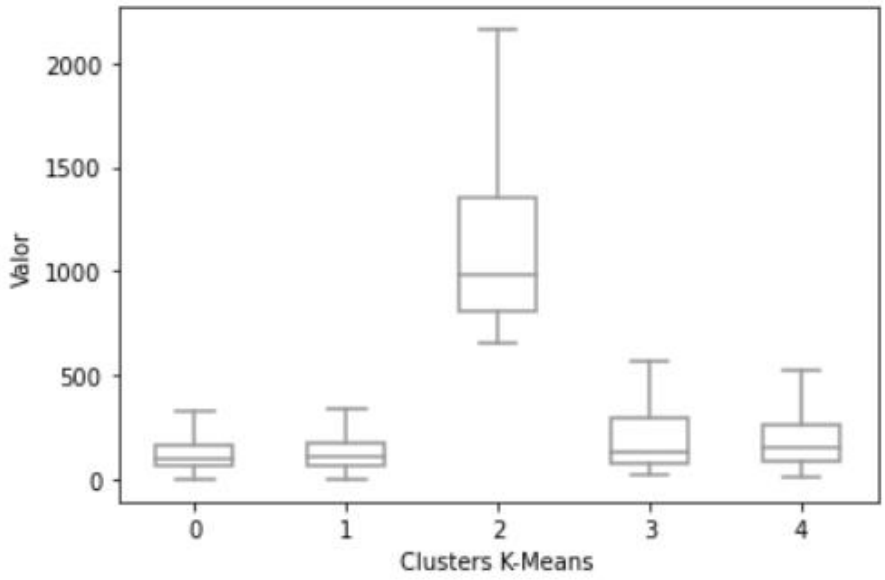
Tabela 6 – Sumário da métrica de frequência para cada *cluster* para o modelo *k-means*

Cluster	Média	Mínimo de frequência	Máximo de frequência
0	1.0	1	2
1	1.0	1	1
2	1.1	1	5
3	8.8	6	33
4	2.2	2	5

A partir da análise de recência e frequência não foi possível claramente diferen-

ciar os grupos 2 e 4. A Figura 22 apresenta o *boxplot* da métrica de valor. É possível notar que o grupo 2 caracteriza-se por altos valores de compra. Este grupo contempla os 2492 clientes com os valores de pedidos mais altos da base de dados. O grupo 4, no entanto, não representa nenhum dos extremos em termos de frequência, recência e valor. Mas, representa um grupo de clientes que, de maneira geral, realizou mais de uma compra (entre 2 e 5 como visto na Tabela 6) e possui valor acima da média da base de dados.

Figura 22 – *boxplot* da métrica de valor para o modelo *k-means*



5.0.4.2 *K-medoids*

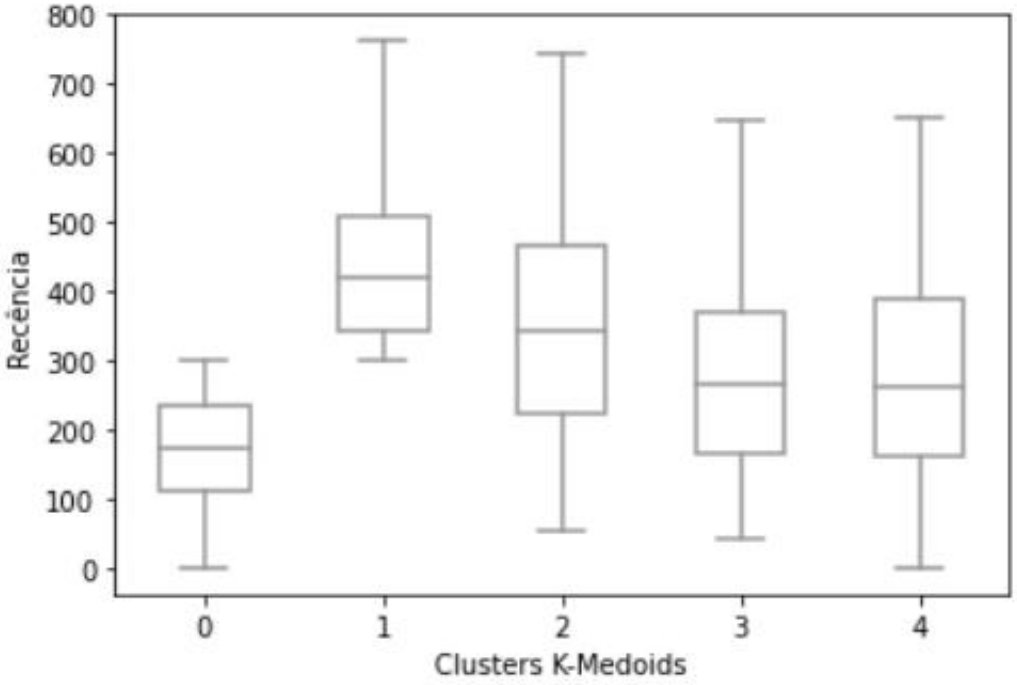
Replicou-se os mesmos procedimentos adotados na seção anterior para a análise dos resultados obtidos com o método *k-medoids*. Dessa forma, a Tabela 7 apresenta um sumário das métricas para cada *cluster* identificado. Da mesma forma como no método anterior, é possível identificar dois *clusters* (0 e 1) que representam os extremos em termos da média de recência. Os grupos citados somados abrangem 85.487 dos 96.096 clientes contidos na base de dados, o que representa aproximadamente 89% do total.

Tabela 7 – Média das métricas RFV para o modelo *k-medoids*

Cluster	Média de recência	Média de frequência	Média de valor	Número de clientes
0	173.6	1.0	122.6	47850
1	433.8	1.0	123.6	37637
2	345.8	5.0	192.4	571
3	280.0	1.1	857.1	5153
4	281.5	2.1	197.6	4885

De fato, é possível visualizar no gráfico abaixo que os grupos 0 e 1 diferenciam-se justamente através da métrica de recência. Para os *clusters* 2,3 e 4 a mesma conclusão não pode ser obtida uma vez que observa-se valores comuns de recência entre esses grupos (Figura 23).

Figura 23 – Média das métricas RFV para o modelo *k-medoids*



Quanto à métrica de frequência, observa-se que os *clusters* 2 e 4 destacam-se devido à quantidade mínima de pedidos superior a 1. O *cluster* 2 apresenta a maior média e abrange os clientes mais assíduos da empresa. Já os grupos 3 e 4, não são claramente diferenciáveis em relação aos seus pares nesta análise (Tabela 8).

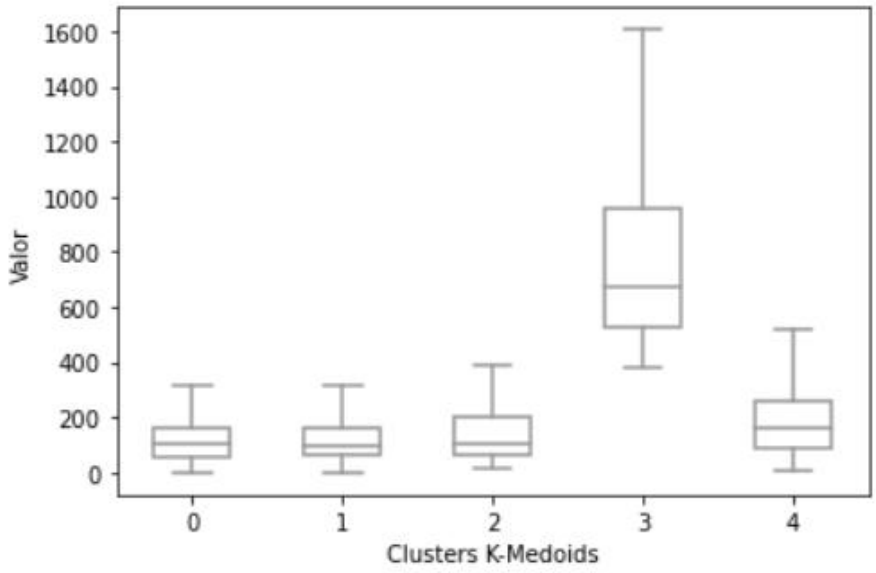
Tabela 8 – Sumário da métrica de frequência para cada *cluster* para o modelo *k-medoids*

Cluster	Média	Mínimo de frequência	Máximo de frequência
0	1.0	1	1
1	1.0	1	2
2	5.0	3	33
3	1.1	1	4
4	2.1	2	3

Para identificação dos principais atributos dos *clusters* 3 e 4, realiza-se a análise da métrica de valor. Abaixo (Figura 24) é possível observar que o *cluster* 3 possui

os maiores números para a métrica de valor. Neste grupo estão contemplados os clientes com os maiores pedidos em termos de valor financeiro. Em média, cada cliente deste *cluster* gastou 857 reais no período ante 127 reais em média dos clientes contemplados nos outros *clusters*. Novamente, observa-se um *cluster* (grupo 4) que não representa nenhum dos extremos em termos de recência, frequência e valor, mas representa um grupo de clientes que possui pelo menos 2 pedidos realizados e valor gasto ligeiramente acima da média.

Figura 24 – *boxplot* da métrica de valor para o modelo *k-medoids*



5.0.5 Comparação dos resultados

Após aplicação dos modelos e análise dos resultados, foi possível observar que ambos os métodos retornaram resultados virtualmente equivalentes. Como observado nas tabelas 9 e 10, dado o valor pré-definido de 5 *clusters*, ambos os métodos foram capazes de identificar grupos similares de clientes. Como forma de facilitar a análise, cada *cluster* foi nomeado de acordo com seu principal atributo. Os clientes de baixo valor e baixa frequência mas diferenciáveis pela recência do pedido foram chamados de clientes ocasionais. Esses clientes podem ser ativos quando a recência do pedido é baixa, ou inativos quando a recência é alta. Os clientes que possuem os pedidos de maior valor foram chamados de clientes de alto valor. Já o grupo de clientes recorrentes representa, em ambos modelos, a composição das observações as quais a frequência foi superior a um mas inferior ao grupo de clientes assíduos, estes são os clientes com os maiores números de pedidos da base de dados.

Tabela 9 – Categorias de clientes *k_means*

Cluster	Característica	Número de clientes	Porcentagem dos clientes
1	Cliente ocasional ativo	50511	52,56%
0	Cliente ocasional inativo	37570	39,10%
4	Cliente recorrente	5380	5,60%
2	Cliente de alto valor	2492	2,59%
3	Cliente assíduo	143	0,15%

Tabela 10 – Categorias de clientes *k_medoids*

Cluster	Característica	Número de clientes	Porcentagem dos clientes
0	Cliente ocasional ativo	47850	49,79%
1	Cliente ocasional inativo	37637	39,16%
3	Cliente de alto valor	5153	5,36%
4	Cliente recorrente	4885	5,08%
2	Cliente assíduo	571	0,59%

Os *clusters* com características similares entre os métodos foram então comparados em relação às suas métricas RFV. Inicialmente, analisou-se os grupos de clientes ocasionais ativos (Figura 25). É possível observar que a distribuição dos dados é bastante similar entre os modelos, no entanto, o método *k-means* abrange um número de clientes ligeiramente maior que seu par. Em ambos os modelos a partição entre os *clusters* de clientes ocasionais ativos e inativos foi realizada para um valor de recência próximo à 300. Da mesma forma para o grupo de clientes ocasionais inativos, é possível observar as distribuições das métricas RFV bastante similares entre os modelos (Figura 26).

Figura 25 – Distribuição das métricas RFV para o *cluster* de clientes ocasionais ativos

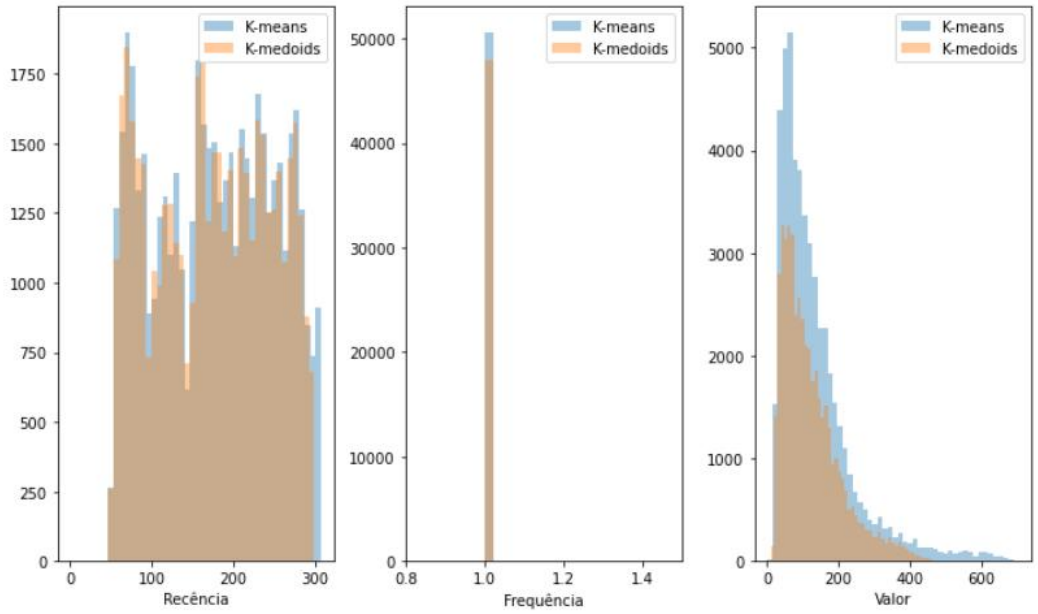
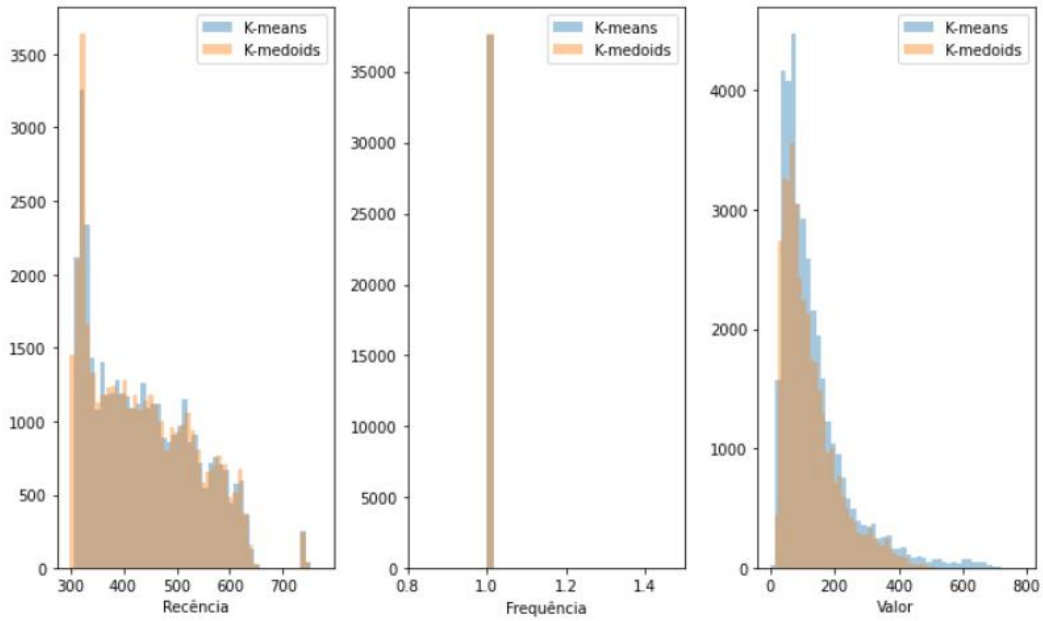
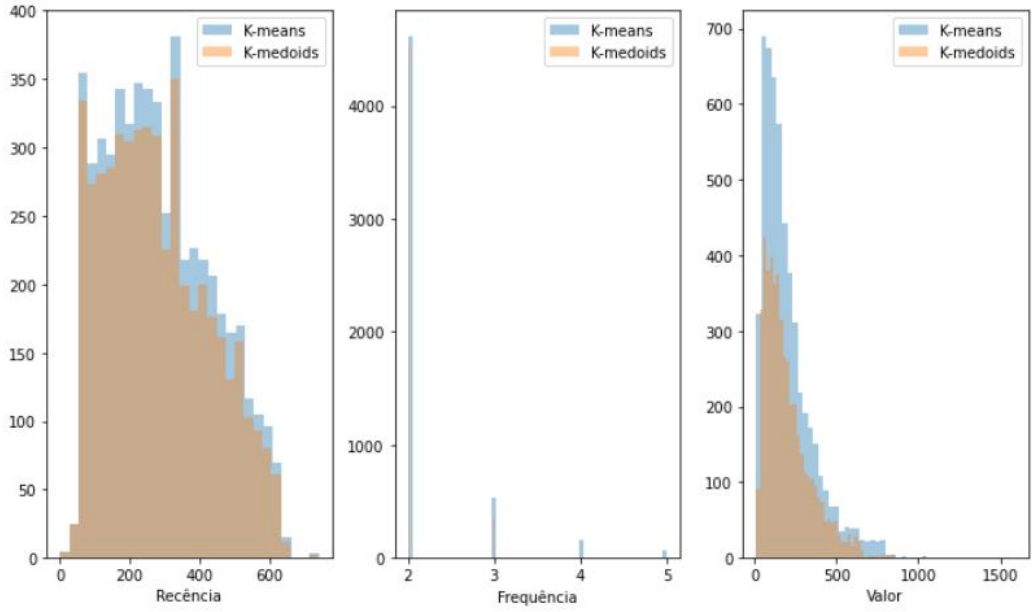


Figura 26 – Distribuição das métricas RFV para o *cluster* de clientes ocasionais inativos



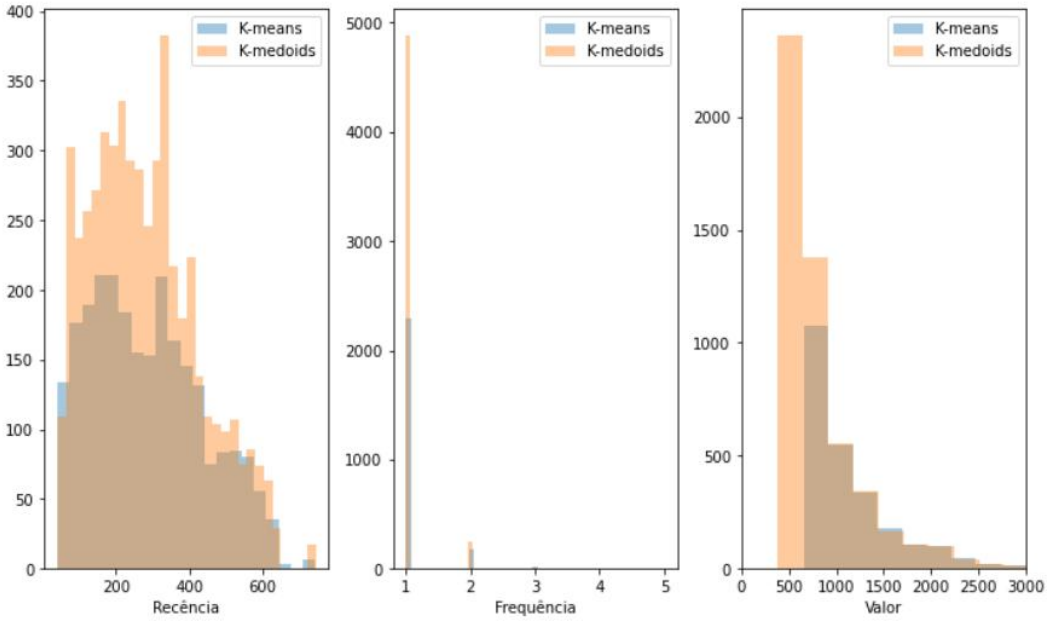
Já os clientes recorrentes, em ambos modelos, formam um *cluster* composto por indivíduos que realizaram mais de um pedido no período. Para o algoritmo *k-means*, os clientes recorrentes possuem entre 2 e 5 pedidos, já para o algoritmo *k-medoids* este grupo possui entre 2 e 3 pedidos. Apesar do primeiro modelo abranger um número maior de clientes em relação ao segundo modelo, é possível novamente observar bastante similaridade entre as distribuições dos dados (Figura 27).

Figura 27 – Distribuição das métricas RFV para o *cluster* de clientes recorrentes



Quanto aos clientes de alto valor, este grupo é composto pelos indivíduos com gastos de em média de 1195,7 reais para o modelo *k-means* e 857,1 reais para o modelo *k-medoids*. Como o modelo *k-medoids* criou uma partição mais ampla, há 4.885 clientes neste grupo ante 2492 no modelo *k-means*. No entanto, observa-se na Figura 34 que o comportamento das distribuições é bastante similar entre os modelos (Figura 28).

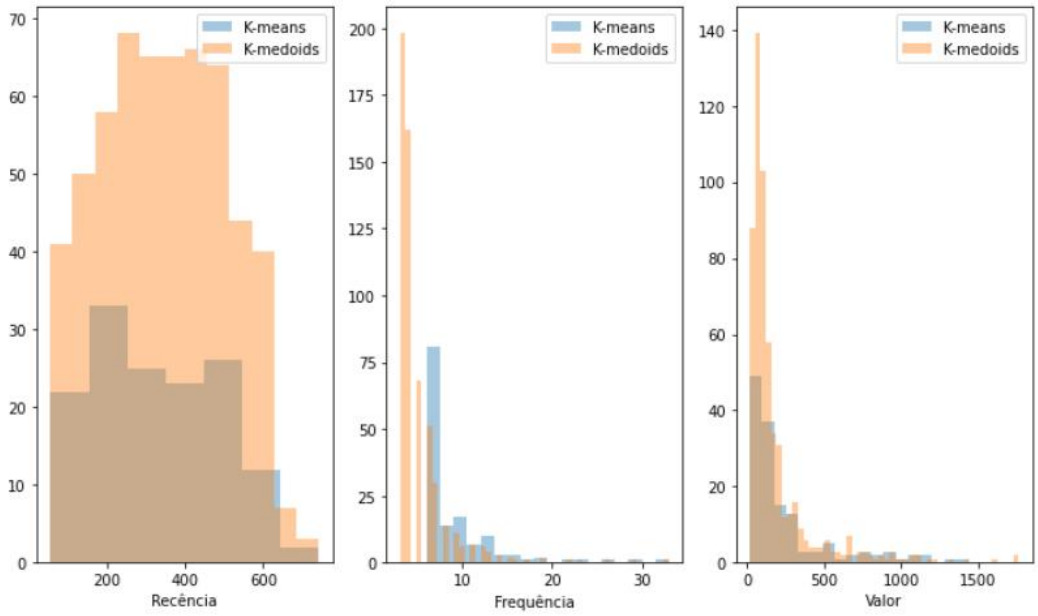
Figura 28 – Distribuição das métricas RFV para o *cluster* de clientes de alto valor



Os indivíduos que possuem a quantidade de pedidos superior aos clientes recorrentes foram chamados de clientes assíduos. Este grupo representa menos de 1%

do total de clientes da base de dados. Neste *cluster* observa-se a maior diferença entre os métodos empregados. Como observado durante a análise exploratória, 94% da base de clientes possui apenas um pedido e, a medida que o número de pedidos cresce, a quantidade de clientes diminui exponencialmente. Dessa forma, pequenas variações nos limites das partições dos *clusters* em relação à frequência causam grandes impactos na quantidade de clientes contemplados pelo respectivo *cluster*. Como pode ser visto nas tabelas 10 e 11 e na Figura 29, o método *k-medoids* contempla aproximadamente 230% a mais de clientes em relação ao *k-means* uma vez que seu intervalo de frequências é ligeiramente mais amplo.

Figura 29 – Distribuição das métricas RFV para o *cluster* de clientes assíduos



Tanto o modelo *k-means* quanto o modelo *k-medoids* produziram resultados bastante similares. A partir da pré-definição de uma quantidade de *clusters*, ambos métodos construíram grupos equivalentes quanto às suas principais métricas. O modelo *k-means*, no entanto, foi computacionalmente mais eficiente, o que permite que um número maior número de modelos sejam treinados com diferentes parâmetros em um menor intervalo de tempo em relação ao modelo *k-medoids*. Como apontado por Han & Kamber (2011, p.454), a vantagem do modelo *k-medoids* sobre o *k-means*, é sua capacidade de lidar com *outliers*. No entanto, o conjunto de dados utilizado neste estudo é advindo de um banco de dados transacional, onde armazena-se a informação mais confiável da empresa. Acredita-se que por este motivo não foram detectados *outliers* no esquema de dados disponibilizado.

6 CONCLUSÕES

Este trabalho surge da necessidade de fornecer embasamento para o direcionamento de estratégias comerciais focalizadas em determinados segmentos de clientes. Sabe-se que os mercados são heterogêneos e, portanto, o processo de segmentação emerge como o primeiro passo para o desenvolvimento de uma estratégia de *marketing* eficaz. Este estudo almejou identificar os principais segmentos de clientes de uma empresa de *e-commerce* baseado em métricas de recência, frequência e valor, objetivo que foi alcançado ao final da análise. Optou-se pela utilização dos métodos de clusterização particionais *k-means* e *k-medoids* devido à facilidade de implementação, escalabilidade, interpretabilidade dos resultados e ambos os métodos foram extensivamente explorados na literatura neste tipo de problema produzindo resultados satisfatórios.

Os métodos de clusterização adotados demandam a pré-definição do número de *clusters*. Este parâmetro foi definido a partir da aplicação extensiva de diferentes modelos com diferentes parâmetros. Em ambos os casos, com auxílio do método do cotovelo, cinco *clusters* foram capazes de produzir resultados satisfatórios, isto é, alta diferenciação entre *clusters* e alta similaridade entre observações dentro de um mesmo grupo. Os resultados obtidos foram então analisados com intuito de aprofundar o conhecimento acerca do conjunto de métricas que define cada partição encontrada. Neste momento, identificou-se que ambos os métodos produziram *clusters* equivalentes:

a) Clientes ocasionais recentes e de baixo valor: grupo composto por indivíduos que realizaram apenas um pedido no período. Dentre os clientes de pedido único de baixo valor, os métodos produziram dois grupos de acordo com a recência do pedido. Assim, este grupo contempla os clientes de pedido mais recente, o que corresponde à aproximadamente 50% do total de clientes para ambos algoritmos utilizados.

b) Clientes ocasionais antigos e de baixo valor: este *cluster* também é composto por indivíduos de pedido único de baixo valor. No entanto, o grupo contempla os clientes com pedidos realizados há mais de 300 dias aproximadamente para ambos os métodos. Este grupo corresponde a 39% dos clientes, o que somado ao grupo de clientes ocasionais recentes e de baixo valor resulta em 90% da base de dados alocadas em apenas dois *cluster*.

c) Clientes de alto valor: *cluster* composto pelos indivíduos com os pedidos de maiores valores. Esses clientes são aqueles cujos valores dos pedidos são maiores que aproximadamente 90% dos pedidos contidos na base.

d) Clientes recorrentes: clientes cujo número de pedidos é superior a 2, inferior a 3 para *k-medoids*, inferior a 5 para *k-means* e não é um cliente de alto valor. Como aproximadamente 94% dos clientes possui apenas um pedido,

é possível compreender que naturalmente os clientes com quantidade de pedidos superior a um correspondam a um grupo distinto de indivíduos.

e) Clientes assíduos: é composto por clientes cuja quantidade de pedidos é superior aos demais *clusters*. As partições foram realizadas para clientes com no mínimo 3 e 6 pedidos para *k-medoids* e *k-means* respectivamente. Esse valor gera o menor dos *clusters* em número de clientes nos dois casos. São 571 clientes assíduos no primeiro caso e 143 no segundo.

Ainda que haja uma ligeira diferença entre os resultados alcançados por cada algoritmo, mais importante que o valor específico das partições, é a intuição gerada pelos métodos. O grande valor da análise deriva da interpretação dos resultados obtidos em vista de responder a pergunta de pesquisa. Como a análise de *clusters* é uma atividade de caráter exploratório, a utilização de mais de um método auxilia a compreensão profunda do conjunto de dados.

Este trabalho representa apenas um ciclo no epiciclo da análise de dados proposto por Peng & Matsui (2015). Os resultados alcançados não apenas buscam responder a pergunta de pesquisa como também auxiliar o estabelecimento de novos questionamentos. Ainda que os objetivos propostos tenham sido alcançados, há muito espaço para aprofundar a análise. Este estudo limitou-se a segmentar os clientes em relação às métricas RFV. No entanto, múltiplas bases de segmentação podem ser utilizadas para obter grupos cada vez mais bem definidos, o que viabiliza a construção de estratégias de negócio.

REFERÊNCIAS

AGGARWAL, Charu C *et al.* Fast algorithms for projected clustering. **ACM SIGMoD record**, ACM New York, NY, USA, v. 28, n. 2, p. 61–72, 1999.

ANITHA, Palaksha; PATIL, Malini M. RFM model for customer purchase behavior using K-Means algorithm. **Journal of King Saud University-Computer and Information Sciences**, Elsevier, 2019.

BALAKRISHNAN, PV Sundar *et al.* Comparative performance of the FSCL neural net and K-means algorithm for market segmentation. **European journal of operational research**, Elsevier, v. 93, n. 2, p. 346–357, 1996.

BERTOLINI, Massimo *et al.* Machine Learning for industrial applications: A comprehensive literature review. **Expert Systems with Applications**, v. 175, p. 114820, mar. 2021. DOI: 10.1016/j.eswa.2021.114820.

BLASHFIELD, Roger K; ALDENDERFER, Mark S. Computer programs for performing iterative partitioning cluster analysis. **Applied Psychological Measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 2, n. 4, p. 533–541, 1978.

BLOCKER, Christopher P; FLINT, Daniel J. Customer segments as moving targets: integrating customer value dynamism into segment instability logic. **Industrial Marketing Management**, Elsevier, v. 36, n. 6, p. 810–822, 2007.

BRITO, Pedro Quelhas *et al.* Customer segmentation in a large database of an online customized fashion business. **Robotics and Computer-Integrated Manufacturing**, Elsevier, v. 36, p. 93–100, 2015.

CHRISTY, A. Joy *et al.* RFM Ranking – An Effective Approach to Customer Segmentation. **J. King Saud Univ. Comput. Inf. Sci.**, Elsevier Science Inc., USA, v. 33, n. 10, p. 1251–1257, dez. 2021. ISSN 1319-1578. DOI: 10.1016/j.jksuci.2018.09.004. Disponível em: https://doi.org/10.1016/j.jksuci.2018.09.004.

DING, Chris *et al.* Adaptive dimension reduction for clustering high dimensional data. *In:* IEEE. 2002 IEEE International Conference on Data Mining, 2002. Proceedings. [*S.l.*: *s.n.*], 2002. P. 147–154.

ESTER, Martin *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. *In:* 34. KDD. [*S.l.*: *s.n.*], 1996. P. 226–231.

EZENKWU, Chinedu Pascal; OZUOMBA, Simeon; KALU, Constance. **Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services**. [*S.l.*: *s.n.*].

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

GARETH, James *et al.* **An introduction to statistical learning: with applications in R**. [*S.l.*]: Spinger, 2013.

GARTNER. **Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics**. [*S.l.*: *s.n.*], mai. 2008. https://www.gartner.com/en/newsroom/press-releases/2018-02-05-gartner-survey-shows-organizations-are-slow-to-advance-in-data-and-analytics.

GREEN, Paul E. A new approach to market segmentation. **Business Horizons**, Elsevier, v. 20, n. 1, p. 61–73, 1977.

HAN, Jiawei; KAMBER, Micheline. Data mining: concepts and techniques, 2nd. **University of Illinois at Urbana Champaign: Morgan Kaufmann**, 2006.

HBR. **Critical Success Factors to Achieve a Better Enterprise Data Strategy in a Multi-cloud Environment**. [*S.l.*: *s.n.*], set. 2019. https://hbr.org/sponsored/2019/09/critical-success-factors-to-achieve-a-better-enterprise-data-strategy-in-a-multi-cloud-environment.

HBR. **Using Data to Deliver Customer Delight: Six Modern Marketing Lessons from Asia Pacific**. [*S.l.*: *s.n.*], dez. 2021. https://hbr.org/sponsored/2021/12/using-data-to-deliver-customer-delight-six-modern-marketing-lessons-from-asia-pacific.

HUGHES, Arthur Middleton. **Strategic database marketing : the masterplan for starting and managing a profitable, customer-based marketing program**. [*S.l.*]: Chicago, Ill. : Probus Pub. Co., ©1994., 1994.

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. Machine Learning sob a ótica estatıstica. **Ufscar/Insper**, 2018.

KAHAN, Ron. Using database marketing techniques to enhance your one-to-one marketing initiatives. **Journal of Consumer Marketing**, MCB UP Ltd, 1998.

KANSAL, Tushar *et al.* Customer Segmentation using K-means Clustering. *In:* 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). [*S.l.*: *s.n.*], 2018. P. 135–139. DOI: 10.1109/CTEMS.2018.8769171.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. [*S.l.*]: John Wiley & Sons, 2009.

KHAJVAND, Mahboubeh *et al.* Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. **Procedia Computer Science**, Elsevier, v. 3, p. 57–63, 2011.

KOTLER, P. **Principles of Marketing**. [*S.l.*]: Prentice-Hall, 1980. (The Prentice-Hall series in marketing). ISBN 9780137015573. Disponível em: https://books.google.com.br/books?id=%5C_HrT6bficY4C.

KRIEGEL, Hans-Peter *et al.* Density-based clustering. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 3, p. 231–240, 2011.

LI, Yue *et al.* Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. **Applied Soft Computing**, v. 113, p. 107924, 2021. ISSN 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2021.107924. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494621008462.

MANYIKA, James *et al.* **Big data: The next frontier for innovation, competition, and productivity**. [*S.l.*]: McKinsey Global Institute, 2011.

MARCUS, Claudio. A practical yet meaningful approach to customer segmentation. **Journal of consumer marketing**, MCB UP Ltd, 1998.

MCKINNEY, Wes. **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**. [*S.l.*]: "O'Reilly Media, Inc.", 2012.

MONALISA, Siti; NADYA, Putri; NOVITA, Rice. Analysis for customer lifetime value categorization with RFM model. **Procedia Computer Science**, Elsevier, v. 161, p. 834–840, 2019.

MOSADDEGH, Abdolreza *et al.* Dynamics of customer segments: A predictor of customer lifetime value. **Expert Systems with Applications**, Elsevier, v. 172, p. 114606, 2021.

MOUNT, John; ZUMEL, Nina. **Practical data science with R**. [*S.l.*]: Simon e Schuster, 2019.

NISBET, Robert; ELDER, John; MINER, Gary D. **Handbook of statistical analysis and data mining applications**. [*S.l.*]: Academic press, 2009.

OLIST; SIONEK, André. **Brazilian E-Commerce Public Dataset by Olist**. [*S.l.*: *s.n.*], 2018. https://www.kaggle.com/dsv/195341.

PENG; MATSUI. The Art of Data Science, 2015.

RAHIM, Mussadiq Abdul *et al.* RFM-based repurchase behavior for customer classification and segmentation. **Journal of Retailing and Consumer Services**, Elsevier, v. 61, p. 102566, 2021.

SHIN, HW; SOHN, So Young. Segmentation of stock trading customers according to potential value. **Expert systems with applications**, Elsevier, v. 27, n. 1, p. 27–33, 2004.

SMITH, Wendell R. Product differentiation and market segmentation as alternative marketing strategies. **Journal of marketing**, SAGE Publications Sage CA: Los Angeles, CA, v. 21, n. 1, p. 3–8, 1956.

STAHL, Heinz K; MATZLER, Kurt; HINTERHUBER, Hans H. Linking customer lifetime value with shareholder value. **Industrial Marketing Management**, Elsevier, v. 32, n. 4, p. 267–279, 2003.

JO-TING, Wei; SHIH-YEN, Lin; HSIN-HUNG, Wu. A review of the application of RFM model. **African Journal of Business Management**, Academic Journals, v. 4, n. 19, p. 4199–4206, 2010.

TYNAN, A Caroline; DRAYTON, Jennifer. Market segmentation. **Journal of marketing management**, Taylor & Francis, v. 2, n. 3, p. 301–335, 1987.

VELOSO, Andres Rodriguez. **Estratégias de segmentação e posicionamento direcionadas para o mercado infantil**. 2008. Tese (Doutorado) – Universidade de São Paulo.

VERHOEF, Peter C *et al.* The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. **Decision Support Systems**, Elsevier, v. 34, n. 4, p. 471–481, 2003.

WEI, Jo-Ting; LIN, Shih-Yen; WU, Hsin-Hung. A review of the application of RFM model. **African Journal of Business Management**, v. 4, p. 4199–4206, 2010.

WIND, Yoram. Issues and advances in segmentation research. **Journal of marketing research**, SAGE Publications Sage CA: Los Angeles, CA, v. 15, n. 3, p. 317–337, 1978.

WITTEN, Ian H; FRANK, Eibe. Data mining: practical machine learning tools and techniques with Java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.

XU, Dongkuan; TIAN, Yingjie. A comprehensive survey of clustering algorithms. **Annals of Data Science**, Springer, v. 2, n. 2, p. 165–193, 2015.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. **IEEE Transactions on neural networks**, Ieee, v. 16, n. 3, p. 645–678, 2005.

ZAIANE, Osmar R *et al.* On data clustering analysis: Scalability, constraints, and validation. *In:* SPRINGER. PACIFIC-ASIA Conference on Knowledge Discovery and Data Mining. [*S.l.*: *s.n.*], 2002. P. 28–39.

ZENONE, Luiz Claudio. **CRM-Customer Relationship Management: gestão do relacionamento com o cliente e a competitividade empresarial**. [*S.l.*]: Novatec Editora, 2007.