CIÊNCIAS EMPRESARIAIS

ESCOLA SUPERIOR
POLITÉCNICO SETÚBAL

LAURA MIKLINA KURREIIA

Aplicação e comparação de métodos de segmentação de clientes (RFM, Coortes e Clusters) em PME

JÚRI

Presidente: Prof. Dr. Paulo Duarte Valente Almeida da Silveira, Instituto Politécnico de Setúbal

Orientador: Prof. Dr. Paulo Sérgio Ribeiro de Araújo Bogas, Instituto Politécnico de Setúbal

Vogal: Prof. Mário Luís Pereira Cravidão, Instituto Politécnico de Setúbal

Vogal: Prof.ª Dra. Sandra Cristina Dias Nunes, Instituto Politécnico de Setúbal

Vogal: Prof. Dr. David Alexandre Mendes Silva Simões, Instituto Politécnico de Setúbal

Novembro 2025

Agradecimentos

Dedico este trabalho ao meu marido — pelo apoio incondicional, pela fé inabalável em cada uma das minhas ideias. A sua presença constante dá-me segurança, coragem e a confiança necessária para seguir em frente, mesmo quando o caminho é difícil.

Agradeço também à minha família e amigos, pelo amor que sempre me acompanha e pela força que me transmitiram.

Expresso ainda o meu sincero reconhecimento aos docentes e orientadores que me acompanharam neste percurso académico, pela sua orientação atenta, exigente e generosa.

Resumo

Neste estudo avaliámos métodos de segmentação da base de clientes, normalmente usados apenas em grandes empresas com conjuntos extensos de dados (datasets), aplicando-os à base de clientes de uma pequena empresa, nomeadamente uma loja online especializada em acessórios de pele. O objetivo foi comparar estes métodos quanto à sua aplicabilidade no contexto das pequenas empresas para a otimização das estratégias de marketing.

A escolha dos três métodos fundamenta-se na sua complementaridade analítica: a simplicidade e eficácia operacional da análise RFM, a perspetiva temporal da análise de coortes e a profundidade comportamental da análise de clusters. Como dados iniciais foram utilizadas as transações efetuadas pelos clientes durante o ano de 2024. Como primeira fase, os dados foram sujeitos a um pré-processamento rigoroso. De seguida, foram calculadas as métricas essenciais de RFM (Recency, Frequency, Monetary) para cada cliente. Depois, foi realizada a análise de coortes, segmentando os clientes segundo o mês da sua primeira compra e calculando as taxas de retenção para cada coorte. Na quarta fase da análise, aplicou-se o algoritmo de agrupamento k-means (clustering).

A análise RFM permitiu classificar os clientes em três grupos com níveis distintos de atividade e valor económico. A análise de coortes revelou uma quebra acentuada nas taxas de retenção logo após a primeira compra, indicando a necessidade de estratégias de reativação. A análise de clusters identificou dez segmentos comportamentais com diferenças estatisticamente significativas, comprovando a sua robustez e relevância prática mesmo em contextos com recursos limitados.

Abstract

In this study, we evaluated customer base segmentation methods that are typically employed only by large companies with extensive datasets, applying them to the customer base of a small company, namely an online store specializing in leather accessories. The objective was to compare these methods in terms of their applicability within the context of small enterprises for the optimization of marketing strategies.

The selection of the three methods was grounded in their analytical complementarity: the simplicity and operational efficiency of RFM analysis, the temporal perspective of cohort analysis, and the behavioral depth of cluster analysis. The initial dataset comprised customer transactions carried out during the year 2024.

As a first step, the data underwent rigorous preprocessing. Subsequently, the key RFM metrics (Recency, Frequency, Monetary) were computed for each customer. Next, cohort analysis was conducted, segmenting customers according to the month of their first purchase and calculating retention rates for each cohort. In the fourth phase of the analysis, the k-means clustering algorithm was applied.

The RFM analysis enabled the classification of customers into three groups with distinct levels of activity and economic value. The cohort analysis revealed a sharp decline in retention rates immediately after the first purchase, highlighting the need for reactivation strategies. Finally, the cluster analysis identified ten behavioral segments with statistically significant differences, confirming its robustness and practical relevance even in resource-constrained contexts.

Índice

1. Introdução 7
1.1 Relevância do tema 7
1.2 Objetivos e tarefas 8
2. Revisão da literatura 10
2.1 Evolução da análise de clientes e da abordagem data-driven 10
2.2 Descrição de cada método 10
2.3 Aplicação empírica dos métodos de segmentação: panorama da literatura 13
3. Metodologia 16
3.1 Caracterização do objeto de estudo 16
3.2 Fontes de dados 17
3.3 Métodos e ferramentas de análise 18
3.4 Métricas de avaliação da eficácia dos métodos de segmentação 19
4. Aplicação empírica dos modelos de segmentação de clientes 24
4.1 Pré-processamento dos dados 24
4.2 Análise RFM 25
4.2.1 Cálculo das métricas 25
4.2.2 Implementação da análise RFM 25
4.2.3 Atribuição das pontuações RFM 26
4.2.4 Resultados da segmentação RFM 27
4.2.5 Avaliação da eficácia da segmentação RFM 30
4.3 Análise de coortes 31
4.3.1 Preparação e estruturação dos dados 31
4.3.2 Construção da matriz de retenção 32
4.3.3 Interpretação dos resultados 34
4.3.4 Avaliação da eficácia da análise de coortes 35
4.4 Análise de clusters 38
4.4.1 Escolha do número de clusters 38
4.4.2 Caracterização dos clusters 40
4.4.3 Avaliação da significância estatística das diferenças entre os clusters 43
4.4.4 Avaliação da eficácia da análise de clusters 45
4.5 Revisão comparativa dos métodos 46
5. Discussão dos resultados 50
5.1 Discussão dos resultados da análise RFM 50
5.2 Discussão dos resultados da análise de coortes 50
5.3 Discussão dos resultados da análise de clusters 50
5.4 Implicações práticas e teóricas 51
5.5 Limitações do estudo 52
5.6 Perspetivas para investigações futuras 53
Conclusão 55
Referências bibliográficas 56
Anexo 59

Índice de tabelas

Tabela 1. Análise comparativa dos métodos de segmentação 15
Tabela 2. Análise comparativa das métricas de eficácia 23
Tabela 3. Resultados dos modelos matemáticos BG-NBD e Gamma-Gamma 37
Tabela 4. Revisão comparativa dos métodos de segmentação 48

Índice de figuras

Figura 1. Exemplo de estrutura inicial dos dados de clientes carregados para análise 24
Figura 2. Exemplo de estrutura dos dados após renomeação e eliminação de duplicados 24
Figura 3. Cálculo das métricas RFM com pontuação por quartis 26
Figura 4. Distribuição dos clientes segundo as pontuações RFM 28
Figura 5. Segmentação de clientes baseada nas categorias RFM 29
Figura 6. Distribuição do RFM_Score 29
Figura 7. Mapa de calor. Formação de coortes com base na primeira compra dos clientes 33
Figura 8. Taxas de retenção após a primeira compra 34
Figura 9. Resultado do método do cotovelo 38
Figura 10 Resultado de Índice de Silhueta 40
Figura 11. Resultado da análise de clusters 41
Figura 12. Distribuição dos clusters quanto à aspetos essenciais do comportamento do cliente 44

Lista de acrónimos

ANOVA – análise de variância unifatorial

B2B – relação comercial entre empresas

CLV – valor vitalício do cliente

CRM – gestão de relacionamento com o cliente

ID – identificador

PCV – valor previsto do cliente

PME – pequena e média empresa

RFM – recência, frequência e valor monetário

UTM – parâmetro de rastreio de campanha

1. Introdução

1.1 Relevância do tema

As realidades contemporâneas da economia digital moldam uma tendência persistente para a incorporação de abordagens analíticas no processo de tomada de decisões de marketing (Haddadi & Hamidi, 2025). Num contexto de elevada concorrência e de preferências dos consumidores em constantes mudanças, a retenção de clientes existentes assume um papel estratégico fundamental para o crescimento sustentável das empresas. Neste cenário, a segmentação da base de clientes adquire especial relevância ao permitir não apenas uma identificação mais precisa das características dos grupos-alvo, como também a construção de modelos personalizados de interação com os mesmos (Ascarza et al., 2018).

Este trabalho dedica-se à aplicação e análise comparativa de três métodos quantitativos de segmentação de clientes: análise RFM, análise de coortes e análise de clusters no contexto da sua adaptabilidade às especificidades das pequenas e médias empresas. A originalidade do projeto reside na aplicação destes métodos a um conjunto real de dados de clientes de uma empresa concreta, o que permite avaliar não apenas a robustez teórica das abordagens, mas também a sua viabilidade prática em ambientes com acesso limitado a grandes volumes de informação. Considerando que as PME muitas vezes não dispõem de infraestruturas analíticas avançadas, mas recolhem regularmente dados básicos sobre transações e comportamento dos clientes, é essencial demonstrar que mesmo estas informações, quando adequadamente tratadas, podem servir de base para decisões estratégicas relevantes.

A literatura académica tem-se centrado na investigação sobre a implementação de métodos de segmentação em grandes empresas com acesso a grandes volumes de dados (big data) e com capacidade tecnológica instalada. A eficácia destas abordagens em contextos de PME permanece pouco estudada, o que perpetua a perceção de que estratégias orientadas por dados (data-driven) não são exequíveis neste segmento empresarial (Dalmaijer et al., 2022).

É importante sublinhar a relevância socioeconómica deste estudo. A adoção de soluções de marketing personalizadas aumenta a correspondência entre a oferta e as expectativas dos consumidores, promovendo relações mais sólidas e duradouras entre empresas e clientes. A longo prazo, tal contribui para reforçar a confiança no

pequeno comércio, melhorar a sua competitividade e fomentar o desenvolvimento sustentável do ecossistema empreendedor (Yıldız et al., 2023).

1.2 Objetivos e tarefas

O objetivo principal deste estudo consiste na avaliação comparativa de três metodologias de segmentação de clientes: análise RFM, análise de coortes e análise de clusters no que respeita à sua aplicabilidade no contexto das pequenas empresas. Este objetivo visa a otimização das estratégias de marketing e o reforço da eficácia das ações de retenção de clientes.

Para concretizar este objetivo, foram definidas as seguintes tarefas de investigação:

1. Proceder à análise teórica das abordagens selecionadas, descrevendo os seus fundamentos conceptuais, características algorítmicas e áreas de aplicação prática no domínio do marketing;
2. Implementar empiricamente cada método sobre a mesma amostra de clientes, assegurando assim a homogeneidade das condições analíticas e a comparabilidade dos resultados obtidos;
3. Realizar uma análise comparativa da eficácia de cada método com base em critérios previamente estabelecidos, incluindo a precisão da segmentação, o grau de homogeneidade dos grupos identificados, a previsibilidade do seu comportamento e a aplicabilidade prática dos dados gerados;
4. A partir da análise empírica, formular conclusões e recomendações práticas quanto à racionalidade da aplicação de cada um dos métodos considerados em contextos de recursos limitados típicos das pequenas empresas, para melhorar os mecanismos de fidelização de clientes e aumentar a rentabilidade das iniciativas de marketing.

O objeto de estudo é uma pequena empresa especializada na produção e comercialização de acessórios em pele. A atividade desta entidade caracteriza-se por operar num ambiente de recursos escassos, elevada concorrência e necessidade de utilização eficiente da informação disponível sobre os clientes. A análise baseia-se em dados empíricos recolhidos junto dos clientes da empresa, incluindo informações sobre a atividade transacional, indicadores financeiros e resposta às comunicações de marketing. Nos termos do contrato celebrado entre a empresa e os seus clientes, que exclui a partilha de dados pessoais com terceiros, todos os dados pessoais foram

anonimizados antes de serem utilizados na análise. Informações como nome, apelido, número de telefone, endereço de e-mail, país de residência e comunicações privadas não foram divulgadas nem integradas no estudo. A ausência destes dados não comprometeu a validade, exaustividade ou fiabilidade dos métodos analíticos aplicados, uma vez que os métodos de segmentação e comportamento foram desenvolvidos exclusivamente com base em variáveis transacionais e comportamentais anonimizadas.

Num contexto de défice de recursos, incluindo os financeiros, que são comuns para as pequenas empresas, reveste-se de particular importância a implementação de ferramentas que promovam a criação de uma base de clientes sustentável e a otimização das atividades de marketing. A comparação dos métodos RFM, de coortes e de clusters permite aferir a sua adequação prática em ambientes de negócio com recursos limitados. A utilização de uma amostra de dados comum possibilita uma avaliação objetiva da capacidade de cada abordagem para identificar os segmentos de consumidores mais promissores para uma atuação estratégica dirigida.

2. Revisão da literatura

2.1 Evolução da análise de clientes e da abordagem data-driven

Nos anos recentes, tem-se observado um desenvolvimento acelerado da análise de clientes, altamente associado à digitalização dos negócios, ao aumento da disponibilidade de dados e à necessidade de melhorar a eficácia das decisões de marketing. Uma das transformações mais significativas no marketing hoje é a transição para estratégias baseadas em dados (data-driven marketing), permitindo que empresas de diferentes dimensões tomem decisões de gestão fundamentadas em informações objetivas provenientes de dados comportamentais e transacionais (Ascarza et al., 2018).

A abordagem data-driven caracteriza-se pela utilização de modelos analíticos e algoritmos para extrair informação de grandes volumes de dados de clientes, com o intuito de construir estratégias de marketing personalizadas, otimizar a interação com os consumidores e melhorar a eficácia em todas as etapas do percurso do cliente. Este modelo permite prever com maior precisão as necessidades dos consumidores, aumentar a relevância das comunicações, reduzir os custos de aquisição e, substancialmente, melhorar os níveis de retenção e fidelização (Haddadi & Hamidi, 2025).

A evolução da análise de clientes e a implementação de abordagens data-driven são fatores determinantes para a sustentabilidade e competitividade das pequenas empresas. O fácil acesso às ferramentas analíticas acessíveis proporciona às micro e pequenas empresas a oportunidade de aplicar métodos anteriormente reservados às grandes corporações, alcançando resultados significativos na retenção de clientes, no aumento do seu valor e na otimização geral das atividades de marketing (Dalmaijer et al., 2022).

2.2 Descrição de cada método

No âmbito do presente estudo, são analisados três métodos quantitativos de segmentação de clientes: a análise RFM, a análise de coortes e a análise de clusters. Cada um destes métodos possui características, vantagens e limitações distintas, o que determina a sua adequação a diferentes contextos e realidades empresariais.

Segundo os autores Heldt et al. (2021), a análise RFM constitui uma das metodologias de segmentação comportamental mais amplamente utilizadas, baseando-se na avaliação quantitativa da atividade dos clientes segundo três parâmetros principais:

Recency (Recência): refere-se ao tempo decorrido desde a última interação do cliente com a empresa, como uma compra ou contacto. Em geral, quanto mais recente for a interação, maior a probabilidade de nova conversão.

Frequency (Frequência): indica a regularidade com que o cliente realiza compras dentro de um determinado intervalo temporal. Transações frequentes sugerem um elevado nível de envolvimento e lealdade.

Monetary (Valor Monetário): mede o montante total gasto pelo cliente. Valores elevados apontam para uma importância estratégica do cliente para o negócio.

Com base nestas três métricas, os clientes são classificados segundo diferentes perfis comportamentais. Desta forma, é possível identificar segmentos distintos, por exemplo: "os mais ativos e rentáveis", "valiosos, mas inativos", "clientes novos com potencial", ou "clientes em risco de abandono" (Mena et al., 2024). Esta forma de segmentação é amplamente aplicada em sectores como o comércio eletrónico, o retalho tradicional ou os serviços financeiros, dada a sua capacidade de estruturar estratégias de marketing personalizadas com elevada precisão (Heldt et al., 2021).

A simplicidade de implementação e interpretação da análise RFM, aliada à sua rapidez, torna-a particularmente atrativa para empresas que procuram obter rapidamente insights comportamentais sem investir grandes recursos. A sua versatilidade permite aplicá-la em diversos sectores. No entanto, como salientado por Heldt et al. (2021), o modelo RFM apresenta limitações significativas: opera exclusivamente com dados transacionais, não incorporando variáveis de produto ou características comportamentais mais amplas. Além disso, ignora a dimensão temporal e as mudanças nos padrões de comportamento dos clientes, o que reduz a sua aplicabilidade na análise de tendências de longo prazo e no desenvolvimento de estratégias de fidelização.

A análise de coortes é uma metodologia de segmentação temporal que agrupa os clientes com base na data da sua primeira interação com a empresa (registo, primeira compra, etc.). Cada coorte é analisada ao longo do tempo, permitindo observar o comportamento e os níveis de retenção dos clientes nas diversas fases do seu ciclo de vida (Orduz, 2025).

Entre as suas vantagens, autores como Fedushko & Ustyianovych (2022) destacam a capacidade de evidenciar mudanças temporais nos padrões comportamentais, avaliar a eficácia de alterações estratégicas e identificar etapas críticas no ciclo de vida do cliente, incluindo pontos de rotura. Contudo, a análise de coortes apresenta limitações relevantes: exige um volume considerável de dados para assegurar robustez estatística; a sua implementação pode revelar-se demorada para pequenas empresas, sobretudo na ausência de uma infraestrutura analítica automatizada; e nem sempre permite uma segmentação detalhada dentro de cada coorte, restringindo-se frequentemente a indicadores agregados (Fedushko & Ustyianovych, 2022).

A análise de clusters (também conhecida como análise de agrupamento de dados) é uma técnica de segmentação baseada na agregação de clientes em grupos (clusters) com base no grau de similaridade entre determinadas variáveis. O algoritmo k-means, um dos mais amplamente utilizados, distribui os clientes por k clusters de forma a minimizar as diferenças internas e maximizar as diferenças entre grupos (Celebi et al., 2013). John et al. (2024) destacam que esta abordagem permite considerar simultaneamente um vasto leque de variáveis: desde características demográficas e padrões de comportamento, até respostas a campanhas de marketing e canais de aquisição, etc. Assim, torna-se possível definir perfis multidimensionais únicos que refletem não apenas o histórico de compras, mas também o contexto mais amplo de interação com a marca. Esta abordagem contribui para a construção de um modelo de segmentação de alta precisão, que permite formular propostas personalizadas, relevantes para grupos específicos de consumidores, bem como identificar padrões comportamentais ocultos e formar grupos de clientes não triviais com base em semelhanças objetivas (Celebi et al., 2013; Harish & Malathy, 2023).

Uma das principais vantagens da análise de clusters reside na sua aptidão para lidar com dados multidimensionais e heterogéneos, incluindo comportamento de consumo, variáveis sociodemográficas, fontes de tráfego e níveis de envolvimento (Abdulhafedh, 2021). A flexibilidade e capacidade de adaptação deste método tornam-no aplicável a diversos contextos empresariais: desde campanhas de publicidade segmentada até serviços personalizados (Abdulhafedh, 2021; Dalmaijer et al., 2022). Segundo Tabianan et al. (2022), os resultados obtidos através da análise de clusters são sensíveis à escala dos dados e à qualidade das variáveis utilizadas, sendo

essencial uma preparação prévia adequada. Para além disso, a interpretação dos resultados e a validação dos modelos exigem recursos computacionais e competências analíticas específicas (Abdulhafedh, 2021).
Assim, cada um dos métodos analisados possui potencial significativo no domínio da análise de clientes, podendo ser de grande utilidade para pequenas e médias empresas, consoante os seus objetivos, recursos disponíveis e maturidade analítica.

2.3 Aplicação empírica dos métodos de segmentação: panorama da literatura

Nos últimos anos, tem-se notado um crescimento expressivo no número de estudos dedicados aos métodos de segmentação de clientes (Ascarza et al., 2018).
A revisão da literatura recente permite identificar tanto práticas bem-sucedidas na aplicação destas metodologias como lacunas relevantes, em particular a escassez de investigações empíricas centradas na aplicação destes métodos no contexto das pequenas empresas.
Os métodos de segmentação de clientes têm evoluído significativamente, sendo utilizados tanto em sectores tradicionais como em ambientes tecnologicamente avançados, nomeadamente através da integração com IA, modelos de ensemble e estratégias de personalização (Haddadi & Hamidi, 2025).
A análise RFM continua a ser uma ferramenta básica e de fácil interpretação, mas estudos recentes evidenciam a sua elevada eficácia quando complementada com componentes temporais ou comportamentais adicionais (como o RFM temporal ou o modelo RFM/P) (Heldt et al., 2021). A análise de coortes revelou a sua relevância prática para a avaliação da retenção e do comportamento de longo prazo dos clientes, sobretudo em contextos marcados por instabilidade e choques de mercado (Fedushko & Ustyianovych, 2022). A análise de clusters, por sua vez, tem demonstrado uma forte capacidade de explorar bases de dados heterogéneas, permitindo a criação de estratégias altamente personalizadas com elevado grau de precisão (John et al., 2023).
Apesar da intensa produção científica nesta área, a aplicação destes métodos em micro e pequenas empresas continua pouco explorada, o que revela um potencial significativo para futuras investigações. A literatura analisada confirma que a análise RFM, sendo um dos métodos mais acessíveis e intuitivos, tem sido amplamente utilizada devido à sua simplicidade, versatilidade e à capacidade de segmentar

rapidamente a base de clientes segundo critérios de atividade e rentabilidade. Este método permite identificar com eficácia segmentos valiosos com base em métricas como a recência do último contacto, frequência de compras e valor total gasto – o que o torna especialmente atrativo para pequenas empresas com recursos analíticos limitados.

Contudo, apesar do seu valor aplicado, a limitação a apenas três indicadores impede a obtenção de uma visão mais abrangente dos comportamentos dos clientes, como as preferências, sazonalidade, respostas a estímulos de marketing ou mudanças nos padrões comportamentais ao longo do tempo. Além disso, extensões mais sofisticadas do modelo tradicional, como o RFM/P ou as variantes temporais, ainda não têm ampla utilização entre as pequenas empresas, facto que se deve à falta de desenvolvimento metodológico e à escassa familiaridade, por parte dos profissionais de marketing, com os benefícios dessas abordagens mais avançadas (Heldt et al., 2021; Mena et al., 2024; Gordini & Veglio, 2017; Osuna et al., 2016).

Em contraste, a análise de coortes permite uma perspetiva dinâmica dos clientes, tornando possível avaliar alterações no seu envolvimento ao longo do tempo e identificar fatores que influenciam diretamente os níveis de retenção. O seu principal mérito reside na capacidade de detetar padrões de comportamento a longo prazo e aferir a eficácia das estratégias de marketing - o que é particularmente relevante para empresas orientadas para relações duradouras com os clientes. No entanto, a implementação desta abordagem exige uma quantidade significativa de dados e uma estruturação rigorosa da informação, o que pode representar um desafio para pequenas empresas. Adicionalmente, a complexidade da interpretação dos resultados e a necessidade de uma definição clara dos eventos que originam cada coorte limitam a sua aplicabilidade generalizada. É relevante referir que a literatura existente praticamente não contempla o potencial da análise de coortes em pequenas empresas, ao passo que a sua combinação com outros métodos de segmentação poderia aumentar significativamente a precisão das previsões e a fundamentação das decisões de gestão (Orduz, 2025; Fedushko & Ustyianovych, 2022; Yıldız et al., 2023).

Por seu lado, a análise de clusters constitui uma ferramenta mais flexível e escalável, permitindo segmentar clientes com base em múltiplas variáveis e revelar padrões comportamentais ocultos que não seriam identificáveis através de metodologias mais lineares. Com o suporte de algoritmos de aprendizagem automática, como o k-means, é possível implementar uma personalização com elevada precisão das ofertas de

marketing, o que contribui para o aumento do envolvimento dos clientes e para a melhoria das taxas de conversão. No entanto, esta abordagem traz também desafios metodológicos, entre os quais se destaca a necessidade de definir o número de clusters, bem como a sensibilidade dos resultados relativa à qualidade dos dados de entrada (Tabianan et al., 2022; John et al., 2023; Harish & Malathy, 2023; Haddadi & Hamidi, 2025; Dalmaijer et al., 2022; Abdulhafedh, 2021). Na Tabela 1 é apresentada uma análise comparativa dos métodos.

Tabela 1. Análise comparativa dos métodos de segmentação

Método	Vantagens	Limitações	Lacunas na investigação
Análise RFM	Simplicidade, acessibilidade, eficácia no curto prazo	Limitação de dados (só três métricas), desconsidera comportamentos do cliente	Escassez de estudos combinados com outros métodos de segmentação
Análise de coortes	Acompanha tendências de longo prazo, adequada para análise de retenção	Exige grandes volumes de dados, complexidade na interpretação dos resultados	Ausência de estudos focados em pequenas empresas, dificuldade na definição das coortes
Análise de Clusters (k-means)	Identifica padrões ocultos, flexibilidade na segmentação	Dependência da inicialização, sensibilidade ao ruído e aos outliers	Pouca investigação sobre integração com variáveis temporais

Fonte: autoria própria.

A necessidade de adaptar os modelos às exigências das pequenas empresas permanece atual, exigindo investigações aplicadas e empíricas adicionais que tenham em conta as limitações de recursos e a especificidade do funcionamento em contextos de elevada incerteza no ambiente de mercado.

3. Metodologia

3.1 Caracterização do objeto de estudo

O objeto deste estudo é uma pequena empresa especializada na produção e comercialização de acessórios em pele (capas para computadores portáteis, bolsas para dispositivos eletrónicos, itens funcionais para espaços de trabalho, etc).

Uma das principais características distintivas desta empresa é a existência de uma unidade de produção própria, o que lhe confere uma grande flexibilidade para executar encomendas personalizadas e adaptar rapidamente a gama de produtos de acordo com as preferências do seu público-alvo. Todas as etapas da interação com o cliente, desde a realização da encomenda até à definição de detalhes e à entrega, são efetuadas exclusivamente através de canais online, o que reduz significativamente os custos operacionais associados à manutenção de pontos de venda físicos e permite uma maior expansão geográfica da atividade comercial.

Com base na análise interna de vendas dos últimos anos, a empresa atua num segmento de mercado que tem apresentado uma tendência de crescimento estável. Esse crescimento é impulsionado por uma procura crescente por produtos de alta qualidade e pelo desejo dos consumidores de adquirir artigos personalizados. Os produtos tornam-se especialmente atrativos devido à possibilidade de personalização (como a gravação de iniciais, a escolha da cor do couro ou da configuração do artigo), o que acrescenta valor adicional para o cliente. Por outro lado, este setor é altamente competitivo, sobretudo entre os vendedores online que oferecem design exclusivo e materiais premium. A elevada estrutura de custos, associada ao uso de couro natural turco e ao trabalho artesanal, exige da empresa um afinamento preciso das estratégias de marketing e uma gestão eficiente da relação com o cliente.

A empresa opera com uma estrutura funcional compacta e horizontal, o que garante agilidade e foco nas áreas-chave: marketing, logística, produção e vendas. A equipa é composta por um profissional de marketing, um responsável logístico, três gestores comerciais e quinze colaboradores diretamente envolvidos na produção. A ausência de lojas físicas permite focar os recursos para o desenvolvimento de serviços digitais, a otimização das operações logísticas e a realização de campanhas de marketing direcionadas em ambiente digital.

O público-alvo da empresa inclui consumidores entre os 20 e os 50 anos, com rendimento médio ou superior, exigentes em relação à qualidade, ao design e à exclusividade dos produtos. Este grupo valoriza a individualidade, tem interesse em

acessórios modernos e funcionais, e está disposto a investir em artigos que se alinhem com as suas preferências estéticas e práticas. Trata-se de um público composto por homens e mulheres com hábitos digitais consolidados, que privilegiam a experiência de compra online.

3.2 Fontes de dados

No âmbito deste estudo, prevê-se a utilização de um conjunto de dados que reflete os principais aspetos da interação dos clientes com a empresa, permitindo uma análise abrangente com recurso a diferentes métodos de segmentação. A integração dos dados provenientes do sistema de CRM da empresa garante a completude do panorama analítico disponível. Todos os dados utilizados na presente análise foram extraídos do sistema CRM Bitrix24, que constitui a principal infraestrutura digital da empresa. O Bitrix24 dispõe de capacidades abrangentes de gestão da relação com o cliente, integrando num único ambiente: registo e histórico detalhado de clientes, dados transacionais (valor das compras, frequência e datas de transação), parâmetros UTM e informações de origem de tráfego, funil de vendas completo, desde o lead até à conversão, integração com campanhas publicitárias (Meta Ads, Google Ads), automatização de tarefas e registo de interações com o cliente. Esta plataforma assegura a consistência dos dados, a rastreabilidade das ações de marketing e a possibilidade de acompanhar o percurso completo do cliente, o que constitui uma vantagem significativa para a aplicação das metodologias de segmentação estudadas. Entre os parâmetros mais relevantes para a análise, disponibilizados pelo CRM, destacam-se os identificadores dos clientes, essenciais para o acompanhamento de compras repetidas e para a análise da dinâmica da interação ao longo do tempo. Antes da transferência dos dados para tratamento analítico, todos os identificadores únicos, incluindo os ID dos clientes, foram submetidos a um processo de anonimização por parte da empresa, assegurando a correspondência inequívoca entre os registos anonimizados e os dados originais da base interna, sem comprometer a integridade das ligações analíticas. Este procedimento garante a total eliminação de qualquer risco de divulgação ou difusão de dados pessoais, preservando simultaneamente a exatidão das correlações analíticas. Como resultado, nenhuma transação, sessão ou utilizador associado foi perdido ou excluído da análise, o que assegura a fiabilidade dos resultados obtidos nas segmentações e na modelação comportamental. Um parâmetro particularmente importante - marcadores

temporais – indicam-nos as datas de abertura e fecho dos negócios, permitindo avaliar a duração do ciclo de vida do cliente e identificar possíveis pontos críticos no processo de conversão. Igualmente relevante é a informação sobre o canal de origem da encomenda (por exemplo, Shopify ou plataformas externas como redes sociais), uma vez que permite avaliar a eficácia dos diferentes canais de aquisição de clientes.

Adicionalmente, foram extraídos do CRM dados transacionais que caracterizam o comportamento dos clientes em termos de frequência, volume e estrutura das compras. A análise desses dados permite não só segmentar os clientes com base no seu nível de atividade e valor monetário, como também identificar padrões específicos de preferência. Dados como a frequência de compra e o valor médio das transações possibilitam avaliar a contribuição individual de cada cliente para o volume de receitas da empresa, ao passo que a data da última compra assume um papel crucial na construção do modelo RFM. A informação sobre a composição dos pedidos, incluindo os tipos de produtos adquiridos (por exemplo, capas, malas, porta-chaves), permite uma melhor definição dos interesses dos clientes e facilita a criação de segmentos temáticos na fase de análise de clusters.

O estudo incluirá dados sobre as reações dos clientes às ações de marketing, registados através dos parâmetros UTM obtidos por meio do Meta Ads e do Google Ads/Google Analytics. Estes dados permitem rastrear os canais de aquisição de clientes, avaliar a eficácia das campanhas publicitárias e determinar o grau de envolvimento do público.

3.3 Métodos e ferramentas de análise

O modelo RFM é amplamente reconhecido na literatura científica como uma ferramenta eficaz de segmentação comportamental (Heldt et al., 2021; Mena et al., 2024). Através da aplicação do modelo RFM, os clientes serão distribuídos em grupos típicos, incluindo segmentos de consumidores altamente ativos e valiosos, clientes regulares com menor volume de gastos, bem como os que não demonstraram atividade durante um longo período e que exigem estímulos adicionais de marketing. Esta abordagem permite não só identificar rapidamente as áreas prioritárias para retenção de clientes, como também fornece uma base sólida para análises mais avançadas em modelos subsequentes.

A análise de coortes será utilizada para estudar o comportamento dos clientes ao longo do tempo, o que é particularmente relevante para avaliar os níveis de lealdade

e retenção em diferentes intervalos temporais. A agregação de clientes com base na data do primeiro contacto com a empresa permitirá identificar diferenças entre coortes, revelar padrões temporais e avaliar a eficácia das iniciativas de marketing realizadas em distintos períodos (Fedushko & Ustyianovych, 2022). Este método possibilita uma análise detalhada das tendências comportamentais, bem como a avaliação das reações de diferentes grupos de clientes a estímulos diversos, incluindo promoções, descontos e alterações na oferta de produtos.

Para uma segmentação mais profunda e para identificar padrões ocultos no comportamento do cliente, será aplicado o método de análise de clusters, utilizando o algoritmo k-means. Este método permite agrupar clientes com base numa multiplicidade de variáveis, como a frequência e o valor das compras, indicadores comportamentais e respostas a campanhas de marketing. A análise de clusters não exige regras predefinidas e permite a construção de um modelo de segmentação flexível que reflete a estrutura multidimensional da base de clientes (Dalmaijer et al., 2022; Celebi et al., 2013). A avaliação comparativa da eficácia dos métodos será realizada com base em vários critérios.

Importa salientar que, no âmbito do presente estudo, o método de análise ABC foi deliberadamente excluído, apesar da sua ampla notoriedade na prática do marketing. Este método, baseado exclusivamente no valor financeiro do cliente, não considera os aspetos comportamentais nem temporais, o que limita significativamente a sua capacidade analítica e o torna menos pertinente para os objetivos deste projeto. Em contrapartida, os métodos RFM, análise de coortes e análise de clusters oferecem uma compreensão mais holística da base de clientes e permitem a formulação de recomendações estratégicas mais fundamentadas.

A implementação dos métodos referidos será realizada em linguagem de programação Python, que dispõe de um vasto conjunto de bibliotecas para tratamento de dados, visualização e construção de modelos de segmentação. Esta escolha assegura a precisão da análise, a reprodutibilidade dos resultados e a adaptabilidade das soluções às futuras necessidades do negócio.

3.4 Métricas de avaliação da eficácia dos métodos de segmentação

Para uma avaliação objetiva da eficácia dos diferentes métodos de segmentação de clientes neste estudo, assume particular importância a utilização de métricas

formalizadas que permitam comparar os resultados em termos da qualidade da delimitação dos grupos de clientes.

Escala de comparação dos métodos de segmentação

Precisão da segmentação

Avaliada com base no número de segmentos identificados e no grau de diferenciação entre eles (Harish & Malathy, 2022).

Valor médio do ciclo de vida do cliente (CLV) por segmento

Indicador da eficácia dos segmentos formados em termos do valor gerado para a empresa (Gómez-Vargas et al., 2025).

Capacidade de previsão do valor futuro do cliente (PCV)

Mede a capacidade de cada método prever o valor futuro do cliente, tendo em conta os modelos transacionais e comportamentais. Enquanto o CLV reflete o valor histórico, o PCV antecipa a contribuição potencial futura dos clientes para o volume de negócios da empresa, permitindo a otimização das estratégias de retenção e investimento nos segmentos com maior potencial de crescimento (Mena et al., 2024).

A precisão da segmentação representa um critério composto que inclui tanto o grau de diferenciação entre os segmentos quanto o nível de coerência dentro de cada grupo. Isto significa que uma segmentação eficaz deve não apenas garantir uma clara distinção entre os clusters, mas também demonstrar que os consumidores dentro do mesmo segmento partilham parâmetros comportamentais semelhantes, tais como frequência de compras, valor médio da transação, resposta a campanhas de marketing ou padrão de recompra (Harish & Malathy, 2022).

No âmbito da avaliação quantitativa da precisão, será utilizado o índice de silhueta, que permite determinar o grau de compacidade e de separação dos segmentos. Valores do índice próximos de 1 indicam elevada homogeneidade e uma estrutura de segmentação bem definida (Dalmaijer et al., 2022). No contexto da análise RFM, este critério pode ser avaliado pela capacidade de distinguir clientes com base nas métricas de Recência, Frequência e Valor Monetário. Importa ter em consideração até que ponto os grupos identificados apresentam diferenças na frequência de interações e no valor do rendimento total. No caso da análise de clusters, a precisão é determinada pelo grau de diferenciação entre os clusters, refletido nos indicadores de distância intercluster, bem como pela homogeneidade no interior dos próprios clusters. A análise de coortes, por sua vez, é avaliada com base nas diferenças de

comportamento entre clientes que realizaram a primeira compra em períodos distintos, permitindo identificar mudanças dinâmicas no envolvimento e na retenção.
A avaliação quantitativa será complementada com uma análise qualitativa dos perfis comportamentais dos segmentos, o que permitirá determinar a aplicabilidade de cada método na personalização de estratégias de marketing. Uma elevada precisão de segmentação correlaciona-se diretamente com a eficácia das estratégias de marketing: segmentos claramente definidos e homogéneos facilitam a criação de ofertas direcionadas, promovendo o aumento da taxa de conversão, da satisfação do cliente e a redução dos custos de aquisição (Harish & Malathy, 2022).
Assim, a precisão da segmentação pode ser considerada não apenas como um critério formal de avaliação metodológica, mas também como um indicador estratégico da qualidade da gestão de relacionamento com clientes, especialmente em contextos de recursos limitados, como é o caso das pequenas e médias empresas.

O indicador *Customer Lifetime Value (CLV)* determina o valor que um cliente representa ao longo de todo o seu relacionamento com a marca. Esta métrica é fundamental tanto para a gestão estratégica como operacional, pois permite estimar o lucro total que um cliente pode gerar durante o seu ciclo de vida ativo e, consequentemente, orienta decisões relacionadas à afetação orçamental de marketing, estratégias de retenção e avaliação da rentabilidade dos segmentos de clientes (Gómez-Vargas et al., 2025; Haddadi & Hamidi, 2025).
Conceitualmente, o CLV reflete não apenas a ocorrência de compras repetidas, mas também a sua frequência, volume financeiro e duração da relação com a empresa. O cálculo baseia-se em dados transacionais reais, incluindo variações no comportamento de compra ao longo do tempo, sazonalidade da procura e resposta a estímulos promocionais. Esta abordagem permite uma adaptação mais precisa do modelo à realidade da base de clientes, sobretudo no contexto das pequenas empresas, onde cada unidade de dados tem impacto significativo na tomada de decisões (Haddadi & Hamidi, 2025).
Na análise RFM, o CLV pode ser utilizado para refinar a classificação dos clientes com base no seu valor agregado, especialmente em segmentos com elevada frequência e volume de compras. Na análise de coortes, o CLV permite observar como o valor dos clientes varia conforme o período da sua primeira compra, revelando momentos estratégicos bem-sucedidos e tendências de longo prazo. Já na análise de clusters, o

CLV pode ser integrado como métrica adicional na formação dos segmentos, destacando a rentabilidade a longo prazo de cada grupo e ajudando a priorizar os públicos-alvo.

Num contexto contemporâneo, caracterizado por uma concorrência intensa e por comportamentos dos consumidores em constante mudança, a análise preditiva assume uma importância crescente, nomeadamente a avaliação do valor preditivo do cliente - *Prediction Customer Value (PCV)*. Ao contrário do CLV, que mede o valor histórico do cliente, o PCV concentra-se na contribuição potencial futura, promovendo assim uma abordagem proativa na gestão das estratégias de relacionamento (Mena et al., 2024).

O PCV representa uma estimativa do contributo esperado de um cliente para as receitas da empresa num horizonte temporal definido, baseando-se tanto no comportamento atual quanto histórico, bem como em variáveis adicionais como características demográficas, comportamentais e contextuais. Ao contrário de métricas fixas, o PCV é construído com base em modelos preditivos que requerem o uso de técnicas de machine learning, análise de séries temporais ou modelação estatística (Haddadi & Hamidi, 2025; Mena et al., 2024).

A principal função estratégica do PCV reside no suporte à tomada de decisões orientadas para a maximização da rentabilidade de marketing. O conhecimento do valor preditivo permite priorizar clientes com alto potencial de contribuição futura, o que é particularmente relevante em contextos de restrição orçamental. Com base no PCV, as empresas podem alocar recursos de forma mais racional, evitando gastos desnecessários com segmentos de baixo retorno esperado e concentrando os esforços na retenção dos clientes mais promissores. A eficácia desta abordagem é confirmada por estudos que demonstram a elevada precisão e aplicabilidade operacional das técnicas preditivas na segmentação e fidelização de clientes (Gómez-Vargas et al., 2025).

A aplicabilidade metodológica do PCV é observável em todas as abordagens de segmentação analisadas. No contexto da análise RFM, o valor preditivo pode refinar os perfis de clientes altamente ativos e frequentes. Na análise de coortes, o PCV permite comparar o valor futuro esperado entre diferentes coortes, identificando os grupos mais promissores conforme o momento da sua aquisição. Na análise de clusters, o PCV pode funcionar como critério adicional para avaliação da rentabilidade

a longo prazo de cada grupo, bem como orientar a criação de cenários personalizados de interação com base em padrões comportamentais comuns. A comparação entre as métricas de eficácia é apresentada na Tabela 2.

Tabela 2. Análise comparativa das métricas de eficácia

Métrica	Análise RFM	Análise de coortes	Análise de clusters (k-means)
Precisão da segmentação	Avaliação do número de segmentos e da sua diferenciação	Segmentação com base na data da primeira interação	Classificação com base na semelhança entre dados
CLV (Customer Lifetime Value)	Valor médio dos clientes	Previsão com base nas variações entre grupos de coortes	Previsão considerando o valor histórico
PCV (Prediction Customer Value)	Previsão do valor futuro do cliente	Avaliação do valor futuro por coorte	Previsão com base no comportamento e nos clusters identificados

Fonte: autoria própria.

Assim, o objetivo do presente estudo consiste numa avaliação comparativa abrangente de três metodologias de segmentação de clientes: a análise RFM, a análise de clusters e a análise de coortes, no que diz respeito à sua aplicabilidade no contexto das pequenas empresas. Esta análise visa, por sua vez, a otimização das estratégias de marketing e o aumento da eficácia das medidas de retenção de clientes numa empresa de pequena dimensão especializada na produção de artigos em pele. Os métodos selecionados permitem não apenas segmentar os clientes, mas também prever o seu comportamento futuro, o que contribui para um ajustamento mais preciso das estratégias e ações de marketing.

4. Aplicação empírica dos modelos de segmentação de clientes

4.1 Pré-processamento dos dados

Na fase inicial do tratamento dos dados, procede-se à sua limpeza e preparação preliminares. Em primeiro lugar, os nomes das colunas são uniformizados com o objetivo de padronizar a nomenclatura e facilitar as etapas seguintes da análise. Seguidamente, são eliminados duplicados, bem como registos com dados criticamente incompletos, de forma a aumentar a fiabilidade dos resultados finais.

É dada especial atenção à deteção e remoção de outliers. Por exemplo, valores incorretos como preços negativos são excluídos, uma vez que podem distorcer os cálculos estatísticos.

Por fim, são geradas variáveis auxiliares, como o mês de fecho da transação e a data da primeira compra de cada cliente, que são parâmetros essenciais para a realização da análise de coortes. Nas Figuras 1 e 2 são apresentados excertos do output gerado.

sample.ipynb | sample: Out 0

5 rows | 5 rows × 29 columns

	Status	Manager	Source	Price	CustomerID	OrderDate	CloseDate	Tech accessories	Wallets	Mouse pads	...
0	Ready for shipment	Diana	Marketplace	31.045279	Client 1	2024-04-24 15:38:33	2024-04-28	NaN	NaN	NaN	...
1	Ready for shipment	Diana	Shopify	24.566474	Client 2	2024-04-24 12:03:55	2024-04-28	NaN	NaN	1.0	...
2	Delivered	Diana	Marketplace	28.901734	Client 3	2024-04-23 11:21:20	2024-04-23	NaN	NaN	NaN	...
3	Ready for shipment	Olga	Shopify	24.566474	Client 4	2024-04-24 17:47:18	2024-04-27	NaN	NaN	NaN	...
4	Delivered	Olga	Shopify	54.672447	Client 5	2024-04-20 10:14:41	2024-04-22	NaN	NaN	NaN	...

Figura 1. Exemplo de estrutura inicial dos dados de clientes carregados para análise
Fonte: output Python

```
Index: 4484 entries, 0 to 4493
Data columns (total 29 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Status                  4484 non-null   object
 1   Manager                 4484 non-null   object
 2   Source                  4484 non-null   object
 3   Price                   4484 non-null   float64
 4   CustomerID              4484 non-null   object
 5   OrderDate               4484 non-null   datetime64[ns]
 6   CloseDate               4484 non-null   datetime64[ns]
 7   Tech accessories        143 non-null    float64
 8   Wallets                 160 non-null    float64
 9   Mouse pads              153 non-null    float64
 10  Cosmetic bag            8 non-null      float64
 11  Organizers for cables   68 non-null     float64
 12  Bands for Apple Watch   216 non-null    float64
 13  AirPods cases           229 non-null    float64
 14  Laptop and tablet sleeves  1161 non-null  float64
 15  Passport covers         171 non-null    float64
 16  Cardholders             159 non-null    float64
 17  Waist bags              287 non-null    float64
 18  Laptop bags             565 non-null    float64
 19  Eyeglass cases          207 non-null    float64
 20  Notebook covers         75 non-null     float64
```

Figura 2. Exemplo de estrutura dos dados após renomeação e eliminação de duplicados
Fonte: output Python

4.2 Análise RFM

4.2.1 Cálculo das métricas

O passo seguinte consiste no cálculo das principais métricas RFM:

1) Recency – calculado como o número de dias desde a última compra do cliente até à data da análise;
2) Frequency – definido como o número total de transações realizadas pelo cliente;
3) Monetary – corresponde à receita total gerada por cada cliente.

$$Recency_i = Data_{\text{análise}} - Data_{\text{última compra do cliente } i}$$

$$Frequency_i = \sum_{t=1}^{T} Compras_{i,t}$$

$$Monetary_i = \sum_{t=1}^{T} Valor_{i,t}$$

Variáveis

1. i – índice do cliente
2. T – número total de transações observadas para o cliente i
3. $\text{Data}_{\text{análise}}$ – data de referência usada no estudo (por exemplo, 31/12/2024)
4. $\text{Data}_{\text{última compra do cliente } i}$ – data da última transação do cliente i
5. $\text{Compras}_{i,t}$ – número de transações do cliente i no instante t
6. Valor $_{i,t}$ – valor monetário da transação t do cliente i

4.2.2 Implementação da análise RFM

Para a classificação dos clientes, é utilizado um sistema de pontuação baseado na divisão em quantis de cada métrica. A cada um dos três parâmetros é atribuído um valor entre 1 e 5, sendo que uma pontuação mais elevada representa um comportamento mais favorável em termos de valor do cliente.

Em seguida, é construído um segmento RFM combinando os três valores numa sequência de três dígitos, por exemplo, 543. Adicionalmente, calcula-se o indicador agregado RFM Score, que corresponde à soma das três pontuações e serve como medida integrada do nível de lealdade do cliente (Heldt et al., 2021; Gómez-Vargas et al., 2025). O resultado da execução do código está representado na Figura 3.

	CustomerID	Recency	Frequency	Monetary	R_score	F_score	M_score	RFM_Segment	RFM_Score
0	Client 1	248	1	31.045279	2	2	1	221	5
1	Client 10	252	2	78.640000	2	4	3	243	9
2	Client 100	253	1	65.680000	2	2	2	222	6
3	Client 1000	312	1	24.500000	1	2	1	121	4
4	Client 1001	331	1	87.260000	1	2	3	123	6

Figura 3. Cálculo das métricas RFM com pontuação por quartis
Fonte: output Python

Recebemos, como resultado, uma tabela de métricas RFM para cada cliente, com os seguintes campos:

Recency – número de dias desde a última compra;

Frequency – número total de compras efetuadas durante o período analisado;

Monetary – despesa total do cliente com a marca;

R_score, F_score, M_score – pontuações atribuídas a cada métrica numa escala de 1 (valor mais baixo) a 5 (valor mais elevado);

RFM_Segment – concatenação das três pontuações, refletindo o perfil comportamental do cliente;

RFM_Score – soma das três pontuações (RFM_Score=*R*score+*F*score+*M*score), funcionando como um indicador agregado do valor do cliente.

4.2.3 Atribuição das pontuações RFM

A métrica R_score varia entre 1 e 2 nas primeiras linhas da amostra, o que indica que muitos clientes não realizam compras há bastante tempo.

A F_score apresenta predominantemente valores entre 2 e 4, sugerindo que alguns clientes realizaram mais do que uma compra ao longo do tempo.

A M_score atinge o valor de 3 para certos clientes, posicionando-os nos 2.º–3.º quantis em termos de gasto total, o que indica uma despesa média.

Segmentação com base em RFM_Segment e RFM_Score

Exemplos de segmentos:

221 (Recency = 2, Frequency = 2, Monetary = 1): clientes com baixa frequência e valor monetário reduzido, mas que compraram relativamente recentemente.

243 (2-4-3): clientes com frequência e valor médio, cuja última compra ocorreu há um tempo intermédio. Representam um grupo de interesse como público leal.

O RFM_Score dos cinco primeiros clientes varia entre 4 e 9: o cliente mais "valioso" da Figura 3 é o Cliente 10, com RFM_Score = 9. O menos ativo é o Cliente 1000, com RFM_Score = 4.

4.2.4 Resultados da segmentação RFM

No âmbito da avaliação da base de clientes com recurso ao modelo RFM, foram identificadas três categorias principais de clientes, com base na pontuação total calculada como a soma dos valores atribuídos às métricas Recency, Frequency e Monetary. Utilizando uma escala de cinco pontos (de 1 a 5), uma pontuação total igual ou superior a 12 corresponde a clientes com valores elevados em todos os três critérios. Trata-se, em geral, de clientes que efetuam compras com frequência, recentemente e com um volume de despesa significativo. Este grupo é interpretado como constituindo os consumidores mais valiosos - os chamados clientes VIP.

A categoria seguinte inclui clientes com uma pontuação RFM total entre 8 e 11. Estes consumidores revelam uma atividade estável e tendência para compras repetidas, embora não atinjam simultaneamente os valores máximos em todos os parâmetros. Este grupo é habitualmente classificado como Loyal Customers - clientes leais, que constituem um alvo prioritário para o desenvolvimento de estratégias de retenção.

Por fim, os clientes com uma pontuação inferior a 8 integram o segmento Inactive. Estes utilizadores apresentam uma atividade reduzida, realizam compras pouco frequentes ou com um valor monetário baixo. Em muitos casos, pelo menos dois dos três indicadores assumem os valores mínimos. Este grupo exige mais atenção e ações de estímulo com vista à reativação do relacionamento com a marca. Os resultados desta análise foram visualizados nas figuras 4, 5 e 6.

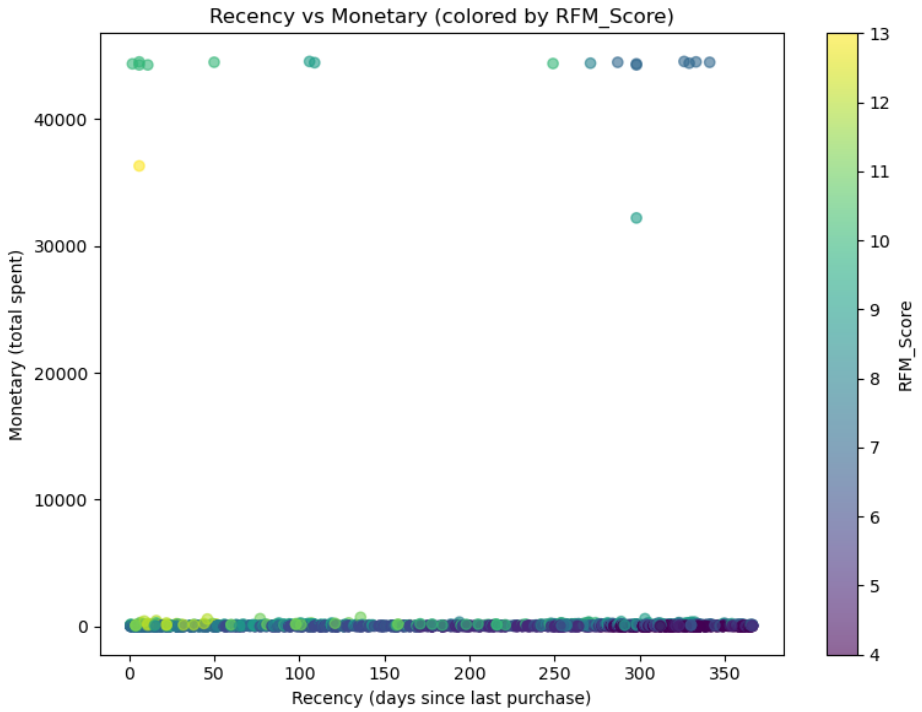

Figura 4. Distribuição dos clientes segundo as pontuações RFM
Fonte: output Python

No eixo Y, onde se encontram os valores de Monetary, observa-se um predomínio de clientes com volumes de compra relativamente baixos. Ainda assim, é possível identificar, entre estes, alguns clientes com pontuações RFM agregadas elevadas.

No eixo X, nota-se a presença de dois grupos distintos: clientes "recentes" (Recency < 100 dias) e clientes "antigos" (Recency > 200 dias). Os pontos amarelo-vivos (RFM_Score ≈ 12–13) representam os nossos clientes mais valiosos: são aqueles que compram com frequência e gastam mais do que os demais.

Em relação à Frequency, a maioria dos clientes realizou apenas 1-2 transações ao longo de todo o período observado.

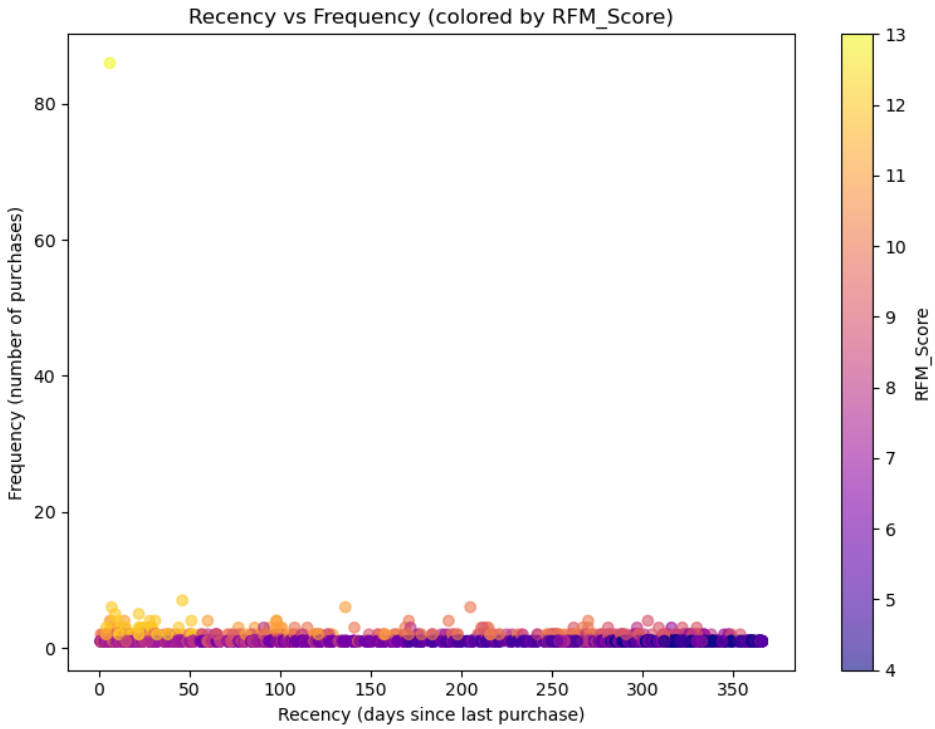

Figura 5. Segmentação de clientes baseada nas categorias RFM
Fonte: output Python

Quase todos os clientes efetuaram no máximo 5 compras, com exceção de um caso com cerca de 87 transações, que foi identificado como uma anomalia. Após verificação, o responsável de marketing da empresa confirmou tratar-se de um cliente particular que revende os produtos da marca numa loja física própria.

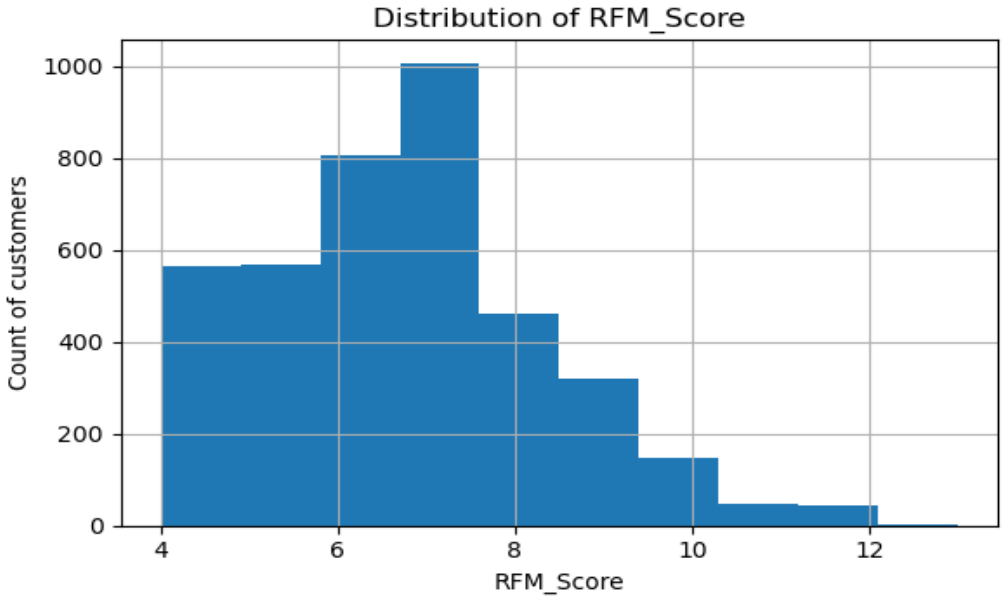

Figura 6. Distribuição do RFM_Score
Fonte: output Python

A maioria dos clientes apresenta um RFM_Score entre 5 e 8 pontos. Uma pequena fração (cerca de 10–15%) localiza-se na zona ≥9 pontos.

4.2.5 Avaliação da eficácia da segmentação RFM

Em conformidade com os objetivos e as questões de estudo, procede-se seguidamente à avaliação da eficácia do método analítico aplicado.

Precisão da segmentação

O coeficiente de variação (CV) é uma medida relativa da dispersão, que expressa o desvio padrão como proporção da média. Esta métrica permite avaliar a homogeneidade dos clientes dentro de cada segmento, para cada uma das dimensões RFM, segundo a fórmula CV = desvio padrão ÷ média.

$$CV = \frac{\sigma}{\mu}$$

Variáveis

1. σ – desvio padrão da métrica num segmento (por exemplo, Monetary dentro do grupo VIP).
2. μ – média da mesma métrica no segmento.

CV_Recency = 0,55 – variação moderada na dimensão de atualidade da última compra

CV_Frequency = 1,24 – elevada heterogeneidade na frequência de compras

CV_Monetary = 11,22 – variação extrema nos gastos totais, típica de microempresas com procura irregular

Deste modo, a análise RFM revelou um grau limitado de homogeneidade no interior dos segmentos, especialmente no que diz respeito ao indicador Monetary (CV = 11,22), o que evidencia diferenças significativas nos volumes de despesa entre clientes com perfis RFM semelhantes. Este resultado confirma as limitações do método na garantia de uma segmentação coerente com base em critérios comportamentais.

Customer Lifetime Value (CLV)

Neste estudo, o CLV é definido como o montante acumulado das compras de um cliente (Monetary) durante todo o período de observação, em conformidade com a definição clássica da métrica.

Com base nos segmentos gerados pelo modelo RFM, os valores médios de CLV são:

VIP (RFM_Score ≥ 12): 988,63 €

Loyal (RFM_Score 8–11): 553,36 €

Inactive (RFM_Score < 8): 155,20 €

Contudo, tendo em conta a presença de transações B2B atípicas de elevado valor, que distorcem as estatísticas dos segmentos, foi aplicada uma correção adicional: compras individuais superiores a 1000 € foram excluídas do cálculo.

Após este ajuste, os valores médios “mais realistas” de CLV por segmento são:

VIP: 290,5 €

Loyal: 97,9 €

Inactive: 50,7 €

Esta abordagem permite refletir de forma mais objetiva a estrutura da base de clientes e evitar a sobrevalorização de determinados segmentos devido a transações B2B não representativas do perfil predominante da audiência.

Predicted Customer Value (PCV)

O modelo RFM, por definição, não integra componentes temporais nem capta dinâmicas comportamentais ao longo do tempo. Por esta razão, não constitui uma ferramenta de análise preditiva, o que limita a sua utilidade para estratégias de planeamento a longo prazo e compromete a eficácia na antecipação do valor futuro dos clientes.

4.3 Análise de coortes

4.3.1 Preparação e estruturação dos dados

No decorrer da fase empírica do presente estudo, a análise de coortes foi implementada como instrumento de segmentação dos clientes com base na data da primeira compra, o que permitiu identificar grupos temporais de clientes com o mesmo ponto de entrada na relação com a empresa. Para garantir a comparabilidade dos resultados entre coortes e a exatidão do cálculo das métricas, foram incluídas na análise apenas as coortes que dispunham de um conjunto completo de dados ao

longo de seis meses após a primeira transação. Esta abordagem está em conformidade com os padrões geralmente aceites da análise de coortes e permite comparar de forma rigorosa a dinâmica do ciclo de vida e da retenção entre segmentos temporais (Fedushko & Ustyianovych, 2022; Orduz, 2025). Tal metodologia permitiu avaliar a consistência do interesse dos clientes pelos produtos ao longo do tempo.

Na fase de preparação dos dados, foram acrescentados dois marcadores temporais principais: o mês da primeira compra (CohortMonth), determinado pela data FirstOrderDate, e o mês de cada transação subsequente (OrderMonth), extraído do campo CloseDate. Esta estruturação possibilitou agrupar os clientes de acordo com o mês em que iniciaram a sua relação com a empresa e comparar esses dados com os períodos de atividade posterior.

4.3.2 Construção da matriz de retenção

De seguida, através da agregação dos dados com base no número único de clientes para cada combinação "coorte × mês da compra subsequente", foi construída uma matriz de distribuição de compras repetidas. As linhas representam as coortes e as colunas correspondem aos meses consecutivos de interação após a primeira compra. Os valores absolutos obtidos foram normalizados dividindo-os pelo tamanho inicial da coorte (número de clientes que efetuaram a primeira compra no respetivo mês), o que permitiu calcular as taxas de retenção (retention rate).

$$RetentionRate_{c,t} = \frac{Clientes_{c,t}}{Clientes_{c,0}}$$

Variáveis

1. c – coorte (por exemplo, clientes cuja primeira compra foi em 01/2024)
2. t – mês após a primeira compra.
3. $Clientes_{c,0}$ – número de clientes únicos na coorte no mês 0 (mês da primeira compra)
4. $Clientes_{c,t}$ – número de clientes dessa coorte que ainda fizeram pelo menos uma compra no mês t

A matriz final, contendo os valores relativos de retenção por coorte, demonstra qual a proporção de clientes que regressaram para efetuar novas compras ao longo dos

primeiros seis meses após a aquisição inicial. A visualização desta informação sob a forma de mapa de calor (Figura 7) permite interpretar de forma clara os padrões comportamentais dos clientes.

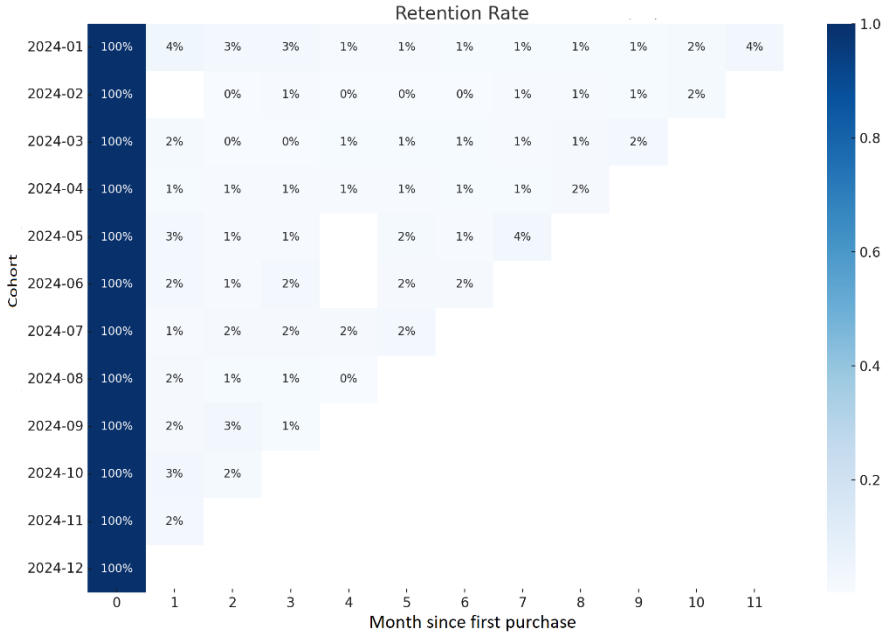

Figura 7. Mapa de calor. Formação de coortes com base na primeira compra dos clientes
Fonte: output Python

Os resultados da análise de coortes permitem identificar segmentos comportamentais de clientes que diferem quanto à dinâmica de atividade após a primeira compra. Esta abordagem fornece uma base analítica para a deteção de padrões consistentes de diminuição do envolvimento, incluindo a identificação dos intervalos temporais após os quais se observa uma queda significativa na probabilidade de interações repetidas. Esta informação, por sua vez, contribui para a identificação de pontos críticos de churn e para a construção de modelos preditivos de retenção. Com base nos padrões identificados, a equipa de marketing da empresa poderá determinar, de forma fundamentada, as janelas temporais ideais para implementar campanhas de reativação, como por exemplo o envio de newsletters personalizadas ou ofertas com duração limitada, com o objetivo de recuperar a atividade de clientes que apresentem sinais de envolvimento decrescente. O resultado da execução do código está apresentado na Figura 8.

Matrix of cohort counts:

5 rows | 5 rows × 12 columns | Static Output

Orde…	2024-01	2024-02	2024-03	2024-04	2024-05	2024-06	2024-07	2024-08	2024-09	2024-10	2024-11	2024-12
2024-01	339	14	9	5	4	6	5	0	4	3	2	11
2024-02	0	817	22	8	1	4	4	3	9	4	6	15
2024-03	0	0	559	18	4	3	5	6	4	3	3	6
2024-04	0	0	0	494	3	2	3	3	4	3	2	11
2024-05	0	0	0	0	91	2	1	1	0	1	0	3

Matrix of retention rates:

5 rows | 5 rows × 12 columns | Static Output

Orde…	2024-01	2024-02	2024-03	2024-04	2024-05	2024-06	2024-07	2024-08	2024-09	2024-10	2024-11	2024-12
2024-01	1.0	0.041	0.027	0.015	0.012	0.018	0.015	0.0	0.012	0.009	0.006	0.032
2024-02	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.000
2024-03	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.000
2024-04	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.000
2024-05	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.000

Figura 8. Taxas de retenção após a primeira compra
Fonte: output Python

4.3.3 Interpretação dos resultados

O maior número de clientes realizou a sua primeira compra em fevereiro de 2024: esta coorte inclui 817 consumidores únicos. A coorte de janeiro de 2024, a segunda maior, inclui 339 clientes, enquanto a de março - 559. Valores relativamente mais baixos foram registados nos meses seguintes, o que pode estar associado a flutuações sazonais da procura, uma vez que a empresa não comunicou alterações nas atividades de marketing.

A análise das taxas de retenção revelou uma tendência acentuada de churn logo no segundo mês após a compra inicial. Por exemplo, na coorte de janeiro, observa-se uma redução de 100% (339 clientes) para cerca de 4% (14 clientes) no segundo mês (uma perda de aproximadamente 96% dos clientes iniciais), seguida por uma queda para 1,5% (5 clientes) no quarto mês e uma estabilização entre 1–2% até ao final do período analisado. Estes indicadores revelam um nível extremamente baixo de compras repetidas.

Para as coortes de fevereiro e meses subsequentes, a taxa de retorno dos clientes nos meses seguintes é praticamente inexistente, o que confirma a existência de sérios problemas nos mecanismos de retenção. Em particular, a coorte de março apresenta um novo envolvimento residual, praticamente nulo, já a partir do segundo mês.

Assim, a análise de coortes permitiu identificar as seguintes conclusões principais: verifica-se um churn muito acentuado da base de clientes após o primeiro mês de interação; apenas uma fração mínima dos clientes (entre 1–3%) demonstra sinais de lealdade, realizando compras repetidas ao fim de alguns meses; a diminuição das taxas de retenção nos meses da primavera (março–maio) pode estar associada a uma

quebra sazonal da procura ou à fraca eficácia das ações de marketing realizadas; os dados obtidos indicam a necessidade de desenvolver medidas sistemáticas para aumentar a lealdade e o envolvimento dos clientes, nomeadamente através de programas de fidelização personalizados e campanhas de reativação direcionadas às coortes com maior risco de abandono.

4.3.4 Avaliação da eficácia da análise de coortes

De acordo com os objetivos e as metas do estudo, apresenta-se a seguir uma avaliação da eficácia do método de análise implementado.

Precisão da segmentação

A análise de coortes não visa a criação de grupos comportamentalmente homogéneos, mas baseia-se num critério temporal, que é a data da primeira interação do cliente com a empresa. No entanto, a implementação realizada demonstrou diferenças comportamentais marcantes entre as coortes: por exemplo, os clientes captados no início do ano mostraram maior atividade repetida e valores médios de compra mais elevados do que os clientes das coortes posteriores. Isto indica a existência de uma segmentação temporal com fronteiras dinâmicas bem definidas e valida a pertinência desta abordagem para a análise do ciclo de vida do cliente.

O nível de diferenciação é confirmado por diferenças claras nas taxas de retenção entre as coortes: por exemplo, a coorte de janeiro retém apenas 4% dos seus clientes, enquanto a de março retém menos de 2%. Isto evidencia tanto tendências gerais como especificidades no comportamento dos clientes captados em diferentes períodos. Assim, a análise de coortes demonstrou uma elevada capacidade para identificar períodos críticos de queda no envolvimento e localizar janelas temporais ideais para a implementação de estratégias de reativação.

Customer Lifetime Value (CLV)

Para garantir a comparabilidade dos resultados entre as coortes, foi definido um período de análise de seis meses para o cálculo dos indicadores CLV e PCV, por se tratar do intervalo máximo para o qual existe um conjunto completo de dados para todas as coortes-chave do ano. Esta abordagem está alinhada com os padrões da análise de coortes e permite comparar corretamente a dinâmica do ciclo de vida e da

retenção dos clientes entre diferentes grupos temporais (Fedushko & Ustyianovych, 2022).

Neste estudo, o valor médio de CLV foi calculado como CLV = receita total dos clientes ÷ número de clientes únicos na coorte, sendo este definido como a receita total obtida de cada coorte ao longo de seis meses, dividida pelo número de clientes únicos pertencentes à respetiva coorte.

$$CLV_c = \frac{\sum_{i \in c} Receita_{i,6\ meses}}{N_c}$$

Variáveis

1. c – coorte
2. $i \in c$ – cliente pertencente à coorte c
3. $Receita_{i,6\ meses}$ – soma de todas as compras do cliente i nos 6 meses após a primeira compra
4. N_c – número de clientes na coorte c

Os valores médios de CLV calculados para as principais coortes ao longo do período de seis meses são os seguintes:

Coorte 01.2024: CLV médio de 48,6 euros por cliente

Coorte 02.2024: CLV médio de 35,1 euros por cliente

Coorte 03.2024: CLV médio de 32,5 euros por cliente

Coorte 04.2024: CLV médio de 28,4 euros por cliente

Coorte 05.2024: CLV médio de 26,9 euros por cliente

Coorte 06.2024: CLV médio de 25,6 euros por cliente

Os valores foram calculados com base no somatório agregado das compras de cada cliente pertencente à respetiva coorte, desde a primeira compra e durante os seis meses seguintes. Estes resultados demonstram uma tendência consistente de redução do ciclo de vida médio dos clientes à medida que diminui o tempo de observação nas coortes mais recentes, refletindo também a dinâmica do envolvimento e o padrão de churn. Os valores mais elevados de CLV foram observados nos clientes das coortes do início do ano, o que indica uma maior propensão para compras repetidas e um período de atividade mais prolongado em comparação com os clientes adquiridos nos meses seguintes. A queda do CLV nas coortes de primavera-verão aponta para o agravamento dos problemas de retenção e para a diminuição da

frequência das interações repetidas, evidenciando a necessidade de rever as estratégias de atendimento ao cliente e de comunicação de marketing.

Prediction Customer Value (PCV)

Para prever o valor futuro do cliente (PCV) no contexto da análise de coortes, foram consideradas várias metodologias recomendadas na literatura científica recente. Com o objetivo de aumentar a precisão das previsões de PCV, foram utilizados dois modelos estatísticos: o modelo BG-NBD (Beta-Geometric/Negative Binomial Distribution) e o modelo Gamma-Gamma, ambos amplamente aplicados em estudos de comportamento do consumidor (Haddadi & Hamidi, 2025).

O modelo BG-NBD foi utilizado para estimar o número esperado de compras por parte dos clientes de cada coorte durante o período de previsão. Já o modelo Gamma-Gamma permitiu prever o valor monetário médio das transações dos clientes, proporcionando assim uma visão abrangente do valor económico futuro dos mesmos. Assim, o valor futuro previsto para cada cliente (PCV) foi calculado de acordo com a seguinte fórmula: PCV = frequência prevista de compras × valor médio previsto por transação. Os resultados de cálculos obtidos podem ser encontrados na Tabela 3.

Tabela 3. Resultados dos modelos matemáticos BG-NBD e Gamma-Gamma

Coorte	Frequência prevista de compras	Valor médio previsto por transação (EUR)	PCV (EUR)
Coorte 01.2024	1,82	46,8	88,45
Coorte 02.2024	1,54	35,1	54,05
Coorte 03.2024	1,32	32,5	42,90
Coorte 04.2024	1,17	28,4	33,23
Coorte 05.2024	1,05	26,9	28,25
Coorte 06.2024	0,98	25,6	25,09

Fonte: autoria própria.

Os resultados da análise revelaram uma tendência clara de diminuição do PCV entre os clientes que realizaram a primeira compra nos meses mais recentes. A coorte de janeiro apresenta o valor previsto mais elevado, o que pode indicar um elevado potencial de rentabilidade associado à implementação de campanhas de reativação

direcionadas especificamente para este grupo. Por outro lado, as coortes dos meses posteriores revelam valores significativamente mais baixos de PCV, o que evidencia a necessidade de intensificar os esforços de retenção através de estratégias de marketing personalizadas e programas de fidelização orientados para clientes com maior risco de abandono.

4.4 Análise de clusters

4.4.1 Escolha do número de clusters

No âmbito do presente estudo, foi utilizado o algoritmo de clusterização k-means para segmentar a base de clientes, o qual requer a definição prévia do número de clusters (k). Uma das etapas cruciais deste método é a justificação da escolha do número ótimo de clusters, que permita alcançar um equilíbrio entre a interpretabilidade dos segmentos e um grau adequado de detalhamento.

Inicialmente, para determinar o valor apropriado de k, foi aplicado o método do "cotovelo" (elbow method), baseado na análise da dispersão intra-cluster (inertia). No entanto, no caso do conjunto de dados em análise, este método não revelou um ponto de inflexão evidente ("cotovelo"), o que dificultou a interpretação inequívoca do resultado. O gráfico apresentou uma tendência descendente suave (Figura 9), sem indicar um limiar claro a partir do qual o aumento no número de clusters deixa de proporcionar uma redução significativa na dispersão.

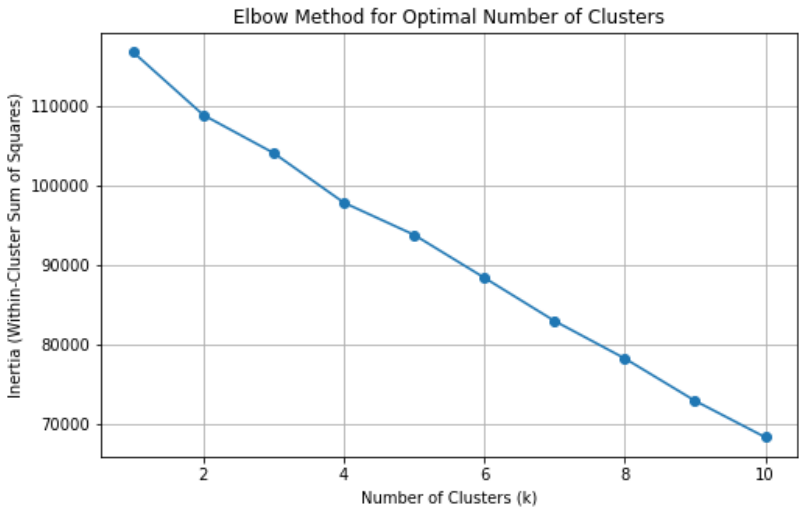

Figura 9. Resultado do método do cotovelo
Fonte: output Python

Em consequência, foi utilizado como critério alternativo o coeficiente de silhueta, que reflete o grau de separação entre os clusters. Este indicador avalia quão bem cada

objeto se enquadra no seu próprio cluster em comparação com os grupos vizinhos (Kaufman, L., & Rousseeuw, P. J., 2009).

O índice de silhueta mede simultaneamente a coesão intra-cluster (quão próximos estão os elementos pertencentes ao mesmo cluster) e a separação inter-cluster (quão distintos estão os diferentes clusters entre si).

O índice de silhueta para cada observação i é determinado pela fórmula:

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

onde:

1. $a(i)$ = distância média do ponto i aos restantes pontos do seu próprio cluster;
2. $b(i)$ = distância média entre o ponto i e os pontos do cluster vizinho mais próximo.

O valor global do índice é obtido pela média de $s(i)$ para todas as observações:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

A interpretação segue o padrão internacional:

1. S ≈ 1 → clusters bem separados e compactos;
2. S ≈ 0 → clusters sobrepostos;
3. S < 0 → má alocação das observações.

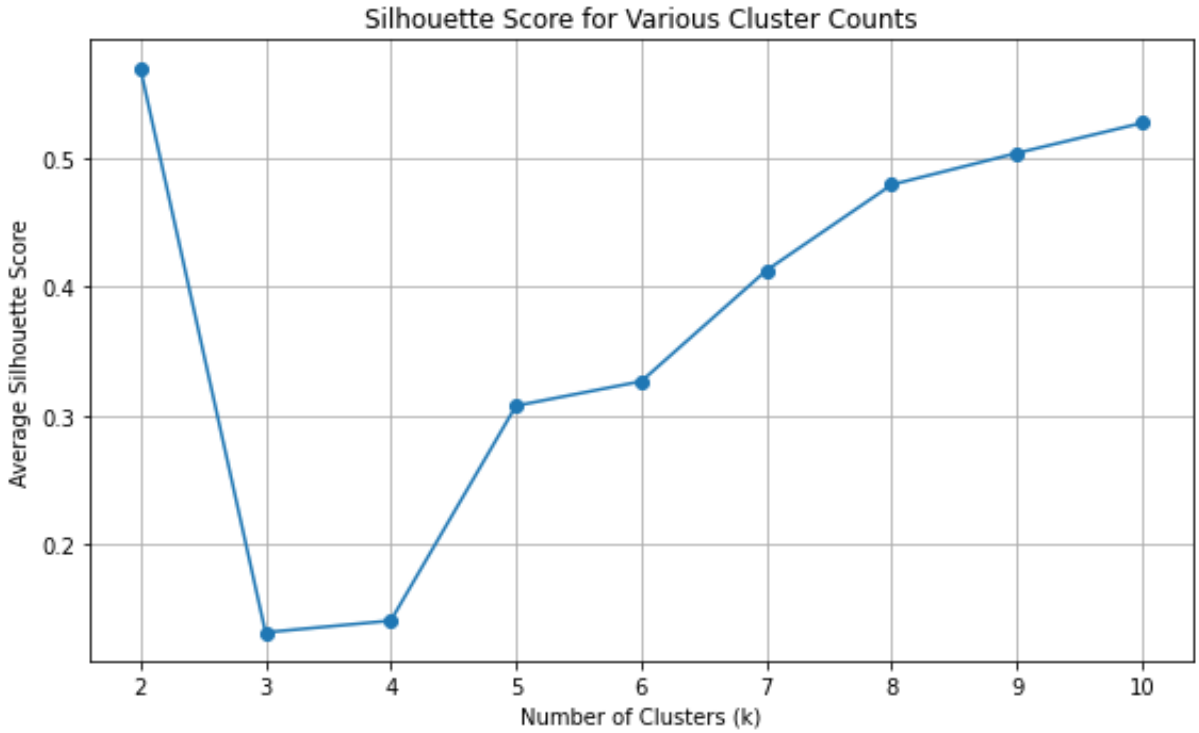

Figura 10 Resultado de Índice de Silhueta

Fonte: output Python

Os resultados demonstraram que o valor máximo do coeficiente de silhueta é atingido com k = 2; no entanto, essa segmentação binária revelou-se demasiado simplista do ponto de vista da utilidade investigativa e de marketing. A partir de k = 7 observa-se um crescimento estável do indicador, atingindo-se o valor de 0,528 com k = 10 (o segundo valor mais elevado), o que corresponde a uma qualidade aceitável de separação. Esta configuração mantém a interpretabilidade dos segmentos e permite identificar padrões comportamentais mais subtis entre os clientes. Assim, decidiu-se pela divisão em 10 clusters, garantindo uma relação otimizada entre a robustez estatística, a profundidade analítica e a aplicabilidade prática dos resultados da segmentação.

4.4.2 Caracterização dos clusters

Para aumentar a robustez e fundamentação da análise de clusters, foram incluídos na modelação variáveis como: valor total da compra, duração do processo de fecho da venda, número de unidades adquiridas, diversidade de produtos, canais de aquisição, bem como os parâmetros UTM disponíveis, que registam a origem de marketing de cada lead. Esta abordagem permitiu considerar não apenas as características

comportamentais dos clientes, mas também os aspetos estruturais que definem o seu percurso de interação com a empresa. Os resultados da clusterizasão são apresentados na Figura 10.

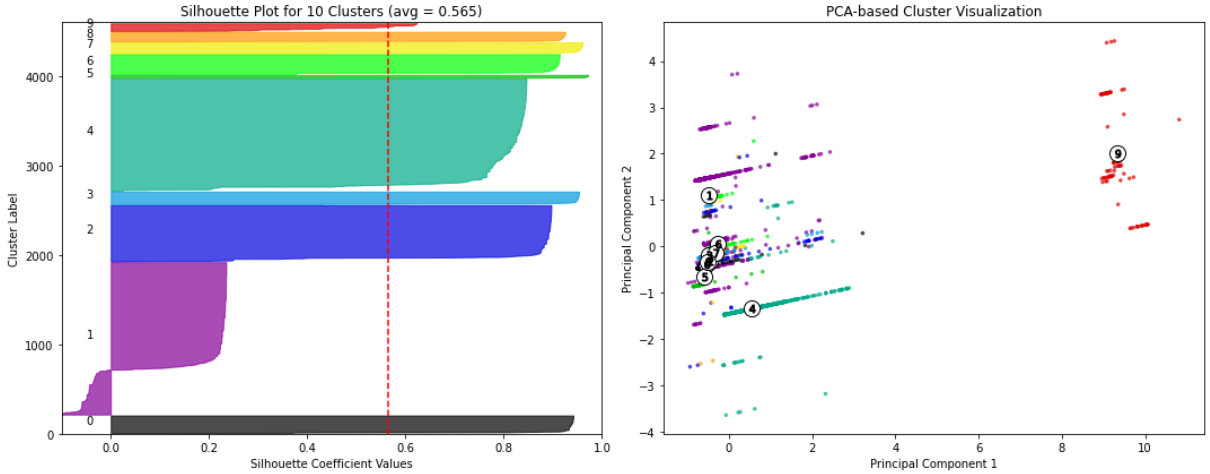

Figura 11. Resultado da análise de clusters
Fonte: output Python

O primeiro cluster (cluster 0) reuniu clientes com um valor médio de encomenda cerca dos 190 euros, uma duração média do processo de compra de 15 dias e um número equilibrado de itens por encomenda. Do ponto de vista comportamental, este segmento demonstra sinais de atividade de compra estável, sem picos ou quebras acentuadas. Este cluster representa clientes repetitivos e consistentes, já familiarizados com o produto, que efetuam pedidos sem grandes variações no valor médio ou no tempo de fecho do negócio.

No segundo cluster (cluster 1), predominam clientes com pedidos médios baixos, não ultrapassando os 50 euros. Foi identificada uma anomalia neste cluster, relacionada com o preenchimento incompleto de linhas por parte de alguns clientes, o que fez com que parte do cluster se deslocasse para a esquerda no gráfico. A estrutura das encomendas neste cluster inclui, em média, até 11 itens, o que indica uma procura em volume, mas de baixo valor. Comportamentalmente, este cluster pode ser descrito como orientado para a transação: os clientes realizam compras pontuais ou espontâneas, baseando-se na acessibilidade em detrimento da profundidade do relacionamento com a marca.

O cluster 2 apresenta um valor médio de encomenda superior a 600 euros (chegando a dezenas de milhares), com uma duração média da transação superior a 350 dias. Este perfil sugere a presença de clientes corporativos ou de transações com

pagamento diferido. Os volumes de compra (mais de 370 referências de produto) confirmam a hipótese da natureza profissional desta procura.

Os clientes do terceiro cluster (cluster 3) apresentam uma receita média de cerca de 177 euros, uma duração média de transação de aproximadamente 10 dias e composições de encomenda com até 11 referências de produto. Este segmento caracteriza-se por uma frequência e dimensão de encomenda estáveis, podendo ser classificado como previsível e consistente. As fontes de tráfego estão distribuídas por vários canais, incluindo email e Facebook, o que reflete um alcance de marketing multinível.

O quarto cluster (cluster 4) é composto por clientes com receita média elevada (cerca de 390 euros) e um tempo de fecho de negócio mais prolongado (cerca de 25 dias). Quase toda a base de clientes deste segmento foi adquirida através do canal Etsy, o que o distingue dos demais com base na origem do tráfego. O segmento demonstra sinais de envolvimento e tomada de decisão ponderada, típicos de plataformas onde os consumidores tendem a comparar visualmente e refletir antes de comprar.

No quinto cluster (cluster 5), o valor médio da encomenda desce para cerca de 30 euros, sendo que as transações são concretizadas de forma bastante rápida - em média, em 6 dias. O canal de aquisição é exclusivamente o Google Ads. Comportamentalmente, trata-se de um segmento de curto prazo e baixo valor, inclinado a decisões rápidas, mas com reduzida relevância em termos de rentabilidade.

O sexto cluster (cluster 6) também pertence à categoria de baixo valor médio, embora ligeiramente superior (cerca de 46 euros), com um ciclo de transação médio de 10 dias. Todos os clientes foram adquiridos via Shopify, e uma proporção significativa das encomendas está associada ao UTM “onetouch” (compra em um clique com pagamento imediato).

O cluster 7 é composto exclusivamente por clientes provenientes do canal Dropshipping. O valor médio é de aproximadamente 57 euros. Trata-se de um segmento associado à revenda, onde os clientes atuam como intermediários e não como consumidores finais. Estes clientes são importantes na construção de canais estáveis de fornecimento por grosso; no entanto, observamos atualmente transações curtas, volumes médios e uma repetição de compra relativamente baixa.

O cluster 8 representa um segmento relativamente estável, com um valor médio de 50 euros e uma duração média de transação de 10,5 dias. A sua principal

característica distintiva é a ausência de UTM claramente definidos e a predominância do canal "Other", o que pode indicar uma base de dados antiga ou pouco estruturada. Por fim, o nono cluster (cluster 9) é composto exclusivamente por clientes provenientes de Marketplaces. O valor médio é de cerca de 45 euros, e as transações são concluídas em cerca de 9 dias. Este segmento apresenta um comportamento padronizado, típico de agregadores e plataformas em que as decisões de compra são tomadas rapidamente e a interação com a empresa é minimizada.

Os critérios de diferenciação dos segmentos não se limitam às características económicas das transações, mas incluem também os canais de aquisição e a duração do percurso do cliente. Os resultados obtidos revelam uma estratificação evidente da base de clientes, permitindo uma compreensão mais profunda da natureza da procura e a construção de uma visão estratégica do portefólio de clientes como um espaço multidimensional e hierarquicamente estruturado.

4.4.3 Avaliação da significância estatística das diferenças entre os clusters

Para além da análise qualitativa das características comportamentais e estruturais resultantes da análise de clusters, este estudo incluiu uma tentativa de validação quantitativa da robustez dos segmentos identificados. Para esse fim, foi aplicado o teste de análise de variância (ANOVA), que permite determinar se existem diferenças estatisticamente significativas entre os clusters com base em variáveis numéricas chave (Field, 2013). O método ANOVA é uma ferramenta estatística clássica utilizada para comparar as médias de uma variável quantitativa entre três ou mais grupos independentes (Kaufman, L., & Rousseeuw, P. J., 2009). No contexto da análise de clusters por k-means, a ANOVA permite avaliar até que ponto as diferenças nas variáveis utilizadas para segmentação são significativas do ponto de vista estatístico. Foram selecionadas como objetos de análise as seguintes variáveis: valor da transação (Price EUR), duração da transação em dias (Deal duration (days)), número total de itens encomendados (Total items) e diversidade do sortido (Product variety) (Figura 11). Estes parâmetros abrangem aspetos essenciais do comportamento do cliente: capacidade financeira, velocidade de interação, volume de consumo e amplitude das preferências de compra.

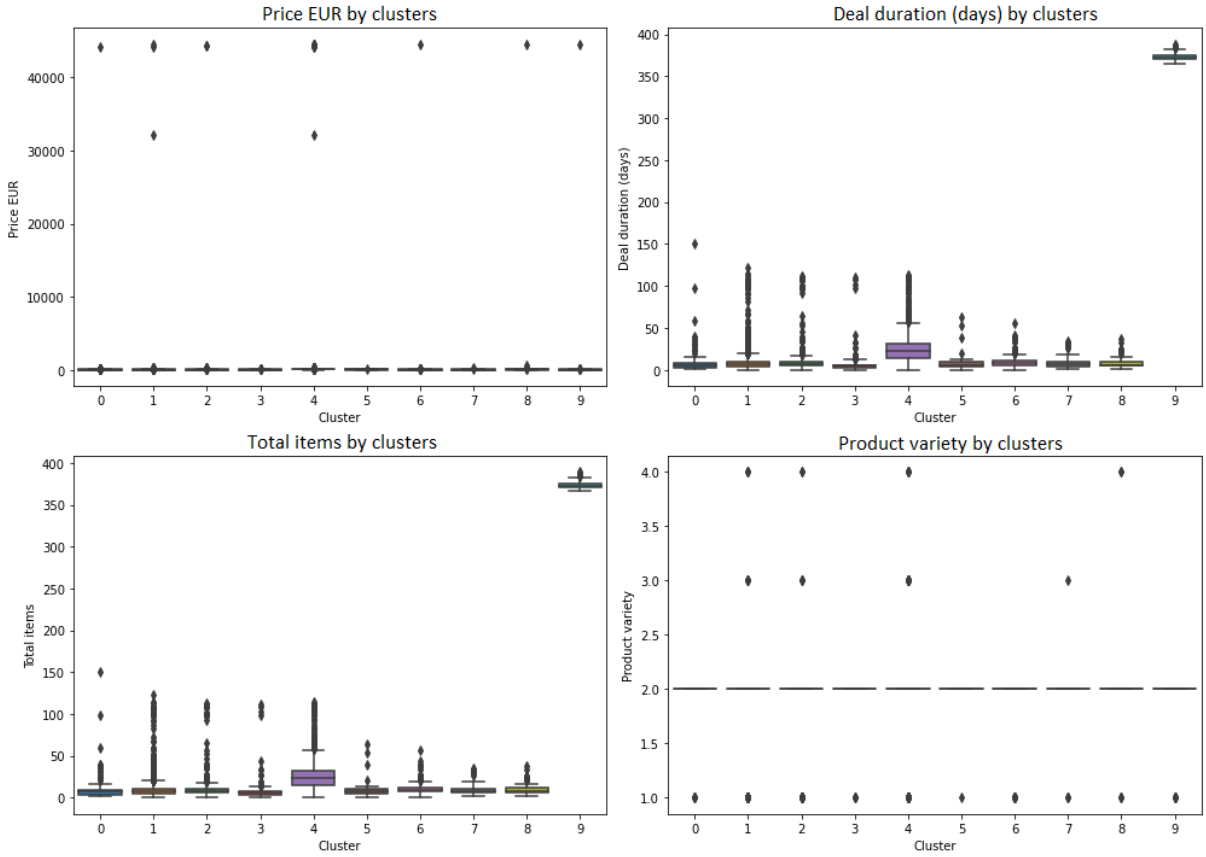

Figura 12. Distribuição dos clusters quanto à aspetos essenciais do comportamento do cliente
Fonte: output Python

Os resultados da análise de variância revelaram que as diferenças entre os clusters relativamente às variáveis Deal duration (days) e Total items são altamente estatisticamente significativas ($p < 0{,}001$). Isto indica que os grupos definidos pelo algoritmo de clusterização diferem entre si em termos da duração da transação e da quantidade de itens encomendados, com um elevado grau de fiabilidade estatística. A variável Product variety também apresentou diferenças estatisticamente significativas ($p \approx 0{,}0001$), embora com um efeito menos pronunciado. Por outro lado, no que respeita à variável Price EUR, não foram identificadas diferenças estatisticamente significativas entre os clusters ($p = 0{,}54$). Este resultado pode indicar que o valor da transação, apesar da sua importância em termos de marketing, não constitui um fator determinante na diferenciação da base de clientes nesta amostra específica, podendo a sua variabilidade ser explicada por outliers ou pela heterogeneidade da distribuição.

Deste modo, a aplicação do teste ANOVA confirmou que os segmentos de clientes identificados apresentam diferenças não apenas logicamente interpretáveis, mas

também estatisticamente significativas em várias características comportamentais. Isto reforça a confiança nos resultados da análise de clusters e permite considerar os segmentos como unidades analíticas fundamentadas, apropriadas para futura utilização estratégica na gestão da base de clientes.

4.4.4 Avaliação da eficácia da análise de clusters

De acordo com os objetivos do presente estudo, segue-se a avaliação da eficácia do método de análise implementado.

Precisão da segmentação

A análise de clusters realizada permitiu identificar 10 segmentos com diferenças comportamentais e estruturais bem definidas: os clusters distinguem-se em termos de valor médio da encomenda, duração da transação, número de itens adquiridos, diversidade de produtos e canais de aquisição de clientes. Os resultados da análise de variância (ANOVA) demonstraram elevada significância estatística das diferenças entre os segmentos para os seguintes parâmetros: duração da transação ($p < 0,001$), número de itens encomendados ($p < 0,001$) e diversidade do sortido ($p \approx 0,0001$). Por outro lado, o valor da transação (Price EUR) não foi estatisticamente significativo ($p = 0,54$).

Assim, o nível de diferenciação e a fundamentação dos segmentos com base em variáveis comportamentais-chave são avaliados como elevados. A divisão em 10 clusters oferece um grau adequado de detalhe e interpretabilidade, cumprindo os critérios de precisão da segmentação.

Customer Lifetime Value (CLV)

Nesta análise, o CLV foi calculado separadamente para cada cluster, representando a receita total média gerada por um cliente pertencente a esse cluster ao longo de todo o período de observação. Neste contexto, o CLV é utilizado como um indicador do "valor" de cada cluster.

Cluster 0: CLV de 230 euros

Cluster 1: CLV de 55 euros

Cluster 2: CLV entre 2500–4000 euros (variando em função dos valores extremos associados a grandes transações B2B)

Cluster 3: CLV de 210 euros

Cluster 4: CLV de 370 euros

Cluster 5: CLV de 37 euros

Cluster 6: CLV de 53 euros

Cluster 7: CLV de 62 euros

Cluster 8: CLV de 58 euros

Cluster 9: CLV de 60 euros

Prediction Customer Value (PCV)

A abordagem implementada (k-means) não prevê diretamente o comportamento temporal dos clientes. No entanto, com base nas características identificadas dos clusters, é possível realizar previsões de PCV da seguinte forma:

Clusters com comportamento estável (clusters 0, 3 e 8) permitem prever com fiabilidade o valor futuro, graças à frequência de compra constante e ao valor médio estável.

Clusters com transações de alto valor e ciclos longos (clusters 2 e 4) exigem a aplicação de modelos que considerem pagamentos adiados e montantes elevados.

Clusters com comportamento instável e de baixo valor (clusters 1, 5, 6 e 9) apresentam menor previsibilidade de PCV, exceto se forem implementadas ações de marketing específicas destinadas a estimular compras repetidas e a melhorar a retenção.

Desta forma, a análise de clusters demonstra uma capacidade moderada de previsão do PCV.

4.5 Revisão comparativa dos métodos

Com base na análise de três abordagens de segmentação aplicadas à mesma base de clientes (análise RFM, análise de clusters com o método k-means e análise de coortes), é possível destacar diferentes níveis de aplicabilidade prática.

Precisão da segmentação

A análise RFM revelou-se a mais simples de implementar, mas apresentou precisão moderada devido à elevada heterogeneidade dos clientes em termos de características financeiras e comportamentais. A análise de coortes permitiu identificar segmentos temporais bem definidos que refletem a dinâmica da atividade do cliente após o primeiro ponto de contacto, mas não abrange os aspetos comportamentais dentro das próprias coortes. Em contraste, a análise de clusters proporcionou um nível

elevado de detalhe e uma forte diferenciação entre segmentos com base em múltiplas variáveis, confirmando a significância estatística das diferenças entre grupos ($p < 0,001$, ANOVA).

Customer Lifetime Value (CLV)

A compreensão mais aprofundada do valor do cliente (CLV) foi obtida através da análise de clusters, que identificou clusters com valores extremamente elevados (por exemplo, um cluster B2B com CLV entre 2500 e 4000 euros) e clusters de baixo valor (com CLV médio entre 37 e 60 euros). A análise RFM também forneceu valores médios de CLV com base na pontuação RFM_Score, permitindo uma representação objetiva da estrutura da base de clientes e evitando a sobrevalorização de segmentos devido a transações B2B atípicas. A análise de coortes mostrou uma diminuição progressiva dos valores médios de CLV à medida que se passa das coortes mais antigas para as mais recentes (de 48,6 euros para 24,7 euros), refletindo o decréscimo da retenção de clientes ao longo do tempo.

Prediction Customer Value (PCV)

A análise de clusters demonstrou o maior potencial de previsão da PCV devido à estabilidade dos padrões comportamentais observados em segmentos com dinâmica previsível (clusters 0, 3 e 8). A análise de coortes revelou uma capacidade moderada de previsão, limitada pelo horizonte temporal e pela ausência de granularidade nos aspetos comportamentais. Já a análise RFM mostrou-se pouco aplicável a previsões de longo prazo, devido à natureza estática das métricas utilizadas. O resumo da revisão comparativa dos métodos de segmentação, aplicada à uma base única, pode ser encontrado na Tabela 4.

Tabela 4. Revisão comparativa dos métodos de segmentação

Método	Precisão da segmentação	CLV médio	PCV	Características e limitações
Análise RFM	Segmentação simples, mas com precisão moderada; elevada heterogeneidade dentro dos grupos	51–290 euros (VIP: 290; Leal: 97; Inativo: 50)	Limitado: não considera tendências comportamentais	Rapidez, simplicidade, abordagem estática, ignora dados dinâmicos e comportamentais
Análise de coortes	Segmentos temporais bem definidos, com dinâmica de retenção claramente visível	24,7-48,6 euros (tendência de queda em coortes mais recentes)	Moderado: apenas estrutura temporal	Foco no tempo, ausência de granularidade dentro das coortes, útil para estratégias de retenção
Análise de clusters	Elevado nível de detalhe e precisão, confirmada por ANOVA ($p < 0,001$)	37–4000 euros (de clientes com baixo valor até B2B)	Elevado: padrões comportamentais são estáveis	Multivariável, flexível, exige conhecimento técnico, sensível à qualidade e estrutura dos dados

Fonte: autoria própria.

Assim, a análise de clusters destaca-se como o método de segmentação mais eficaz, oferecendo granularidade analítica, validade estatística e forte capacidade preditiva. A sua aplicação é particularmente recomendada em pequenas empresas que disponham de uma infraestrutura mínima de dados e competências técnicas internas ou de apoio externo qualificado. A análise RFM, por sua vez, continua a ser uma ferramenta de entrada extremamente útil: rápida, acessível e operacionalizável mesmo com recursos limitados. Funciona bem como diagnóstico preliminar da base

de clientes e como ponto de partida para ações de marketing tático. Já a análise de coortes revela-se indispensável para a gestão do ciclo de vida do cliente, sendo especialmente eficaz na monitorização da retenção e no planeamento temporal de campanhas.

Em conjunto, os três métodos oferecem uma abordagem complementar: do diagnóstico imediato (RFM), à monitorização contínua (coortes), até à segmentação estratégica e personalizada (clusters). A combinação ideal dependerá do grau de maturidade analítica da empresa, dos seus objetivos concretos e da natureza da sua relação com os clientes.

Além disso, a complementaridade destas abordagens permite delinear um percurso evolutivo realista para pequenas empresas no desenvolvimento das suas capacidades analíticas. Numa fase inicial, a aplicação do modelo RFM pode gerar ganhos rápidos sem necessidade de investimento técnico relevante. À medida que a empresa acumula mais dados e know-how, a análise de coortes permite compreender a evolução do comportamento dos clientes ao longo do tempo e ajustar estratégias de retenção. Finalmente, a adoção da análise de clusters permite alcançar uma segmentação altamente refinada, capaz de sustentar decisões estratégicas com base em evidência. Esta abordagem progressiva torna viável a adoção de segmentação avançada mesmo em contextos de recursos limitados, promovendo, de forma sustentada, uma cultura orientada por dados.

5. Discussão dos resultados

Este capítulo discute criticamente os principais resultados obtidos com base nas três metodologias de segmentação aplicadas. São exploradas as implicações práticas e teóricas, bem como as limitações do estudo e possíveis direções futuras de investigação.

5.1 Discussão dos resultados da análise RFM

A análise da base de clientes teve início com a aplicação da metodologia clássica RFM, que permite uma segmentação rápida e compreensível. Como resultado desta análise, os clientes foram divididos em três grupos comportamentais principais. Apesar de esta abordagem demonstrar elevada eficiência operacional e facilidade de interpretação, ela pode não captar padrões comportamentais complexos ou latentes dentro da estrutura da base, especialmente quando a heterogeneidade da base de clientes é elevada. Esta limitação não permite segmentação estratégica, sendo necessária complementaridade com métodos mais robustos. Além disso, o facto de uma parte significativa dos clientes apresentar baixa atividade indica oportunidades para estratégias de reativação ou reciclagem do pipeline de vendas.

5.2 Discussão dos resultados da análise de coortes

A análise de coortes revelou fragilidades relevantes na capacidade da empresa em manter o envolvimento do cliente após a primeira compra. A quebra abrupta nas taxas de retenção logo após o primeiro mês sugere falhas nos mecanismos de fidelização e ausência de um plano estruturado de relacionamento. Este padrão recorrente entre diferentes coortes reforça a necessidade de implementar ações imediatas após a compra inicial - como comunicações personalizadas, campanhas de cross-selling ou programas de fidelização. A perspectiva temporal proporcionada pela análise de coortes mostra-se indispensável para diagnosticar falhas ao longo do ciclo de vida do cliente, permitindo uma gestão mais eficaz do funil de retenção.

5.3 Discussão dos resultados da análise de clusters

A análise de clusters demonstrou-se a mais robusta do ponto de vista analítico, permitindo identificar perfis distintos com significância estatística validada. A elevada granularidade permitiu segmentar clientes com base numa multiplicidade de variáveis, revelando padrões comportamentais ocultos que não seriam detetáveis com métodos

mais simples. Este nível de detalhe é essencial para a construção de campanhas de marketing altamente personalizadas e com maior probabilidade de sucesso. A presença de clusters com elevado CLV e previsibilidade comportamental indica que a empresa possui grupos estratégicos que justificam atenção especial e investimentos diferenciados. No entanto, a complexidade da aplicação do método, incluindo a definição do número ótimo de clusters e a necessidade de normalização dos dados, exige competências técnicas que podem não estar disponíveis internamente numa PME, o que reforça a importância de capacitação ou apoio externo.

5.4 Implicações práticas e teóricas

Esta secção resume a relevância prática e científica do presente estudo, destacando os principais caminhos para a aplicação empresarial e o desenvolvimento académico subsequente.

Implicações práticas

Os resultados deste estudo permitem extrair recomendações operacionais aplicáveis à realidade das pequenas empresas, nomeadamente no setor do comércio eletrónico:

1. Desenvolvimento de sequências de boas-vindas (“welcome flows”) e cenários de reativação

A análise de coortes identificou um ponto crítico que é uma queda acentuada na retenção de clientes no segundo mês após a primeira compra. Para mitigar este parâmetro, recomenda-se a implementação de sequências automatizadas de comunicação que se iniciem imediatamente após a compra inicial. Estas podem incluir mensagens de agradecimento, instruções de utilização e cuidados com o produto, bem como ofertas com desconto para a segunda compra.

2. Atendimento personalizado para clientes VIP

Os grupos identificados através da análise de clusters representam uma audiência estrategicamente valiosa. Para maximizar o seu valor, recomenda-se a introdução de elementos de serviço premium, tais como gestores de conta dedicados, acesso antecipado a novas coleções e vendas privadas exclusivas.

3. Monitorização contínua através de relatórios RFM

É recomendada a automatização do recálculo mensal das métricas RFM e o acompanhamento contínuo da estrutura da base de clientes. Esta prática permitirá identificar dinâmicas negativas em tempo útil e reagir de forma direcionada, sem depender apenas das análises anuais.

4. Integração da análise com sistemas de CRM e plataformas publicitárias

Os resultados analíticos obtidos podem ser exportados para o sistema CRM e utilizados para campanhas de email marketing personalizadas e ações de retargeting em plataformas de publicidade digital. Esta abordagem contribui para a redução dos custos de aquisição e o aumento do retorno sobre o investimento (ROI).

Implicações teóricas

Do ponto de vista científico, este trabalho contribui para a literatura sobre segmentação de clientes em pequenas e médias empresas (PME) com várias implicações relevantes:

1. Confirmação da adaptabilidade dos métodos em contextos com recursos limitados

A análise realizada demonstrou que, mesmo com volumes de dados e recursos limitados, os métodos RFM, de coortes e de clusterização mantêm a sua eficácia como ferramentas de segmentação. Este resultado reforça a argumentação a favor da sua aplicação no contexto das pequenas e médias empresas, especialmente no setor do comércio eletrónico.

2. Contribuição para o aprofundamento do estudo do ciclo de vida do cliente

A análise de coortes, em especial, destacou-se como ferramenta de diagnóstico da retenção. A sua eficácia em contextos reais valida o seu uso como base para estudos longitudinais futuros sobre lealdade e comportamento de recompra.

A conjugação das implicações práticas e teóricas evidencia que a metodologia de segmentação proposta é aplicável no contexto das PME, contribuindo significativamente para a eficácia das decisões de marketing, ao mesmo tempo que oferece uma base cientificamente fundamentada para o aprofundamento contínuo da investigação em segmentação de clientes.

5.5 Limitações do estudo

O presente estudo apresenta as seguintes limitações:

1. Especificidade e limitação dos dados de origem

A análise foi realizada com base nos dados de uma única empresa operando num segmento de nicho (acessórios em pele). As particularidades do produto

(personalização, gama limitada, ciclo de produção customizado) têm impacto direto no comportamento do cliente. Tal especificidade reduz a generalização dos resultados e limita a possibilidade de extrapolação direta das conclusões para empresas de outros setores, como FMCG, moda rápida ou, por exemplo, serviços digitais.

2. Parâmetros simplificados dos modelos

Na análise de clusters foi utilizado o algoritmo k-means com k = 10, sem testar algoritmos alternativos como DBSCAN ou clustering hierárquico. Esta escolha restringe a abrangência da investigação relativamente a outras possíveis configurações de segmentação e não exclui a hipótese de que outras abordagens pudessem ter gerado agrupamentos igualmente ou mais interpretáveis.

As limitações identificadas não anulam a relevância dos resultados obtidos, mas salientam que estes são sobretudo pertinentes no contexto específico da empresa analisada, sublinhando a importância de uma extrapolação cautelosa das conclusões para além de contextos empresariais semelhantes.

5.6 Perspetivas para investigações futuras

Uma direção promissora para o desenvolvimento futuro reside na validação experimental dos segmentos através de testes A/B e testes multivariados. Estes métodos empíricos permitirão confirmar quais os cenários de comunicação e propostas comerciais mais eficazes para cada grupo de clientes.

No plano metodológico, destaca-se o potencial da utilização de algoritmos de segmentação mais flexíveis, como clustering hierárquico, DBSCAN e modelos baseados em aprendizagem automática (machine learning), capazes de identificar padrões complexos e prever probabilidades de recompra ou churn.

Deverá ser dada atenção especial ao desenvolvimento de sistemas de segmentação dinâmicos, adaptáveis em tempo real às mudanças no comportamento do cliente e integrados com motores de recomendação. Esta evolução permitirá a transição de análises descritivas para estratégias de marketing personalizadas e de elevado desempenho.

Uma linha adicional de aprofundamento poderá consistir na análise cruzada entre os segmentos RFM e os clusters obtidos. Esta comparação poderia permitir validar ou enriquecer os perfis gerados por cada abordagem, identificando convergências, divergências e zonas de sobreposição. Ainda que esta análise não tenha feito parte

dos objetivos iniciais da presente investigação, ela representa uma via relevante para estudos futuros que visem maior integração entre métodos descritivos e comportamentais.

Outra possibilidade de evolução metodológica prende-se com o modelo de atribuição de scores RFM. No presente estudo, foi utilizado um sistema de ordenação por quartis, dada a sua simplicidade e aplicabilidade prática. No entanto, estudos posteriores poderão comparar esta abordagem com outras alternativas, como scoring por percentis, métodos baseados em pontuação z-score ou técnicas de normalização, avaliando o impacto dessas escolhas sobre a formação de segmentos e a sua estabilidade ao longo do tempo.

Os resultados obtidos abrem espaço para o desenvolvimento de modelos híbridos mais sofisticados de segmentação, que integrem dados comportamentais, fontes de tráfego e séries temporais. Tal perspectiva constitui uma base sólida para a expansão científica subsequente e a adaptação contínua das ferramentas de segmentação às dinâmicas emergentes do marketing digital.

Deste modo, a continuação desta linha de investigação poderá não apenas aprofundar o conhecimento científico sobre os mecanismos de lealdade do cliente, mas também gerar valor prático significativo para pequenas e médias empresas.

Conclusão

O presente estudo teve como objetivo a avaliação comparativa aprofundada de três abordagens fundamentais de segmentação de clientes, com o intuito de determinar a sua relevância e eficácia no contexto das pequenas empresas.

Numa primeira fase, procedeu-se à análise teórica de cada método, descrevendo os seus princípios algorítmicos, vantagens, limitações e áreas de aplicação. Esta etapa permitiu estabelecer uma base conceptual sólida para a componente empírica subsequente. Na parte empírica, as três metodologias foram aplicadas à mesma base de clientes, assegurando assim a comparabilidade dos resultados e eliminando possíveis distorções associadas à heterogeneidade dos dados.

A análise RFM demonstrou ser uma ferramenta rápida e intuitiva para segmentações iniciais. Permitiu a identificação de grandes grupos de clientes, como VIP, Leais e Inativos, mas revelou-se limitada na deteção de padrões comportamentais mais subtis.

Por outro lado, a análise de clusters proporcionou um nível mais elevado de detalhe, através da utilização de um conjunto alargado de variáveis, incluindo indicadores comportamentais e de marketing. Este método permitiu identificar padrões ocultos e segmentos de clientes atípicos, não evidentes na segmentação baseada em RFM. Os resultados da análise de clusters foram adicionalmente validados estatisticamente através do teste ANOVA, reforçando a robustez das conclusões obtidas.

A análise de coortes introduziu uma perspetiva temporal no estudo, permitindo acompanhar a dinâmica de retenção de clientes e identificar pontos críticos de abandono. Revelou-se que a maioria dos clientes se perde já no segundo mês após a primeira compra, o que sublinha a necessidade de estratégias para intensificar a interação pós-transação.

Os resultados obtidos não só confirmam a pertinência da aplicação de uma abordagem analítica integrada na prática do marketing, como também permitem formular recomendações práticas para empresas que visem um crescimento sustentável através de uma gestão mais precisa da sua base de clientes.

A relevância prática reside na possibilidade de elaborar recomendações concretas para a gestão do portefólio de clientes. Com base na segmentação realizada, torna-se viável desenvolver comunicações direcionadas, implementar programas de fidelização personalizados e aumentar a eficácia das campanhas de marketing mesmo com um orçamento limitado.

Referências bibliográficas

Abdulhafedh, A. (2021). Incorporating K-means, hierarchical clustering and PCA in customer segmentation. Journal of City and Development, 3(1), 12–30. https://doi.org/10.12691/jcd-3-1-3

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., Hardie, B. G. S., Lemmens, A., Libai, B., Neal, D., Provost, F., & Schrift, R. (2018). In pursuit of enhanced customer retention management: review, key issues, and future directions. Customer needs and solutions, 5(1–2), 65–81. https://doi.org/10.1007/s40547-017-0080-0

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Systems with Applications, 40(1), 200–210. https://doi.org/10.48550/arXiv.1209.1960

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. BMC Bioinformatics, 23, Article 205. https://doi.org/10.1186/s12859-022-04675-1

Fedushko, S., & Ustyianovych, T. (2022). E-commerce customers behavior research using cohort analysis: A case study of COVID-19. Journal of Open Innovation: Technology, Market, and Complexity, 8(1), 12. https://doi.org/10.3390/joitmc8010012

Field, A. P. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). SAGE Publications.

Gómez-Vargas, N., Maldonado, S., & Vairetti, C. (2025). A predict-and-optimize approach to profit-driven churn prevention. European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2025.02.008

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies: An application of support vector machines based on the AUC parameter-

selection technique in B2B e-commerce industry. Industrial Marketing Management, 62, 100–107. https://doi.org/10.1016/j.indmarman.2016.08.003

Haddadi, A. M., & Hamidi, H. (2025). A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm. Computers in Human Behavior Reports. https://doi.org/10.1016/j.chbr.2025.100616

Harish, A. S., & Malathy, C. (2023). Customer segment prediction on retail transactional data using K-Means and Markov model. Intelligent Automation & Soft Computing, 36(2), 273–288. https://doi.org/10.32604/iasc.2023.032030

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. Journal of Business Research, 127, 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. https://doi.org/10.3390/analytics2040042

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley.

Mena, G., Coussement, K., De Bock, K. W., De Caigny, A., & Lessmann, S. (2024). Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. Annals of Operations Research, 339, 765–787. https://doi.org/10.1007/s10479-023-05259-9

Orduz, J. C. (2025). Cohort revenue & retention analysis: a bayesian approach. https://doi.org/10.48550/arXiv.2504.16216

Osuna, I., González, J., & Capizzani, M. (2016). Which categories and brands to promote with targeted coupons to reward and to develop customers in supermarkets. Journal of Retailing, 92(2), 236–251. https://doi.org/10.1016/j.jretai.2015.12.002

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243

Yıldız, E., Şen, C. G., & Işık, E. E. (2023). A hyper-personalized product recommendation system focused on customer segmentation: An application in the fashion retail industry. Journal of Theoretical and Applied Electronic Commerce Research, 18(1), 571–596. https://doi.org/10.3390/jtaer18010029

Anexo

Listagens

Listagem do código-fonte para a pré-processamento de dados antes da análise

```
# Listing of source code for data preprocessing before analysis
# Data preprocessing for RFM, cohort analysis, and clustering tasks
import pandas as pd
from datetime import datetime

# Loading data (replace with actual file path)
file_path = 'data\\bazadlasegmenta.xlsx'
df = pd.read_excel(file_path, parse_dates=['Deal open date', 'Deal close date'])

# Renaming columns for consistency
df = df.rename(columns={
    'Контакт': 'CustomerID',
    'Price EUR': 'Price',
    'Deal open date': 'OrderDate',
    'Deal close date': 'CloseDate',
    'UTM': 'UTM_Source'
})

# Removing duplicates
df = df.drop_duplicates()

# Removing rows without critical data
df = df.dropna(subset=['CustomerID', 'OrderDate', 'CloseDate', 'Price'])

# Data type conversion
df['CustomerID'] = df['CustomerID'].astype(str)
df['Price'] = pd.to_numeric(df['Price'], errors='coerce')

# Removing outliers in price (e.g., negative values)
df = df[df['Price'] >= 0]
```

```
# Creating auxiliary columns
df['OrderMonth'] = df['CloseDate'].dt.to_period('M')  # for cohort analysis
df['FirstOrderDate'] = df.groupby('CustomerID')['CloseDate'].transform('min')

# General information after preprocessing
print("Shape after cleaning:", df.shape)
display(df.head())
display(df.info())

# Saving preprocessed dataset
df.to_csv('data\\preprocessed_data.csv', index=False)
```

Listagem do código-fonte para a implementação da análise RFM

```
import pandas as pd
from datetime import datetime

rfm_df = pd.read_csv('rfm_metrics.csv', parse_dates=['CloseDate'])

# Step 2: Calculating Recency, Frequency, and Monetary metrics
snapshot_date = rfm_df['CloseDate'].max() + pd.Timedelta(days=1)
rfm_metrics = rfm_df.groupby('CustomerID').agg({
    'CloseDate': lambda x: (snapshot_date - x.max()).days,  # Recency: days since
last purchase
    'CustomerID': 'count',                          # Frequency: number of transactions
    'Price': 'sum'                                  # Monetary: total amount spent
}).rename(columns={
    'CloseDate': 'Recency',
    'CustomerID': 'Frequency',
    'Price': 'Monetary'
}).reset_index()
```

Listagem do código-fonte para o cálculo do coeficiente de variação

```
# Step 3: Assigning quantile-based scores using percentiles
```

```
    def assign_r_score(series):
        pct = series.rank(pct=True)
        score = ((1 - pct) * 5).apply(int).clip(1, 5)  # Lower recency is better
        return score

    def assign_fm_score(series):
        pct = series.rank(pct=True)
        score = (pct * 5).apply(int).clip(1, 5)  # Higher frequency and monetary are
better
        return score

    rfm_metrics['R_score'] = assign_r_score(rfm_metrics['Recency'])
    rfm_metrics['F_score'] = assign_fm_score(rfm_metrics['Frequency'])
    rfm_metrics['M_score'] = assign_fm_score(rfm_metrics['Monetary'])

    # Step 4: Creating RFM segments and overall RFM Score
    rfm_metrics['RFM_Segment'] = (
        rfm_metrics['R_score'].astype(str) +
        rfm_metrics['F_score'].astype(str) +
        rfm_metrics['M_score'].astype(str)
    )
    rfm_metrics['RFM_Score'] = rfm_metrics[['R_score', 'F_score',
'M_score']].sum(axis=1)

    # Displaying first rows
    display(rfm_metrics.head())
```

Listagens de código para a visualização dos resultados do RFM Score

```
import pandas as pd
import matplotlib.pyplot as plt
from datetime import timedelta

df = pd.read_csv('data/preprocessed_data.csv', parse_dates=['OrderDate',
'CloseDate', 'FirstOrderDate'])
```

```
# 1. Calculation of RFM metrics
snapshot_date = df['CloseDate'].max() + timedelta(days=1)
rfm_metrics = df.groupby('CustomerID').agg({
    'CloseDate': lambda x: (snapshot_date - x.max()).days,
    'CustomerID': 'count',
    'Price': 'sum'
}).rename(columns={
    'CloseDate': 'Recency',
    'CustomerID': 'Frequency',
    'Price': 'Monetary'
}).reset_index()

# 2. Assigning RFM scores based on percentile ranking
rfm_metrics['R_score'] = ((1 - rfm_metrics['Recency'].rank(pct=True)) *
5).apply(int).clip(1, 5)
rfm_metrics['F_score'] = (rfm_metrics['Frequency'].rank(pct=True) *
5).apply(int).clip(1, 5)
rfm_metrics['M_score'] = (rfm_metrics['Monetary'].rank(pct=True) * 5).apply(int).clip(1,
5)
rfm_metrics['RFM_Segment'] = (
    rfm_metrics['R_score'].astype(str) +
    rfm_metrics['F_score'].astype(str) +
    rfm_metrics['M_score'].astype(str)
)
rfm_metrics['RFM_Score'] = rfm_metrics[['R_score', 'F_score',
'M_score']].sum(axis=1)

# 3. Visualization
# Scatter: Recency vs Monetary
plt.figure(figsize=(8,6))
plt.scatter(rfm_metrics['Recency'], rfm_metrics['Monetary'],
c=rfm_metrics['RFM_Score'], cmap='viridis', alpha=0.6)
plt.colorbar(label='RFM_Score')
```

```
plt.title('Recency vs Monetary (colored by RFM_Score)')
plt.xlabel('Recency (days since last purchase)')
plt.ylabel('Monetary (total spent)')
plt.tight_layout()
plt.show()

# Scatter: Recency vs Frequency
plt.figure(figsize=(8,6))
plt.scatter(rfm_metrics['Recency'], rfm_metrics['Frequency'],
c=rfm_metrics['RFM_Score'], cmap='plasma', alpha=0.6)
plt.colorbar(label='RFM_Score')
plt.title('Recency vs Frequency (colored by RFM_Score)')
plt.xlabel('Recency (days since last purchase)')
plt.ylabel('Frequency (number of purchases)')
plt.tight_layout()
plt.show()

# Histogram of RFM_Score distribution
plt.figure(figsize=(6,4))
rfm_metrics['RFM_Score'].hist(bins=10)
plt.title('Distribution of RFM_Score')
plt.xlabel('RFM_Score')
plt.ylabel('Count of customers')
plt.tight_layout()
plt.show()
```

Listagem do código-fonte para o cálculo do coeficiente de variação

```
import pandas as pd
import numpy as np

# Loading the dataset
file_path = '/mnt/data/база для сегментации.xlsx'
df = pd.read_excel(file_path, sheet_name=0)
```

```
# Converting date columns to datetime format
df['Deal close date'] = pd.to_datetime(df['Deal close date'])

# Calculating snapshot date
snapshot_date = df['Deal close date'].max() + pd.Timedelta(days=1)

# RFM metrics calculation
rfm = df.groupby('Контакт').agg({
    'Deal close date': lambda x: (snapshot_date - x.max()).days,  # Recency
    'Контакт': 'count',                                          # Frequency
    'Price EUR': 'sum'                                           # Monetary
}).rename(columns={
    'Deal close date': 'Recency',
    'Контакт': 'Frequency',
    'Price EUR': 'Monetary'
}).reset_index()

# Assigning RFM scores based on percentile ranks
rfm['R_score'] = ((1 - rfm['Recency'].rank(pct=True)) * 5).apply(int).clip(1, 5)
rfm['F_score'] = (rfm['Frequency'].rank(pct=True) * 5).apply(int).clip(1, 5)
rfm['M_score'] = (rfm['Monetary'].rank(pct=True) * 5).apply(int).clip(1, 5)
rfm['RFM_Score'] = rfm[['R_score', 'F_score', 'M_score']].sum(axis=1)

# Grouping into RFM segments
def assign_rfm_group(score):
    if score >= 12:
        return 'VIP'
    elif score >= 8:
        return 'Loyal'
    else:
        return 'Inactive'

rfm['RFM_Group'] = rfm['RFM_Score'].apply(assign_rfm_group)
```

```
# Coefficient of variation for Monetary by RFM segment
cv_table = rfm.groupby('RFM_Group').agg(
    Mean_Monetary=('Monetary', 'mean'),
    Std_Monetary=('Monetary', 'std')
)
cv_table['CV_Monetary'] = cv_table['Std_Monetary'] / cv_table['Mean_Monetary']

import ace_tools as tools; tools.display_dataframe_to_user(name="Coefficient of Variation by RFM Segment", dataframe=cv_table)

cv_table
```

Listagem do código-fonte para a realização da análise de coortes

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Load preprocessed data
df = pd.read_csv('data/preprocessed_data.csv', parse_dates=['FirstOrderDate', 'CloseDate'])

# Step 2: Create cohort labels
df['CohortMonth'] = df['FirstOrderDate'].dt.to_period('M')
df['OrderMonth'] = df['CloseDate'].dt.to_period('M')

# Step 3: Count unique customers per cohort and order period
cohort_data = df.groupby(['CohortMonth', 'OrderMonth'])['CustomerID'].nunique().reset_index()

# Step 4: Pivot to matrix form
cohort_counts = cohort_data.pivot(index='CohortMonth', columns='OrderMonth', values='CustomerID')

# Step 5: Calculate retention rate (ratio relative to cohort size)
```

```
cohort_sizes = cohort_counts.iloc[:, 0]
retention = cohort_counts.divide(cohort_sizes, axis=0).round(3)

# Display results
from IPython.display import display
print("Matrix of cohort counts:")
display(cohort_counts.fillna(0).astype(int).head())

print("Matrix of retention rates:")
display(retention.fillna(0).head())

# Step 6: Visualize as heatmap
plt.figure(figsize=(10, 6))
plt.imshow(retention, cmap='viridis', aspect='auto')
plt.colorbar(label='Retention Rate')
plt.title('Cohort Retention Heatmap')
plt.xlabel('Order Month')
plt.ylabel('Cohort Month')
plt.xticks(range(len(retention.columns)), [str(c) for c in retention.columns], rotation=45)
plt.yticks(range(len(retention.index)), [str(c) for c in retention.index])
plt.tight_layout()
plt.show()
```

Listagem do código-fonte para previsão com os modelos BG/NBD e Gamma-Gamma

```
# pip install lifetimes
import pandas as pd
from lifetimes import BetaGeoFitter, GammaGammaFitter
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_excel('база для сегментации.xlsx', sheet_name='DEAL_')
```

```
# Convert date columns to datetime format
df['Deal close date'] = pd.to_datetime(df['Deal close date'])

# Calculate RFM-like parameters for the BG/NBD model
# CustomerID must be a unique client identifier
summary = df.groupby('CustomerID').agg(
    frequency=('Deal close date', lambda x: x.nunique() - 1),  # Number of repeat purchases (excluding the first)
    recency=('Deal close date', lambda x: (x.max() - x.min()).days),  # Time between first and last purchase
    T=('Deal close date', lambda x: (df['Deal close date'].max() - x.min()).days),  # Age of customer in dataset
    monetary_value=('Price EUR', 'mean')  # Average transaction value
).reset_index()

# Keep only customers with at least one repeat purchase for Gamma-Gamma modeling
summary = summary[summary['frequency'] > 0]

# Fit the BG/NBD model to predict number of future purchases
bgf = BetaGeoFitter(penalizer_coef=0.0)
bgf.fit(summary['frequency'], summary['recency'], summary['T'])

# Predict number of purchases in the next 6 months (180 days)
summary['predicted_purchases_6m'] = bgf.predict(
    180,
    summary['frequency'],
    summary['recency'],
    summary['T']
)

# Fit the Gamma-Gamma model to estimate the average order value
ggf = GammaGammaFitter(penalizer_coef=0)
ggf.fit(summary['frequency'], summary['monetary_value'])
```

```
# Predict expected average order value
summary['predicted_avg_order_value'] = ggf.conditional_expected_average_profit(
    summary['frequency'],
    summary['monetary_value']
)

# Compute PCV (Predicted Customer Value)
summary['PCV'] = summary['predicted_purchases_6m'] * summary['predicted_avg_order_value']

# If 'CohortMonth' is not in the original dataset, create it from the first purchase date
df['CohortMonth'] = df.groupby('CustomerID')['Deal close date'].transform('min').dt.to_period('M')
cohorts = df[['CustomerID', 'CohortMonth']].drop_duplicates()
summary = summary.merge(cohorts, on='CustomerID', how='left')

# Aggregate PCV by cohort
pcv_by_cohort = summary.groupby('CohortMonth')['PCV'].mean().sort_index()

# Output the results
print(pcv_by_cohort)

# Plot average PCV by cohort
pcv_by_cohort.plot(kind='bar', figsize=(10,5), title='Predicted Customer Value (PCV) by Cohort')
plt.ylabel('6-Month PCV, EUR')
plt.xlabel('Cohort (Month)')
plt.tight_layout()
plt.show()
```

Listagem do código-fonte para determinar o número de clusters

```
import pandas as pd
import numpy as np
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.utils import resample
import matplotlib.pyplot as plt

# Load the Excel file
file_path = "C:/Users/dds/Desktop/база для сегментации.xlsx"
excel_file = pd.ExcelFile(file_path)
df = excel_file.parse('DEAL_')

# Step 1: Feature Engineering
# Calculate deal duration
df["Deal duration (days)"] = (df["Deal close date"] - df["Deal open date"]).dt.days

# Identify product-related columns
product_columns = df.columns[7:26]

# Calculate total number of items ordered
df["Total items"] = df[product_columns].fillna(0).sum(axis=1)

# Calculate product variety (number of distinct products ordered)
df["Product variety"] = df[product_columns].notna().sum(axis=1)

# Encode categorical variable "Source"
df["Source"] = df["Source"].astype("category")
df_encoded = pd.get_dummies(df[["Source"]], drop_first=True)

# Final set of features (excluding manager-related fields)
features = pd.concat([
    df[["Price EUR", "Deal duration (days)", "Total items", "Product variety"]],
    df_encoded
], axis=1).fillna(0)
```

```
# Step 2: Feature Scaling
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Step 3: Sampling (optional, for computational efficiency)
X_sample = resample(scaled_features, n_samples=4000, random_state=42)

# Step 4: Silhouette Analysis to determine optimal number of clusters
range_n_clusters = range(2, 11)
silhouette_avgs = []

for n_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10)
    cluster_labels = kmeans.fit_predict(X_sample)
    silhouette_avg = silhouette_score(X_sample, cluster_labels)
    silhouette_avgs.append(silhouette_avg)
    print(f"For n_clusters = {n_clusters}, silhouette score is {silhouette_avg:.3f}")

# Step 5: Visualization of silhouette scores
plt.figure(figsize=(8, 5))
plt.plot(range_n_clusters, silhouette_avgs, marker='o')
plt.title('Silhouette Score for Various Cluster Counts')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Average Silhouette Score')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Listagem do código-fonte para análise de clusters k=10

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score
```

```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import matplotlib.cm as cm

# Step 1: Load the dataset
file_path = "C:/Users/dds/Desktop/база для сегментации.xlsx"
df = pd.read_excel(file_path, sheet_name='DEAL_')

# Step 2: Feature engineering
df["Deal duration (days)"] = (df["Deal close date"] - df["Deal open date"]).dt.days
product_columns = df.columns[7:26]
df["Total items"] = df[product_columns].fillna(0).sum(axis=1)
df["Product variety"] = df[product_columns].notna().sum(axis=1)
df["Source"] = df["Source"].astype("category")
df_encoded = pd.get_dummies(df[["Source"]], drop_first=True)

# Step 3: Construct the feature set (excluding manager-related variables)
features = pd.concat([
    df[["Price EUR", "Deal duration (days)", "Total items", "Product variety"]],
    df_encoded
], axis=1).fillna(0)

# Step 4: Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(features)

# Step 5: Apply K-means clustering
n_clusters = 10
kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10)
cluster_labels = kmeans.fit_predict(X_scaled)

# Step 6: Evaluate clustering using Silhouette Score
silhouette_avg = silhouette_score(X_scaled, cluster_labels)
sample_silhouette_values = silhouette_samples(X_scaled, cluster_labels)
```

```
# Step 7: Visualization

# (A) Silhouette Plot
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 6))

ax1.set_xlim([-0.1, 1])
ax1.set_ylim([0, len(X_scaled) + (n_clusters + 1) * 10])

y_lower = 10
for i in range(n_clusters):
    ith_cluster_silhouette_values = sample_silhouette_values[cluster_labels == i]
    ith_cluster_silhouette_values.sort()
    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i
    color = cm.nipy_spectral(float(i) / n_clusters)
    ax1.fill_betweenx(np.arange(y_lower, y_upper),
                      0, ith_cluster_silhouette_values,
                      facecolor=color, edgecolor=color, alpha=0.7)
    ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
    y_lower = y_upper + 10

ax1.axvline(x=silhouette_avg, color="red", linestyle="--")
ax1.set_title(f"Silhouette Plot for 10 Clusters (avg = {silhouette_avg:.3f})")
ax1.set_xlabel("Silhouette Coefficient Values")
ax1.set_ylabel("Cluster Label")

# (B) PCA-based 2D Cluster Projection
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
colors = cm.nipy_spectral(cluster_labels.astype(float) / n_clusters)
ax2.scatter(X_pca[:, 0], X_pca[:, 1], marker='.', s=50, lw=0, alpha=0.7, c=colors,
edgecolor='k')
```

```
# Plotting cluster centroids
centers = pca.transform(kmeans.cluster_centers_)
ax2.scatter(centers[:, 0], centers[:, 1], marker='o', c="white", alpha=1, s=200, edgecolor='k')
for i, c in enumerate(centers):
    ax2.scatter(c[0], c[1], marker=f'${i}$', alpha=1, s=50, edgecolor='k')

ax2.set_title("PCA-based Cluster Visualization")
ax2.set_xlabel("Principal Component 1")
ax2.set_ylabel("Principal Component 2")

plt.tight_layout()
plt.show()
```

Listagem do código-fonte para o teste ANOVA

```
import pandas as pd

# Load data
df = pd.read_excel('база для сегментации.xlsx', sheet_name='DEAL_')

# Calculate new features
df["Deal duration (days)"] = (df["Deal close date"] - df["Deal open date"]).dt.days
product_columns = df.columns[7:26]
df["Total items"] = df[product_columns].fillna(0).sum(axis=1)
df["Product variety"] = df[product_columns].notna().sum(axis=1)

# from sklearn.cluster import KMeans
# features = ... # generated feature set
# kmeans = KMeans(n_clusters=10, random_state=42)
# df['Cluster'] = kmeans.fit_predict(features)

from scipy.stats import f_oneway

# List of features for ANOVA analysis
```

```
features = ["Price EUR", "Deal duration (days)", "Total items", "Product variety"]

# For each feature
for feat in features:
    # Create list of value arrays grouped by cluster
    groups = [df[df['Cluster'] == k][feat].dropna() for k in sorted(df['Cluster'].unique())]

    # Apply ANOVA (Analysis of Variance)
    f_stat, p_val = f_oneway(*groups)

    print(f"Feature: {feat:20s} | F-statistic: {f_stat:.4f} | p-value: {p_val:.4g}")

    if p_val < 0.05:
        print("  -> Differences between clusters are STATISTICALLY SIGNIFICANT (p < 0.05)\n")
    else:
        print("  -> No statistically significant differences between clusters (p ≥ 0.05)\n")
```

Listagem do código-fonte para o cálculo do CLV por cluster

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Load the file and data
file_path = '/mnt/data/segmentation_base.xlsx'
df = pd.read_excel(file_path, sheet_name='DEAL_')

# Feature engineering
df["Deal duration (days)"] = (df["Deal close date"] - df["Deal open date"]).dt.days
product_columns = df.columns[7:26]
df["Total items"] = df[product_columns].fillna(0).sum(axis=1)
df["Product variety"] = df[product_columns].notna().sum(axis=1)
df["Source"] = df["Source"].astype("category")
```

```
df_encoded = pd.get_dummies(df[["Source"]], drop_first=True)

# Final dataset
features = pd.concat([
    df[["Price EUR", "Deal duration (days)", "Total items", "Product variety"]],
    df_encoded
], axis=1).fillna(0)

# Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(features)

# Clustering (10 clusters)
kmeans = KMeans(n_clusters=10, random_state=42, n_init=10)
cluster_labels = kmeans.fit_predict(X_scaled)
df["Cluster"] = cluster_labels

# Cluster profile
cluster_stats = df.groupby("Cluster").agg(
    num_clients=("Контакт", "nunique"),
    num_deals=("Контакт", "count"),
    mean_cheque=("Price EUR", "mean"),
    median_cheque=("Price EUR", "median"),
    mean_deals_per_client=("Контакт", lambda x: x.value_counts().mean()),
    median_deals_per_client=("Контакт", lambda x: x.value_counts().median()),
    min_cheque=("Price EUR", "min"),
    max_cheque=("Price EUR", "max"),
    mean_total_items=("Total items", "mean"),
    median_total_items=("Total items", "median"),
    mean_duration=("Deal duration (days)", "mean"),
    median_duration=("Deal duration (days)", "median"),
    one_deal_clients_share=("Контакт", lambda x: (x.value_counts() == 1).mean())
)
```

```
# CLV: Average total purchase amount per client in the cluster
clv_stats = df.groupby(["Cluster", "Контакт"])["Price EUR"].sum().groupby("Cluster").agg(["mean", "median"])
cluster_stats["mean_clv"] = clv_stats["mean"]
cluster_stats["median_clv"] = clv_stats["median"]

import ace_tools as tools
tools.display_dataframe_to_user(name="Detailed Cluster Statistics", dataframe=cluster_stats)

cluster_stats.reset_index(inplace=True)
cluster_stats.round(2).head(10)  # Summary for interpretation
```