RELATÓRIO FINAL – SEGMENTAÇÃO DE CLIENTES

Autor: Leonardo Nassib Fonseca
Projeto: Análise e modelagem de dados de clientes para identificação de segmentos com características semelhantes.
Objetivo: Identificar segmentos distintos de clientes com base em características comportamentais e de consumo, visando compreender padrões de compra e apoiar a tomada de decisão estratégica, com foco em ações de maior rentabilidade financeira.

1. Resumo Executivo

Este projeto teve como objetivo realizar a segmentação de clientes a partir de dados comportamentais e transacionais, visando identificar grupos com características semelhantes de consumo, engajamento e valor financeiro. A análise envolveu etapas de limpeza, engenharia de variáveis, exploração dos dados, preparação, redução de dimensionalidade e aplicação de diferentes algoritmos de clusterização.

Foram testados métodos como K-Means, Gaussian Mixture Models (GMM) e Hierarchical Clustering (HC), avaliados por métricas como Silhouette Score, além do uso de técnicas de redução de dimensionalidade (PCA, UMAP, t-SNE e abordagem baseada em árvores). Os resultados indicaram que o K-Means com três clusters apresentou o melhor equilíbrio entre qualidade estatística, estabilidade e interpretabilidade para o negócio.

A segmentação final revelou perfis bem definidos de clientes , incluindo grupos de alto valor, baixo engajamento e consumo moderado, gerando insights acionáveis para direcionamento de campanhas, estratégias de retenção, upsell e alocação mais eficiente de recursos.

2. Problema de Negócio

A empresa enfrenta o desafio de compreender melhor o comportamento de seus clientes, que apresentam diferentes níveis de consumo, engajamento e resposta a campanhas de marketing. Tratar todos os clientes de forma homogênea pode resultar em desperdício de recursos, baixa efetividade de campanhas e perda de oportunidades de maximizar o valor do relacionamento com o cliente.

Diante disso, surge a necessidade de responder perguntas-chave como:

- Existem grupos distintos de clientes com padrões de compra semelhantes?
- Quais segmentos concentram maior valor financeiro e engajamento?
- Quais perfis apresentam potencial para retenção, reativação ou crescimento de receita?

A segmentação de clientes torna-se essencial para direcionar ações estratégicas, personalizar campanhas de marketing, priorizar investimentos e apoiar decisões orientadas por dados, aumentando a rentabilidade e a eficiência das iniciativas comerciais.

3. Dados Utilizados

O dataset utilizado contém informações estruturadas o comportamento dos clientes, incluindo:

- Ano de nascimento do cliente
- Nível educacional
- Estado civil
- Renda anual do cliente
- Número de dependentes em casa
- Número de dias desde a última compra
- Valor gasto em geral
- Número de compras em diversos canais
- Aceitação em campanhas

O dataset utilizado é composto por 2.240 clientes e 29 variáveis, reunindo informações demográficas (idade, estado civil, escolaridade, renda), estrutura familiar (presença de crianças e adolescentes), comportamento de compra, histórico de gastos

por categoria de produto, interação com canais de venda e resposta a campanhas de marketing.

Embora seja uma base relativamente rica em atributos comportamentais e transacionais, o conjunto de dados apresenta limitações típicas de bases de marketing, como a ausência de informações externas (ex.: contexto econômico, concorrência ou histórico longitudinal detalhado). Ainda assim, o dataset é adequado para análises exploratórias, segmentação de clientes e identificação de padrões de consumo, oferecendo insumos relevantes para a construção de estratégias orientadas por dados.

4. Estratégia de Solução

A solução foi estruturada em nove etapas principais:

1. **Descrição dos dados:** Identificação da estrutura da base, tipos de variáveis, valores ausentes, duplicidades e aderência ao problema de negócio.
2. **Engenharia de variáveis:** Criação, transformação e recodificação de variáveis para melhorar a representatividade dos dados.
3. **Filtragem das variáveis:** Remoção de atributos redundantes, irrelevantes ou indisponíveis em ambiente produtivo.
4. **Análise exploratória:** Investigação de distribuições, correlações e padrões de comportamento para orientar a modelagem.
5. **Preparação dos dados:** Codificação de variáveis categóricas, padronização, tratamento de valores ausentes e aplicação de técnicas de redução de dimensionalidade.
6. **Seleção de variáveis:** Identificação do conjunto mais relevante de variáveis para representar o comportamento dos clientes.
7. **Modelagem e avaliação:** Aplicação de diferentes algoritmos de clusterização e comparação por métricas como Silhouette Score.
8. **Avaliação do modelo final:** Escolha do modelo com melhor equilíbrio entre qualidade estatística, estabilidade dos clusters e interpretabilidade para o negócio.
9. **Análise dos clusters:** Interpretação dos segmentos formados a partir das variáveis originais, identificando perfis distintos de clientes e seus comportamentos.

5. Principais Insights

1. **Clientes com filhos compram mais frequentemente que clientes sem filhos?**

Os dados indicam que clientes sem filhos apresentam uma frequência de compra ligeiramente maior em comparação aos clientes com filhos. Isso sugere que a presença de dependentes pode influenciar o padrão de consumo, reduzindo a recorrência das compras (figura 1).

2. Clientes casados compram mais que clientes solteiros?

A análise mostra que clientes casados não compram mais do que clientes solteiros, apresentando comportamentos muito semelhantes. Curiosamente, o grupo de clientes viúvos é o que apresenta a maior frequência de compras, indicando um padrão de consumo distinto dentro do estado civil (figura 2).

3. Clientes digitais respondem menos às campanhas?

Contrariando a hipótese inicial, os clientes digitais respondem significativamente mais às campanhas do que os clientes não digitais. Esse resultado evidencia a eficácia dos canais digitais como meio de comunicação e conversão para ações de marketing (figura 3).

6. Resultados da Modelagem

Foram avaliados diferentes algoritmos de clusterização não supervisionados (K-Means, GMM e Hierarchical Clustering) ao longo do projeto. Inicialmente, os modelos foram aplicados sem técnicas de redução de dimensionalidade, o que resultou em baixos valores do coeficiente Silhouette, indicando dificuldade na separação consistente dos grupos e alta sobreposição entre os clusters.

Na sequência, foram testadas técnicas de redução dimensional como PCA, UMAP e t-SNE. Embora essas abordagens tenham melhorado a visualização dos dados em duas dimensões, os resultados mostraram limitações quando poucos componentes explicavam uma parcela restrita da variância total, o que impactou negativamente a qualidade estatística da clusterização em alguns cenários.

Como alternativa, foi adotada uma estratégia baseada em modelos de árvore para seleção das variáveis mais relevantes, utilizando a variável monetary como target. Essa abordagem permitiu preservar relações não lineares entre as variáveis e reduziu ruídos do conjunto de dados, resultando em ganhos significativos na qualidade da clusterização, com destaque consistente para o algoritmo K-Means.

A avaliação final, combinando o coeficiente Silhouette e projeções em PCA para fins de visualização, indicou que o K-Means com k = 3 oferece o melhor equilíbrio entre qualidade estatística, estabilidade dos clusters e interpretabilidade para o negócio. O modelo identificou três perfis distintos de clientes — um grupo mais coeso e homogêneo, um grupo intermediário com maior variabilidade e um grupo mais disperso — refletindo um

cenário realista de segmentação de clientes e fornecendo insights acionáveis para estratégias comerciais e de marketing.

7. Conclusões

Este projeto de segmentação de clientes demonstrou que é possível identificar perfis claramente distintos de comportamento de consumo, engajamento e valor financeiro a partir dos dados disponíveis. A aplicação estruturada de técnicas de preparação dos dados, análise exploratória, avaliação de variância, correlação, seleção de variáveis e algoritmos de clusterização permitiu construir uma segmentação consistente, interpretável e orientada ao negócio.

Embora análises iniciais e métricas estatísticas sugerissem a existência de um número maior de clusters, a escolha estratégica por três clusters mostrou-se mais adequada do ponto de vista prático. Essa decisão priorizou simplicidade, clareza na interpretação e facilidade de uso pelos times de negócio, sem comprometer a capacidade do modelo de capturar diferenças relevantes entre os clientes.

Os clusters identificados refletem um padrão comum em bases de comportamento de consumidores, no qual existem gradientes contínuos de valor e engajamento, e não fronteiras rígidas. Ainda assim, os grupos formados apresentam perfis suficientemente distintos para embasar decisões estratégicas, especialmente em ações de marketing, retenção e priorização comercial.

Dessa forma, o modelo final baseado em K-Means com três clusters atingiu um equilíbrio adequado entre qualidade estatística, estabilidade do agrupamento e valor estratégico, cumprindo o objetivo central do projeto de apoiar decisões orientadas por dados na gestão de clientes.

8. Recomendações

Com base na segmentação obtida, recomenda-se a adoção de estratégias diferenciadas para cada perfil de cliente:

- **Clientes Premium de Alto Valor (Cluster 1):**
  Devem ser o principal foco da estratégia de retenção. Ações como programas VIP, benefícios exclusivos, ofertas personalizadas e comunicação direcionada são fundamentais para manter o relacionamento e maximizar o LTV desse grupo, que representa a maior fatia da receita.
- **Clientes de Consumo Moderado (Cluster 3):**
  Representam uma base sólida para estratégias de crescimento. Recomenda-se investir em ações de **upsell e cross-sell**, campanhas personalizadas e estímulo ao

consumo de categorias premium, aproveitando o bom nível de engajamento e frequência de compra.

- **Clientes de Baixo Valor e Baixo Engajamento (Cluster 2):**
  Devem ser abordados com cautela. Estratégias de ativação pontuais, campanhas de baixo custo e comunicação mais seletiva podem ser testadas, sempre considerando o baixo potencial de monetização. Em alguns casos, a redução de investimentos nesse grupo pode ser uma decisão eficiente.

Por fim, a segmentação construída oferece uma base sólida para evoluções futuras, como o monitoramento dinâmico dos clusters ao longo do tempo, testes de campanhas direcionadas por segmento e integração do modelo a sistemas de recomendação e CRM.

9. Limitações

Este projeto apresenta algumas limitações inerentes ao escopo e à base de dados utilizada. A segmentação foi realizada a partir de variáveis comportamentais e demográficas disponíveis, não contemplando informações externas relevantes, como histórico completo de marketing, dados de concorrência.

Além disso, por se tratar de um problema não supervisionado, a definição do número de clusters envolve decisões estratégicas e interpretação humana, o que pode variar conforme o contexto de negócio. Apesar dessas limitações, os resultados obtidos são consistentes, interpretáveis e oferecem valor prático para apoio à tomada de decisão.

Anexos – Figuras

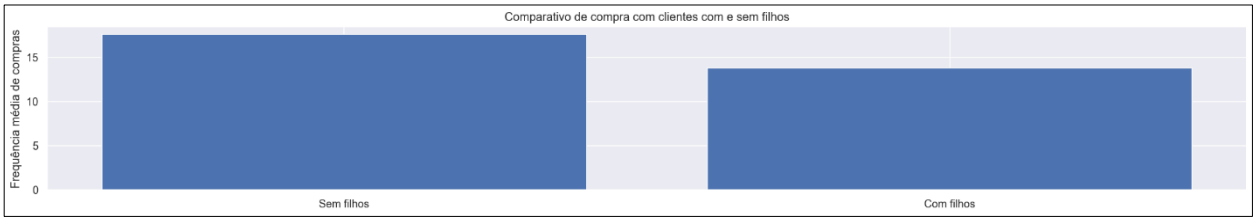

Figura 1 – Frequência de compras por presença de filhos.

A figura 1 indica que clientes sem filhos apresentam uma frequência média de compras ligeiramente superior à dos clientes com filhos. Esse resultado sugere que a ausência de dependentes pode estar associada a maior disponibilidade financeira ou maior recorrência de consumo.

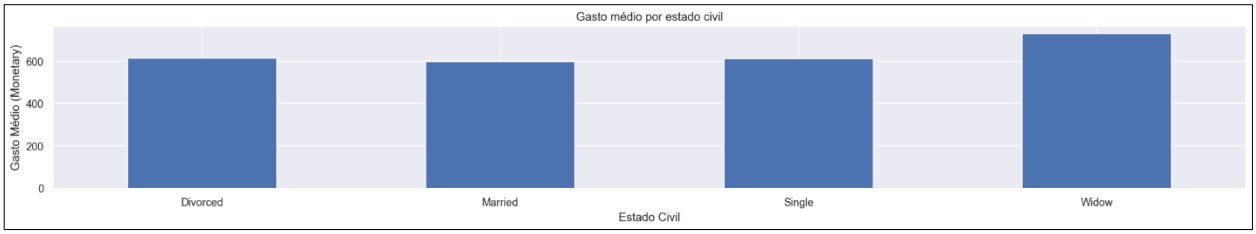

Figura 2 – Frequência de compras por estado civil.

Observa-se na figura 2 que clientes viúvos apresentam o maior gasto médio, seguidos por solteiros, divorciados e casados. A diferença entre casados e solteiros é pequena, indicando que o estado civil, isoladamente, não é um fator determinante do gasto, com exceção do grupo de viúvos.

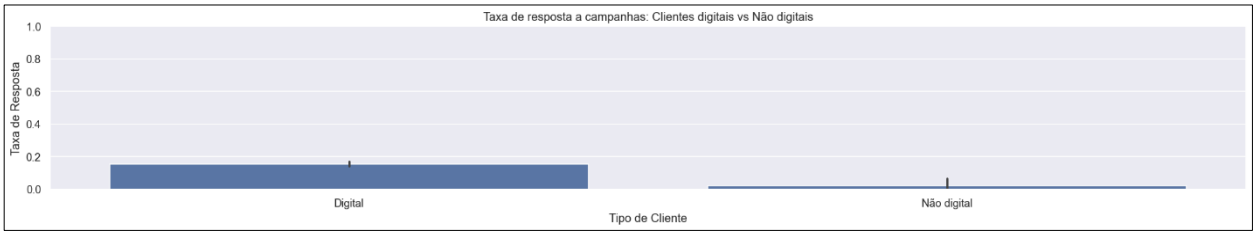

Figura 3 – Taxa de resposta a campanhas por perfil digital.

Conforme figura 3 acima, é possível identificar que clientes digitais respondem significativamente mais às campanhas em comparação aos clientes não digitais. Esse resultado reforça a importância dos canais digitais como meio mais eficiente para ações de marketing e comunicação com os clientes.

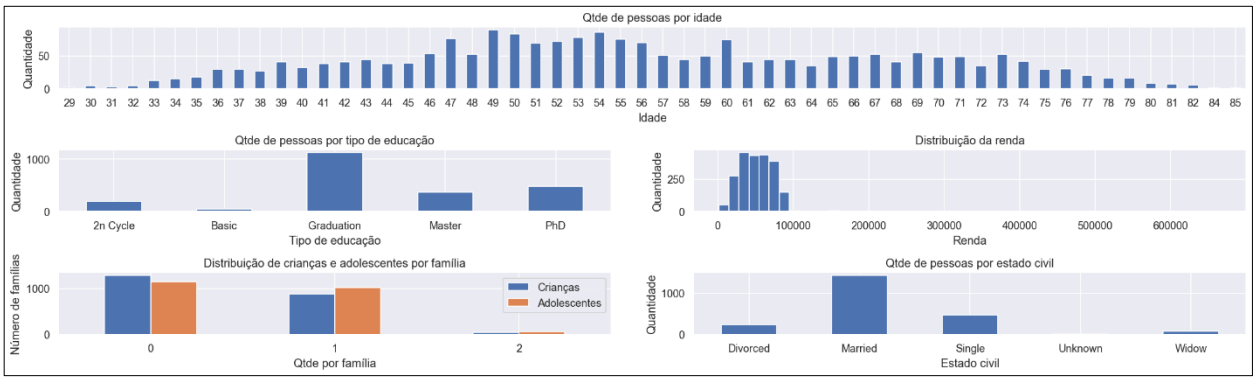

Figura 4 – Comportamento dos clientes.

A figura 4 acima mostram 5 gráficos, ou seja, quantidade de pessoas por idade, estado civil e também por educação, distribuição de renda e dependentes por familia.

- Tipo de educação, a maioria dos clientes tem Graduação, seguido por Master e PhD. O nível Basic tem poucos casos, e 2n Cycle aparece como um intermediário menor.

- Distribuição da renda, a renda é bastante concentrada entre 20.000 e 80.000, com poucos clientes acima de 100.000. Há alguns valores extremos (outliers), mas a grande massa está até 100.000.
- Crianças e adolescentes por família, a maioria das famílias não tem filhos, tanto crianças quanto adolescentes. Entre os que têm, o padrão mais comum é 1 filho e raramente 2.
- Estado civil, a maior parte dos clientes é casada (Married). Após isso aparecem Single e Divorced em proporções menores. A categoria Unknown é pequena e pode indicar dados ausentes.

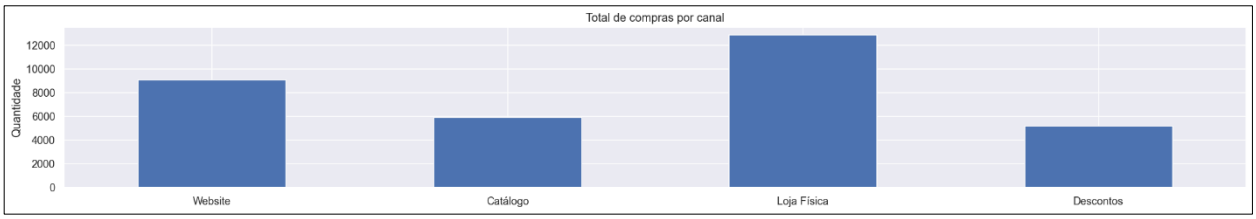

Figura 5 – Total de compras por canal.

É possível observar na figura 5 que a loja física é o principal canal de vendas, pois com mais de 12.000 compras, a loja física é claramente o canal dominante. Isso indica forte presença do público no ponto físico, provável preferência por atendimento presencial e também a importância estratégica das lojas como principal motor de receita.

Website é o segundo canal mais relevante, onde cerca de 9.000 compras, o site também é um canal importante, mostrando boa digitalização do negócio, embora ainda distante da loja física, e por fim, catálogo tem desempenho intermediário e compras por desconto são os menores volumes.

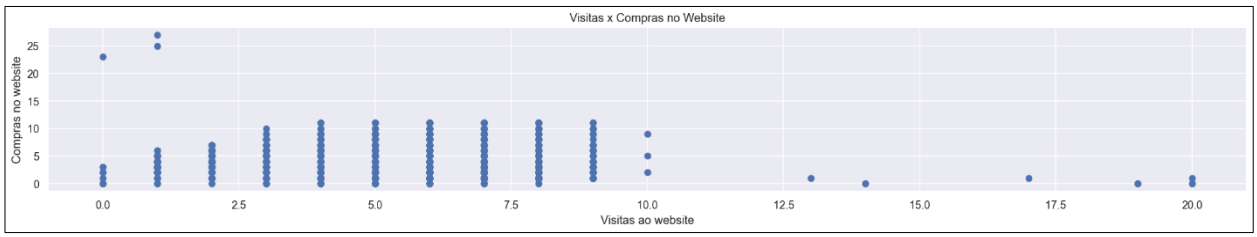

Figura 6 – Visitas vs Compras no website.

Conforme figura 6, não há uma relação linear clara, os pontos estão espalhados de forma bem aleatória, indicando que mais visitas não significa necessariamente mais compras, sugerindo baixa conversão digital, visitas não qualificadas, usuários navegando sem intenção de compra e possível problema de usabilidade ou de preços.

Também existem outliers importantes, pois pontos muito altos (ex.: 23, 26 compras) que são exceções possivelmente clientes VIP ou de comportamento atípico.

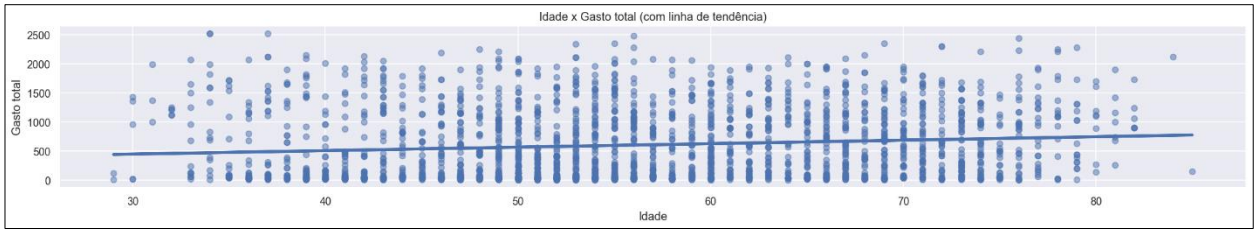

Figura 7 – Idade vs Total de gastos.

A figura 7 acima mostra uma correlação linear positiva, mas extremamente fraca, entre Idade e Gasto total, com a maioria das observações concentradas em baixos valores de Gasto Total, independentemente da idade.

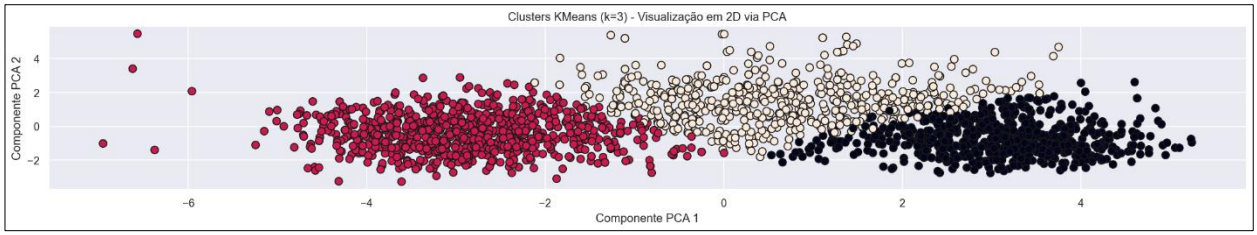

Figura 8 – Visualização dos clusters em 2D.

A análise combinando a projeção em PCA (para fins de visualização) e o coeficiente Silhouette indica que o algoritmo KMeans com k = 3 oferece uma segmentação estável, coerente e compatível com a estrutura encontrada nos dados. Observa-se que o cluster central apresenta a melhor coesão interna, enquanto os outros dois grupos revelam comportamentos distintos, um mais disperso e outro mais concentrado, sugerindo padrões bem diferenciados entre os clientes (figura 8).

Cluster 01: (Clientes Premium de Alto Valor)

- Características principais:
  - Renda: Alta com média de R$ 73.294 (a maior entre os grupos)
  - Idade: Média de 57 anos
  - Gasto: Médio de R$ 1.270,65
  - Consumo: Elevado de vinhos, carnes e produtos premium
  - Frequência de compras: Alta com média de 20,9 compras
  - Dependentes: Baixa presença de dependentes
  - Engajamento: Maior engajamento com campanhas
  - Canal de compra: Multicanal, especialmente loja física, catálogo e website

Cluster 02: (Clientes de Baixo Valor e Baixo Engajamento)

- Características principais:
  - Renda: Baixa com média de R$ 33.262
  - Idade: Média de 53 anos
  - Gasto: Médio de R$ 71,42
  - Consumo: Menor consumo em todas as categorias
  - Frequência de compra: Baixa com média de 7,08 compras
  - Dependentes: Maior número de dependentes
  - Engajamento: Quase nulo com campanhas
  - Canal de compra: Presenciais e poucas compras online

Cluster 03: (Clientes de Consumo Moderado e Perfil Tradicional)

- Características principais:
  - Renda: Moderada com média de R$ 55.496
  - Idade: Média de 59 anos
  - Gasto: Moderado com média de R$ 617,09
  - Consumo: Preferencialmente vinhos e carnes
  - Frequência de compras: Moderada com média de 19,39 compras
  - Dependentes: Maior quantidade de adolescentes como dependentes
  - Engajamento: Moderada às campanhas
  - Canal de compra: Boa atividade multicanal

Figura 9 – Características dos clusters.

A figura 9 é possível identificar quais são as características de cada cluster:

Cluster 01: (Clientes Premium de Alto Valor), este é o grupo mais valioso da base, com maior gasto total e comportamento de compra mais consistente. Este cluster representa clientes de alto poder aquisitivo, maduros, fiéis e com alto valor para a empresa. Eles respondem bem a campanhas e realizam compras frequentes e de produtos de maior ticket.

Para estes clientes, deve-se ter um foco maior na retenção, programas VIPs e campanhas exclusivas, pois representa a maior fatia de receita.

Cluster 02: (Clientes de Baixo Valor e Baixo Engajamento), este grupo com menor renda, baixo gasto e reduzido envolvimento com campanhas da empresa. Este cluster representa clientes sensíveis a preço, pouco fiéis e de baixo LTV. São esporádicos e dificilmente respondem a campanhas.

Para este grupo, seriam necessárias mais ações de ativação e reengajamento, mas possui baixo potencial de monetização.

Cluster 03: (Clientes de Consumo Moderado e Perfil Tradicional), grupo intermediário, com poder aquisitivo médio e comportamento de compra consistente, porém menos intenso que o cluster premium. Este cluster é composto por clientes estáveis, com boa renda e comportamento relativamente previsível. Apesar de não serem tão valiosos quanto o cluster 0, apresentam um bom potencial para campanhas personalizadas.

Para estes consumidores demonstram uma base sólida para upsell, cross-sell e campanhas personalizadas, especialmente em categorias premium de menor consumo.