FIAP
DATA SCIENCE

Grupo Octant

Projeto BeeHive

São Paulo
2025

Sumário

Resumo

Este projeto teve como objetivo segmentar a base de clientes da TOTVS por meio de técnicas de clusterização aplicadas sobre variáveis RFM (Recência, Frequência e Monetário), combinadas a indicadores de NPS, Tickets e Faixa de Faturamento.
A partir da clusterização e dos insights gerados temos o objetivo de auxiliar a TOTVS a conseguir atuar em sua base de clientes, prevenindo churn aumentando a retenção de clientes estratégicos e reativando clientes perdidos. segue abaixo um resumo dos clusters encontros

Cluster -1 – Ruídos

Este cluster reúne apenas 13 clientes sem padrões consistentes de comportamento. As variáveis de faturamento, NPS e tickets não apresentam consistência analítica, caracterizando o grupo como um conjunto de ruídos.

Cluster 0 – Inativos (Long Tail)

Formado por clientes com recência muito alta (cerca de 550 a 700 dias), frequência e monetário praticamente nulos, o que indica ausência de receita recorrente.Este cluster representa clientes cancelados ou contratos antigos, sem relacionamento ativo, mas que podem ser alvo de estratégias de reativação.

Cluster 1 – Ativos de Médio Porte

Com recência longa (550 a 725 dias), frequência baixa e receita quase nula, este grupo reúne clientes que já adquiriram produtos, mas não se engajaram plenamente.São clientes que possivelmente tiveram baixa adesão ou insatisfação, necessitando estratégias de recuperação de relacionamento.

Cluster 2 – Recentes com Baixo Ticket

Concentrado em clientes mais recentes (150 a 300 dias), com contratos únicos ou poucos (1 a 3), mas praticamente sem monetário (MRR) registrado.Este grupo representa clientes novos ou reativados, de baixo valor, que ainda não consolidaram receita recorrente, mas com potencial de evolução.

Cluster 3 – Clientes Recentes e Estratégicos

Com recência baixa (175 a 225 dias) e frequência maior (até 20 contratos), este cluster reúne clientes presentes em faixas de faturamento elevadas, os clientes apresentam estabilidade operacional e alto potencial de expansão, sendo considerados estratégicos.

Cluster 4 – Key Accounts (Alto Valor)

Agrupa clientes ativos com recência entre 200 e 300 dias, frequência baixa (1 contrato), mas com valores monetários muito elevados em casos específicos, chegando a bilhões. Há representatividade em todas as faixas de faturamento, especialmente acima de 75 milhões. Os NPS apresentam comportamento bimodal, com notas neutras ou muito altas (9 e 10), e os *tickets* variam de baixos até casos extremos acima de 1.500. Este cluster representa clientes de alto valor, estratégicos para retenção e com grande potencial de expansão via cross-sell e up-sell.

Descrição do Problema

Com uma base de clientes extensa e heterogênea a TOTVS tem uma grande dificuldade em compreender a sua carteira de clientes e ajustar sua abordagem/jornada para cada cliente, tornando-se necessário identificar padrões de comportamento que permitam:

- Reter clientes estratégicos.
- Reativar clientes inativos.
- Identificar oportunidades de expansão (cross-sell, up-sell).

Metodologia

RFM

A segmentação RFM é utilizada para o ranqueamento e segmentação de consumidores com base no seu comportamento de compras. Esse método é especialmente útil em setores onde há alto número de clientes realizando transações, como no varejo e no e-commerce Christy et al. 2021]. Nesse método, os clientes são agrupados em três dimensões.

Recência

Refere-se ao número de dias desde que um consumidor realizou sua última compra. Para o ranqueamento, quanto menor for esse número, maior é a pontuação de recência. A base de consumidores é dividida em quintis, onde os 20% de clientes mais recentes recebem a pontuação máxima de 5, e os demais Recebem pontuações decrescentes até 1 [Hughes 1994, Wei et al. 2010].

Frequência

É definida como o número de compras que um consumidor fez dentro de um período. Também é classificada de 1 a 5 [Hughes 1994, Wei et al. 2010].

Valor Monetário

Corresponde ao total de dinheiro gasto pelo consumidor em um
período. Nesse caso, a pontuação varia de 1 a 5, com critério análogo à dimensão
de Recência [Hughes 1994, Wei et al. 2010]

Algoritmo

O HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) é uma evolução do DBSCAN, projetada para lidar melhor com conjuntos de dados complexos, distribuídos de forma irregular e com densidades diferentes.

Enquanto o DBSCAN encontra clusters baseando-se em regiões de alta densidade separadas por regiões de baixa densidade, ele exige dois parâmetros principais:

- eps (raio de vizinhança)
- minPts (número mínimo de pontos para formar um cluster)

O problema é que o DBSCAN assume densidade uniforme entre clusters, o que nem sempre acontece. O HDBSCAN resolve isso criando uma hierarquia de clusters baseada em densidade variável e selecionando a solução mais estável dessa hierarquia.

Fonte e Preparação dos Dados

Fonte de Dados

Todos os dados usados no projeto foram cedidos pela TOTVS

Variáveis RFM

Monetário (M)

MRR médio dos últimos 12 meses. (Arquivo: MRR)

Recência (R)

Data da assinatura do contrato mais recente (Arquivo: dados_clientes)

Frequência (F)

Número de contratações nos últimos 12 meses (Arquivo: Contratacoes_ultimos_12_meses)

Distribuições RFM

Distribuições fortemente assimétricas que dificultam a separação de clusters, aumentando a sensibilidade a outliers.

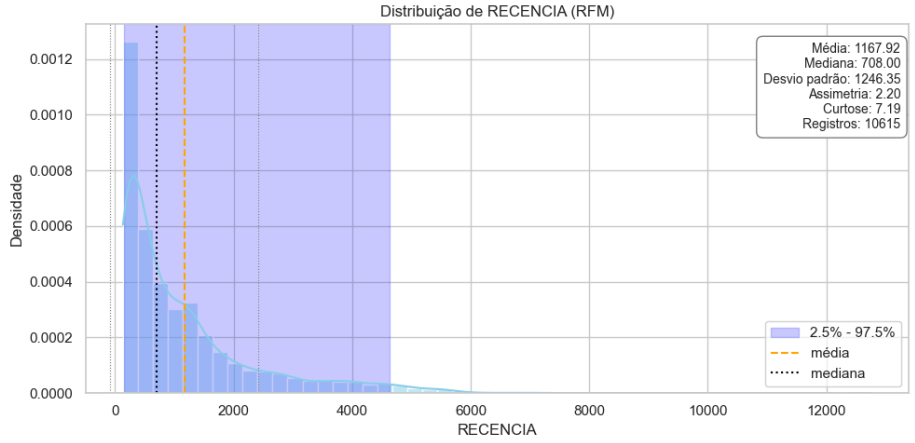


Figura 1: A maioria dos clientes tem valores baixos de recência existe um número significativo de clientes antigos sem novas contratações, o que estende a cauda para a direita

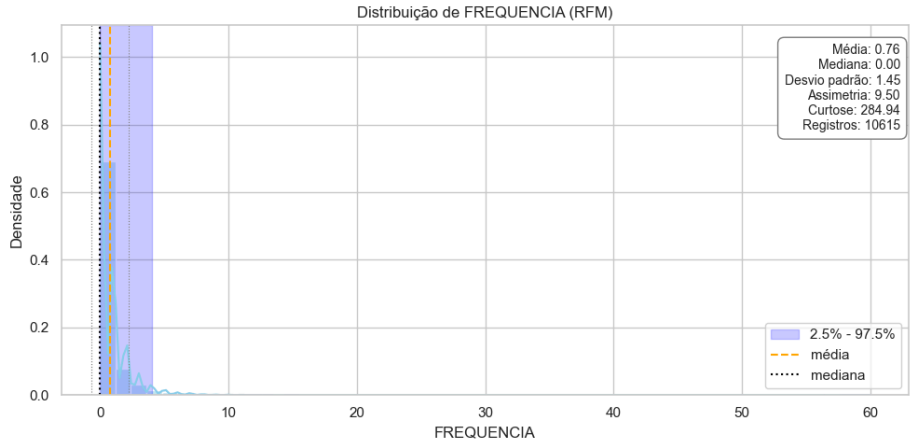


Figura 2: A distribuição está fortemente concentrada à esquerda (próximo de zero) e tem uma cauda longa à direita

Figura 3: A maioria dos clientes tem MRR médio muito baixo e poucos clientes têm MRR muito altos, o que “puxa” a cauda da distribuição.

Solução: Tornar as distribuições mais próximas de normal, melhorando a separação e estabilidade, por uma seleção automática por coluna

Transformações Aplicadas

- Box-Cox e Yeo-Johnson para aproximar distribuições da normalidade
- Padronização com StandardScaler

O código implementa uma pipeline de segmentação de clientes utilizando métricas **RFM** e clusterização com **HDBSCAN Flat**.

Aplicamos um limite superior de 720 dias na Recência — isso significa que, mesmo que existam clientes inativos há mais tempo, eles são tratados como se estivessem nesse teto. Assim, evitamos que esses casos extremos prejudiquem o cálculo das distâncias do algoritmo de clusterização.”

Efeito: melhorar estabilidade e separação dos clusters, mantendo a interpretação “2 anos ou mais” como um nível de inatividade.

As variáveis Recência, Frequência e Monetário passam por transformações estatísticas: Box-Cox é aplicado em variáveis estritamente positivas e Yeo-Johnson nas demais. O objetivo é reduzir a assimetria e aproximar as distribuições da normalidade. Depois, o StandardScaler padroniza todas as variáveis, garantindo que estejam na mesma escala.

Com os dados preparados, aplica-se o HDBSCAN Flat, configurado para 5 clusters, min_cluster_size=120 e min_samples=10.
Esse algoritmo é robusto a ruídos, encontra clusters de diferentes densidades e gera rótulos para cada cliente. Por fim, os resultados são adicionados ao DataFrame, que contém CD_CLIENTE, as métricas RFM e o cluster atribuído.

Resultados

Cluster -1 - **Ruídos** (13 clientes)

O Cluster -1 representa um grupo com apenas 13 clientes distribuídos em diferentes segmentos. Não apresenta padrões consistentes de comportamento

Variáveis de NPS, Tickets e Faixa de Faturamento:

- **Faixa de Faturamento:** valores distribuídos em diferentes faixas, sem concentração dominante.
- **NPS Relacional e Transacional:** poucos registros, sem consistência para análise.
- **Tickets:** alguns clientes com valores altos, mas a maioria sem movimentação relevante.

Cluster 0 - **Clientes inativos- Long Tail** (2485 clientes)

Variáveis RFM:

- **R :** Concentrada em valores altos (entre ~550 e 700 dias), indicando clientes sem contato há muito tempo.
- **F :** Constante e próxima de 0, clientes sem interações no período.
- **M :** Também constante em 0, sugerindo ausência de receita recorrente registrada neste cluster.

Variáveis de NPS, Tickets e Faixa de Faturamento:

Faixa de Faturamento: Concentrada em valores baixos (faixas iniciais), com longa cauda de poucos clientes em faixas mais altas.

NPS Relacional e NPS Transaciona: sem registros

Tickets: Constante em 304

Top Segmentos:

1. Serviços
2. Varejo
3. Manufatura

Dado esse cenário, é possível que sejam clientes cancelados há muito tempo que ainda estão na base mas sem relacionamento ativo, sem geração de receita recorrente e com baixo engajamento.

Esses clientes podem representar:

- Histórico de base legada (contratos antigos não atualizados).
- Oportunidade de reativação (se houver estratégia de campanhas de recuperação específicas para baixo ticket).

Cluster 1 - **Clientes Ativos de Médio Porte** (3816 clientes)

Variáveis RFM:

- R: concentrado em 725, mas inicia em 550
- F: constante em 0
- M: concentrado em 0, mas que vai até 4

Variáveis de NPS, Tickets e Faixa de Faturamento:

Faixa de Faturamento: a maioria sem informações; entre os que possuem registros, destaque para: *Faixa 00 – Até 4,5M / Faixa 03 – De 15M até 25M / Faixa 09 – De 300M até 500M*

NPS Relacional: concentrado em 0, com poucos clientes acima de 8.

NPS Transacional: concentrado em 0, com alguns valores acima de 8.

Tickets: concentrados em 0, poucos registros relevantes.

Top Segmentos:

1. Manufatura
2. Serviços
3. Varejo

Top Subsegmentos:

1. Atacadista e Distribuidor
2. Provedor de Serviços
3. Hospedagem

Dado este cenário, é possível que esses clientes tenham comprado algum produto Totvs, e não gostaram ou a adesão não foi completa por parte da empresa, isso porque o tempo de recência é muito longo, a frequência de aquisição é ruim, e apesar das notas de NPS terem uma concentração em 0, os tickets são poucos também

Cluster 2 - **Recentes com Baixo Ticket** (827 clientes)

Variáveis RFM:

- **R :** concentrada em valores mais baixos (150 a 300 dias), ou seja, clientes com interações relativamente recentes em comparação com outros clusters.
- **F** : Variando entre 1 e 3, com destaque para frequência 1 (contrato único) e alguns casos até 4 ou 5.
- **M:** constante em 0

Variáveis de NPS, Tickets e Faixa de Faturamento:

Faixa de Faturamento: concentrada em valores baixos, mas com alguns clientes em faixas intermediárias.

NPS Relacional e Transacional: poucos registros, com dispersão entre 6 e 8 → percepção neutra/levemente positiva.

Tickets: dispersão ampla, com alguns clientes registrando valores altos, mas a maioria próxima do padrão.

Top Segmentos:

1. Serviços
2. Varejo
3. Manufatura

Top Subsegmentos:

1. Provedor de Serviços
2. Viagens
3. Construtoras

O Cluster 2 concentra clientes mais recentes em relação a outros grupos, mas com baixa frequência de interações e praticamente sem monetário registrado. Os principais segmentos são Serviços e Varejo, com grande peso em Provedores de Serviços.

Sendo assim, um grupo de clientes novos ou reativados, de baixo valor, que ainda não consolidaram receita recorrente, mas têm potencial de evolução no relacionamento.

Cluster 3 - **Clientes Recentes e Estratégicos** (1689 clientes)

Variáveis RFM:

- **R :** concentrada em valores mais baixos (175 e 225 dias), ou seja, clientes com interações relativamente recentes em comparação com outros clusters.
- **F** : Variando um pouco, entre 1 a 20 contratos feitos por cliente
- **M:** constante em 0

Variáveis de NPS, Tickets e Faixa de Faturamento:

Faixa de Faturamento: Há clientes de todas as faixas, destaque para: "Faixa 07 - De 75 M até 150 M", "Faixa 08 - De 150 M até 300 M"

NPS Relacional: Concentração em 0, mas muitas notas acima de 7

NPS Transacional: Concentração em valores maiores que 8

Tickets: Concentrado em 0, poucos tickets

Top Segmentos:

1. Manufatura
2. Varejo
3. Serviços

Top Subsegmentos:

1. Atacadista e Distribuidor
2. Bens de Consumo
3. Bens de Capital

Nos indicadores de satisfação, o NPS Relacional apresenta notas acima de 7 em boa parte dos casos, enquanto o NPS Transacional é consistente em valores acima de 8. O baixo volume de tickets sugere que a operação é mais estável ou menos demandante de suporte.

Trata-se de um cluster com alto potencial de expansão, reunindo clientes relevantes em porte e com boa percepção de valor

Cluster 4 - **Key Accounts** (1785 clientes)

Variáveis RFM:

- **R :** distribuição entre 150 e 550 dias, com maior concentração entre 200 e 300 dias, mostrando clientes que ainda têm interações relativamente recentes.
- **F:** constante em 1
- **M:** fortemente concentrado próximo de 0, mas com cauda longa até valores muito altos (acima de 8 bilhões)

Variáveis de NPS, Tickets e Faixa de Faturamento:

Faixa de Faturamento: Forte concentração em faixas baixas, mas com representatividade em faixas elevadas, especialmente acima de 75M.

NPS Relacional: Bimodal, com concentração em 0 (clientes neutros/inativos na pesquisa) e notas altas, especialmente **acima de 8**.

NPS Transacional: Comportamento similar, com destaque para notas **9 e 10**, sinalizando boa experiência de uso.

Tickets: Concentrados em valores baixos, mas com cauda longa que chega a mais de 1.500 tickets.

Top Segmentos:

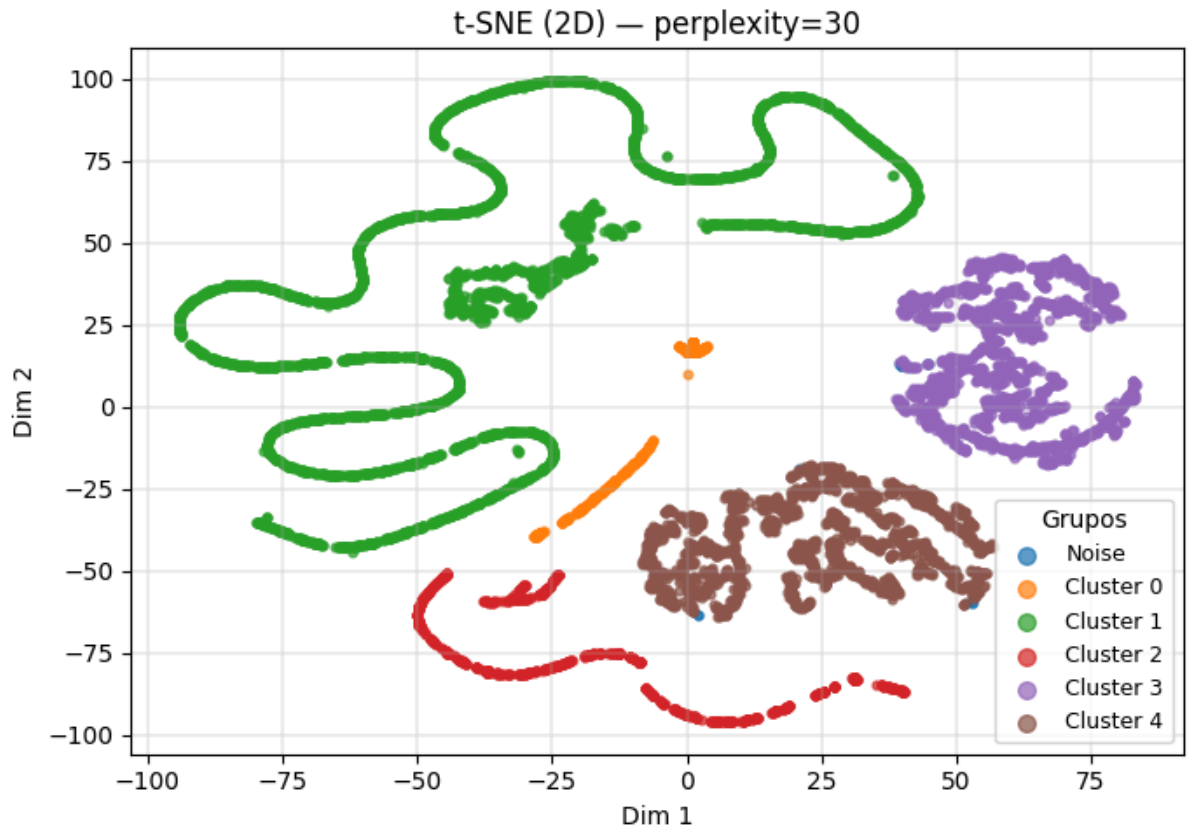
1. Varejo
2. Manufatura
3. Serviços

Top Subsegmentos:

1. Atacadista e Distribuidor
2. Provedor de Serviços
3. Hospedagem

O Cluster 4 representa clientes ativos e heterogêneos, que vão desde empresas menores até grandes grupos com contratos bilionários.

Esse cluster pode ser visto como um grupo de Clientes Ativos de Alto Valor, com potencial tanto de retenção (fidelizar clientes estratégicos) quanto de expansão (cross-sell e up-sell em clientes que já possuem alta relevância).



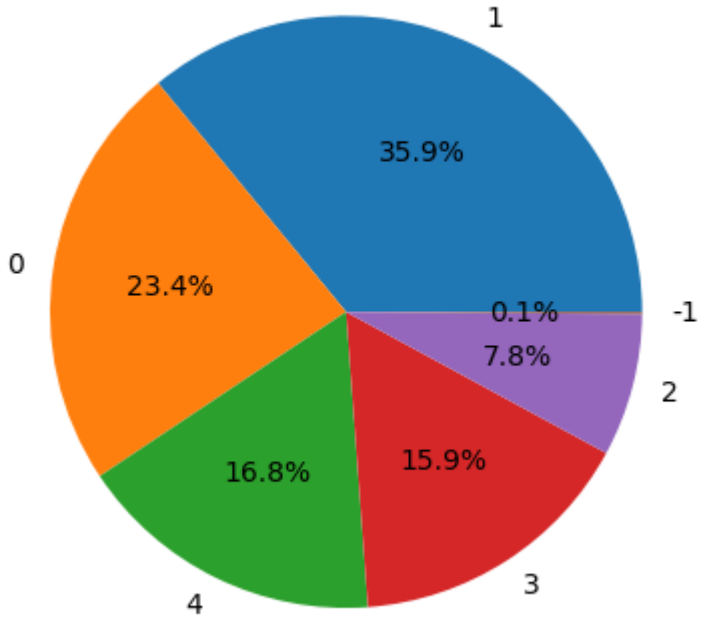
visualização dos clusters usando T-SNE para uma melhor visualização em 2D

Visualização

[Power BI]

Métricas e Avaliação

Distribuição dos Clusters



pie chart para visualizar a distribuição entre os clusters.

Métricas de Avaliação

Silhouette Score: 0,67
Indicando boa separação

Davies-Bouldin: 0,74
Indicando boa coesão entre os clusters

Calinski-Harabasz: 45.970
Indicando boa definição entre os clusters

Estabilidade (ARI bootstrap) 0,97 ± 0,016
indicando excelente reprodutibilidade

Conclusão

A segmentação realizada permitiu identificar perfis distintos de clientes da TOTVS, desde grupos inativos e de baixo engajamento até contas estratégicas e de alto valor. Enquanto os Clusters 0 e 1 representam clientes inativos ou com baixo relacionamento, os Clusters 2 e 3 destacam oportunidades de evolução e expansão em clientes mais recentes e engajados. Já o Cluster 4 reúne as key accounts, fundamentais para retenção e crescimento via cross-sell e up-sell. Assim, os resultados fornecem uma base sólida para direcionar ações específicas de reativação, retenção e expansão, maximizando valor.