Segmentação de Clientes e Análise de Risco de crédito em Pernambuco

Paulo Antônio Martins Silva

2025

Visão Geral do Projeto

No contexto da disciplina de Mineração de Dados do curso de Estatística, desenvolvi uma série de projetos práticos com foco em problemas reais enfrentados por profissionais da área de dados. Entre esses trabalhos, destaco com grande satisfação o projeto de Segmentação de Clientes e Análise de Risco de Crédito, que simula processos adotados por equipes de risco, crédito e ciência de dados no mercado financeiro.

O desenvolvimento foi estruturado em três etapas principais, cada uma representando uma fase essencial do ciclo analítico:

- **Etapa 1 - Preparação da Base:** acesso, leitura e filtragem dos dados brutos do SCR.data utilizando DuckDB e SQL.
- **Etapa 2 - Pré-processamento:** limpeza, padronização, transformação de variáveis e construção de uma base adequada para modelagem.
- **Etapa 3 - Segmentação de Clientes:** aplicação do algoritmo K-Prototypes para identificar perfis distintos de clientes, análise dos grupos resultantes e interpretação dos riscos associados.

Este documento reúne e organiza essas etapas, apresentando o fluxo completo do projeto, desde a construção do dataset até a análise dos perfis e recomendações estratégicas.

Etapa 1 – Preparação da Base de Dados

Nesta etapa inicial, o objetivo foi construir um dataset filtrado a partir do **SCR.data**, disponibilizado pelo Banco Central do Brasil. O SCR.data é um conjunto de dados públicos que reúne informações agregadas sobre operações de crédito de pessoas físicas e jurídicas, permitindo análises de risco sem expor dados pessoais.
Para isso, foi utilizada a combinação de **DuckDB** e **SQL**, permitindo consultar arquivos CSV brutos de grande volume de forma eficiente, sem necessidade de carregá-los completamente na memória.

Fonte de dados

Os arquivos foram obtidos em:

https://dadosabertos.bcb.gov.br/dataset/scr_data

A tarefa consistiu em baixar os dados referentes ao ano de 2025 e selecionar o mês mais recente disponível (agosto de 2025).

Escopo definido

Duas possibilidades de estudo foram propostas dentro do tema de risco de crédito:

1. Extração de Micro-Segmentos de Risco (Pessoa Física)
2. Extração de Risco Setorial (Pessoa Jurídica)

Para este projeto, foi escolhida a **opção 1**, com foco exclusivamente em clientes **Pessoa Física** do estado de **Pernambuco (PE)**.

Objetivo da etapa

Preparar um dataset contendo apenas registros de clientes PF do estado de PE, reduzindo o volume e mantendo apenas o que é relevante para análises futuras, como pré-processamento, modelagem estatística e segmentação.

Filtro aplicado

- cliente = `"PF"`
- uf = `"PE"`

Processo de Extração com DuckDB e SQL

A consulta foi estruturada da seguinte forma:

- Conexão com o DuckDB em memória
- Definição do caminho do arquivo CSV bruto
- Construção de uma consulta SQL para filtrar clientes PF de Pernambuco
- Exportação do resultado para um arquivo `.csv` usado na etapa seguinte

Etapa 2 – Pré-processamento dos Dados

Após a preparação da base filtrada do SCR.data, iniciou-se o processo de pré-processamento, etapa essencial para garantir a qualidade dos dados antes de qualquer modelagem estatística ou aplicação de algoritmos de machine learning. O objetivo desta fase foi limpar, padronizar, transformar e estruturar o dataset para deixá-lo pronto para a etapa de segmentação.

A base utilizada contém **5.086 operações de crédito** referentes ao estado de Pernambuco, no mês de agosto de 2025, distribuídas inicialmente em **23 variáveis** categóricas e numéricas.

Carregamento da base e inspeção inicial

A base `analise_pe.csv` foi importada e avaliada com funções de inspeção como `glimpse()` e `summary()`. Essa etapa permitiu verificar tipos incorretos, presença de texto onde deveria haver números, e identificar variáveis redundantes.

Limpeza inicial das variáveis

Após essa inspeção, algumas variáveis foram removidas por não contribuírem para análises futuras:

- **data_base**, **uf** e **cliente** – valores repetidos e sem variância.
- **cnae_secao** e **cnae_subclasse** – referentes a Pessoas Jurídicas (PJ), não sendo relevantes para a análise exclusiva de PF.

Esse procedimento reduz ruído e agiliza o processamento dos dados.

Padronização de valores textuais

Algumas variáveis categóricas continham o prefixo `"PF - "` no início das categorias (como porte, ocupação e modalidade). Para facilitar leitura e interpretação dos dados, esse prefixo foi removido. Isso deixa as categorias mais limpas e evita duplicações acidentais no momento da análise.

Tratamento e transformação das variáveis numéricas

Ajuste da variável `numero_de_operacoes`

Essa variável continha valores no formato `<= 15`, impedindo conversão direta para tipo numérico. Para padronizar, valores desse tipo foram substituídos pelo ponto médio (**8**), arredondado para cima. Em seguida, toda a coluna foi convertida para numérica.

Conversão de colunas financeiras

Muitas variáveis numéricas eram importadas como texto devido ao uso de vírgula como separador decimal. Para corrigir, todas as vírgulas foram substituídas por pontos antes da conversão para número. Esse procedimento foi aplicado a:

- vencido_acima_de_15_dias
- carteira_ativa
- carteira_inadimplida_arrastada

- ativo_problematico
- todas as variáveis que começam com "a_vencer_"

Essa etapa garantiu a consistência dos valores financeiros presentes na base.

Verificação das categorias e dados ausentes

As variáveis categóricas foram inspecionadas para verificar distribuição e possíveis inconsistências. Em seguida:

- Foi analisado o resumo estatístico das variáveis
- Verificada a presença de valores faltantes

Após as transformações, a base ficou **sem valores ausentes**, ideal para modelagem.

Padronização Z-Score

Para evitar que variáveis com escalas diferentes influenciassem a segmentação, foi realizada a padronização Z-Score utilizando o pacote `tidymodels`:

- O **recipe** foi definido aplicando `step_normalize()` a todas as variáveis numéricas.
- Em seguida, a receita foi **treinada** com `prep()`.
- Por fim, a padronização foi **aplicada** à base completa via `bake()`.

Essa etapa garantiu que todas as variáveis numéricas ficassem:

- Média = 0
- Desvio padrão = 1

A inspeção final confirmou que a padronização foi aplicada corretamente, deixando o dataset pronto para a etapa de segmentação de clientes.

Etapa 3 – Segmentação de Clientes

Após o pré-processamento e padronização da base de dados, iniciou-se a etapa de segmentação, cujo objetivo foi identificar grupos distintos de clientes com características semelhantes. A partir dessa análise, foi possível compreender diferentes perfis de risco, comportamento financeiro e potenciais oportunidades comerciais dentro do portfólio de crédito.

1. Sumário Executivo

A segmentação revelou **seis perfis distintos de clientes**, cada um com padrões específicos de renda, comportamento de pagamento, modalidade de crédito e risco associado.

Entre os grupos identificados, destacam-se:

- **Perfil 3 – Clientes com Alto Uso de Crédito**, que apresenta maior necessidade de ações preventivas e renegociação.
- **Perfil 5 – MEIs de Alta Renda com Baixo Risco**, que representa forte potencial para expansão de crédito produtivo e oferta de novos produtos.

Os demais perfis também trazem importantes indicações de risco e oportunidades, detalhadas ao longo desta seção.

2. Introdução e Objetivo

O portfólio analisado reúne clientes com características heterogêneas, variando amplamente em termos de renda, nível de endividamento, modalidade de crédito utilizada e comportamento de pagamento. Tratar todos esses clientes da mesma forma pode levar a decisões ineficientes, maior exposição ao risco e perda de oportunidades comerciais.

Diante disso, o objetivo desta etapa foi **identificar perfis acionáveis de clientes**, utilizando técnicas de clusterização para revelar grupos com padrões semelhantes. Com esses perfis, torna-se possível:

- Direcionar estratégias comerciais,
- Definir políticas de crédito mais eficientes,
- Antecipar riscos comportamentais,
- Melhorar a alocação de recursos de cobrança e retenção.

3. Metodologia de Segmentação

3.1. Fonte de Dados

A base utilizada na segmentação corresponde à versão final tratada e padronizada na etapa anterior, contendo **5.086 observações** e **14 variáveis selecionadas**, todas referentes a operações de crédito agregadas de Pernambuco (agosto de 2025), disponibilizadas pelo Banco Central do Brasil.

3.2. Justificativa do Algoritmo

A base combina variáveis **numéricas** (ex.: carteira ativa, ativo problemático, valores em atraso) e **categóricas** (ex.: ocupação, porte, modalidade). Métodos tradicionais de clusterização apresentam limitações:

- **K-Means** exige apenas variáveis numéricas;
- **K-Modes** funciona somente com variáveis categóricas.

Para lidar simultaneamente com ambos os tipos, foi adotado o **K-Prototypes**, algoritmo híbrido que utiliza:

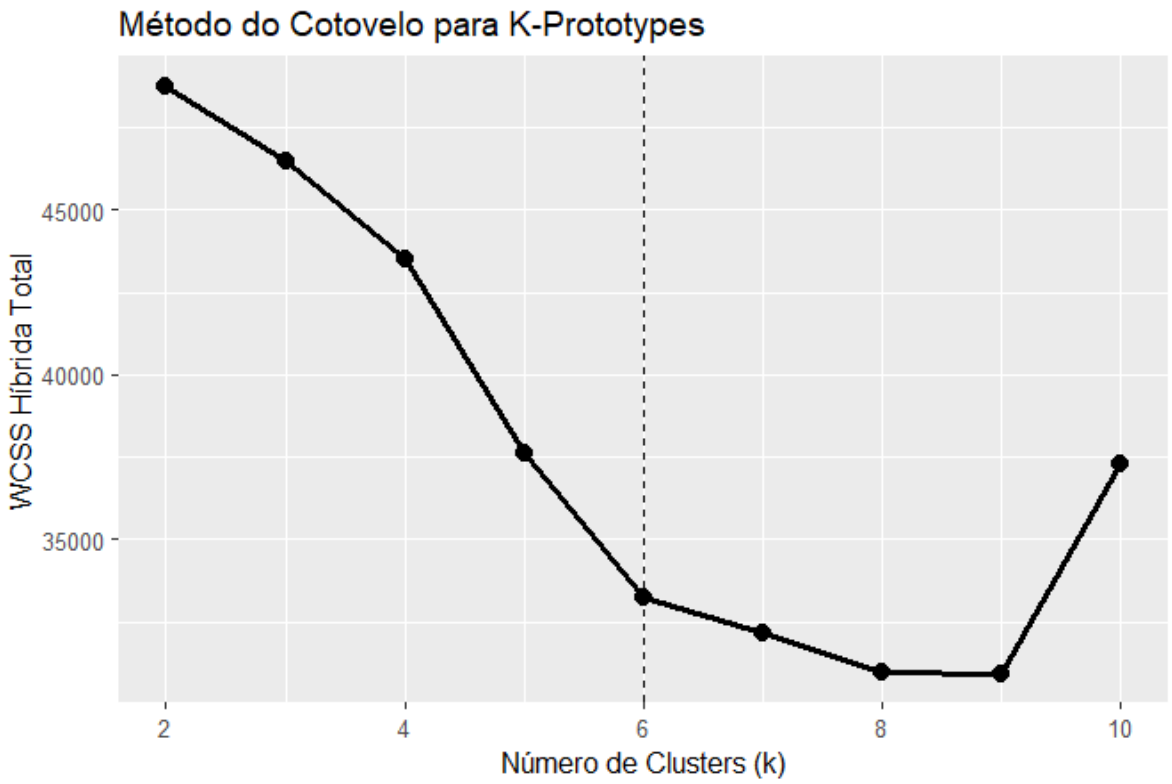
- Distância Euclidiana para variáveis numéricas padronizadas,
- Distância de Hamming para variáveis categóricas.

Essa combinação permite criar clusters coerentes e representativos da realidade do portfólio de crédito.

Essa abordagem torna o K-Prototypes especialmente adequado para bases financeiras que misturam atributos demográficos, comportamentais e monetários.

3.3. Definição do Número de Grupos (k)

O método do cotovelo foi utilizado como referência para avaliar como o desempenho da segmentação varia conforme diferentes valores de k. Com base nessa análise e considerando a diversidade do portfólio, optou-se por trabalhar com **6 clusters**, pois esse número permite distinguir melhor os diferentes perfis de clientes e revelar diferenças importantes para a tomada de decisão em risco e estratégia comercial.



4. Resultados: Perfis Identificados

A seguir, são apresentados os seis perfis encontrados, acompanhados de interpretações e pontos de destaque para cada grupo.

Perfil 1 – Cliente de Baixo Risco e Baixo Volume

Quem é: Baixa renda (1–2 salários), operações pequenas e baixo endividamento.
Ocupação: Predominância de "Outros".
Modalidade: Outros créditos (baixa complexidade).

Comportamento financeiro:

- Volume médio muito baixo (aprox. 366 mil operações).
- Valores reduzidos em faixas de atraso.
- Inadimplência baixa e proporcional ao porte.
- Carteira ativa relevante (cerca de R$ 169 milhões), porém com risco mínimo.
- Segmento massivo (2.400 clientes), baixo tíquete.

Perfil 2 – Super-Endividados de Alta Renda e Alto Risco

Quem é: Renda elevada (5–10 salários), porém com risco extremamente alto.
Ocupação: "Outros" e Servidor Público.
Modalidade: Habitacional.

Comportamento financeiro:

- Volumes altíssimos (média acima de 95 mil operações).
- Valores expressivos em todas as faixas de atraso (centenas de milhões).
- Inadimplência acumulada muito alta (R$ 29 milhões).
- É o grupo mais crítico, apesar de possuir poucos clientes (17).

Perfil 3 – Clientes com Alto Uso de Crédito

Quem é: Baixa renda (até 1 salário), mas utilização muito intensa do crédito.
Ocupação: "Outros".
Modalidade: Cartão de crédito.

Comportamento financeiro:

- Elevado volume médio (872 mil operações).
- Atrasos concentrados em faixas antigas (R$ 103 milhões).
- Ativo problemático elevado (R$ 137 milhões).
- Histórico de alta exposição ao risco ao longo do tempo.

Perfil 4 – Clientes de Risco Moderado com Forte Uso de Cartão

Quem é: Renda média (3–5 salários).
Ocupação: "Outros".
Modalidade: Cartão de crédito.

Comportamento financeiro:

- Volume médio de 82 mil operações.
- Atrasos significativos, embora inferiores aos perfis 2 e 3.
- Carteira ativa de cerca de R$ 34 bilhões.
- Sinais de risco crescente (R$ 143 milhões em vencido 15+).

Perfil 5 – MEIs de Alta Renda com Baixo Risco

Quem é: Renda elevada (10–20 salários).
Ocupação: MEI.
Modalidade: Cartão de crédito.

Comportamento financeiro:

- Baixo volume médio (7 mil operações).
- Baixos níveis de atraso e inadimplência.
- Carteira ativa de R$ 2,4 bilhões.
- Segmento numeroso (1.117), com alta capacidade de pagamento.

Perfil 6 – Autônomos de Médio Risco e Baixa Renda

Quem é: Renda entre 2 e 3 salários.
Ocupação: Autônomos.
Modalidade: Empréstimo sem consignação.

Comportamento financeiro:

- Baixa movimentação (1.270 operações).
- Atrasos moderados distribuídos entre as faixas.
- Inadimplência média (R$ 341 mil).
- Segmento relevante (1.444 clientes).

5. Recomendações Estratégicas

As recomendações abaixo foram elaboradas com foco em ações acionáveis, priorização de risco e oportunidades comerciais.

Perfil 1 – Baixo Risco e Baixo Volume

- Oferecer microcrédito simplificado.
- Enviar comunicações educativas.

Perfil 2 – Alto Risco e Super-Endividamento

- Renegociar com foco em prazos longos.
- Encaminhar para equipe especializada de recuperação.

Perfil 3 – Alto Uso de Crédito

- Reduzir limites e oferecer acordos imediatos.
- Implementar alertas de uso excessivo.

Perfil 4 – Risco Moderado

- Incentivar migração para produtos com juros menores.
- Monitorar atrasos acima de 15 dias.

Perfil 5 – MEIs de Baixo Risco

- Aumentar limites e oferecer capital de giro.
- Ofertar produtos de valor agregado (maquininha, seguros, antecipação).

Perfil 6 – Autônomos de Médio Risco

- Oferecer crédito com parcelas flexíveis.
- Incluir seguro prestamista e reforçar análise comportamental.

Descrição das Tabelas

Tabela 1 - Perfil financeiro médio dos grupos de clientes

Variável	Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5	Perfil 6
Operacoes	365,67	95.788,65	872.343,85	82.037,21	7.082,91	1.270,20
Ate90	281.530,16	83.885.380,04	392.425.023,48	82.984.901,17	5.705.720,08	608.655,47
De91a360	402.853,28	218.780.265,93	176.766.519,16	73.996.025,17	5.642.011,82	1.041.823,94
De361a1080	470.057,85	450.792.007,71	15.302.734,19	71.080.067,87	6.039.616,49	1.199.236,26
De1081a1800	186.689,31	283.279.647,81	3.724.653,07	32.967.255,67	2.445.239,26	435.989,60
De1801a5400	189.284,74	468.226.097,05	1.630.530,07	50.786.649,72	2.599.286,76	384.994,65
Acima5400	40.785,52	192.681.687,61	406.717,80	17.902.754,59	630.363,72	80.076,88
Vencido15	127.321,65	7.163.234,28	103.661.333,20	14.308.853,65	1.437.233,28	230.529,56
Ativa	169.852.250,79	170.480.832.043,71	69.391.751.097,54	34.402.650.784,67	2.449.947.140,65	398.130.635,63
Inadimplida	175.832,99	29.252.197,49	89.255.509,50	19.316.761,28	1.797.614,02	341.407,05
AtivoProb	288.308,49	80.617.154,18	137.770.519,55	37.756.920,50	3.085.971,83	591.927,72
Contagem	2.400,00	17,00	13,00	95,00	1.117,00	1.444,00

A Tabela 1 apresenta os valores médios das principais variáveis financeiras para cada um dos seis perfis de clientes identificados na segmentação.
Ela resume como cada grupo se comporta em termos de:

- quantidade média de operações agregadas,
- valores a vencer em diferentes horizontes de tempo,
- valores vencidos,

- carteira ativa,
- inadimplência,
- ativos problemáticos,
- e quantidade de clientes em cada perfil.

Esses indicadores permitem comparar o nível de exposição ao risco, o tamanho das carteiras e o comportamento de pagamento entre os grupos.
A tabela oferece uma visão clara de quais perfis concentram maior volume financeiro, maior risco ou maior representatividade.

Tabela 2 – Características predominantes dos grupos de clientes

A Tabela 2 sintetiza as características sociodemográficas e operacionais predominantes em cada cluster.
Para cada perfil, são destacados três atributos principais:

- **Ocupação**
- **Porte econômico** (faixa de renda estimada)
- **Modalidade predominante de crédito**

Essa tabela complementa a anterior ao mostrar quem são os clientes de cada grupo, permitindo interpretar não apenas o comportamento financeiro, mas também o contexto socioeconômico de cada perfil.

Tabela 2 – Características predominantes dos grupos de clientes

Cluster	Ocupacao	Porte	Modalidade
Perfil 1	Outros	Mais de 1 a 2 salários mínimos	Outros créditos
Perfil 2	Outros	Mais de 5 a 10 salários mínimos	Habitacional
Perfil 2	Servidor ou empregado público	Mais de 5 a 10 salários mínimos	Habitacional
Perfil 3	Outros	Até 1 salário mínimo	Cartão de crédito
Perfil 4	Outros	Mais de 3 a 5 salários mínimos	Cartão de crédito
Perfil 5	MEI	Mais de 10 a 20 salários mínimos	Cartão de crédito
Perfil 6	Autônomo	Mais de 2 a 3 salários mínimos	Empréstimo sem consignação em folha

Conclusão

A segmentação realizada neste projeto permitiu compreender de maneira estruturada a diversidade existente no portfólio de crédito analisado. A partir da preparação, limpeza e padronização dos dados, foi possível aplicar um método de clusterização adequado ao formato misto da base e identificar **seis perfis distintos de clientes**, cada um com padrões próprios de renda, comportamento de pagamento, nível de risco e potencial comercial.

Os resultados mostraram que tratar todos os clientes de forma homogênea não é eficiente. Ao contrário, os perfis revelam diferenças importantes que permitem ações mais direcionadas, seja

para reduzir inadimplência, identificar oportunidades de expansão, ajustar limites de crédito, fortalecer relações com clientes de baixo risco ou atuar preventivamente com grupos mais vulneráveis.

Além disso, o processo reforça a importância do pré-processamento cuidadoso, da escolha apropriada do algoritmo e da interpretação contextual do portfólio para transformar dados brutos em decisões estratégicas.

Em síntese, a segmentação fornece um mapa claro dos diferentes tipos de clientes presentes na carteira e serve como base para políticas mais inteligentes, personalizadas e voltadas para resultados tanto em risco quanto em estratégia comercial.

Dicionário de Variáveis

A tabela abaixo descreve cada variável utilizada no projeto após o pré-processamento, seu tipo e seu significado no contexto de risco de crédito.

Variável	Tipo	Descrição
tcb	Nominal	Tipo de carteira bancária (ex.: bancário, não bancário).
sr	Nominal	Segmento de risco informado pela instituição.
ocupacao	Nominal	Ocupação declarada do cliente (aposentado, autônomo, MEI, etc.).
porte	Ordinal	Porte econômico do cliente, geralmente relacionado à renda.
modalidade	Nominal	Modalidade da operação de crédito (ex.: cartão, empréstimo, habitacional).
origem	Nominal	Origem ou finalidade declarada do crédito.
indexador	Nominal	Tipo de indexador aplicado à operação.
numero_de_operacoes	Numérica	Quantidade de operações de crédito agregadas por cliente.
a_vencer_ate_ 90_dias	Numérica	Valor total das operações a vencer em até 90 dias.
a_vencer_de_91_ ate_360_dias	Numérica	Valor das operações a vencer entre 91 e 360 dias.
a_vencer_de_361_ ate_1080_dias	Numérica	Valor das operações a vencer entre 361 e 1080 dias.
a_vencer_de_1081_ ate_1800_dias	Numérica	Valor das operações a vencer entre 1081 e 1800 dias.
a_vencer_de_1801_ ate_5400_dias	Numérica	Valor das operações a vencer entre 1801 e 5400 dias.

Variável	Tipo	Descrição
a_vencer_acima_ de_5400_dias	Numérica	Valor das operações a vencer acima de 5400 dias.
vencido_acima_ de_15_dias	Numérica	Valor das operações vencidas há mais de 15 dias.
carteira_ativa	Numérica	Total da carteira ativa do cliente.
carteira_inadimplida _arrastada	Numérica	Total da carteira inadimplida acumulada.
ativo_problematico	Numérica	Valor agregado dos ativos problemáticos associados ao cliente.