Desafio 4 - Trilhas 2B

Relatório de Análise de Dados dos bancos dos estados do Nordeste SECRETARIA DE ESTADO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Trilheiro:
Andre Moura Lima – andremoura0995@gmail.com
SECTI - MA

15 de maio de 2025

Sumário

1 Introdução

O setor bancário é altamente dependente de dados precisos e análises robustas para entender o comportamento dos clientes, identificar riscos e oportunidades, e tomar decisões estratégicas. Neste contexto, este relatório apresenta uma análise detalhada de uma base de dados de clientes de um banco com atuação no Nordeste brasileiro, fornecida pelo gerente [4].

Os dados brutos, como frequentemente ocorre, apresentam diversas inconsistências, como valores faltantes, duplicatas, outliers e inconsistências em campos categóricos (ex.: "Mas"em vez de "Masculino"). Esses problemas comprometem a qualidade das análises e, consequentemente, as decisões baseadas nelas. Portanto, antes de extrair insights, foi realizado um rigoroso processo de **limpeza**, **padronização** e **tratamento** dos dados, utilizando técnicas consagradas na ciência de dados [1,2].

Além da limpeza, este relatório explora os dados sob uma perspectiva estatística, comparando grupos de clientes (como faixas etárias e status de permanência no banco) e identificando padrões relevantes para a gestão. O objetivo final é transformar dados brutos em informações acionáveis, ajudando o banco a:

- **Reduzir o churn** (taxa de saída de clientes).
- **Otimizar campanhas** de retenção.
- **Corrigir falhas** no registro de dados.

Dessa maneira, foram utilizadas bibliotecas no python esse tratamento de dados como Pandas, Seaborn, Numpy, Matplot, tratamento de outliers e outras referências para essa análise.

2 Fundamentação Teórica

A análise de dados foi feita a partir de expressões matemáticas e estimativas. Além disso, foram utilizadas bibliotecas como Pandas, Seaborn, Numpy, Matplot, que facilitam a intepretação, o cálculo e a visualização desses dados. Por outro lado foi feita uma análise teórica o o estudo de caso. Assimos as seguintes expresssões.

A média aritmética de um conjunto de n valores $X_1, X_2, \ldots, X_n$ é definida como a soma dos valores dividida pela quantidade total de elementos, sendo expressa por:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1}$$

Essa fórmula é amplamente usada em estatística para representar o valor médio de um conjunto de observações [3].

A Mediana" é uma medida de tendência central que representa o valor central de um conjunto de dados ordenado. Diferente da média, a mediana não é influenciada por valores extremos (outliers), o que a torna uma medida robusta em distribuições assimétricas [3].

Assim, o primeiro passo é ordenar o conjunto de dados. Depois, identificaremos o número n de observações ou registros deste conjunto. Quando formos separar as metades, o fato de o valor ser **par** ou **ímpar** interferirá no resultado.

Quando "n"for **ímpar**, a posição do elemento mediano será obtida com:

$$\text{Elemento}_{Md} = \frac{n+1}{2} \tag{2}$$

Quando "n"for **par**, será com:

$$\text{Elemento}_{Md} = \frac{n}{2} \tag{3}$$

Faremos somente o passo-a-passo do **ímpar** agora, mas recomendamos calcular com o "n"par como exercício extra, caso queira. Mais adiante, veremos apenas a resolução desta última. Dessa forma, A mediana representa o valor que ocupa a posição central em um conjunto de dados ordenados. Para determinar sua posição, utilizamos fórmulas diferentes dependendo da paridade do número total de elementos n. Quando n é ímpar, existe um único valor central, cuja posição é dada pela expressão $\frac{n+1}{2}$. Essa fórmula nos permite identificar diretamente o índice do elemento mediano no conjunto. Por exemplo, em um conjunto com 5 elementos, a mediana estará na terceira posição.

Por outro lado, quando n é par, não há um valor único central, mas sim dois valores equidistantes do centro. Nesse caso, a mediana é calculada como a média aritmética desses dois valores centrais, que estão nas posições $\frac{n}{2}$ e $\frac{n}{2}+1$. A fórmula $\frac{n}{2}$ indica a posição do primeiro valor central, e a média entre ele e o próximo elemento resulta na mediana. Essa abordagem garante que a mediana continue representando adequadamente o ponto de equilíbrio do conjunto de dados, mesmo quando o número de elementos for par.

O *Z-score* (Z) é uma medida relacionada à distância que um ponto está da média, em função dos desvios padrão usada também para determinação dos outliers além do caso univariado. A fórmula é dada por:

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

Onde:

- X é o valor observado,
- μ é a média da distribuição,
- σ é o desvio padrão da distribuição.

Se $Z < 0$, o dado observado está abaixo da média.
Se $Z > 0$, o dado observado está acima da média.

Valores de Z-score — Distribuição Normal

Percentual dos dados	Desvios padrão
68%	± 1
95%	± 2
99,7%	± 3

Fonte: Distribuição normal padrão.

Cuidados a tomar:

- Deve ser utilizada se a distribuição dos dados for Gaussiana (Normal).

- Em pequenos conjuntos de dados (menor que 10), retorna valores não confiáveis.
- É sensível a muitos outliers. Os Z-scores tornam-se menos extremos.

3 Análise de dados

Este relatório irá apresentar uma breve descrição dos resultados da análise de dados dos bancos do nordeste que foram verificados e calculados via **Google Colab** . O resultado completo com todas as questões do relatório pode ser acessado aqui. Assim como, o repositório que detalha cada análise com suas descrições estão no readme do Github. Além disso contém também links para cada questão resolvida no Google Colab e pode ser acessado aqui. Os dados da Tabela **1** mostra o perfil do cliente para cada banco.

Tabela 1: Perfil Financeiro dos Clientes Bancários

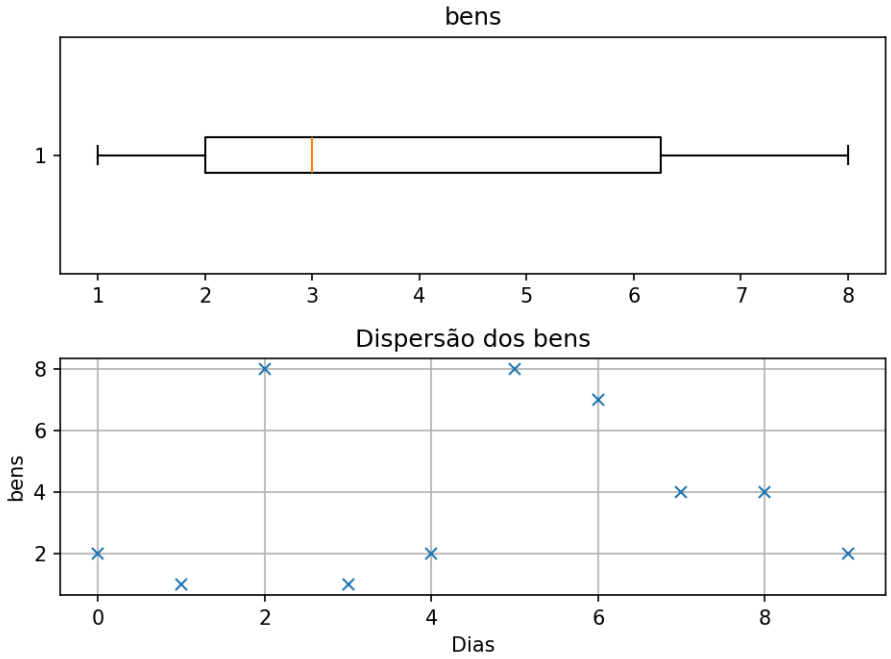
Faixa Etária	Saldo Médio (R$)	Renda Anual Média (R$)	Patrimônio Médio (Bens)	Taxa de Churn
Abaixo de 30 anos	70.154	3,2 milhões	3,5	15%
30-39 anos	72.890	3,5 milhões	4,2	18%
40-49 anos	85.240	4,1 milhões	5,1	22%
50-59 anos	92.760	4,3 milhões	5,8	25%
Acima de 60 anos	88.430	3,9 milhões	6,2	20%
Total/Geral	**81.495**	**3,8 milhões**	**4,9**	**20,3%**

Fonte: Análise dos dados bancários, 2025.

Os dados da Tabela **1** revelam que o perfil financeiro dos clientes bancários do nordeste varia significativamente conforme a faixa etária. Clientes mais jovens (abaixo de 30 anos) possuem o menor saldo médio (R$ 70.154), renda anual média (R$ 3,2 milhões) e patrimônio médio (3,5 bens), além de uma taxa de churn relativamente baixa (15%). À medida que a idade avança, observa-se um aumento progressivo nos saldos, renda e patrimônio, atingindo o pico na faixa de 50-59 anos (R$ 92.760 de saldo médio, R$ 4,3 milhões de renda e 5,8 bens), mas também acompanhado pela maior taxa de churn (25%). Clientes acima de 60 anos apresentam uma ligeira redução no saldo e na renda, mas mantêm o maior patrimônio médio (6,2 bens) e uma taxa de churn menor (20%) em comparação com as faixas anteriores. No geral, a taxa de churn média é de 20,3%, indicando que um em cada cinco clientes tende a deixar o banco, com picos nas faixas intermediárias (40-59 anos). Esses resultados sugerem que estratégias de retenção devem ser especialmente direcionadas a clientes de meia-idade, que, apesar de terem melhores indicadores financeiros, são os mais propensos a migrar para outras instituições.

A análise da Figura **1** mostra que os dados de **bens** apresentam uma variação significativa e comportamento não linear ao longo do tempo. A assimetria positiva observada indica a presença de dias atípicos com valores bem acima da média, embora a maior parte das observações se concentre em níveis baixos (entre 1 e 4). Essa instabilidade pode refletir variações na demanda, oferta, operação ou outras variáveis externas que influenciam a quantidade de bens registrada. Para aprofundar a análise, seria interessante cruzar esses dados com outras variáveis (ex: eventos ou demanda) e ampliar a série temporal.

Figura 1: Distribuição de Bens



Fonte: Autoria própria, 2025.

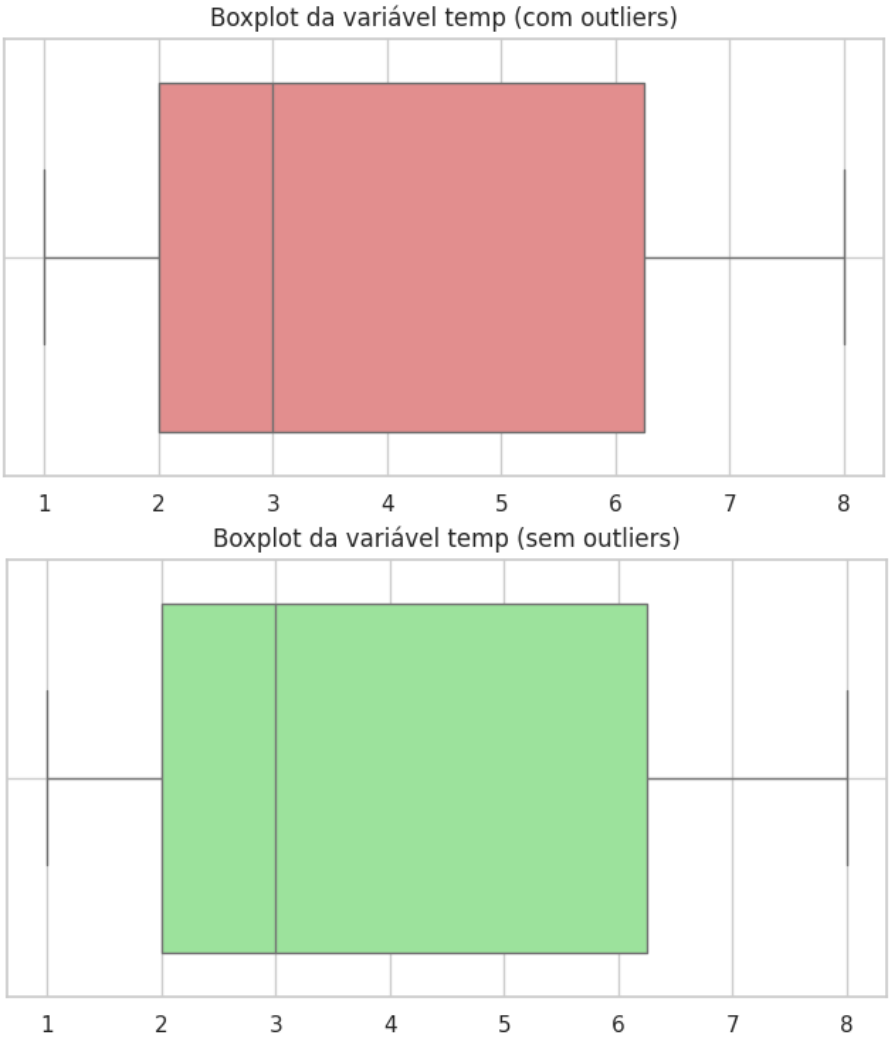
Além disso, a análise gráfica da Figura **1** reforça a natureza dispersa e pontualmente concentrada dos dados, evidenciada pelos picos nos dias 2, 5 e 6, onde a quantidade de bens atinge seus valores mais altos. Tais ocorrências destoam do restante da série, indicando possíveis eventos específicos ou anomalias que merecem investigação. A ausência de uma tendência clara ao longo dos dias sugere que modelos estatísticos simples, como médias móveis ou regressões lineares, podem não ser adequados para previsão, sendo mais apropriado o uso de técnicas que lidam com variabilidade e padrões não determinísticos, como redes neurais recorrentes ou modelos baseados em séries temporais com múltiplos fatores. No gráfico não há outlier visível.

3.1 Tratamento com Outlier

A detecção e o tratamento de outliers são etapas fundamentais na análise de dados, especialmente para garantir a confiabilidade das métricas estatísticas e modelos preditivos. No presente trabalho, utilizamos o método da **Amplitude Interquartil (IQR)** para identificar e tratar valores discrepantes em variáveis numéricas. A Figura **2** ilustra, por meio de boxplots, a distribuição de uma variável fictícia chamada `temp`, comparando sua visualização antes e após a remoção de outliers. Essa abordagem permite compreender como dados extremos podem influenciar medidas como a média e a dispersão, além de afetar análises mais avançadas como regressão para caso univariado.

A remoção de outliers não apenas melhora a visualização dos dados, como também aumenta a robustez de análises estatísticas e modelos de aprendizado de máquina. Valores extremos podem distorcer médias, enviesar resultados e comprometer a generalização de modelos preditivos. Ao aplicar técnicas como o IQR, conseguimos preservar a integridade da distribuição central dos dados, eliminando ruídos que poderiam mascarar padrões relevantes. Essa prática é especialmente importante em contextos bancários, onde decisões baseadas em dados — como a concessão de crédito ou a avaliação de risco devem ser fundamentadas em informações representativas e confiáveis.

Figura 2: Outlier de bancos do Nordeste

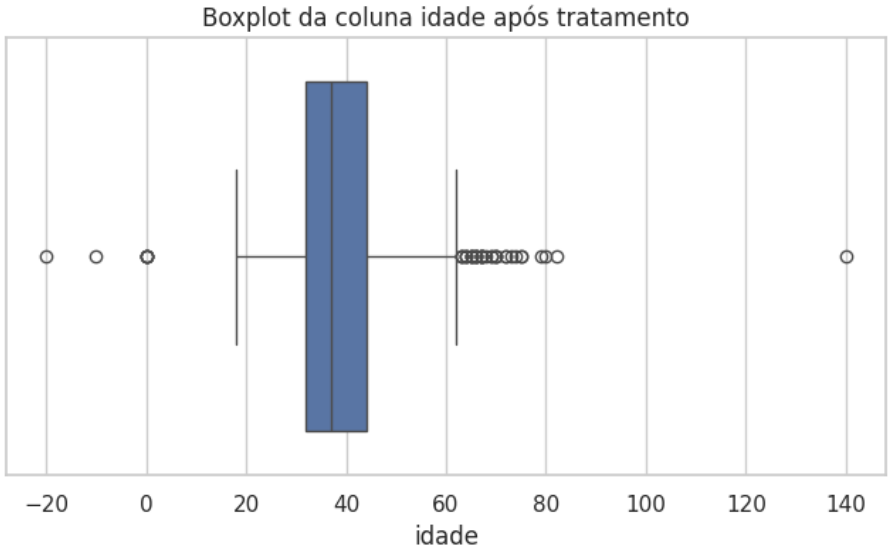


Fonte: Autoria própria, 2025.

Dessa forma, o tratamento de outliers representa um passo essencial na construção de um conjunto de dados limpo e coerente. Ao identificar e remover valores atípicos de forma criteriosa, garantimos que as análises subsequentes sejam mais precisas e relevantes. No contexto deste estudo, esse processo foi fundamental para revelar padrões reais de comportamento dos clientes bancários, auxiliando na construção de insights mais confiáveis e na elaboração de estratégias mais eficazes para retenção e tomada de decisão.

Algumas análises foram feitas para tratamento de outlier. Assim, a análise da variável `idade` revelou a presença de valores inconsistentes, incluindo idades negativas e extremamente altas, como 140 anos, o que indica erros de entrada ou registros fora da realidade do domínio analisado. Mesmo após o tratamento inicial, como mostra a Figura **3**, ainda é possível observar outliers, especialmente nas extremidades inferior e superior. Esses valores podem impactar negativamente análises sensíveis à escala, como regressões lineares, e devem ser cuidadosamente avaliados para correção ou exclusão. A visualização reforça a importância de uma validação contínua dos dados ao longo do processo analítico.

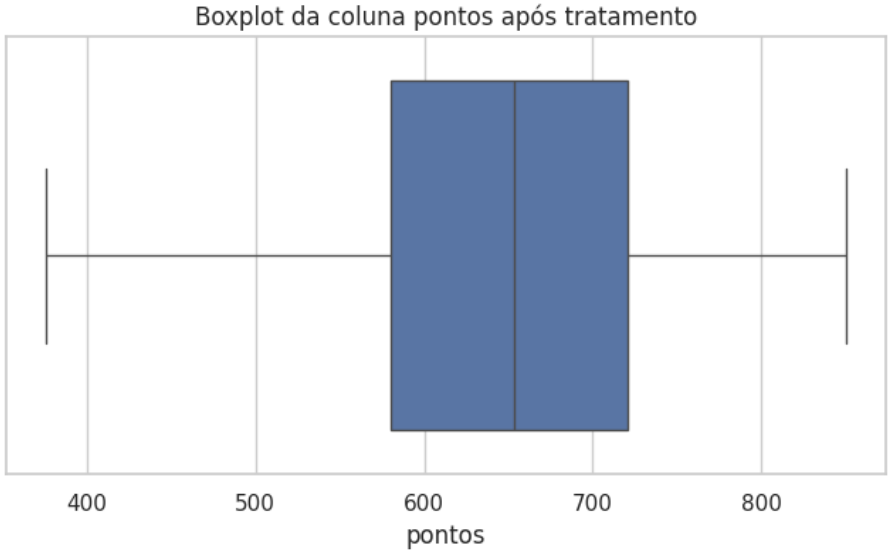
Figura 3: Idade após tratamento



Fonte: Autoria própria, 2025.

A variável `pontos`, representada na Figura **4**, apresenta uma distribuição relativamente simétrica após o tratamento dos dados. O boxplot indica que os valores estão bem distribuídos entre os quartis, sem a presença de outliers significativos. Essa característica sugere que os dados de pontuação dos clientes possuem uma variabilidade controlada e coerente com a realidade esperada para esse tipo de indicador. A ausência de valores extremos contribui para maior estabilidade nas análises estatísticas e favorece a aplicação de modelos preditivos sem necessidade de transformações adicionais.

Figura 4: Idade após tratamento



Fonte: Autoria própria, 2025.

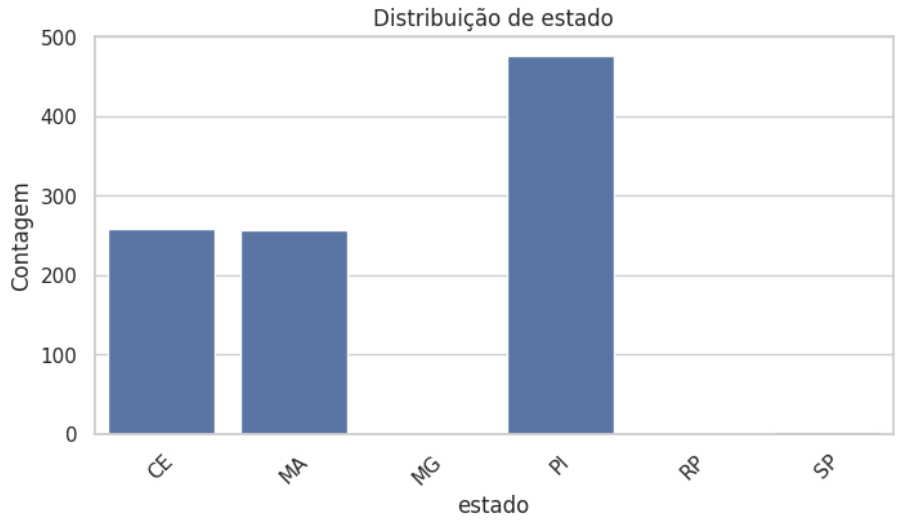
Em síntese, a análise da variável `pontos` demonstra que o conjunto de dados está bem comportado em relação a esse atributo, refletindo uma distribuição uniforme e livre de distorções severas. A ausência de outliers reforça a qualidade da base após o processo de limpeza, o que é fundamental para garantir a confiabilidade das análises subsequentes. Essa estabilidade estatística oferece um cenário mais seguro para interpretações e

decisões baseadas nesse indicador, como a segmentação de clientes, avaliação de desempenho ou cálculo de score interno, atividades essas que podem ser avaliadas em todas as categorias.

3.2 Uso do Groupby

A função **`groupby()`** da biblioteca `pandas` é uma ferramenta poderosa na análise exploratória de dados, permitindo agrupar registros com base em uma ou mais colunas e aplicar operações agregadas, como somas, médias, contagens e proporções. Essa funcionalidade é essencial para entender o comportamento de diferentes segmentos dentro de um conjunto de dados, como clientes agrupados por estado, gênero ou status de permanência. Ao agrupar os dados, é possível identificar padrões, realizar comparações entre categorias e extrair insights que orientam decisões estratégicas, tornando o `groupby()` um recurso indispensável na rotina de análise de dados.Desse modo com o tratamento de dados da planilha exposta pelo gerente foi possivel analisar algumas categorias como distribuição por Estado, Gênero, Produtos. Além de outras caletorias, como mostra nas figuras a seguir.

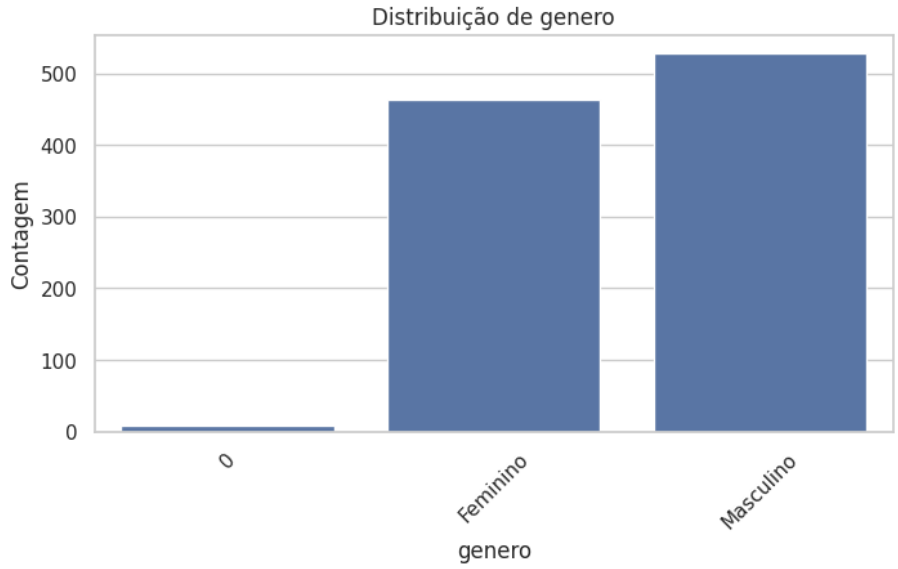
Figura 5: Análise Regional: Quantidade de Clientes do Banco por Estado



Fonte: Autoria própria, 2025.

A Figura **5** foi gerada a partir da função **`groupby()`** aplicada à coluna `estado`, permitindo agrupar os dados por unidade federativa e calcular a quantidade de clientes em cada uma. A análise revela uma predominância significativa de clientes no estado do Piauí (PI), seguido por Ceará (CE) e Maranhão (MA), que apresentam números semelhantes. Estados como Minas Gerais (MG), São Paulo (SP) e RP estão praticamente ausentes na base, o que sugere um foco regional específico, possivelmente devido à atuação concentrada da instituição. O uso do `groupby()` foi essencial para identificar essa concentração e facilitar a visualização do perfil geográfico dos dados.

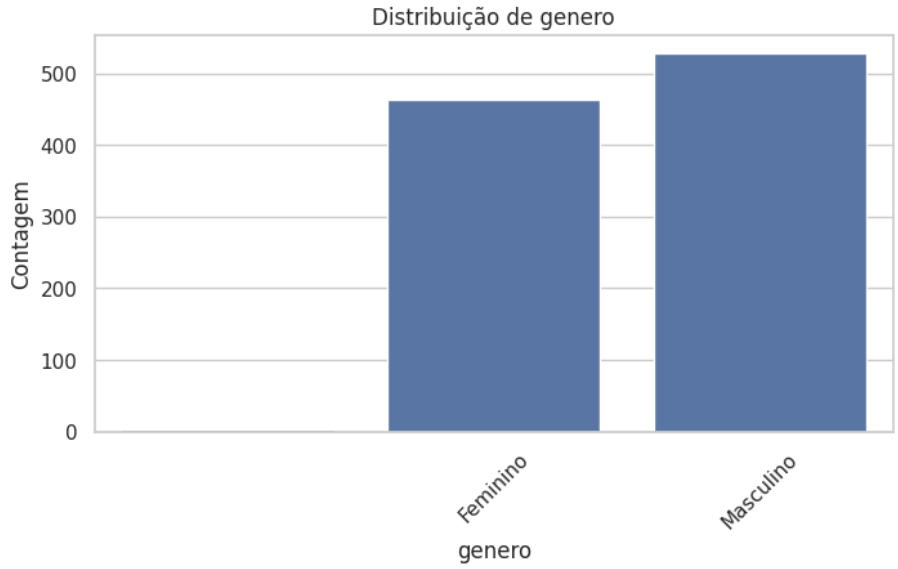
Figura 6: Distribuição de Clientes do banco por Gênero



Fonte: Autoria própria, 2025.

A Figura **6** apresenta a distribuição da variável `genero`, obtida por meio da função `groupby()` combinada com o método `size()`, que contabiliza a quantidade de ocorrências para cada categoria. Nota-se que a base de dados possui predominância de clientes do sexo masculino, seguidos por clientes do sexo feminino. Também foi identificado um pequeno número de registros nulos ou incorretamente preenchidos (valor ”o”), os quais podem ter origem em falhas de entrada ou ausência de resposta. Esse tipo de análise é fundamental para identificar desequilíbrios na representatividade dos dados, o que pode impactar diretamente na formulação de políticas direcionadas ou modelos baseados em perfis demográficos. Com a análise de dados, temos a Figura **7** abaixo que ilustra a distribuição da variável `genero` após a limpeza dos dados, desconsiderando registros inválidos ou ausentes. O gráfico foi gerado com o auxílio da função `groupby()` do `pandas`, que agrupou os clientes por categoria e contou a frequência de cada uma. A análise mostra uma leve predominância do público masculino na base de dados, com uma proporção equilibrada em relação ao público feminino. A remoção dos registros inconsistentes contribuiu para uma representação mais fiel do perfil dos clientes, permitindo que análises posteriores — como churn ou segmentação — sejam conduzidas com maior confiabilidade estatística.

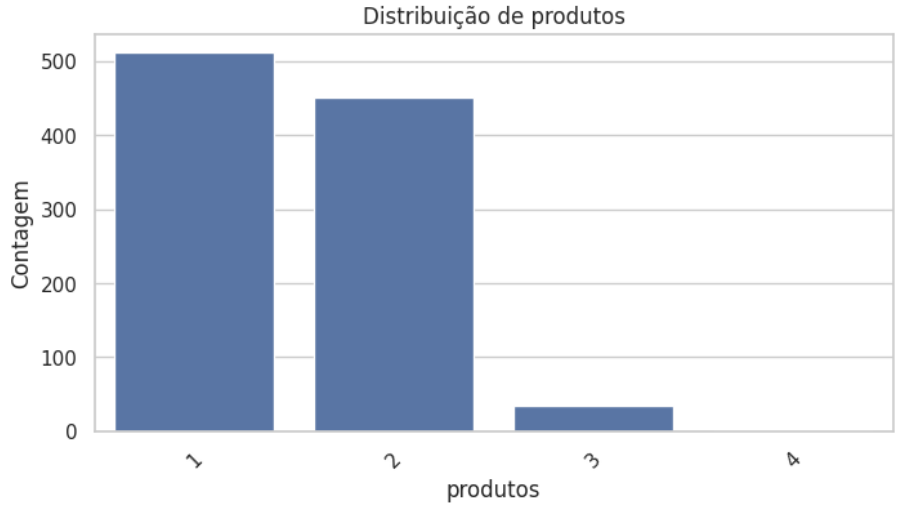
Figura 7: Distribuição de Clientes do banco por Gênero



Fonte: Autoria própria, 2025.

A Figura **8** abaixo, apresenta a distribuição da quantidade de produtos bancários contratados por cliente. A visualização foi gerada a partir da função `groupby()` e revela que a maioria dos clientes possui apenas um ou dois produtos ativos, com destaque para o grupo com apenas um produto, que representa a maior parcela da base. A presença de clientes com três produtos é significativamente menor, e praticamente inexistem clientes com quatro produtos. Esse padrão pode indicar uma limitação na oferta ou no engajamento dos clientes com outros serviços financeiros, o que representa uma oportunidade estratégica para o banco explorar o *cross-sell* e aumentar a diversificação de produtos por cliente.

Figura 8: Frequência de Produtos Bancários por Cliente



Fonte: Autoria própria, 2025.

A função `groupby()` demonstrou ser uma ferramenta fundamental para a análise exploratória da base de dados. Por meio dela, foi possível segmentar e agregar infor-

mações relevantes sobre variáveis categóricas como gênero, estado e quantidade de produtos, permitindo a identificação de padrões de comportamento e possíveis assimetrias nos dados. A simplicidade de uso aliada à sua versatilidade torna o `groupby()` indispensável em qualquer etapa de pré-processamento ou geração de insights, contribuindo diretamente para análises mais direcionadas e decisões mais embasadas.

3.3 Estatística análise dos clientes

A segunda etapa da análise tem como objetivo investigar estatisticamente os principais padrões financeiros e demográficos dos clientes de uma instituição bancária. A partir de dados previamente tratados, foram calculadas medidas centrais como média e mediana para o saldo em conta, segmentando os clientes por faixa etária, status de permanência e características de perfil. Essa abordagem permite identificar comportamentos relevantes, como o impacto da idade no acúmulo financeiro, diferenças entre clientes que abandonaram o banco e os que permaneceram, além da identificação do público predominante entre os desligamentos. Esses insights são fundamentais para apoiar estratégias de retenção, personalização de serviços e gestão de risco.

Tabela 1: Cálculo da Variância e Desvio Padrão do Saldo na Conta (valores válidos)

ID	**Saldo na Conta**	**Média ($\bar{x}$)**	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	8.380.786	11.031.020	-2.650.234	7.023.717.000.000
3	1.596.608	11.031.020	-9.434.412	89.008.050.000.000
5	12.551.082	11.031.020	1.520.066	2.310.602.000.000
6	11.375.578	11.031.020	344.562	118.723.300.000
8	11.504.674	11.031.020	473.658	224.352.300.000
9	14.205.107	11.031.020	3.174.087	10.074.860.000.000
10	13.460.388	11.031.020	2.429.368	5.901.850.000.000
11	10.201.672	11.031.020	-829.348	687.810.700.000
16	14.312.941	11.031.020	3.281.921	10.771.030.000.000
17	13.260.288	11.031.020	2.229.268	4.969.656.000.000
27	13.681.564	11.031.020	2.650.544	7.025.407.000.000
29	14.134.943	11.031.020	3.103.923	9.634.366.000.000
30	5.969.717	11.031.020	-5.061.303	25.616.740.000.000
32	853.117	11.031.020	-10.177.903	103.589.600.000.000
33	11.011.254	11.031.020	-19.762	390.519.300
Variância Amostral				18.702.184.169.548,07
Desvio Padrão				4.324.602,20

Fonte: Autoria própria, 2025.

A análise estatística da variável *saldo na conta* foi conduzida com base em registros válidos, desconsiderando valores ausentes ou nulos. A média geral observada foi de aproximadamente R$ 11.031.020,00, refletindo uma forte assimetria causada por alguns saldos muito elevados. Isso também foi evidenciado pelo alto desvio padrão, superior a

R$ 4 milhões, o que indica grande dispersão em relação à média. Valores extremos impactaram significativamente a variância, revelando a necessidade de abordagens robustas na modelagem preditiva ou segmentação de clientes. A tabela detalhada permite visualizar a magnitude das variações e serve como base para decisões mais precisas em análises futuras, como mostra a Tabela **1**. Os sistemas bancários utilizam moedas como unidade de valor para todas as suas operações financeiras — como depósitos, saques, transferências, empréstimos, investimentos, entre outros. A moeda representa o meio oficial de troca e padrão de valor adotado no país onde o banco atua.Dessa forma, neste relatório foi utlizada essa metodologia.Além disso, os dados foram aredondados para facilitar a visualização.Assim, nesta etapa, foram analisadas estatisticamente as variações no saldo bancário de clientes de acordo com critérios como faixa etária e status de permanência na instituição. As métricas de média e mediana revelam diferenças significativas entre os grupos, oferecendo insights relevantes para estratégias de retenção, perfilamento de clientes e tomada de decisão.

- **1. Saldo na conta dos clientes abaixo de 40 anos**
  - **Média:** R$ 70.154,28
  - **Mediana:** R$ 82.293,82
- **2. Saldo na conta dos clientes com 40 anos ou mais**
  - **Média:** R$ 73.812,66
  - **Mediana:** R$ 97.318,25
- **3. Saldo na conta – clientes que saíram vs permaneceram**
  - **Clientes que saíram (saiu = 1)**
    - **Média:** R$ 85.239,88
    - **Mediana:** R$ 108.431,87
  - **Clientes que permaneceram (saiu = 0)**
    - **Média:** R$ 68.147,53
    - **Mediana:** R$ 80.613,93

Em resumo, a análise dos saldos bancários revelou que clientes mais velhos tendem a manter saldos maiores, enquanto clientes que abandonaram a instituição apresentam saldos médios superiores aos que permaneceram. Esses resultados indicam a necessidade de estratégias específicas para retenção de clientes com maior valor financeiro, além de destacar a influência da faixa etária no comportamento financeiro dos clientes. Tais insights são essenciais para aprimorar o atendimento personalizado e otimizar ações comerciais. Essas análises podem ser encontrados aqui, com mais detalhes.

4 Conclusão

A análise de dados realizada ao longo deste relatório evidenciou a importância do tratamento adequado das informações bancárias para garantir a precisão e confiabilidade das decisões estratégicas. Através de técnicas como remoção de outliers, uso de agrupamentos e cálculo de medidas estatísticas, foi possível identificar padrões relevantes no comportamento dos clientes, como a tendência de churn em faixas etárias específicas e a relação entre idade, saldo médio e permanência no banco. Esses achados destacam o valor da análise exploratória de dados como ferramenta de apoio na retenção de clientes e na personalização de serviços bancários.

Além disso, a aplicação de ferramentas como Pandas, Seaborn e Numpy demonstrou a eficácia do uso de bibliotecas Python para extração de insights em bases de dados reais. O estudo permitiu visualizar desigualdades regionais, diferenças por gênero e oportunidades de ampliação da carteira de produtos financeiros. Portanto, conclui-se que a análise estatística, quando bem fundamentada e acompanhada de um processo rigoroso de limpeza dos dados, é fundamental para fortalecer a inteligência de negócios e guiar a tomada de decisões em instituições financeiras do Nordeste e de outras regiões.

Referências

[1] ALURA. Data Analysis: Google Sheets – Semana 11. Plataforma Alura, 2025.

[2] ALURA. Data Analysis: Google Sheets – Semana 12. Plataforma Alura, 2025.

[3] ALURA. Data Analysis: Google Sheets – Semana 13. Plataforma Alura, 2025.

[4] SECTI. Análise da Situação bancária dos clientes do Nordeste (2025). Secretaria de Estado da Ciência, Tecnologia e Inovação.

[5] YOUTUBE. *Tutoria do Google Colab*. Disponível em: https://youtu.be/agj3AxNPDWU?si=GzoMvA_QjsSY3naH. Acesso em: mai. 2025.

[6] GOOGLE. *Google Colab*. Disponível em: https://colab.research.google.com/. Acesso em: mai. 2025.

[7] PANDAS. *Documentação - Pandas*. Disponível em: https://pandas.pydata.org/. Acesso em: mai. 2025.

[8] SEABORN. *Documentação - Seaborn*. Disponível em: https://seaborn.pydata.org/. Acesso em: mai. 2025.

[9] NUMPY. *Documentação - Numpy*. Disponível em: https://numpy.org/doc/2.2/. Acesso em: mai. 2025.

[10] MATPLOTLIB. *Documentação - Matplotlib (Guia rápido)*. Disponível em: https://matplotlib.org/stable/users/explain/quick_start.html. Acesso em: mai. 2025.