Relatório de Observações ao Gerente

Análise de Dados de Clientes - Estatística Descritiva e Qualidade de Dados

1. Introdução

Este relatório apresenta os principais resultados e observações referentes à análise estatística do saldo na conta dos clientes, divididos por faixas de idade e situação de permanência/saída, bem como as questões de qualidade de dados encontradas durante o processo de limpeza e tratamento.

Objetivos:

- Calcular média e mediana do saldo para clientes abaixo e acima de 40 anos.
- Comparar saldo médio e mediano entre clientes que saíram e permaneceram.
- Perfil demográfico e econômico dos clientes que saíram (gênero, idade, saldo, patrimônio, estado).
- Levantar inconsistências e propor recomendações para melhoria da qualidade dos dados.

2. Resumo da Estampas de Limpeza de Dados

2.1 **Padronização de nomes de colunas:** todas em minúsculas, sem espaços e com underlines (`df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')`).

2.2 **Tratamento de valores faltantes (NaN):**

- Numéricos: preenchidos com mediana da respectiva coluna.
- Categóricos: preenchidos com a moda (valor mais frequente).

2.3 **Tratamento de outliers (IQR Capping):** limites inferior e superior calculados como Q1–1.5×IQR e Q3+1.5×IQR; valores além desses limites foram “capados” nos limites correspondentes.

2.4 **Remoção de duplicados:** eliminação de quaisquer linhas idênticas para garantir unicidade.

2.5 **Padronização de categorias:** uniformização de variações em ‘genero’ (ex.: “mas”, “m” → “Masculino”; “fem”, “f” → “Feminino”).

3. Principais Resultados estatísticos

Segmento	Média do Saldo	Mediana do Saldo
Clientes < 40 anos	R$ 7,015,427.85	R$ 8,229,382.00
Clientes >= 40 anos	R$ 7,381,265.97	R$ 9,731,825.00
Clientes que permaneceram (saiu=0)	R$ 7.162423e+06	R$ 8926348.5
Clientes que saíram(saiu=1)	R$	R$

- **Menores de 40 anos** apresentaram saldo médio de R$ 7,015,427.85 e mediana de R$ 8,229,382.00.
- **Maiores de 40 anos** apresentaram saldo médio de 7,381,265.97 e mediana de R$ 9,731,825.00

  Para a comparação entre **permaneceram vs. saíram**, o grupo que permaneceu teve valores calculáveis; **não foi possível calcular** média/mediana para os que saíram, pois **não havia nenhum registro não nulo** na coluna de saldo para esse grupo.

4. Perfil dos Clientes que Saíram

Durante este segmento, constatamos ausência completa de dados para as variáveis analisadas:

- **Gênero predominante:** não calculável (todos os registros de `gênero` estavam nulos).
- **Idade, saldo e patrimônio (bens):** não calculáveis (todas as entradas eram NaN).
- **Distribuição por estado:** não calculável (nenhum valor registrado).

Implicação:
A total falta de dados para o subconjunto “clientes que saíram” impede qualquer análise de perfil ou inferência sobre os motivos de churn.

5. Observações e Recomendações

Problema Identificado	Impacto	Ação Recomendada
1. Ausência de qualquer valor em colunas-chave (saldo, idade, bens) para clientes que saíram.	Impossibilita análise de churn e segmentação de risco.	Investigar processo de extração / integração para garantir que esses campos sejam coletados ou migrados corretamente.
2. Populações de “saíram” possivelmente codificadas de forma diferente (ex.: “sim”/“não” em vez de 1/0).	Filtro `saiu == 1` não capturou registros.	Padronizar a codificação de “saiu” e auditar valores existentes.
3. Uso de `inplace=True` gerou avisos de futuro descontinuamento.	Risco de falha em atualizações com pandas 3.0+.	Migrar para atribuições diretas (`df[col] = df[col].fillna(...)`).
4. Falta de metadados sobre o processo de coleta / limpeza.	Dificulta rastreabilidade de problemas de qualidade.	Estabelecer documentação de pipeline de dados e auditorias periódicas.

6. Conclusão

Embora tenhamos conseguido calcular médias e medianas para a maioria dos clientes, a ausência total de dados nos registros de clientes que saíram compromete a análise de churn e visão de risco. Recomenda-se priorizar a correção do pipeline de ingestão para essas variáveis, bem como padronizar formatos e implementar validações no processo de ETL.

Com esses ajustes, poderemos gerar relatórios confiáveis, suportar iniciativas de retenção e embasar decisões estratégicas com base em dados completos e consistentes.

Lucas, Cientista de Dados
Data: 16 de maio de 2025