Projeto Final: Análise exploratória de dados

Nome: Fábio Rodrigues de Andrade Santos
dataset Analisado: USA Bank Financial Data

1 Introdução

Aplicar os conhecimentos aprendidos ao longo do curso para resolver um problema real de negócios, utilizando análise exploratória de dados (EDA), estatística descritiva e visualizações.

O *dataset* escolhido para o projeto foi USA Bank Financial Data pelos motivos principais:

1. Um *dataset* contendo os dados de um banco pode ser usado para melhor entender a realidade com caixa, assim como se aproximar da realidade dos funcionários. Mesmo sendo um banco fictício no caso escolhido
2. Dados temporais podem mostrar tendências sazonais, flutuações e padrões ao longo da análise, o que é interessante para a aplicabilidade dos conceitos, entrando em contraste com os dataset*s* utilizados nas aulas que eram mais dados discretos que contínuos.

1.1 Problemas de Negócio

Após a escolha do objeto a ser estudado e um vista sobre a descrição do *dataset* no site onde foi adquirido, foram geradas as seguintes perguntas sobre a performance financeira do banco:

1. Como evoluiu a performance financeira do banco ao longo do tempo?
2. O banco gerencia seus ativos de forma eficiente? O quanto isso afetou a rentabilidade do banco?
3. O quanto os valores de ações do banco refletem em seu rendimento financeiro?

2 Pré-Processamento e limpeza

Inicialmente se realiza um *check-up* dos *dataset* importado, se verificam suas colunas, quais tipos de dados estão sendo abordados, se as variáveis estão de acordo ou deveriam ser transformadas em algo diferente, como de *string* para inteira, se existem colunas duplicadas e se existem dados nulos.

2.1 Sobre o *dataset*

O dataset contém as seguintes colunas:

Table 1: Resumo da Estrutura do Dataset MyUSA Bank

index	**Coluna**	**Não Nulos**	**Tipo**	**Significado**
0	Date	523	object	Período de referência dos dado
1	Interest_Income	471	float64	Despesas com juros pagos a depositantes
2	Interest_Expense	523	int64	Receita com juros de empréstimos e investimentos
3	Average_Earning_Assets	523	int64	Ativos que geram renda para o banco
4	Net_Income	471	float64	Lucro líquido após todas as despesas
5	Total_Assets	523	int64	Valor total de todos os ativos do banco
6	Shareholder_Equity	523	int64	Patrimônio líquido (capital próprio)
7	Operating_Expenses	523	int64	Custos operacionais (exceto juros)
8	Operating_Income	523	int64	Resultado das operações principais
9	Market_Share	523	int64	Participação no mercado bancário
10	Stock_Price	523	int64	Preço da ação no mercado

O *dataset* apresenta dados temporais, isto é, uma evolução ao longo de um certo período contínuo de tempo, neste caso são as análises de 03/01/2022 à 29/12/2023. Com isso em mente, a tabela foi alterada para a

coluna data ser o índice do *dataset* e seu tipo foi alterado para *datetime* e formatado no modelo americano por praticidade. Foram encontradas colunas duplicadas as quais foram removidas, vide imagem 1. Como também os dados foram rearranjados em ordem crescente, do mais novo ao mais antigo, tendo a estrutura como a figura 2 mostra.

```
Número de datas duplicadas: 24

Datas duplicadas encontradas:
DatetimeIndex(['2022-01-03', '2022-02-22', '2022-03-17', '2022-03-18',
               '2022-05-04', '2022-06-28', '2022-08-23', '2022-10-05',
               '2022-10-17', '2022-11-14', '2022-11-28', '2023-01-18',
               '2023-02-07', '2023-03-03', '2023-03-13', '2023-03-23',
               '2023-05-10', '2023-09-13', '2023-09-15', '2023-10-20',
               '2023-11-01', '2023-11-09', '2023-11-20', '2023-12-06'],
              dtype='datetime64[ns]', name='Date', freq=None)
```

Figure 1: Datas Duplicadas

```
df.head()
✓ 0.0s                                                                                    Python
```

Date	Interest_Income	Interest_Expense	Average_Earning_Assets	Net_Income	Total_Assets	Shareholder_Equity	Operating_Expenses	Operating_Income	Market_Share	Stock_Price
2022-01-03	2121958.0	773237	55001045	1572212.0	95404302	23496605	811672	3494178	24	128
2022-01-04	2671155.0	518070	54463418	1336326.0	86440010	24948600	1030817	3231496	20	63
2022-01-05	2131932.0	797921	59771802	1224137.0	88674163	26416438	1345961	3692148	22	196
2022-01-06	2365838.0	556958	54345760	1452268.0	97221407	29694095	1289921	4779685	14	177
2022-01-07	2259178.0	1746222	57153768	3858336.0	98279553	25311499	1432303	4764985	10	103

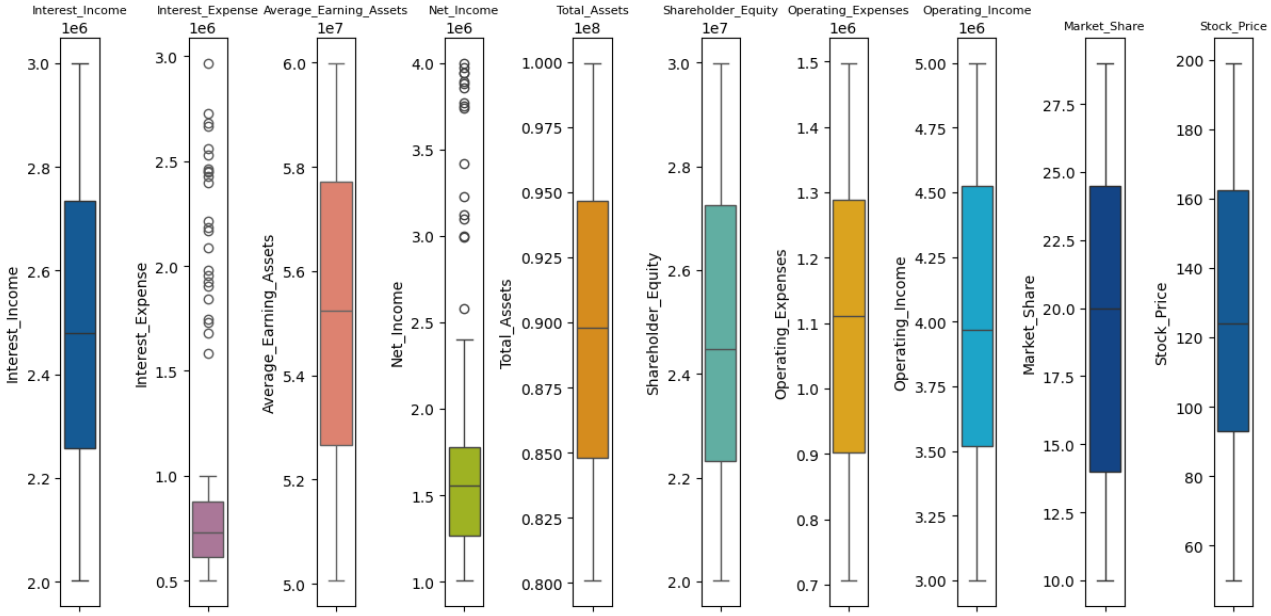
```
df.tail()
✓ 0.0s                                                                                    Python
```

Date	Interest_Income	Interest_Expense	Average_Earning_Assets	Net_Income	Total_Assets	Shareholder_Equity	Operating_Expenses	Operating_Income	Market_Share	Stock_Price
2023-12-22	2445101.0	612816	56790968	1909827.0	94722892	27353172	1324411	4664986	13	195
2023-12-26	2036631.0	628778	53751198	1669648.0	89223644	21437188	1496948	3044760	10	104
2023-12-27	2766577.0	562292	57369434	1613474.0	90758430	21656321	1172501	4047159	22	169
2023-12-28	2072991.0	785977	53138145	1924044.0	99661007	20538645	848586	3285622	29	124
2023-12-29	2004014.0	1905690	51226476	3768400.0	93863529	29960559	1083395	4962889	16	70

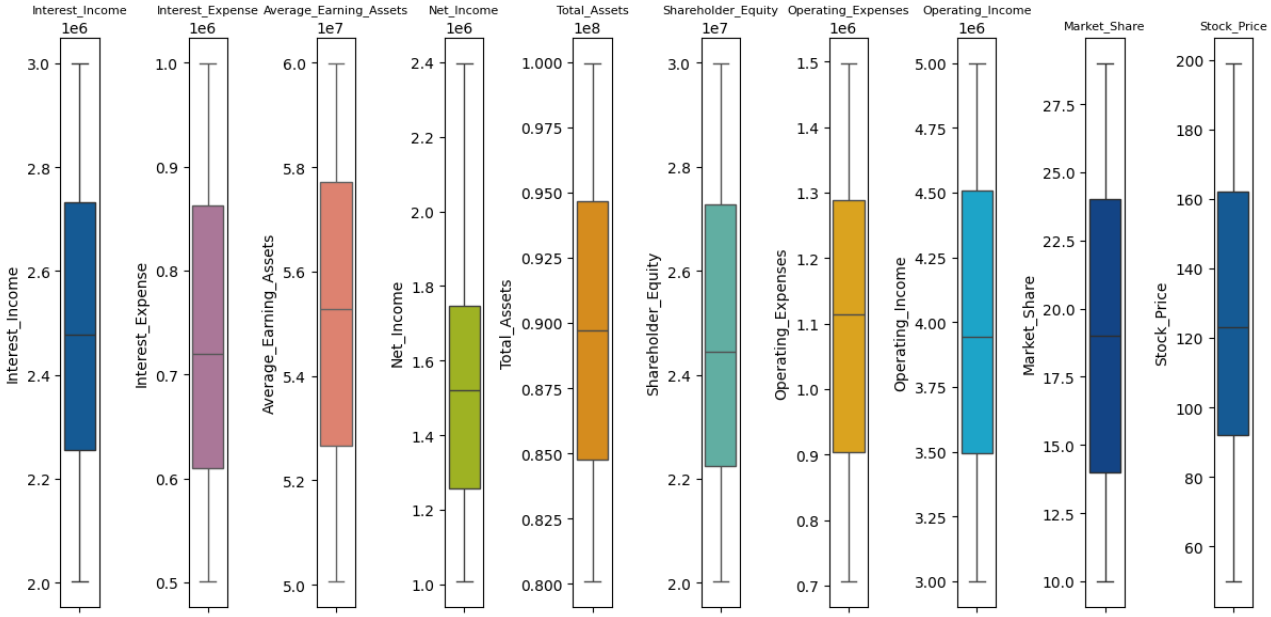
Figure 2: Primeiras 5 e últimas 5 linhas de dados

2.1.1 Dados Nulos e *Outliers*

Como a tabela 1 demonstra, existem dados nulos em duas colunas,*Interest_Income* e *Net_Income* e a averiguação dos *Outliers* foi realizada por meio do boxplot das colunas e como é possível ver na figura 3a, existem duas coluans com *outlires*, então foi com o cálculo do IQR foram impostos limites nestes *outliers* para uma análise uma melhor análise descritiva, a diferença se encontra na imagem 3b



(a) Outliers



(b) Inliners.

Figure 3: Boxplot de Outliners e Inliners.

3 Visualização e Análise de Dados

Após o pré-processamento do *dataset*, foram geradas visualizações com o objetivo de compreender o comportamento das variáveis financeiras ao longo do tempo. A Figura 4 apresenta a distribuição de cada variável e permite uma avaliação geral sobre a estabilidade e a volatilidade de determinados indicadores.

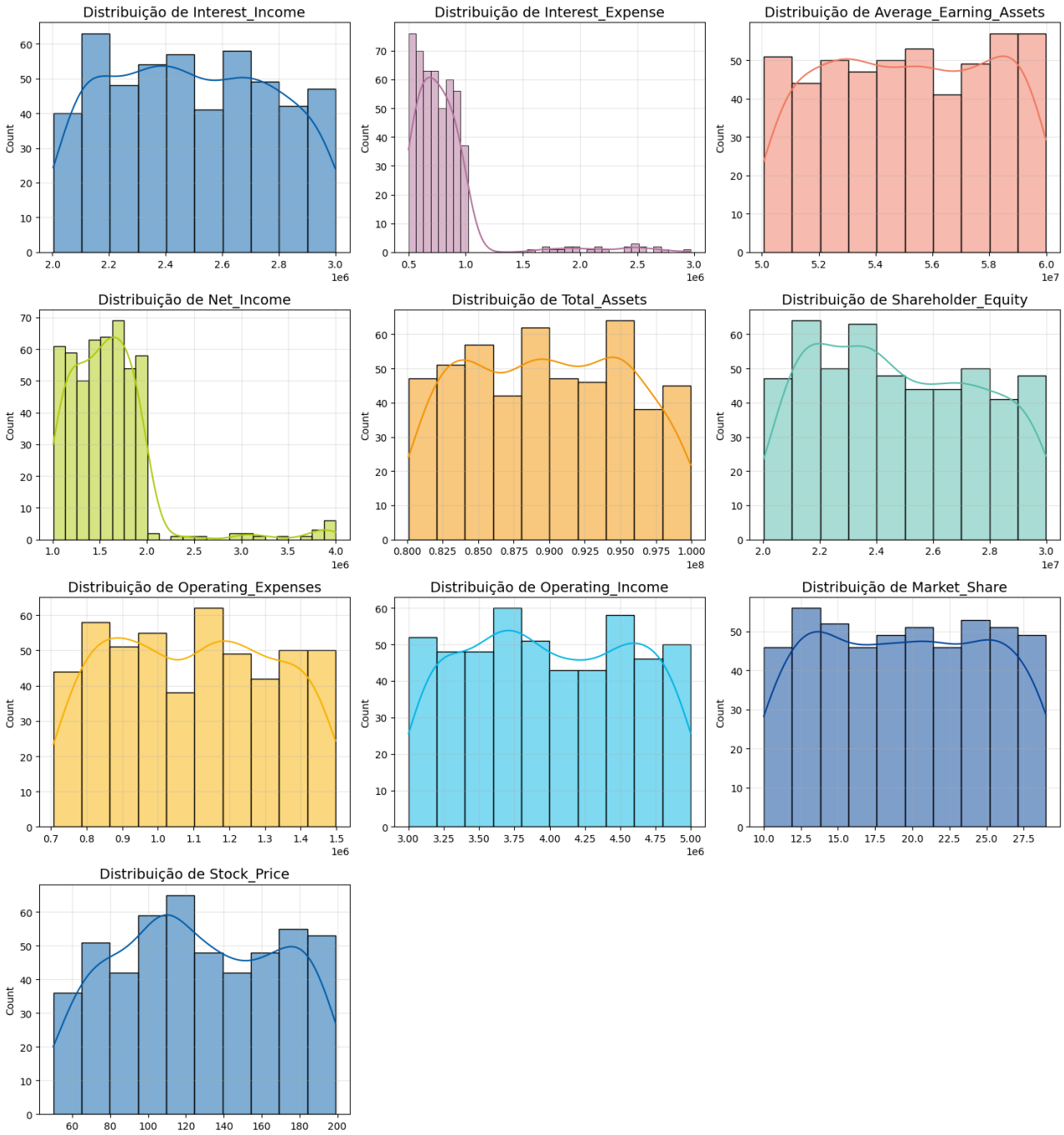


Figure 4: Distribuição das Variáveis ao longo do tempo

De forma geral, observa-se que o banco apresenta uma **estrutura financeira sólida**, evidenciada pela estabilidade em ativos totais, ativos geradores de receita e receitas de juros. Entretanto, algumas variáveis se mostraram mais voláteis, como o lucro líquido, as despesas operacionais e o preço da ação, indicando oscilações no desempenho e na percepção do mercado. Essa combinação sugere que, embora a base financeira do banco seja consistente, existem fatores conjunturais que impactam diretamente sua lucratividade e competitividade.

Complementarmente, foi analisada a relação entre receitas e despesas, apresentada na Figura 5. Nota-se um comportamento relativamente equilibrado entre entradas e saídas, ainda que as despesas apresentem tendência de se manter em patamar superior às receitas, o que é esperado no contexto bancário.

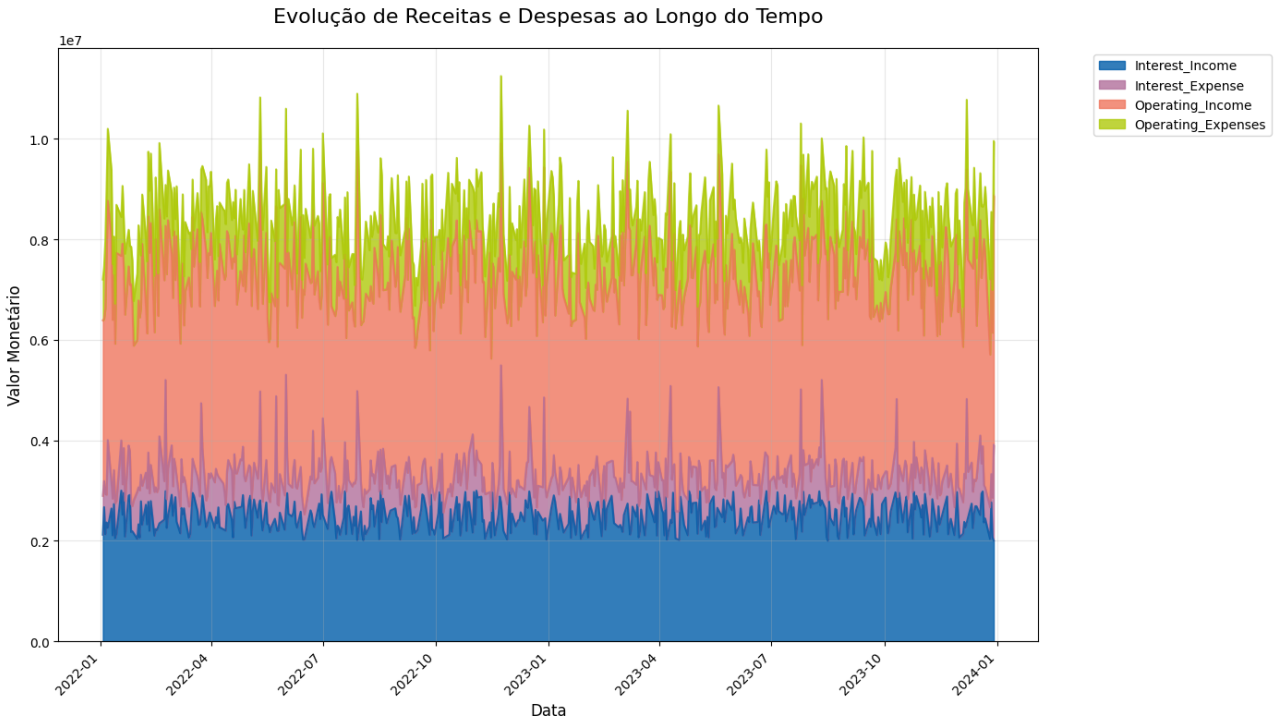


Figure 5: Evolução de Receitas e Despesas ao Longo do Tempo

3.1 Correlação entre variáveis

A matriz de correlação apresentada na Figura 6 evidencia relações moderadas entre algumas variáveis, com destaque para a correlação entre *Net_Income* e *Interest_Expense*, em torno de 0.5. Esse resultado sugere que aumentos nas despesas de juros podem impactar diretamente a lucratividade do banco, embora não haja correlações muito altas que indiquem forte dependência entre variáveis.

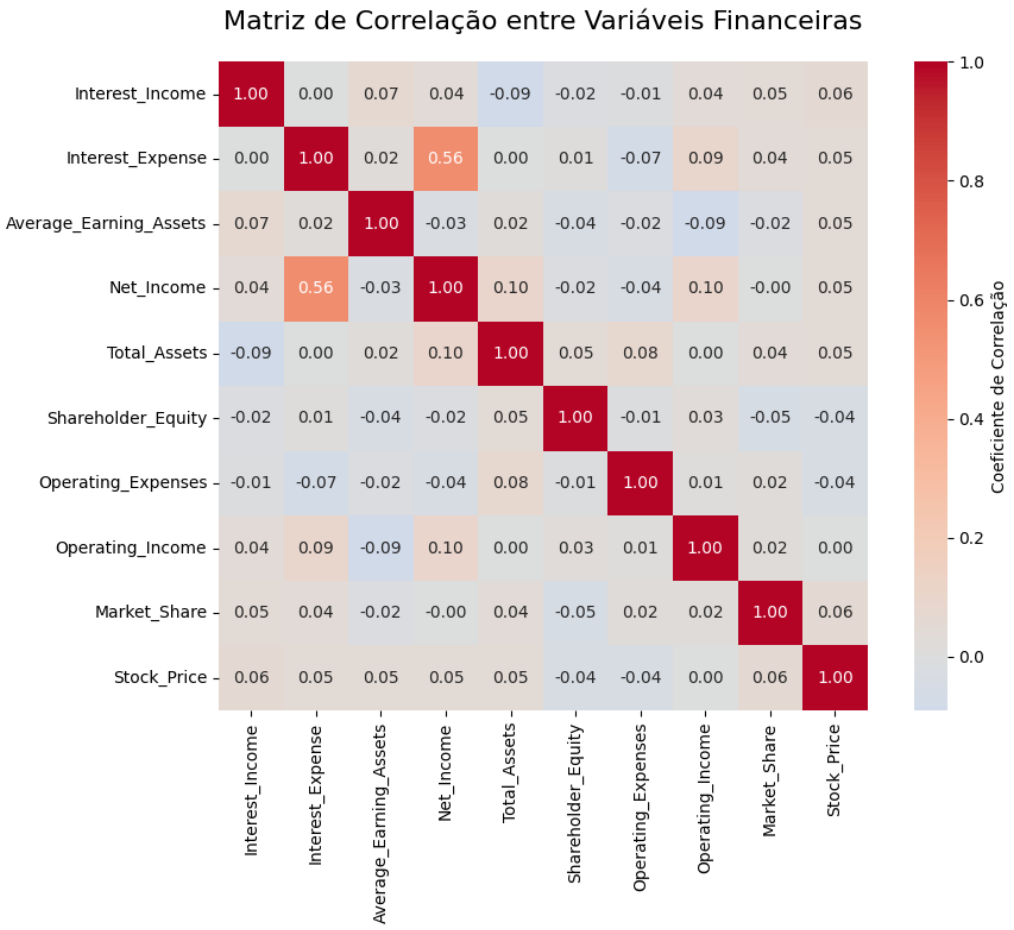


Figure 6: *Heatmap* de Correlação

4 Conclusão

A análise exploratória do *dataset* USA Bank Financial Data permitiu compreender o comportamento histórico das principais variáveis financeiras de um banco específico. As evidências apontam para uma estrutura patrimonial e de ativos consistente, mas com oscilações relevantes em despesas, lucratividade e participação de mercado.

Retomando as perguntas de negócio propostas:

1. **Evolução da performance financeira:** ao longo do período analisado, o banco demonstrou estabilidade em ativos e receitas, mas apresentou variações na lucratividade e no preço das ações.

2. **Eficiência na gestão de ativos:** os ativos geradores de receita mantiveram-se estáveis, indicando boa gestão, ainda que fatores externos tenham impactado a rentabilidade.

3. **Relação entre preço da ação e desempenho financeiro:** verificou-se volatilidade no preço das ações, condizente com as oscilações em despesas e lucros, sugerindo que o mercado reflete essas variações na valorização do banco.

Dessa forma, a análise mostra que, apesar da base financeira sólida, a instituição enfrenta desafios em controlar seus custos e manter estabilidade nos lucros, o que impacta diretamente sua competitividade e a percepção dos investidores.

Referências

- Kaggle. (2024). *USA Bank Financial Data*. Disponível em: https://www.kaggle.com/datasets/joebeachcapital/usa-bank-financial-data