



UNIVERSIDADE FEDERAL DO MARANHÃO

CAMPUS DE SÃO LUÍS - CIDADE UNIVERSITÁRIA

ENGENHARIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS

TURMA 1

**EXPLORAÇÃO E PRÉ-PROCESSAMENTO DE DADOS PARA ANÁLISE DE
SÉRIES TEMPORAIS DE ATIVOS DA B3**

PROFESSOR: THALES LEVI AZEVEDO VALENTE

ALUNOS:

DIOGO BRASIL DA SILVA - 2020010438

EMANUEL LOPES SILVA - 2021017818

MATHEUS COSTA ALVES - 2020002427

São Luís – MA

22/01/2025

DIOGO BRASIL DA SILVA - 2020010438

EMANUEL LOPES SILVA - 2021017818

MATHEUS COSTA ALVES - 2020002427

MINERAÇÃO DE DADOS

TURMA 1

**EXPLORAÇÃO E PRÉ-PROCESSAMENTO DE DADOS PARA ANÁLISE DE
SÉRIES TEMPORAIS DE ATIVOS DA B3**

Relatório referente à segunda
avaliação da disciplina Mineração
de Dados Aplicada a Engenharia,
Turma 1, do período 2024.2,
lecionada pelo Prof. Dr. Thales
Levi Azevedo Valente.

São Luís - MA

29/02/2025

SUMÁRIO

1. INTRODUÇÃO.....	7
1.1 - Contextualização.....	7
1.2 - Desafios na modelagem preditiva de séries temporais financeiras.....	8
1.3 - Importância do pré-processamento e da análise exploratória.....	10
1.4 - Objetivo do artigo.....	11
2. TRABALHOS RELACIONADOS.....	11
2.1-B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis..	12
2.2-Previsão de Preços de Ações Utilizando Inteligência Artificial.....	12
2.3-Predição de Tendências em Séries Financeiras Utilizando Meta Classificadores.....	12
3. FUNDAMENTAÇÃO TEÓRICA.....	13
3.1 - Séries Temporais Financeiras.....	13
3.2 - Análise Exploratória de Dados (EDA).....	14
3.2.1 - Visualização dos Dados.....	14
3.2.1.1 - Análise Combinando Gráfico Violino e Boxplot.....	15
3.2.1.1.1 - Boxplot.....	15
3.2.1.1.2 - Gráfico Violino.....	16
3.2.1.1.3 - Combinação Violino-Boxplot.....	17
3.2.2 - Métricas Estatísticas.....	17
3.2.3 - Detecção de Outliers.....	18
3.2.4 - Definição dos Parâmetros Analisados.....	18
3.2.4.1 - Skewness e Kurtosis.....	18
3.3 - Técnicas de Pré-Processamento em Séries Temporais.....	19
3.3.1 - Normalização e Padronização:.....	19
3.3.2 - Diferenciamento.....	20
3.3.3 - Engenharia de Atributos.....	20
3.4 - Importância do Preparo dos Dados para Modelos Preditivos.....	20
3.4.1 - Impacto na Performance dos Modelos.....	20
3.4.2 - Redução de Ruído e Tratamento de Falhas.....	21
4. MATERIAIS E MÉTODOS.....	21
4.3 - Organização da Base de Dados.....	22
4.3.1 - Definição da Estrutura de Pastas.....	23
4.3.2 - Upload dos Dados para um Repositório em Nuvem.....	23
4.3.3 - Consolidação dos Dados Processados.....	24
4.4 - Análise Exploratória dos Dados (EDA).....	24
4.4.1 - Análise das séries temporais.....	24
4.4.2 - Análise de distribuições.....	25
4.4.3 - Análise de correlação.....	26
4.4.4 - Análise da Kurtosis e Skewness.....	26

4.5 - Normalização dos Dados.....	27
5. RESULTADOS E DISCUSSÃO.....	27
5.1 - Skewness e Kurtosis.....	27
5.2 - Visualizações Utilizando Gráfico Violino e Boxplot.....	32
5.2.1 - Visualizações para todo o período do Dataset (de 2010 até 2022).....	32
5.2.5 - Comparação dos dados a partir de 2019 com anos anteriores.....	35
5.3 - Correlação dos Dados.....	38
5.3.1 - Correlação Forte Positiva.....	39
5.3.3 - Correlação Baixa a Moderada.....	39
5.3.4 - Correlação Baixa.....	39
5.3.5 - Implicações para Diversificação.....	39
5.3.6 - Setores Fortemente Relacionados.....	40
5.4 - Análise da Correlação Para Dados Normalizados.....	40
5.5 - Volatilidade dos dados.....	42
6. CONCLUSÃO.....	46
7. REFERÊNCIAS.....	48

RESUMO:

Este estudo investigou a análise exploratória e o pré-processamento de séries temporais financeiras de ativos negociados na B3 (Bolsa de Valores do Brasil). Dividiu os ativos em três grupos com base em Skewness e Kurtosis, permitindo identificar perfis de risco e retorno para diferentes investidores. Métodos como boxplots, violin plots e bandas de Bollinger foram utilizados para explorar volatilidade, tendências e sazonalidade dos preços. Ademais, foi constatado que eventos macroeconômicos, como a pandemia de COVID-19, impactaram diretamente a dinâmica do mercado, aumentando a volatilidade e alterando padrões de preços em períodos distintos. A matriz de correlação revelou forte relação entre ativos do mesmo setor, como VALE3 e BRAP3, enquanto setores independentes apresentaram baixa correlação, destacando oportunidades de diversificação. A normalização dos dados permitiu comparações mais precisas, revelando tendências semelhantes entre ativos correlacionados. Por fim, o estudo reforça a importância do pré-processamento e da análise exploratória para melhorar a qualidade dos dados e fundamentar estratégias de investimento e modelagem preditiva, fornecendo insights valiosos sobre a dinâmica do mercado financeiro brasileiro.

Palavras-chaves: B3, Pré-Processamento, Séries Temporais Financeiras, Análise Exploratória de Dados

ABSTRACT:

This study analyzed exploratory data analysis (EDA) and preprocessing of financial time series from B3 (Brazilian Stock Exchange) assets. Assets were classified into three groups based on Skewness and Kurtosis, identifying risk and return profiles for different investor types. Methods such as boxplots, violin plots, and Bollinger bands were employed to explore volatility, trends, and seasonality in asset prices. The findings highlighted the impact of macroeconomic events, such as the COVID-19 pandemic, on market dynamics, increasing volatility and altering price patterns across distinct periods. The correlation matrix revealed strong relationships among assets within the same sector, such as VALE3 and BRAP3, while independent sectors showed low correlations, indicating diversification opportunities. Data normalization enabled more precise comparisons, uncovering similar trends among correlated assets. The study emphasizes the importance of preprocessing and exploratory analysis in enhancing data quality and supporting investment strategies and predictive modeling, offering valuable insights into the dynamics of the Brazilian financial market.

Keywords:B3, Exploratory Data Analysis, Data Preprocessing, Financial Time Series

1. INTRODUÇÃO

1.1 - Contextualização

Segundo Carlos de Assis, prever o futuro é uma das maiores ambições do ser humano. Não existe um sistema perfeito para tal finalidade, mas é possível encontrar na literatura abordagens nos mais diversos contextos, várias delas com resultados relevantes (Assis, Carrano e Pereira, 2020, p. 30). A previsão de preços de ações é um desses contextos, e ainda mais, é uma área de extrema importância no mercado financeiro, pois permite que investidores, traders e analistas tomem decisões informadas para conseguir se aproveitar das grandes flutuações do mercado, e conseguir o prezado retorno positivo de capital, maximizando o volume de dinheiro que entra, comparado como o pouco que é gasto. O mercado de ações é altamente dinâmico e influenciado por diversos fatores, incluindo mudanças econômicas, decisões políticas, variações na oferta e demanda, além de eventos inesperados, como crises financeiras e pandemias, e principalmente as variações aleatórias geradas por medo, ou excitação sobre um determinado tópico. Nesse contexto, métodos preditivos tornam-se ferramentas valiosas, ajudando a antecipar movimentos de mercado e a identificar oportunidades de compra e venda com base em padrões históricos.

A crescente disponibilidade de dados financeiros e o avanço das técnicas computacionais revolucionaram a forma como a previsão de preços de ações é realizada. No passado, análises fundamentalistas e técnicas estatísticas básicas eram predominantes, em que práticas que se baseiam no estudo dos preços passados e volumes de negociação para identificar padrões gráficos e tendências de mercado eram amplamente implementadas. No entanto, atualmente, modelos sofisticados de aprendizado de máquina e redes neurais são amplamente utilizados para detectar padrões ocultos e realizar previsões mais precisas. Métodos como redes neurais recorrentes (RNNs) e Long Short-Term Memory (LSTMs) são especialmente populares para lidar com séries temporais financeiras, pois conseguem capturar dependências de longo prazo nos dados.

Entretanto, apesar dos avanços, a previsão de preços de ações continua sendo um desafio devido à natureza estocástica e caótica do mercado. Como foi mencionado anteriormente, movimentos de preço são frequentemente influenciados por fatores imprevisíveis, como anúncios corporativos, mudanças na regulamentação e comportamento dos investidores. Principalmente nesta era de redes sociais, onde a informação pode ser

compartilhada tão facilmente, um post ou pequena mudança de opinião pode acarretar a mudança de poder financeiro na magnitude de bilhões. Além disso, o fenômeno da autocorrelação e a não estacionariedade das séries temporais dificultam a modelagem tradicional, tornando necessário o uso de técnicas avançadas de pré-processamento, normalização e extração de características relevantes para melhorar o desempenho dos modelos preditivos.

A importância da previsão de preços de ações vai além do lucro individual de investidores. Ela também é crucial para instituições financeiras, fundos de investimento e bancos, que utilizam essas técnicas para gerenciar riscos e otimizar suas carteiras de ativos. Além disso, governos e órgãos reguladores monitoram as tendências do mercado para desenvolver políticas econômicas e evitar crises financeiras. Dessa forma, a previsão de preços não apenas impulsiona a eficiência do mercado, mas também contribui para a estabilidade econômica global, tornando-se um campo de estudo fundamental dentro das finanças quantitativas.

1.2 - Desafios na modelagem preditiva de séries temporais financeiras

Primeiramente, a modelagem preditiva de séries temporais financeiras apresenta desafios significativos devido à natureza complexa e volátil dos mercados financeiros. Isso, como já foi mencionado brevemente no tópico anterior, pode gerar consequências desastrosas, como, no caso de uma empresa, um modelo preditivo inadequado pode levar a decisões erradas no gerenciamento de riscos, na alocação de recursos e na tomada de decisões estratégicas. Por exemplo, uma previsão errônea de um mercado subindo ou caindo pode resultar em investimentos mal direcionados, levando à perda de lucros significativos ou até mesmo a falências em casos extremos. Ademais, consumidores comuns também podem sofrer com decisões financeiras baseadas em previsões imprecisas. Por exemplo, um empresário que toma decisões de expansão ou investimento com base em uma previsão fraca pode perder oportunidades de crescimento ou enfrentar dificuldades financeiras se a realidade for diferente do previsto. Igualmente, indivíduos que confiam em conselhos financeiros baseados em modelos preditivos errôneos podem correr o risco de investir mal, comprometendo seu patrimônio líquido e futuros planejamentos, entrando em débito e perdendo seus investimentos.

Outrossim, tendo consciência das consequências dos erros das modelagens, deve-se explorar a ideia de que os modelos devem ser avaliados não apenas pela precisão das previsões, mas também pelo impacto financeiro das decisões tomadas com base nelas (Assis, Carrano e Pereira, 2020, p. 41). E por conseguinte, devemos desenvolver modelos que evitem gerar prejuízos catastróficos para a economia e a sociedade. Para isso, devemos combater os principais desafios para os modelos atuais, que serão descritos abaixo.

Dentre os principais desafios, aquele que pode ser considerado de mister importância é a alta volatilidade, que condiz do fato que os preços dos ativos financeiros estão sujeitos a grandes variações em curtos períodos de tempo. Pequenas mudanças no mercado podem ter impactos significativos e imprevisíveis nas séries temporais. Um ótimo exemplo, apesar de não tão recente, é a crise financeira de 2008, que acabou demonstrando que modelos preditivos falharam ao antecipar a magnitude da queda dos mercados devido a interações complexas entre variáveis econômicas e comportamentais.

Outro fator é a dependência dos dados históricos, pois o passado ajuda a entender o futuro, mas não o prevê. Essa abordagem pressupõe que as relações entre variáveis observadas no passado permanecerão válidas no futuro, o que nem sempre é verdadeiro. Mercados financeiros são influenciados por uma ampla gama de fatores, incluindo mudanças econômicas, políticas e comportamentais, que podem alterar significativamente a dinâmica dos preços dos ativos. Eventos como crises financeiras, pandemias e mudanças regulatórias criam rupturas que tornam os modelos treinados com dados históricos menos eficazes para prever o futuro.

Ademais, a qualidade do dado é um fator determinante, pois a precisão de qualquer modelo preditivo de séries temporais financeiras depende diretamente da qualidade e disponibilidade dos dados utilizados. No entanto, obter dados financeiros limpos e confiáveis pode ser um grande desafio, especialmente em mercados emergentes ou ativos com baixa liquidez. Problemas como valores ausentes, dados inconsistentes e informações manipuladas podem comprometer a eficiência dos modelos e gerar previsões distorcidas. Além disso, a granularidade dos dados também impacta os resultados – modelos que operam com dados diários podem não capturar flutuações de curto prazo, enquanto dados intradiários podem conter excesso de ruído.

Alterando um pouco o aspecto que vem sendo discutido neste documento, se torna necessário verificar algo além dos aspectos puramente estatísticos, que é o fato de que os mercados financeiros são fortemente influenciados por eventos externos e fatores comportamentais, que muitas vezes não podem ser antecipados pelos modelos preditivos tradicionais. Notícias sobre crises econômicas, mudanças em políticas governamentais, desastres naturais, pandemias e até decisões geopolíticas inesperadas podem provocar grandes oscilações nos preços dos ativos de forma abrupta. Esses eventos geralmente causam reações emocionais nos investidores, levando a movimentos irracionais de mercado, como pânico generalizado ou euforia especulativa.

1.3 - Importância do pré-processamento e da análise exploratória

O pré-processamento dos dados é uma etapa fundamental na modelagem preditiva de séries temporais financeiras, pois garante que as informações utilizadas pelos modelos estejam limpas, organizadas e padronizadas. Como os dados financeiros podem conter valores ausentes, ruídos, outliers e escalas diferentes, é essencial aplicar técnicas que minimizem esses problemas antes de alimentar os modelos de previsão. A normalização ou padronização dos dados, por exemplo, evita que variáveis com amplitudes diferentes dominem o aprendizado do modelo, garantindo uma melhor convergência dos algoritmos de machine learning. Além disso, a remoção de outliers é crucial, pois valores extremos podem distorcer padrões e levar o modelo a aprender comportamentos não representativos da realidade.

Além da limpeza e transformação dos dados, a escolha da janela temporal também desempenha um papel crítico no desempenho dos modelos. Utilizar um período muito curto pode resultar em previsões excessivamente sensíveis a flutuações aleatórias, enquanto períodos muito longos podem diluir padrões recentes e relevantes. A técnica de janela deslizante permite que os modelos sejam treinados continuamente com os dados mais atualizados, garantindo maior adaptabilidade às mudanças de mercado. Adicionalmente, a seleção de variáveis e a extração de novas features ajudam os modelos a capturar informações relevantes para a tomada de decisão. Dessa forma, um bom pré-processamento não apenas melhora a qualidade dos dados, mas também reduz o risco de overfitting, tornando as previsões mais robustas e confiáveis.

Ademais, a análise exploratória de dados é um passo essencial antes da construção de modelos preditivos, pois permite uma compreensão aprofundada da estrutura e dos padrões das séries temporais financeiras. Ao permitir a visualização da evolução dos preços ao longo do tempo, estatísticas descritivas como média, mediana, desvio-padrão e assimetria ajudam a identificar tendências e sazonalidades que podem impactar o desempenho do modelo. Além disso, gráficos como séries temporais, histogramas e boxplots auxiliam na detecção de anomalias e padrões ocultos nos dados. Em modelos de aprendizado de máquina, a análise exploratória de dados também ajuda a definir quais variáveis podem ser mais relevantes para a predição e se há necessidade de engenharia de atributos. Em resumo, uma análise exploratória bem-feita reduz incertezas, evita armadilhas na modelagem e melhora significativamente a qualidade das previsões financeiras.

1.4 - Objetivo do artigo

O objetivo deste artigo é explorar e aplicar técnicas de pré-processamento e análise exploratória de dados em séries temporais de ativos da B3, preparando os dados para futuras modelagens preditivas. De forma mais detalhada, o objetivo é apresentar um estudo detalhado sobre as técnicas e desafios envolvidos na preparação de dados financeiros para modelagem preditiva. O foco principal é garantir que os dados históricos de ativos negociados na Bolsa de Valores do Brasil (B3) sejam devidamente tratados antes da aplicação de modelos de previsão, como redes neurais e algoritmos de aprendizado de máquina. Para isso, o documento aborda métodos de exploração, limpeza, transformação e análise exploratória de dados, buscando identificar padrões, remover inconsistências e preparar um conjunto de dados adequado para análise futura. Além disso, através de estatísticas descritivas e visualizações gráficas, como histogramas, boxplots e gráficos de séries temporais, o estudo busca identificar tendências, sazonalidades e anomalias que possam impactar os modelos preditivos.

2. TRABALHOS RELACIONADOS

Nesta seção serão apresentados estudos que possuem relação com a proposta, servindo de base para organização, motivação e definição do tema, a fim de buscar contribuições que auxiliaram no desenvolvimento deste trabalho.

2.1-B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis

Este estudo conseguiu aplicar a forma não usual de integrar dados financeiros históricos com análise de sentimentos extraídos do Twitter, permitindo que o modelo capturasse oscilações emocionais do mercado. O processo de exploração envolveu a coleta de dados históricos de preços de empresas da B3, além da extração e análise de postagens para medir o impacto do sentimento do mercado nos preços das ações. Durante o pré-processamento, os textos das postagens foram tokenizados e limpos por meio de técnicas de Processamento de Linguagem Natural (NLP), convertendo sentimentos qualitativos em variáveis numéricas. Além disso, os preços das ações e os sentimentos foram normalizados para manter a coerência das variáveis e alimentar corretamente a rede neural LSTM. A definição de janelas deslizantes garantiu que o modelo aprendesse padrões em séries temporais de curto e longo prazo, resultando em uma redução significativa do erro de previsão.

2.2-Previsão de Preços de Ações Utilizando Inteligência Artificial

Este estudo apresenta uma abordagem de pré-processamento que se concentra no uso de modelos estatísticos e de aprendizado profundo para prever preços de ações. O processo de exploração de dados incluiu a análise de estatísticas descritivas, gráficos de séries temporais e histogramas para identificar padrões, sazonalidades e volatilidade dos ativos. No pré-processamento, foram aplicadas técnicas como normalização e padronização para evitar discrepâncias numéricas que poderiam impactar o treinamento dos modelos, além da remoção de outliers e preenchimento de dados ausentes por interpolação. A definição da granularidade temporal foi ajustada para otimizar a entrada nos modelos ARIMA, PROPHET e LSTM, garantindo uma melhor adaptação a diferentes horizontes de previsão.

2.3-Predição de Tendências em Séries Financeiras Utilizando Meta Classificadores

Já este estudo adotou uma abordagem baseada em metaclassificação, combinando diferentes algoritmos para aprimorar a previsão de tendências financeiras. A exploração dos dados focou na análise de tendências de mercado, sazonalidades e padrões estatísticos, avaliando a eficácia de diferentes classificadores aplicados a nove ativos da B3.

No pré-processamento, foi realizada uma rigorosa seleção de variáveis para evitar viés e overfitting, além da aplicação de engenharia de atributos para extrair características relevantes dos dados financeiros. A normalização garantiu que os diversos modelos operassem sob as mesmas condições, enquanto a filtragem de ruído e a suavização de séries temporais contribuíram para uma modelagem mais precisa. A divisão dos dados em conjuntos de treinamento e teste utilizou validação cruzada, otimizando o desempenho do metaclassificador.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 - Séries Temporais Financeiras

Uma série temporal financeira é um conjunto de observações ordenadas no tempo, normalmente associadas a ativos como ações, commodities e moedas. Essas séries apresentam características próprias, como tendência, sazonalidade e autocorrelação, tornando a modelagem um desafio para previsões precisas (Box, Jenkins & Reinsel, 2015).

Matematicamente, uma série temporal pode ser representada como:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-n}) + \epsilon_t \quad (1)$$

onde Y_t representa o valor da série no tempo t , f é uma função desconhecida que pode capturar dependências temporais, e ϵ_t é um termo de erro aleatório (Hyndman & Athanasopoulos, 2018).

Em finanças, os principais tipos de séries temporais incluem:

- **Preços de fechamento:** valores finais de uma ação em cada pregão.
- **Retornos logarítmicos:** taxa de variação percentual dos preços, calculada como:

$$R_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (2)$$

- **Volatilidade:** medida de dispersão dos retornos, crucial para avaliação de risco (Engle, 1982).

A modelagem dessas séries requer técnicas avançadas, como modelos estatísticos (ARIMA, GARCH) e aprendizado profundo (redes neurais recorrentes e transformers) (Goodfellow, Bengio & Courville, 2016).

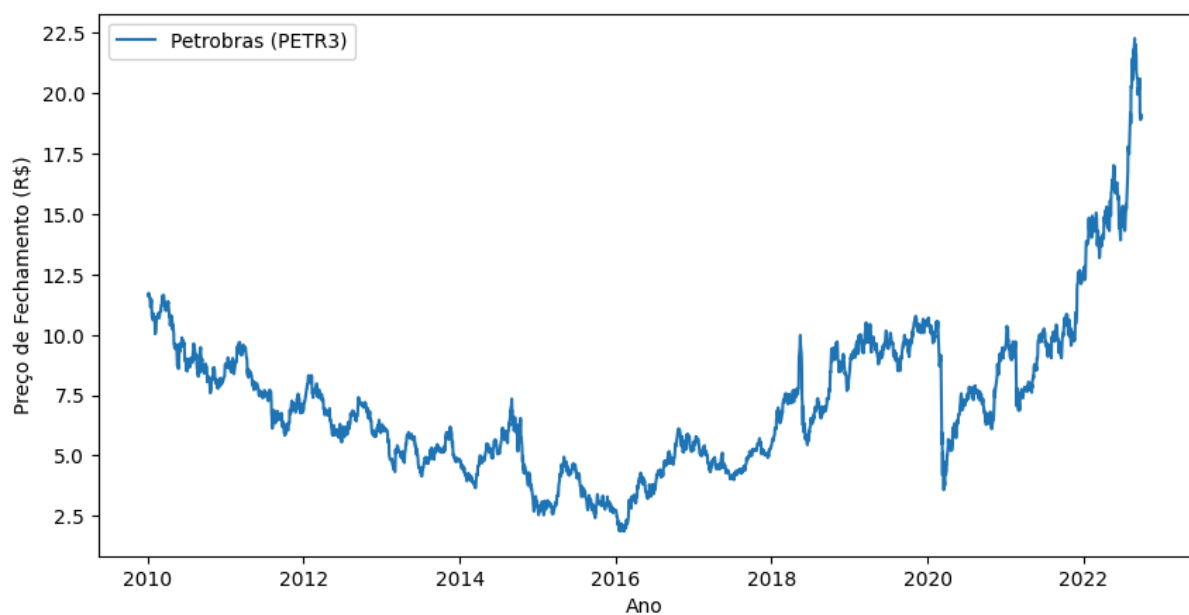
3.2 - Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) é um passo essencial para entender padrões, anomalias e relações entre variáveis antes da modelagem preditiva (Tukey, 1977). No contexto de séries temporais financeiras, algumas etapas fundamentais da EDA incluem:

3.2.1 - Visualização dos Dados

A plotagem dos preços ao longo do tempo permite identificar tendências e sazonalidades. Por exemplo, ao analisar o preço de fechamento de uma ação da B3, podemos usar um gráfico de linhas para observar padrões (Figura 01).

Figura 01 - Preço de fechamento da Petrobras (de 2010 até 2022)



Autoria Própria (2025)

3.2.1.1 - Análise Combinando Gráfico Violino e Boxplot

A análise da distribuição dos preços de ações é uma etapa essencial na exploração de séries temporais financeiras. Compreender como os preços se comportam ao longo do tempo pode revelar padrões importantes, como a presença de outliers, a volatilidade dos ativos e possíveis mudanças estruturais nos preços. Entre as diversas formas de visualizar distribuições de dados, os gráficos de **boxplot** e **violino** são ferramentas poderosas. Cada um fornece informações valiosas individualmente, mas quando combinados, oferecem uma visão ainda mais rica e detalhada sobre o comportamento dos preços de ativos financeiros.

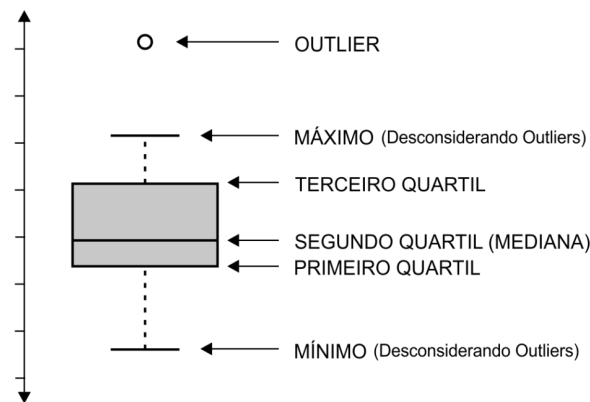
3.2.1.1.1 - Boxplot

O **boxplot** (ou diagrama de caixa) é uma representação gráfica dos dados baseada em seus quartis. Ele fornece informações sobre a dispersão e a assimetria dos preços, destacando a mediana, os quartis inferior e superior e a presença de outliers. A estrutura básica do boxplot é (Figura 02):

- **Linha central (mediana):** Representa o valor central da distribuição.
- **Caixa:** Delimitada pelos quartis inferior (Q1) e superior (Q3), mostrando onde está concentrada a maior parte dos dados.
- **Extremos ("whiskers"):** Indicam a amplitude da distribuição dos dados, excluindo outliers.
- **Outliers:** Representados como pontos fora da caixa, evidenciam variações atípicas nos preços das ações.

O boxplot é uma excelente ferramenta para identificar a dispersão dos preços e detectar possíveis eventos anômalos no mercado financeiro.

Figura 02 : Imagem Ilustrativa Boxplot



Oliveira (2019)

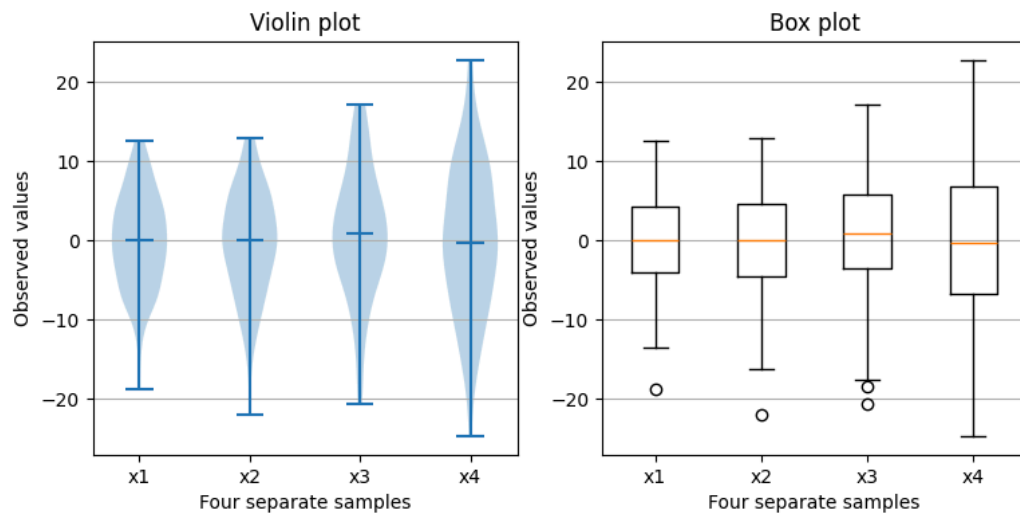
3.2.1.1.2 - Gráfico Violino

O **gráfico de violino** é uma extensão do boxplot que incorpora uma estimativa da densidade dos dados, fornecendo uma visão mais detalhada da distribuição (Figura 03). Ele apresenta um contorno suave que reflete a frequência relativa dos preços em diferentes faixas de valores. Com esse gráfico, é possível identificar:

- **A forma da distribuição:** Se os preços seguem uma distribuição normal, assimétrica ou multimodal.
- **Regiões de maior concentração:** As partes mais largas do gráfico indicam onde os preços ocorreram com maior frequência.
- **Presença de múltiplos picos:** Diferente do boxplot, o gráfico de violino pode evidenciar distribuições bimodais ou multimodais, sugerindo mudanças nos regimes do mercado.

O gráfico de violino permite uma análise mais rica das tendências dos preços das ações, principalmente em ativos com volatilidade significativa.

Figura 03: Comparação entre violino e boxplot



Hunter (2012)

3.2.1.1.3 - Combinação Violino-Boxplot

A combinação do **boxplot** com o **gráfico de violino** proporciona uma visão mais completa dos dados. Enquanto o **boxplot** fornece um resumo estatístico conciso, destacando os quartis e os outliers, o **violino** adiciona uma camada de informações sobre a distribuição completa dos preços. Isso é especialmente útil para:

- **Identificar outliers e analisar sua influência na distribuição geral.**
- **Detectar assimetrias que podem não ser evidentes apenas no boxplot.**
- **Visualizar se a distribuição dos preços é unimodal ou multimodal.**
- **Comparar a volatilidade de diferentes ações ao longo do tempo.**

Essa abordagem combinada permite não apenas um entendimento mais preciso da distribuição dos preços, mas também fornece uma interpretação mais intuitiva para tomadas de decisão no mercado financeiro.

3.2.2 - Métricas Estatísticas

Além da visualização, estatísticas descritivas como média, mediana, desvio padrão e curtose ajudam a entender a distribuição dos dados. O cálculo da autocorrelação também é crucial (Shumway & Stoffer, 2017):

$$ACF(k) = \frac{\sum_{t=1}^{N-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^N (Y_t - \bar{Y})^2} \quad (3)$$

A função de autocorrelação (ACF) ajuda a determinar a dependência entre observações passadas e presentes, auxiliando na escolha de modelos adequados.

3.2.3 - Detecção de Outliers

Valores atípicos podem ser detectados usando métodos estatísticos como o IQR (Intervalo Interquartil) ou técnicas de aprendizado de máquina, como Isolation Forest (Liu, Ting & Zhou, 2008).

3.2.4 - Definição dos Parâmetros Analisados

3.2.4.1 - Skewness e Kurtosis

Para iniciar a discussão sobre os gráficos, é mister a explicação dos parâmetros que serão discutidos, que são as Skewness, conhecida popularmente como a assimetria, e a Kurtosis, que é nomeada como Excentricidade no conceito comum.

A Skewness, também conhecida como assimetria, é um dos quatro momentos da distribuição de probabilidade. Ele descreve a simetria ou assimetria de uma distribuição em relação à média. Ela tem as seguintes possibilidades de classificação:

- Positivo : Quando a distribuição tem um "caimento" mais a esquerda e a cauda se estende para a direita, é chamada de skewness positiva (ou right-skewed). Isso significa que há uma maior frequência de valores baixos com poucos ou nenhum valor ainda mais baixo.
- Negativo : Quando a distribuição tem um "caimento" mais à direita e a cauda se estende para a esquerda, é chamada de skewness negativa (ou left-skewed). Isso significa que há uma maior frequência de valores altos com poucos ou nenhum valor ainda mais alto.
- Nulo : Uma distribuição simétrica tem skewness igual a zero.

Ademais, a Kurtosis , também conhecida como excentricidade, é um parâmetro que descreve a "cauda" ou o "peito" da distribuição em relação à média e à variância.

- Kurtosis de 3 : A kurtosis padrão (ou excedente) subtrai 3 para comparar com a distribuição normal. Assim, uma kurtosis de 0 indica que a distribuição tem o mesmo "peito" e cauda em comparação à distribuição normal.
- Kurtosis maior do que 3 : A distribuição é considerada leptocúrtica (ou mesocúrtica), com uma distribuição mais pesada e caudas mais aguçadas. Isso significa que há uma tendência de valores extremos maiores.
- Kurtosis menor do que 3 : A distribuição é considerada platicúrtica , com uma distribuição mais plana, o que significa que os valores extremos são menos prováveis e a distribuição tem uma cauda mais alongada.

3.3 - Técnicas de Pré-Processamento em Séries Temporais

O pré-processamento é uma etapa crucial para garantir que os dados estejam adequados para a modelagem. Algumas das principais técnicas são:

3.3.1 - Normalização e Padronização:

Redes neurais e modelos estatísticos podem ser sensíveis a escalas diferentes. Métodos comuns incluem:

- **Min-Max Scaling** (Han, Kamber & Pei, 2011):

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (4)$$

- **Padronização (Z-score normalization):**

$$X' = \frac{X - \mu}{\sigma} \quad (5)$$

Onde μ é a média e σ o desvio padrão da série.

3.3.2 - Diferençamento

Para tornar séries estacionárias (sem tendência), aplica-se o diferenciamento (Hamilton, 1994):

$$Y'_t = Y_t - Y_{t-1} \quad (6)$$

Isso é útil em modelos como ARIMA, que assumem estacionariedade nos dados.

3.3.3 - Engenharia de Atributos

Transformar dados crus em representações mais informativas pode melhorar a performance dos modelos. Exemplos incluem:

- **Médias móveis:** suavizam variações e ajudam a capturar tendências.
- **Retornos acumulados:** mostram a evolução percentual de um investimento ao longo do tempo.

3.4 - Importância do Preparo dos Dados para Modelos Preditivos

O sucesso de um modelo preditivo depende fortemente da qualidade dos dados utilizados. Dados financeiros possuem ruído, padrões complexos e dependências temporais que podem comprometer previsões se não forem tratados adequadamente (Chollet, 2021).

3.4.1 - Impacto na Performance dos Modelos

Um conjunto de dados mal preparado pode levar a:

- Overfitting (ajuste excessivo) devido a padrões espúrios.
- Underfitting (ajuste insuficiente) se os dados não capturarem informações relevantes.
- Viés devido a dados inconsistentes ou mal escalonados.

3.4.2 - Redução de Ruído e Tratamento de Falhas

Séries financeiras podem conter valores ausentes devido a feriados ou problemas de coleta. Métodos de imputação, como interpolação linear e modelos baseados em aprendizado de máquina, podem ser aplicados (Hastie, Tibshirani & Friedman, 2009).

4. MATERIAIS E MÉTODOS

4.1 -Softwares Utilizados

Para a realização do projeto, o grupo utilizou três softwares principais, os quais desempenharam papéis cruciais no desenvolvimento, organização e colaboração ao longo do processo. A seguir, são descritos os principais softwares utilizados:

1. Visual Studio Code (VS Code)

O *Visual Studio Code* foi a ferramenta principal utilizada para o desenvolvimento e organização da estrutura de códigos. Este editor de código-fonte, altamente configurável, permite a escrita de código em diversas linguagens, incluindo Python, e oferece funcionalidades como autocompletar, integração com o terminal e extensões específicas para análise de dados e visualização.

2. Anaconda Navigator

O *Anaconda Navigator* foi utilizado para a criação e gerenciamento dos ambientes necessários para o desenvolvimento do projeto. Este software fornece uma interface gráfica para facilitar a instalação e o gerenciamento de pacotes Python, bem como a criação de ambientes virtuais.

3. GitHub

O *GitHub*, em conjunto com o *Git*, foi utilizado para o versionamento do código e para o trabalho colaborativo entre os membros da equipe. O uso de repositórios Git permitiu o acompanhamento das mudanças no código-fonte ao longo do tempo, a integração de contribuições de diferentes membros da equipe e a resolução eficiente de conflitos.

Esses três softwares foram fundamentais para o sucesso do projeto, facilitando o desenvolvimento, organização e colaboração entre os membros da equipe, além de garantir uma estrutura sólida e escalável para o projeto.

4.2 - Obtenção e características dos dados

Os dados utilizados neste projeto foram obtidos diretamente do *Yahoo Finance*, por meio da biblioteca *yfinance* do Python, que permite o download e acesso a informações financeiras de diversas ações listadas nas bolsas de valores. Através dessa ferramenta, foi possível acessar os dados históricos das ações da B3, no período de 04/01/2010 a 29/09/2022, para as seguintes empresas: BBAS3, BRAP3, VALE3, ITUB4, CMIG4, CYRE3, DIRR3, ELET3, EQTL3, TEND3 e PETR3.

Cada ação foi extraída separadamente, sendo que os dados brutos de cada uma foram armazenados em arquivos CSV individuais. Cada arquivo contém as seguintes colunas:

- **Price:** Preço de fechamento ajustado.
- **Adj Close:** Preço de fechamento ajustado.
- **Close:** Preço de fechamento sem ajustes.
- **High:** Maior preço do dia.
- **Low:** Menor preço do dia.
- **Open:** Preço de abertura.
- **Volume:** Volume de ações negociadas.

Cada arquivo CSV contém 3165 linhas e 7 colunas, representando os dados diários das ações ao longo do período mencionado. Essas características permitem a análise de tendências e volatilidade das ações de forma detalhada, com informações sobre os preços de fechamento, abertura e os volumes de negociação.

Esses dados foram processados e utilizados para realizar as análises propostas, como a correlação entre as ações e a avaliação da volatilidade, por meio de gráficos como os boxplots e as bandas de Bollinger.

4.3 - Organização da Base de Dados

A organização da base de dados foi realizada para garantir acessibilidade, integridade e reprodutibilidade das análises. Essa etapa envolveu a definição de uma estrutura de pastas e o upload dos dados para um repositório em nuvem. A seguir, detalhamos cada uma dessas ações.

4.3.1 - Definição da Estrutura de Pastas

Para facilitar o gerenciamento dos dados e scripts, foi criada uma estrutura de pastas bem definida, composta por:

Pasta **raw**: Armazena os dados brutos obtidos diretamente da fonte (Yahoo Finance API), sem qualquer tipo de processamento ou alteração. Essa pasta contém arquivos no formato CSV, com informações históricas de preços, volumes e outros indicadores das ações.

Pasta **processed**: Contém os dados que foram processados e preparados para análise. Aqui são armazenados arquivos resultantes de transformações como limpeza, filtragem e normalização de dados.

Pasta **scripts**: Guarda certos scripts utilizados para manipular, processar e analisar os dados. Isso inclui scripts em Python para download, transformação e visualização dos dados.

4.3.2 - Upload dos Dados para um Repositório em Nuvem

Para garantir a segurança, acessibilidade e escalabilidade dos dados, optou-se por utilizar um repositório em nuvem. A plataforma escolhida foi o Google Drive, devido à sua facilidade de uso, integração com ferramentas de análise de dados e capacidade de armazenamento gratuito. Os passos realizados incluem:

1. Upload dos Dados:

Os arquivos brutos e processados foram carregados para pastas específicas no Google Drive, seguindo a mesma estrutura de pastas local (raw, processed).

A nomenclatura dos arquivos foi padronizada para facilitar a identificação (ex: PETR4.csv).

2. Configuração de Permissões:

As pastas no Google Drive foram configuradas com permissões de acesso restrito, garantindo que apenas membros autorizados do projeto pudessem visualizar ou editar os dados.

4.3.3 - Consolidação dos Dados Processados

Para facilitar a análise e o processamento dos dados, foi criado um dataset único contendo apenas os preços de fechamento das ações. Essa etapa envolveu:

1. Seleção dos Preços de Fechamento:

A partir dos dados brutos, foram extraídos apenas os preços de fechamento de cada ação, descartando outras colunas como preço de abertura, máximo, mínimo e volume. Essa seleção foi feita para simplificar a análise e reduzir a complexidade do dataset.

2. Junção dos Dados:

Os preços de fechamento de todas as ações foram consolidados em um único DataFrame utilizando a biblioteca Pandas. Foi adicionada uma coluna identificadora (ticker) para diferenciar os dados de cada ação (ex: PETR4, VALE3, etc.).

3. Armazenamento do Dataset Consolidado:

O dataset consolidado, contendo apenas os preços de fechamento, foi salvo em um único arquivo CSV na pasta **processed**, facilitando o acesso e a reprodutibilidade das análises subsequentes.

4.4 - Análise Exploratória dos Dados (EDA)

A análise exploratória dos dados (EDA) foi realizada utilizando métodos estatísticos e técnicas de visualização para compreender as características do dataset, identificar padrões, tendências e possíveis inconsistências. As análises foram divididas em quatro categorias principais: análise de séries temporais, análise de distribuições, análise de correlações e análise de normalização. A seguir, descrevemos cada uma dessas abordagens.

4.4.1 - Análise das séries temporais

Para analisar o comportamento dos preços das ações ao longo do tempo, foram aplicadas técnicas de análise de séries temporais com o intuito de identificar tendências, sazonalidades e padrões.

- Gráfico de Linhas de Séries Temporais:

Séries temporais dos preços de fechamento de ações selecionadas foram plotadas, possibilitando a comparação visual de suas trajetórias. Essa abordagem visa identificar tendências de longo prazo e padrões de comportamento similares entre as ações analisadas.

- Gráfico de Bandas de Bollinger:

Bandas de Bollinger foram geradas para ações específicas, representando a volatilidade dos preços ao longo do tempo. Esta técnica foi utilizada para identificar períodos de alta volatilidade e potenciais pontos de reversão de tendência.

- Gráfico de Retornos Percentuais Diários:

Os retornos percentuais diários foram calculados e plotados, mostrando a variação percentual entre o preço atual e o preço anterior, conforme a fórmula abaixo:

$$Retorno = \frac{\text{Preço Atual} - \text{Preço Anterior}}{\text{Preço Anterior}} \times 100 \quad (7)$$

Esta métrica foi empregada para avaliar a volatilidade diária e detectar dias com movimentos atípicos no mercado.

4.4.2 - Análise de distribuições

Técnicas estatísticas foram aplicadas para examinar a distribuição dos preços e retornos das ações, permitindo a identificação de padrões, assimetrias e possíveis valores atípicos.

- Histogramas

Histogramas foram gerados para os preços de fechamento das ações, proporcionando uma visualização clara da distribuição dos dados.

Essa análise possibilita a identificação de assimetrias e padrões na distribuição, auxiliando na compreensão do comportamento dos preços ao longo do tempo.

- Gráficos de boxplot

Boxplots foram construídos e agrupados por ano, além de uma comparação geral entre as ações por meio de seus boxplots, ilustrando a distribuição dos preços de fechamento ao longo dos anos. Essa abordagem permite identificar variações sazonais e comparar a dispersão dos preços entre diferentes ações, além de detectar a presença de valores atípicos que podem indicar eventos anômalos no mercado.

- Gráfico de Violino com Boxplot

Gráficos de violino foram combinados com boxplots em uma única visualização, proporcionando uma análise mais detalhada da distribuição e densidade dos dados.

Esse método foi empregado para comparar a distribuição dos preços ou retornos em diferentes granularidades, como variações mensais ou anuais. Primeiramente visualizou-se todo o período do dataset com as diferentes granularidades. Em seguida, separou-se o dataset em dois períodos: antes de 2019 e após 2019, a fim de comparar o comportamento do mercado de ações nesses dois períodos.

4.4.3 - Análise de correlação

Técnicas estatísticas foram aplicadas para examinar as relações entre os preços das ações, possibilitando a identificação de padrões de dependência entre os ativos.

- Matriz de Correlação e Heatmap:

A matriz de correlação entre as ações foi calculada utilizando o coeficiente de Pearson, permitindo quantificar a relação linear entre os preços dos ativos.

Um heatmap foi gerado para representar visualmente a matriz de correlação, facilitando a interpretação das relações entre as ações.

Essa abordagem permite identificar rapidamente grupos de ativos com alta ou baixa correlação, auxiliando em análises comparativas e estratégias de diversificação.

4.4.4 - Análise da Kurtosis e Skewness.

Esta análise focou na avaliação da curtose e da assimetria (skewness) das distribuições dos preços de fechamento. O objetivo foi identificar distribuições com caudas

pesadas ou assimétricas, características que podem indicar potenciais riscos ou oportunidades no mercado.

Para facilitar a interpretação, foram gerados gráficos, como histogramas com linhas de densidade, permitindo a visualização clara da curtose e assimetria das distribuições..

4.5 - Normalização dos Dados

A normalização dos dados foi aplicada com o objetivo de padronizar as escalas dos preços de fechamento das ações, permitindo comparações visuais mais precisas e facilitando a análise de séries temporais. Essa etapa é fundamental, pois os preços das ações possuem magnitudes distintas, o que pode dificultar a interpretação dos padrões e relações entre diferentes ativos.

Para essa normalização, utilizamos a biblioteca *scikit-learn*, que oferece ferramentas eficientes para a transformação de dados. Foram aplicadas três técnicas distintas de normalização:

- **Min-Max Scaling:** Reescalona os dados para um intervalo entre 0 e 1, preservando a distribuição original e tornando-os diretamente comparáveis.
- **Z-score (Padronização):** Centraliza os dados em torno da média e os escala de acordo com o desvio padrão, sendo útil para identificar desvios significativos em relação ao comportamento médio dos preços.
- **Robust Scaler:** Baseia-se na mediana e no intervalo interquartil (IQR), tornando-se menos sensível a outliers, o que é vantajoso para séries com grande variação nos preços.

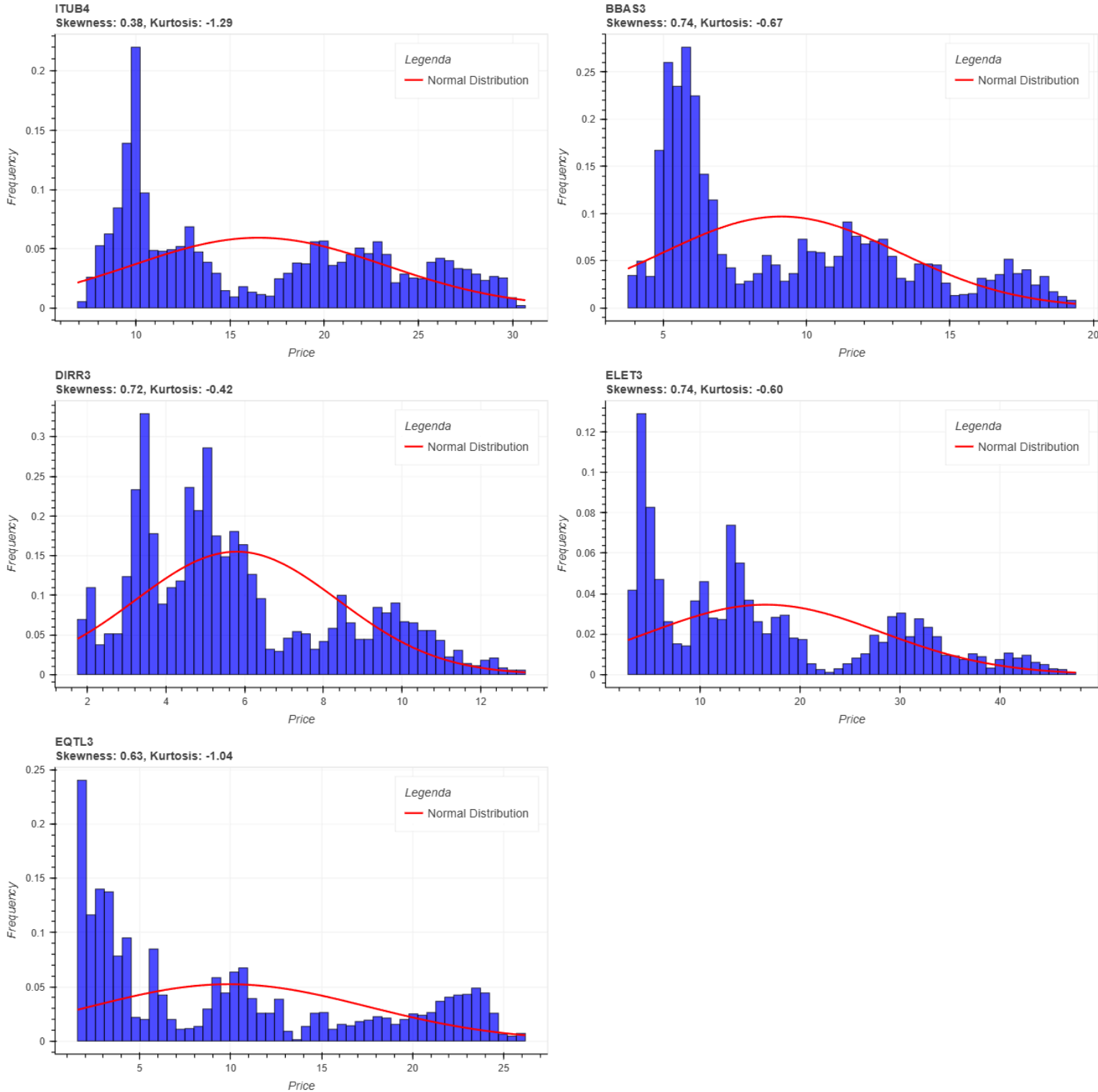
A implementação foi realizada por meio de um script em Python, onde as transformações foram aplicadas às séries históricas dos preços de fechamento das ações analisadas. Os dados normalizados foram posteriormente armazenados em arquivos CSV na pasta *processed*, permitindo fácil reutilização para etapas posteriores da análise

5. RESULTADOS E DISCUSSÃO

5.1 - Skewness e Kurtosis

Os gráficos das ações foram divididos em 3 grupos

Figura 04 - Leve Assimetria



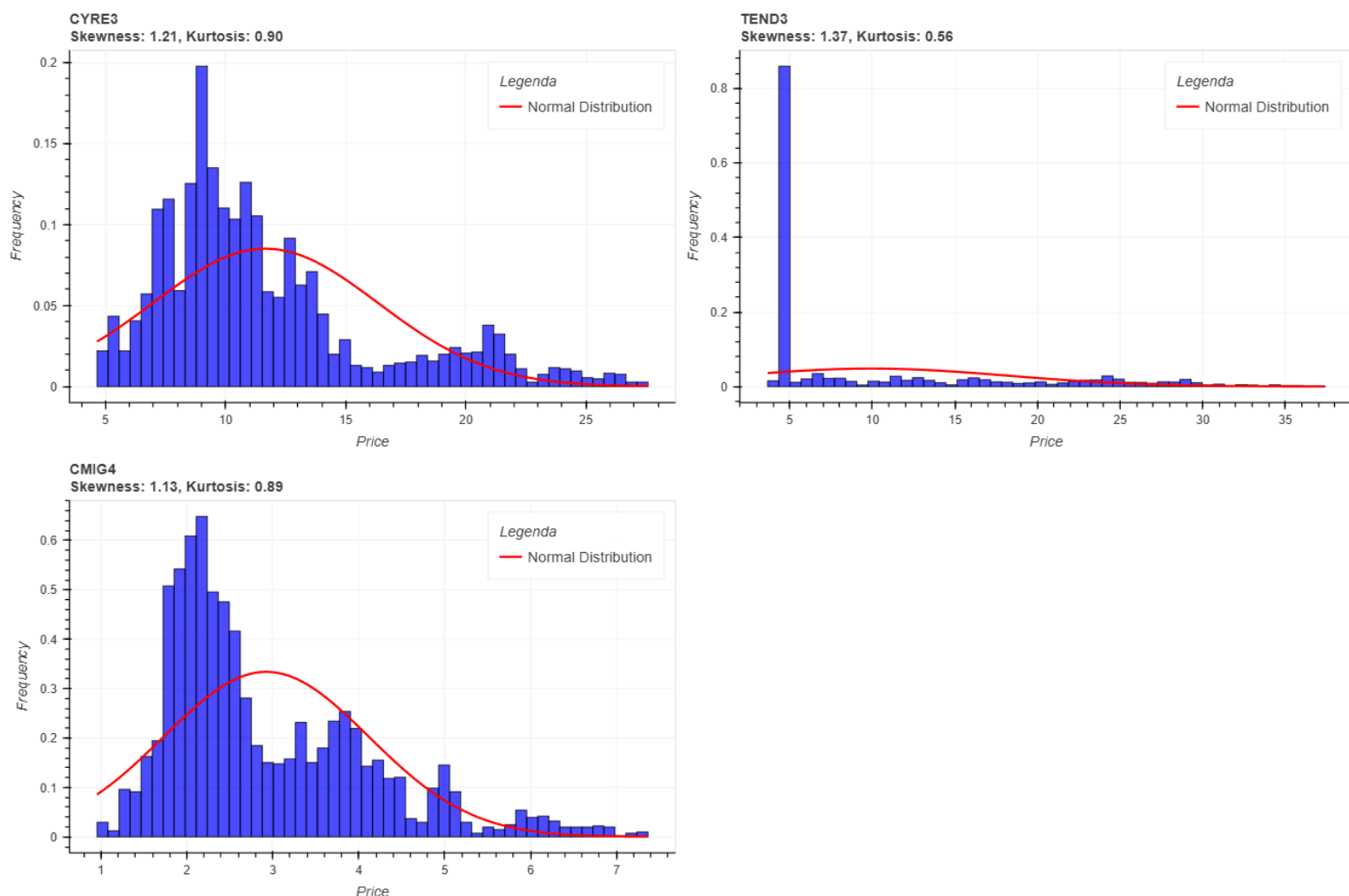
Autoria Própria (2025)

No Grupo 1, representado pela figura 4, temos ativos cujas distribuições apresentam baixa assimetria e curtose negativa ou próxima de zero, como evidenciado pelas Skewness e Kurtosis reportadas. As ações como ITUB4, BBAS3, DIRR3, ELET3 e EQTL3 possuem gráficos de histograma que se aproximam de uma distribuição normal, com caudas leves e simétricas em torno da média. Isso indica que os preços desses ativos são mais estáveis, com variações limitadas ao longo do tempo. Esses ativos são caracterizados por uma previsibilidade maior, tornando-os mais atraentes para investidores conservadores que buscam baixa volatilidade e risco reduzido.

É possível caracterizar esse grupo com base em seus parâmetros, atribuindo as ações presentes nele as seguintes características:

- Baixo Risco: Por serem ativos mais estáveis, são menos suscetíveis a grandes oscilações. Isso reduz as chances de perdas significativas, tornando-os ideais para investidores conservadores.
- Retornos Moderados: Embora apresentem menor risco, os retornos esperados também tendem a ser mais previsíveis e moderados, sem grandes oportunidades de lucros extraordinários.
- Atratividade para Investidores Conservadores: Indicados para investidores que priorizam a segurança e a preservação de capital em vez de altos rendimentos.

Figura 05 - Média Assimetria



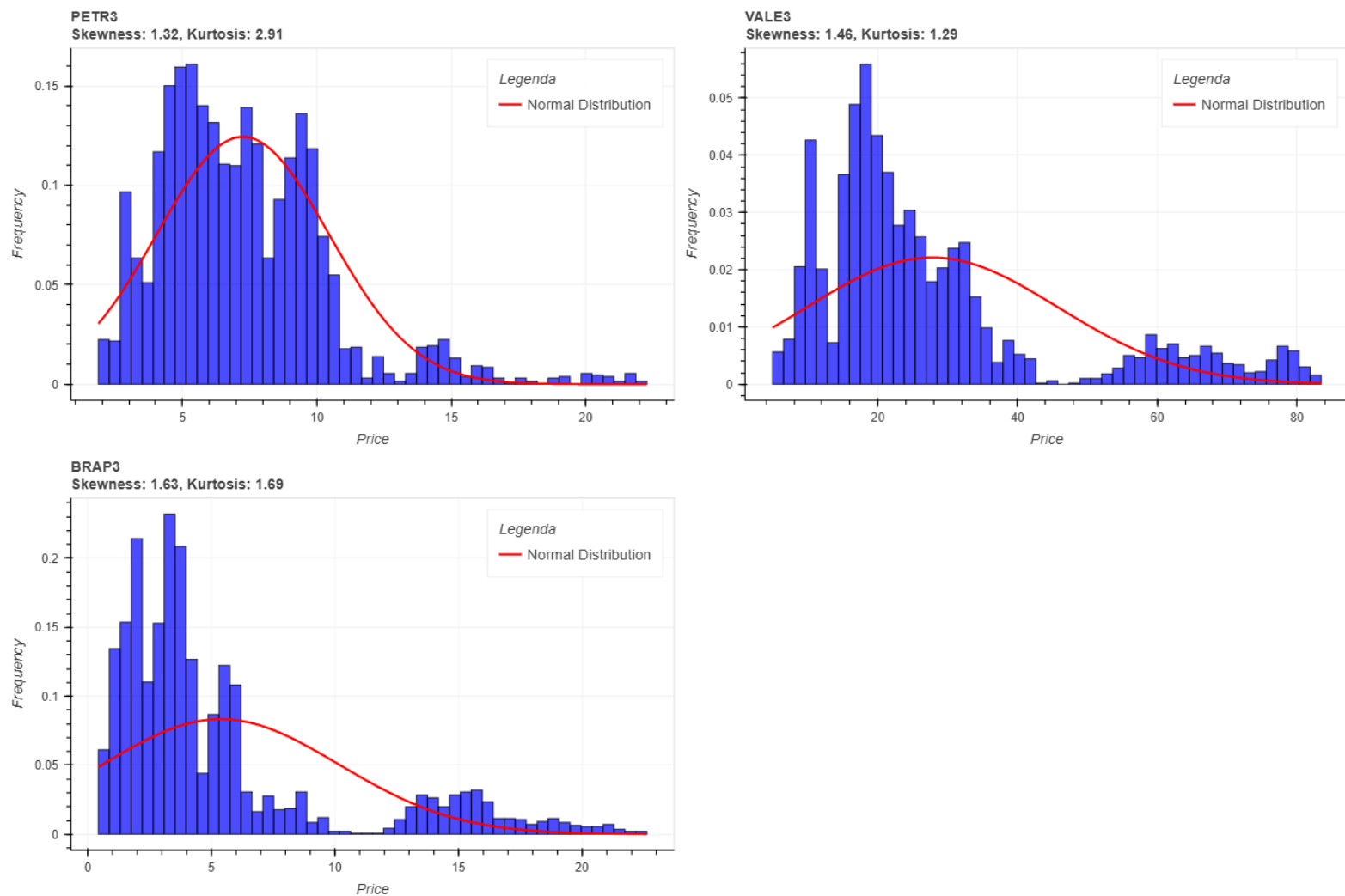
Autoria Própria (2025)

No Grupo 2, representado pela figura 5, os ativos como CYRE3, TEND3 e CMIG4 exibem assimetria moderada, com valores de Skewness entre 0.75 e 1.4. Isso indica que os preços têm uma leve inclinação para a direita, sugerindo uma concentração maior de valores baixos e algumas ocorrências de preços elevados. A curtose próxima de zero sugere caudas mais equilibradas, mas com uma leve propensão a eventos fora do padrão. Os gráficos mostram distribuições que começam a se afastar de uma curva normal, mas sem grandes desvios. Este grupo reflete ativos de risco intermediário, ideais para investidores que buscam um equilíbrio entre estabilidade e possibilidade de ganhos moderados.

Essas ações apresentam as seguintes características:

- Risco Moderado: Investidores podem enfrentar flutuações mais pronunciadas em comparação com o Grupo 1, mas ainda dentro de limites razoáveis.
- Possibilidade de Ganhos Leves: Por apresentarem leve inclinação para valores altos, existe a possibilidade de retornos melhores, porém com maior risco.
- Adequação para Investidores Moderados: Esses ativos são ideais para investidores que estão dispostos a correr algum risco, mas não desejam se expor a oscilações extremas.

Figura 06 - Alta Assimetria



Autoria Própria (2025)

No Grupo 3, representado pela figura 6, os ativos como, PETR3, VALE3 e BRAP3 apresentam características marcantes de alta assimetria e curtose elevada. Com Skewness acima de 1.4 e Kurtosis acima de 1.5, os gráficos mostram distribuições com caudas longas à

direita, indicando alta concentração de valores baixos, mas com grandes oscilações e eventos extremos. Esses ativos são altamente voláteis, propensos a grandes oscilações nos preços e influenciados por fatores externos, como mudanças econômicas ou notícias de mercado. Este grupo representa uma alta oportunidade de retorno, mas com risco significativo, sendo mais adequado para investidores agressivos que aceitam a possibilidade de grandes perdas em troca de lucros potenciais elevados.

Suas características são as seguintes:

- **Alto Risco:** Investidores estão sujeitos a oscilações drásticas, o que pode resultar em grandes perdas financeiras em curto prazo.
- **Oportunidades de Ganhos Elevados:** Por outro lado, a presença de eventos extremos também oferece chances de ganhos extraordinários, atraindo investidores agressivos.
- **Impacto de Fatores Externos:** Esses ativos são altamente influenciados por eventos econômicos, políticos ou notícias de mercado, o que os torna imprevisíveis.
- **Adequação para Investidores Agressivos:** São indicados para quem está disposto a aceitar grandes riscos em troca de altas recompensas potenciais.

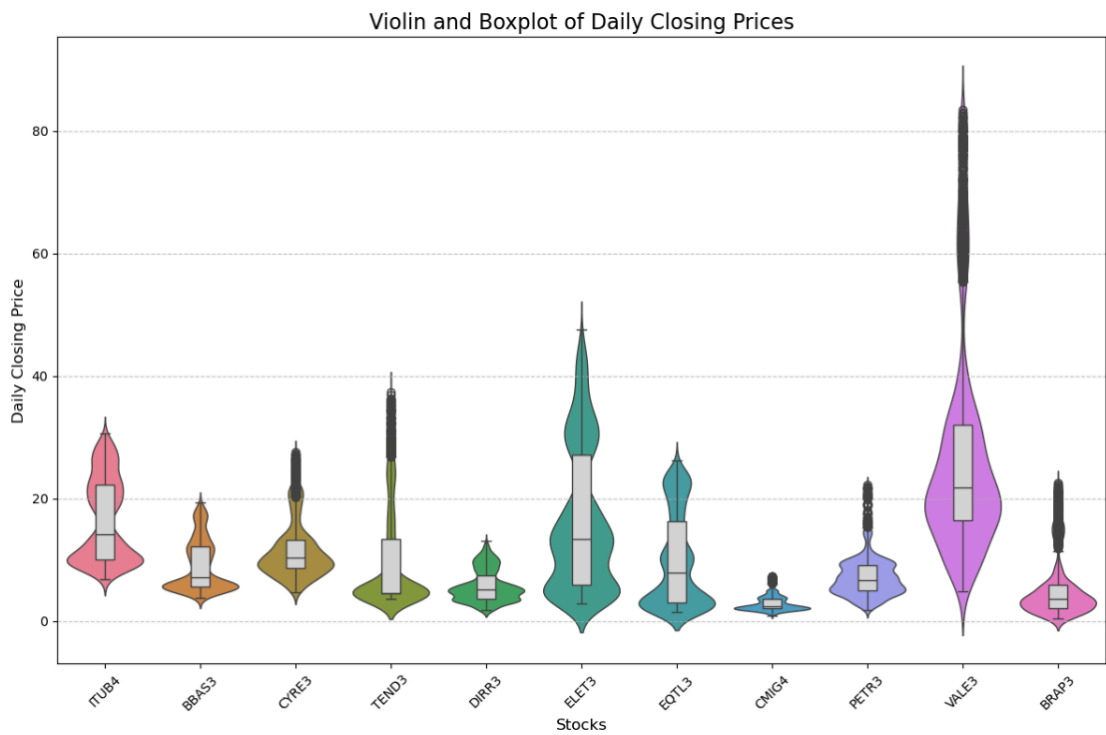
5.2 - Visualizações Utilizando Gráfico Violino e Boxplot

5.2.1 - Visualizações para todo o período do Dataset (de 2010 até 2022)

Os outliers ficam bem evidentes quando utiliza-se a granularidade diária (Figura 07) e vão ficando menos aparentes quando aumenta-se essa granularidade (Figuras 08, 09 e 10). Tais mudanças ficam fáceis de serem observadas com a abordagem gráfica utilizada:

- **Dados diários:**

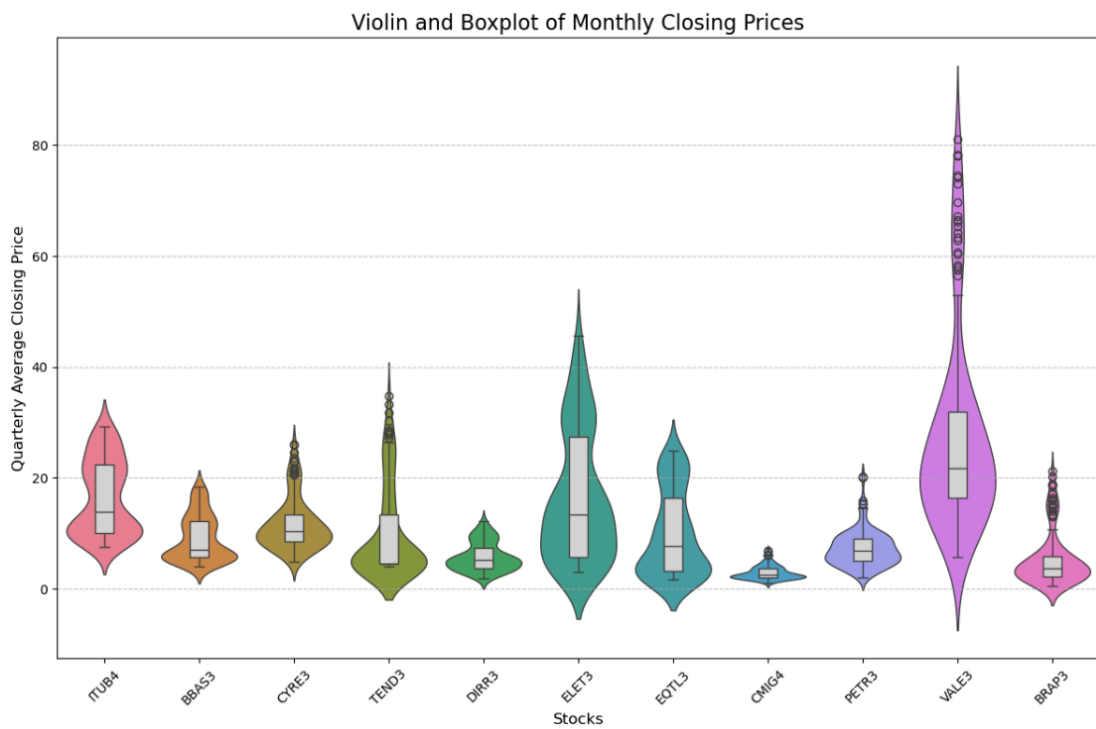
Figura 07: Dados diários das 11 ações utilizando violin-boxplot



Autoria Própria (2025)

- **Dados mensais:**

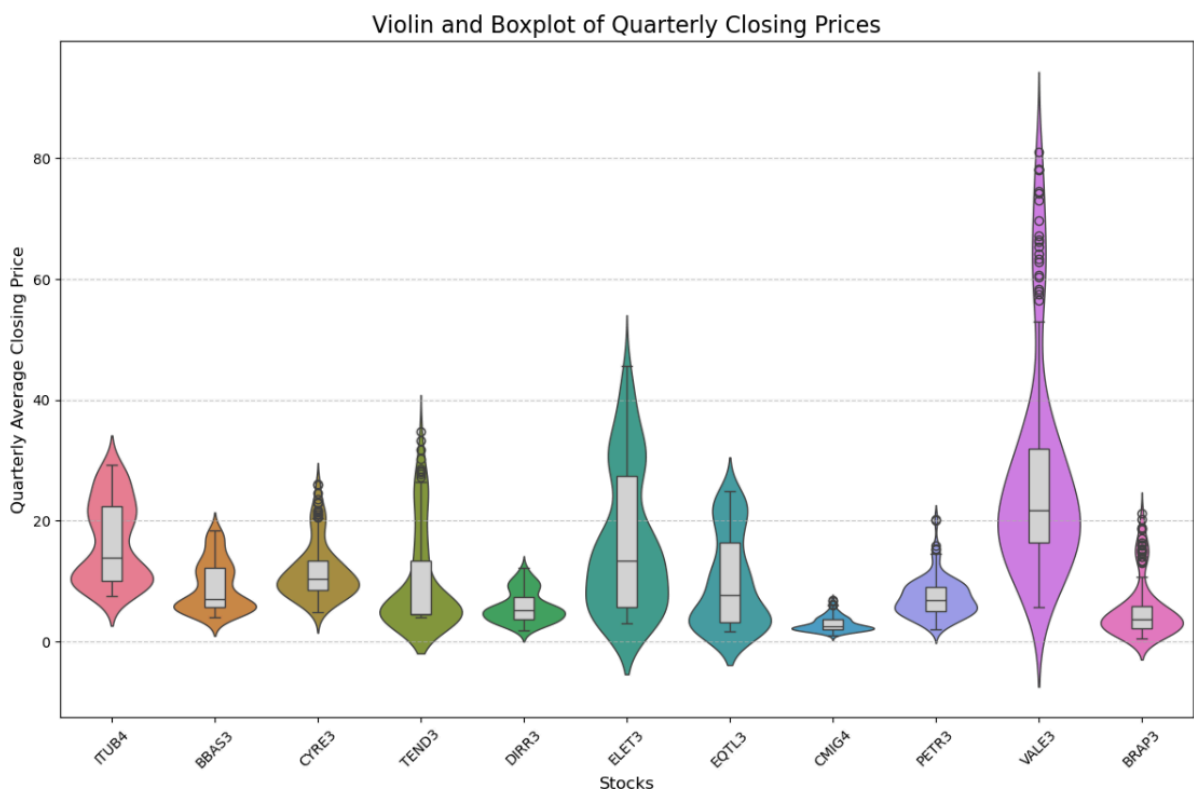
Figura 08: Dados mensais das 11 ações utilizando violin-boxplot



Autoria Própria (2025)

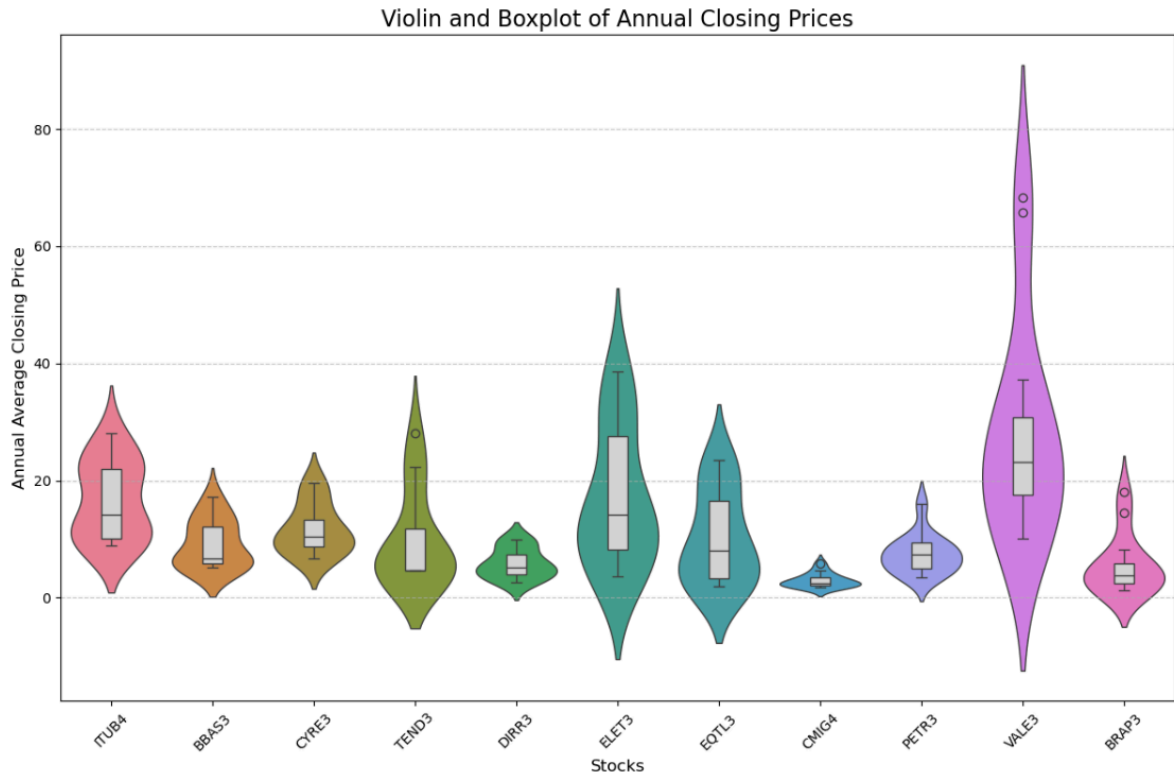
- **Dados trimestrais:**

Figura 09: Dados trimestrais das 11 ações utilizando violin-boxplot



- **Dados anuais:**

Figura 10: Dados anuais das 11 ações utilizando violin-boxplot



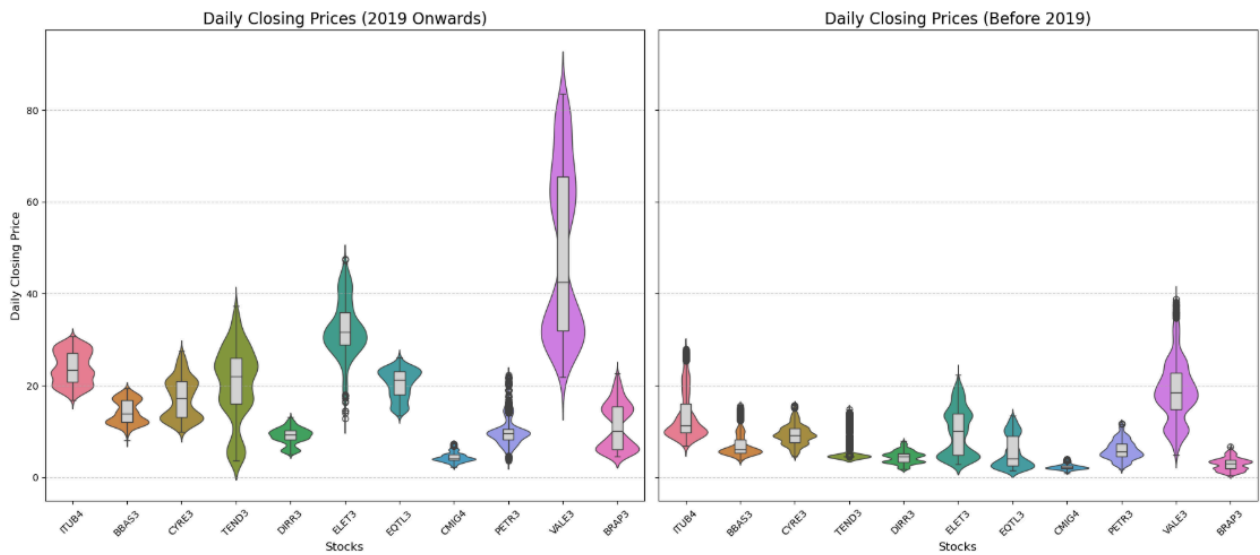
Nas ações CYRE3, TEND3, ELET3 e VALE3 é onde essa mudança de comportamento fica mais evidente. O gráfico violino deixa a análise ainda mais rica, pois fica muito claro em qual região há uma concentração maior dos dados e também a distribuição dos mesmos.

5.2.5 - Comparação dos dados a partir de 2019 com anos anteriores

Os gráficos (Figura 11, 12, 13 e 14) revelam mudanças significativas no comportamento dos preços médios de fechamento das ações antes e após 2019. Após 2019, observa-se maior volatilidade em ativos como VALE3 e PETR3, com distribuições mais amplas e medianas mais elevadas, refletindo o impacto de eventos macroeconômicos, como a pandemia de COVID-19. Antes de 2019, as distribuições são mais concentradas, indicando menor dispersão e maior estabilidade nos preços.

- **Diários:**

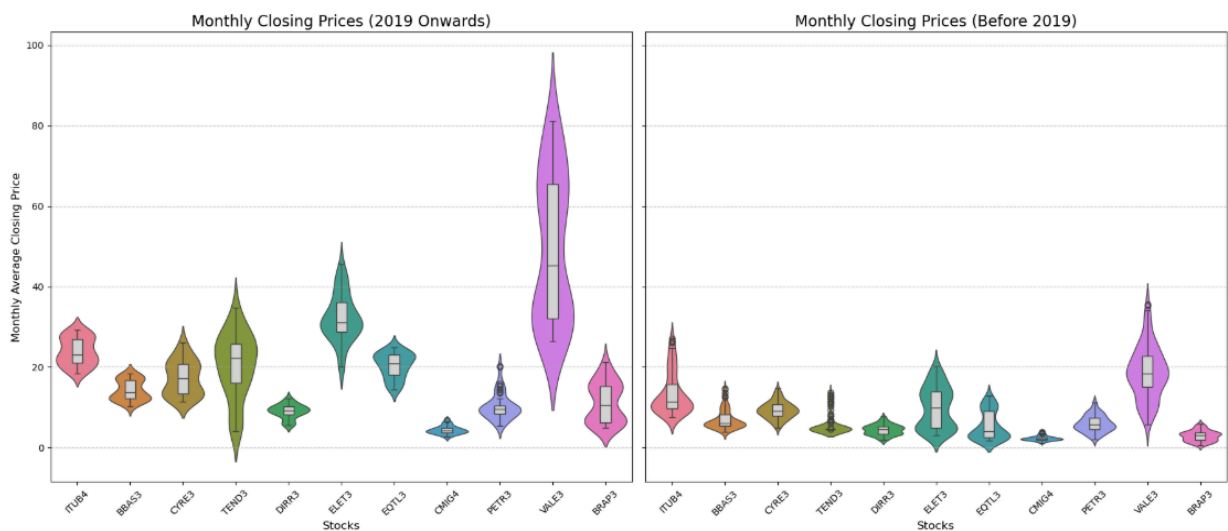
Figura 11: Dados diários das ações antes de 2019 e após 2019



Autoria Própria (2025)

- **Mensais:**

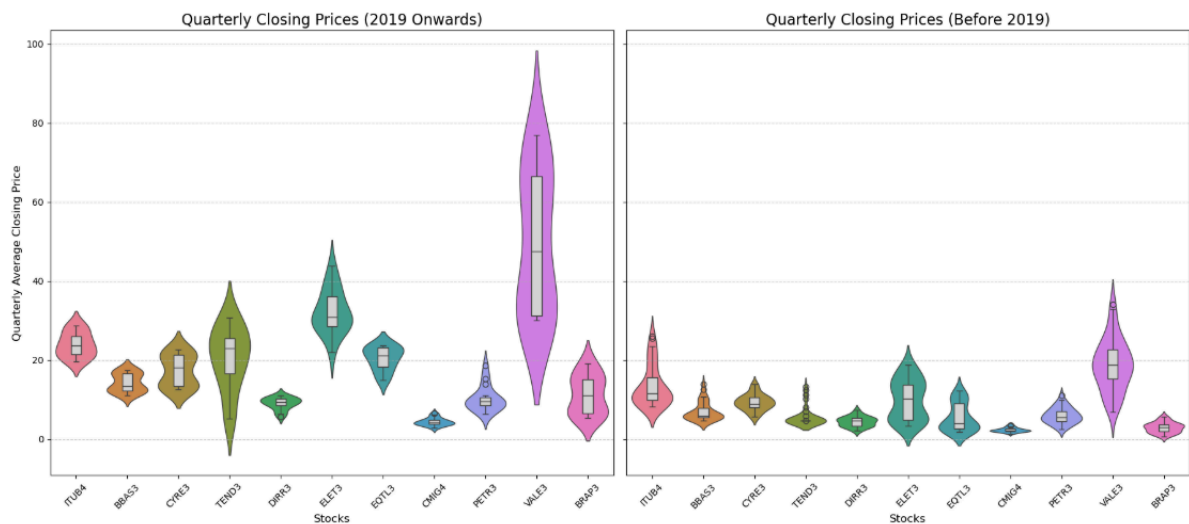
Figura 12: Dados mensais das ações antes de 2019 e após 2019



Autoria Própria (2025)

- **Trimestrais:**

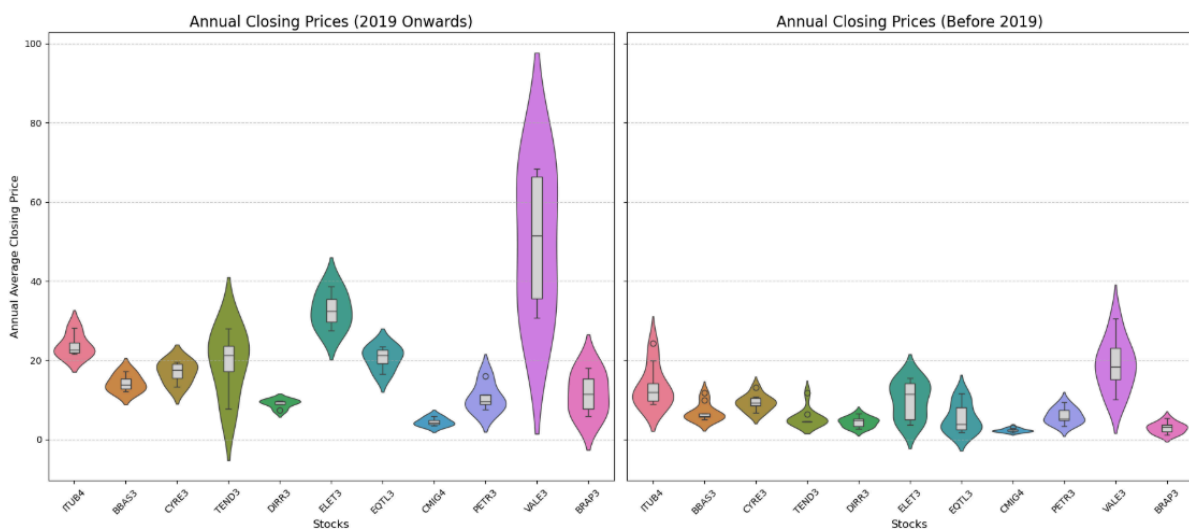
Figura 13: Dados trimestrais das ações antes de 2019 e após 2019



Autoria Própria (2025)

- **Anuais:**

Figura 14: Dados anuais das ações antes de 2019 e após 2019



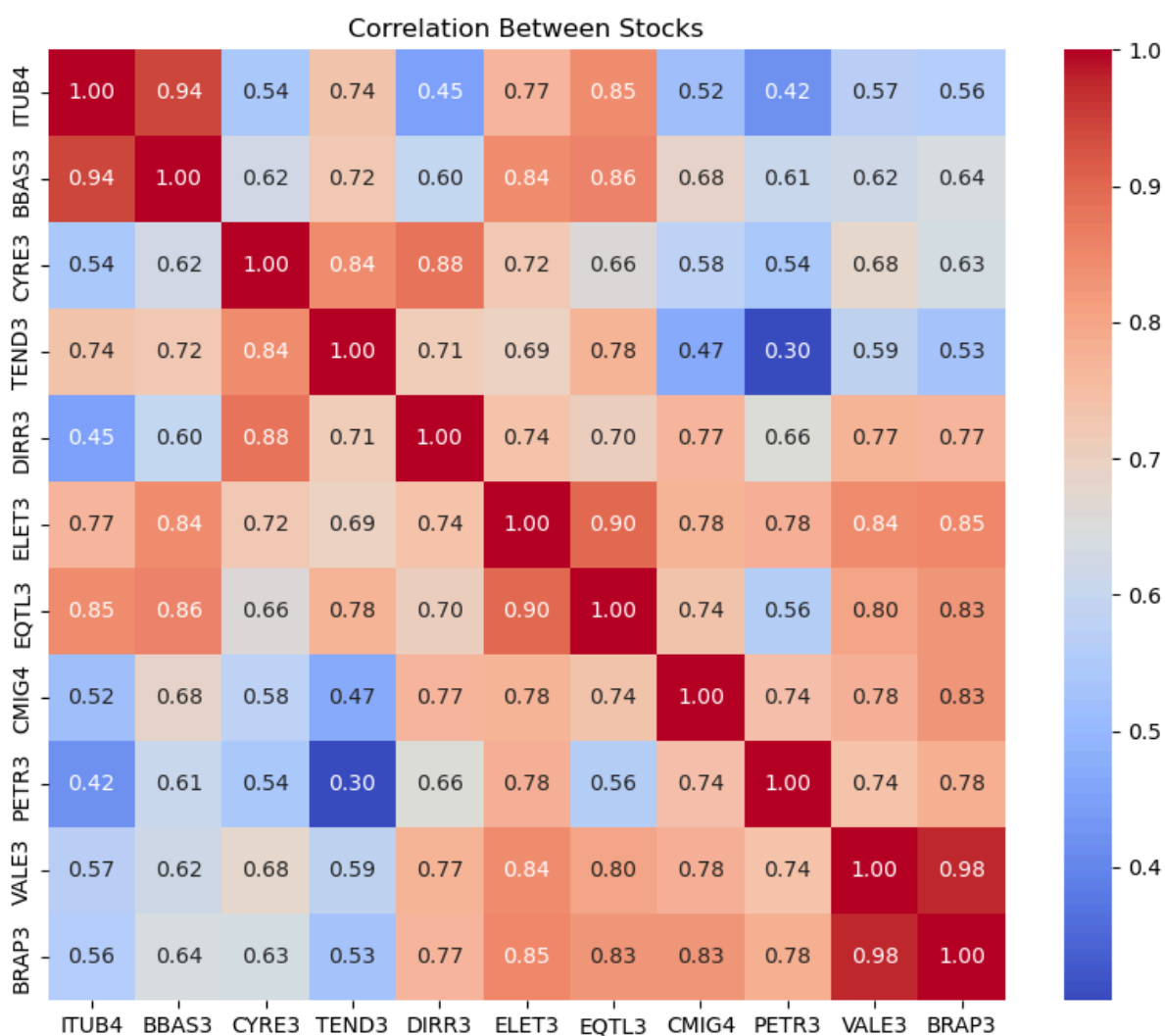
Autoria Própria (2025)

Essa análise evidencia como o comportamento dos ativos é sensível às condições econômicas e setoriais, proporcionando insights valiosos para decisões estratégicas.

5.3 - Correlação dos Dados

A matriz de correlação entre as ações foi calculada para identificar as relações entre os preços de fechamento dos ativos. Abaixo, apresentamos os resultados dessa análise, os quais fornecem informações sobre a força e a direção das correlações entre as diferentes ações.

Figura 15: Matriz de Correlação dos Dados



Autoria Própria (2025)

A análise dos coeficientes de correlação revelou padrões distintos de relacionamento entre os ativos, categorizados da seguinte forma:

5.3.1 - Correlação Forte Positiva

- **BRAP3 e VALE3 (0.98)** apresentam a maior correlação observada, refletindo a forte conexão entre as empresas do setor de mineração, que compartilham fatores macroeconômicos semelhantes.
- **BBAS3 e ITUB4 (0.94)** demonstram correlação significativa, em linha com a expectativa para empresas do setor financeiro, que são influenciadas por condições de mercado similares.

5.3.2 - Correlação Moderada a Alta

- **ELET3 e EQTL3 (0.90)** também exibem alta correlação, consistente com a atuação similar no setor de energia elétrica.
- **CMIG4 e EQTL3 (0.74)** apresentam uma relação moderada, indicando uma conexão no setor elétrico, embora menos intensa do que a observada entre **ELET3 e EQTL3**.
- Empresas como **DIRR3** apresentam uma correlação mais forte com outras ações do setor de construção civil, como **CYRE3 (0.85)**, indicando uma maior sincronia nos movimentos de preços..

5.3.3 - Correlação Baixa a Moderada

- **PETR3** apresenta correlações dispersas, como com **BBAS3 (0.61)** e **VALE3 (0.74)**, possivelmente devido à sua independência setorial (energia versus mineração).
- As ações do Banco do Brasil (**BBAS3**) e da Tenda (**TEND3**) apresentam uma correlação de 0.72, indicando uma relação moderada entre seus preços. Embora pertençam a setores distintos – financeiro e construção civil –, essa correlação pode estar associada a fatores macroeconômicos comuns, como taxas de juros e condições de crédito

5.3.4 - Correlação Baixa

- **ITUB4 e DIRR3 (0.45)** possuem uma das menores correlações, indicando que os fatores que afetam essas ações são bastante distintos.

5.3.5 - Implicações para Diversificação

Os resultados indicam que investidores podem explorar a baixa correlação entre setores, como o financeiro (ITUB4) e a construção civil (DIRR3), para compor carteiras mais

diversificadas e reduzir riscos. Por outro lado, a forte correlação entre empresas do mesmo setor, como VALE3 e BRAP3, pode limitar os benefícios da diversificação, uma vez que esses ativos tendem a se movimentar de forma semelhante.

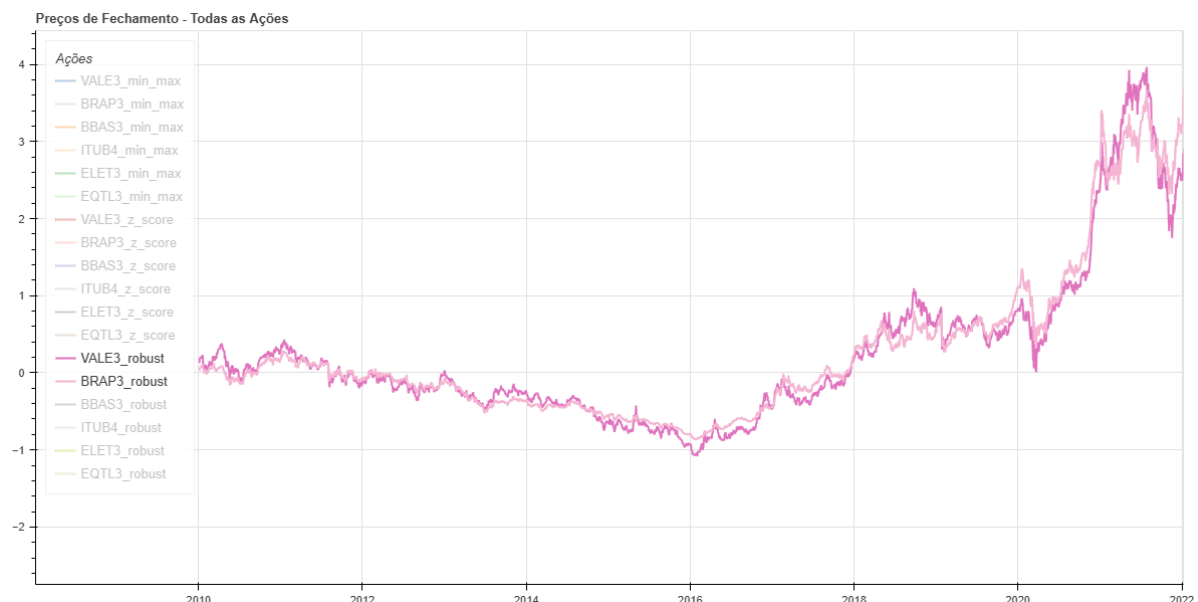
5.3.6 - Setores Fortemente Relacionados

A alta correlação observada em setores como mineração e energia pode ser explicada pela dependência de variáveis macroeconômicas comuns, como o preço de commodities e a regulação governamental. Além disso, a correlação entre PETR3 e VALE3 (0.74) sugere a influência de fatores globais, como taxas de câmbio e crescimento econômico internacional, sobre essas empresas.

5.4 - Análise da Correlação Para Dados Normalizados

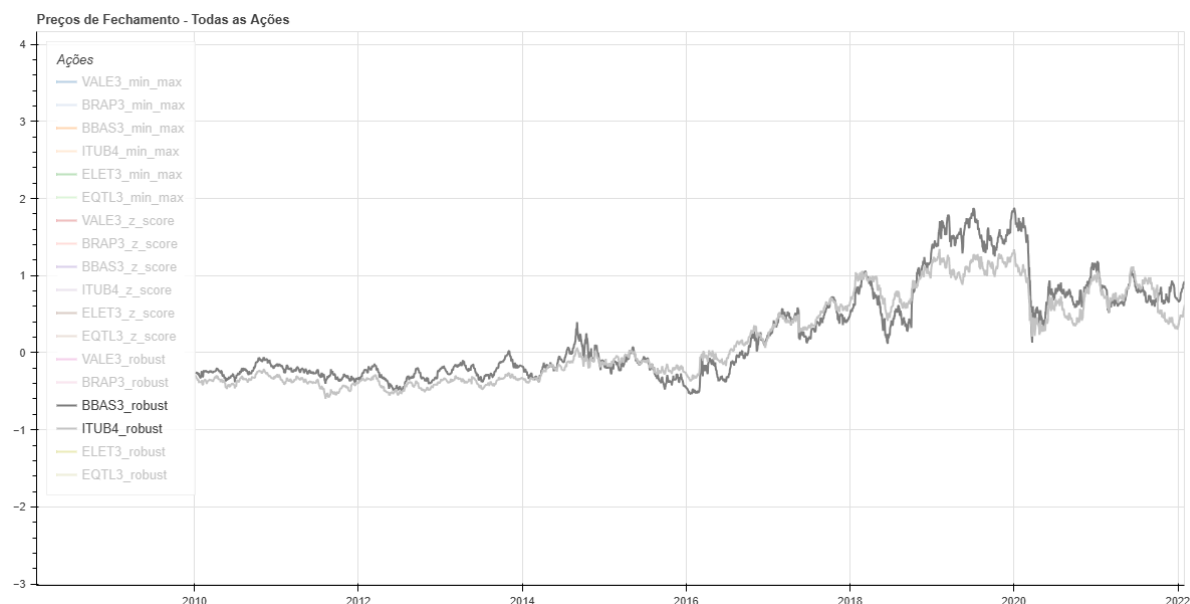
Dado que algumas ações apresentaram alta correlação, aplicamos técnicas de normalização para avaliar se suas tendências ao longo do tempo também eram semelhantes. A seguir, apresentamos os gráficos de linhas de algumas das ações normalizadas para melhor visualização dessas tendências.

Figura 16: Gráfico de Linha da Vale e BRAP



Autoria Própria (2025)

Figura 17: Gráfico de Linha da BBAS3 e ITUB4



Autoria Própria (2025)

Como esperado, as ações com altíssima correlação, VALE3 e BRAP3 (0.98) e ITUB4 e BBAS3 (0.94), mantiveram uma tendência muito similar em seus gráficos de linhas durante o período analisado. Esta similaridade reforça a hipótese de que empresas atuantes no mesmo setor ou que compartilham fatores macroeconômicos comuns tendem a apresentar comportamentos alinhados ao longo do tempo.

A VALE3 e a BRAP3, ambas inseridas no setor de mineração, são influenciadas por variáveis como o preço das commodities, as condições econômicas globais e as políticas regulatórias, resultando em trajetórias de preços que refletem essas condições de forma semelhante. Já as ações ITUB4 e BBAS3, representativas do setor bancário, demonstram que fatores como taxas de juros, políticas monetárias e o desempenho econômico doméstico impactam suas cotações de maneira parecida, contribuindo para a alta correlação observada.

A normalização dos preços permitiu a comparação direta das tendências, evidenciando que, apesar das variações pontuais, as ações correlacionadas tendem a seguir direções semelhantes. Este comportamento sugere que os investidores que buscam

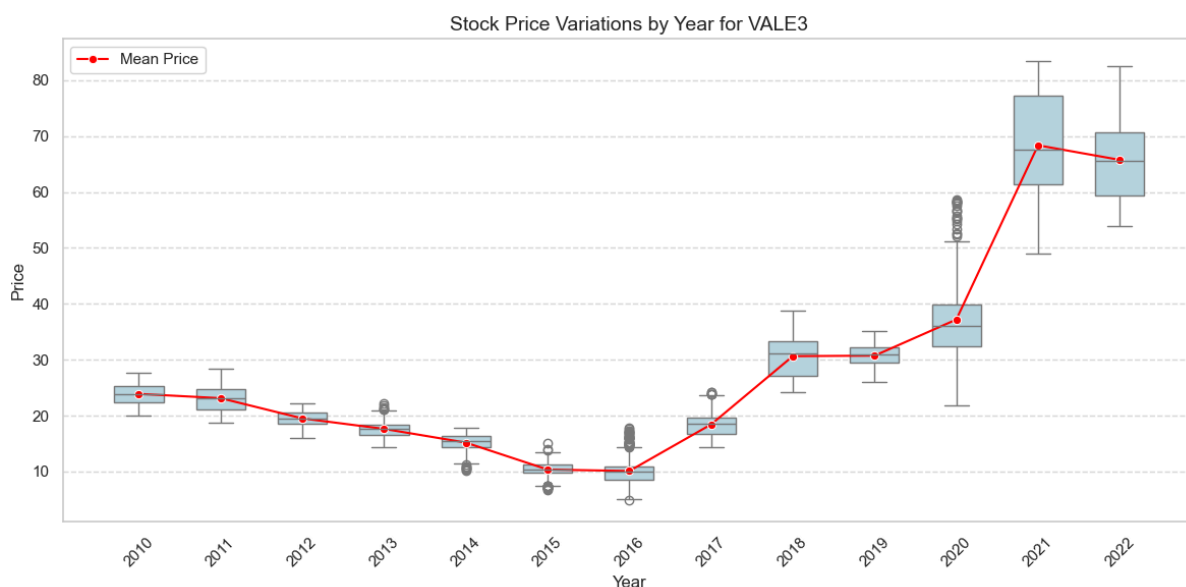
diversificação dentro de um mesmo setor devem estar atentos às limitações impostas pela correlação elevada, uma vez que eventos adversos podem impactar todos os ativos correlacionados simultaneamente.

Além disso, a análise gráfica destacou que, mesmo em períodos de alta volatilidade, as ações correlacionadas apresentaram movimentos convergentes, reforçando a ideia de que a correlação não se limita a períodos de estabilidade, mas também se manifesta em cenários de maior incerteza.

5.5 - Volatilidade dos dados

Os boxplots revelam que algumas ações apresentam uma grande quantidade de outliers e variações nas larguras das caixas. Em particular, VALE3 exibiu um número significativo de outliers, sugerindo alta volatilidade. Para compreender melhor a evolução dessa volatilidade ao longo do tempo, foram analisados boxplots anuais das ações, além da aplicação das Bandas de Bollinger.

Figura 18: Boxplot da VALE3

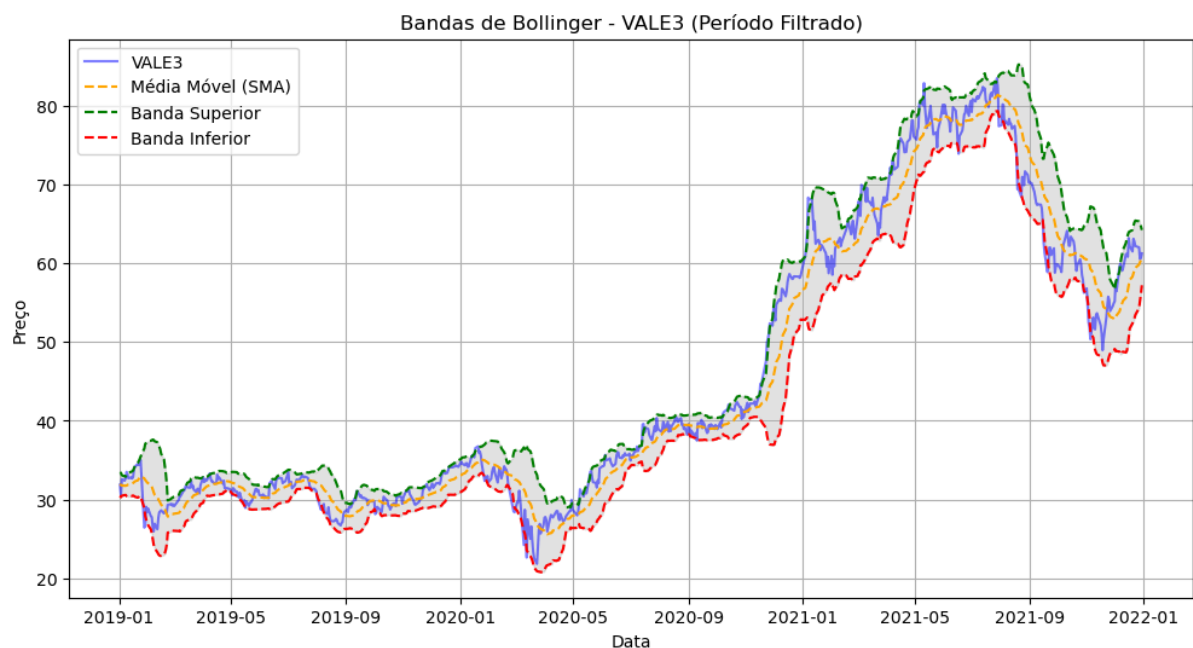


Autoria Própria (2025)

A análise da volatilidade da ação VALE3 ao longo dos anos revela variações significativas em sua dispersão de preços, especialmente a partir de 2016. A evolução dos

boxplots anuais indica um aumento gradual na amplitude dos preços de fechamento, refletindo uma intensificação da volatilidade. Esse efeito se torna particularmente evidente nos anos de 2020 e 2021, que apresentam os maiores intervalos interquartis, possivelmente devido à instabilidade macroeconômica causada pela pandemia de COVID-19. Além disso, observa-se a presença de outliers em determinados períodos, como 2014 e 2016, sugerindo momentos de queda abrupta nos preços, possivelmente decorrentes de choques pontuais no mercado.

Figura 19: Bandas de Bollinger da VALE3

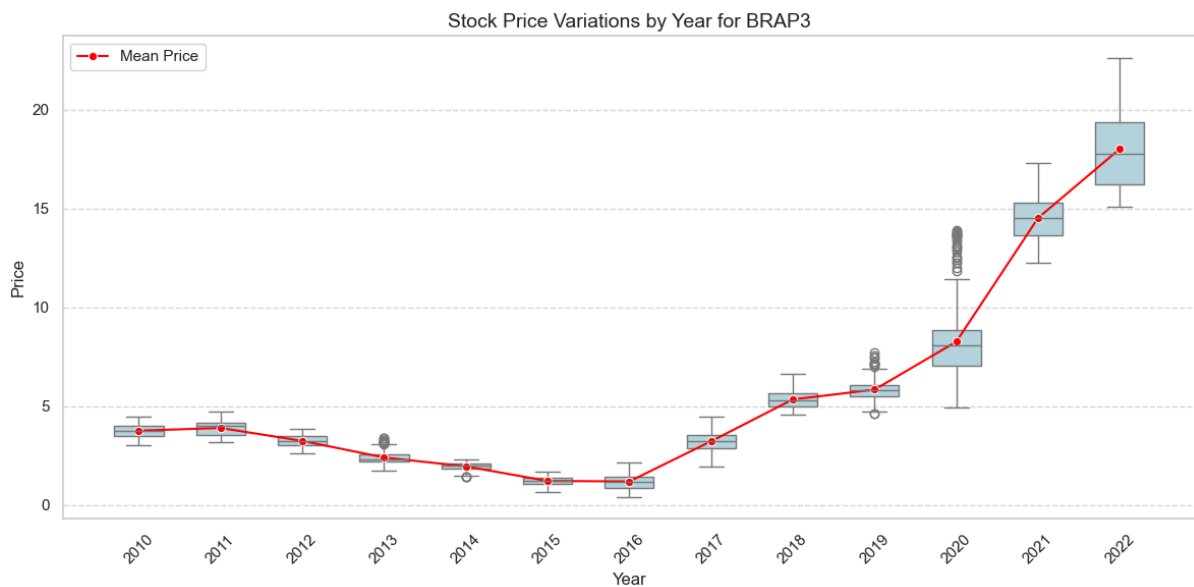


Autoria Própria (2025)

Essa dinâmica de volatilidade também é confirmada pelo comportamento das Bandas de Bollinger (Figura 19), que refletem o alargamento e estreitamento das oscilações ao longo do tempo. Durante os períodos de maior instabilidade, como os picos de 2020 e 2021, as bandas se expandem significativamente, acompanhando as variações expressivas nos preços. Em contraste, em períodos de menor volatilidade, como no início de 2019, as bandas se tornam mais estreitas, sinalizando um comportamento mais estável do ativo. Apesar dessas flutuações, a trajetória geral dos preços de fechamento da VALE3 apresenta uma tendência de crescimento contínuo desde 2016, o que sugere um fortalecimento da confiança do mercado na valorização da empresa, mesmo diante de períodos de intensa oscilação.

É relevante destacar que a BRAP3 apresenta um comportamento muito semelhante ao da VALE3 em relação à volatilidade e dispersão dos preços de fechamento, o que pode ser atribuído à alta correlação entre as duas ações (0.98). Essa forte relação se reflete nos padrões observados ao longo dos anos, conforme ilustrado no gráfico de boxplot a seguir.

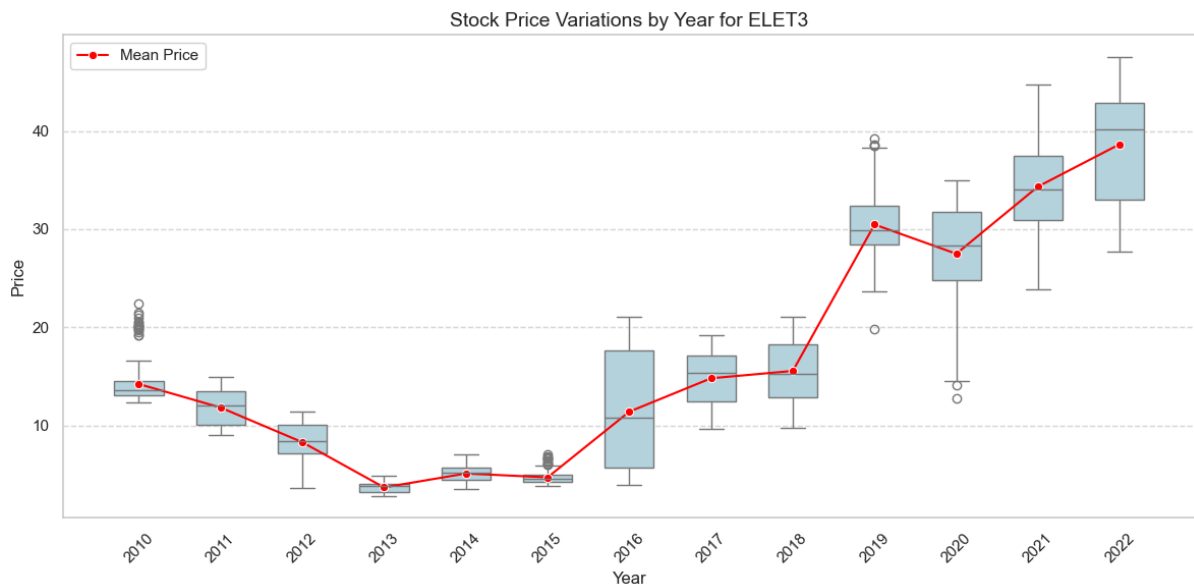
Figura 20: Boxplot da BRAP3



Autoria Própria (2025)

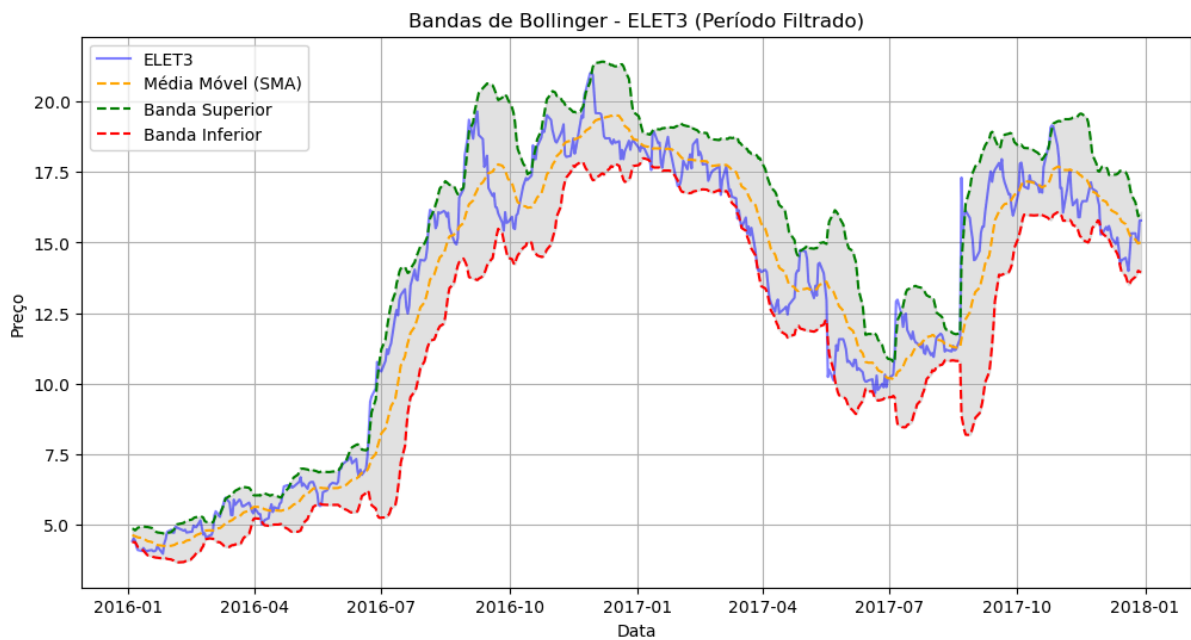
Outro ativo que demonstra volatilidade é a ELET3. Seus gráficos a seguir ilustram essa característica, evidenciando variações significativas nos preços de fechamento ao longo do tempo.

Figura 21: Boxplot da ELET3



Autoria Própria (2025)

Figura 22: Bandas de Bollinger da ELET3



Autoria Própria (2025)

Durante o período analisado, a ação ELET3 apresentou variações significativas em sua volatilidade, refletidas tanto nos boxplots anuais quanto nas Bandas de Bollinger. Em 2016, houve um aumento expressivo na dispersão dos preços de fechamento, evidenciado

pelo alargamento do boxplot. Esse comportamento sugere um período de instabilidade acentuada, possivelmente influenciado por fatores econômicos e políticos que impactaram o setor de energia. O aumento da volatilidade nesse ano pode estar associado a incertezas no mercado ou a um processo de recuperação nos preços da ação.

Entre 2017 e 2019, observa-se uma estabilização, com uma redução na largura dos boxplots e menor dispersão dos preços. Esse período indica uma fase de consolidação, na qual os preços seguiram uma trajetória mais previsível, sem grandes oscilações. A estabilidade relativa sugere que, após as flutuações bruscas de 2016, o mercado ajustou-se a um novo patamar de precificação.

A partir de 2020, a volatilidade volta a crescer, evidenciada pelo aumento no intervalo interquartil dos boxplots e pela presença de um número maior de outliers. Esse comportamento está alinhado com o impacto da pandemia de COVID-19 no mercado financeiro, que trouxe instabilidade e flutuações acentuadas nos preços. O período pós-pandemia também mostra um processo de recuperação, refletido no aumento da dispersão dos preços nos anos seguintes.

A análise das Bandas de Bollinger reforça esses achados. No início de 2016, as bandas eram estreitas, sugerindo baixa volatilidade. No entanto, ao longo do ano, expandiram-se significativamente, acompanhando o aumento da oscilação nos preços. Durante esse movimento de valorização, os preços frequentemente atingiam ou ultrapassavam a banda superior, indicando uma forte tendência de alta e possíveis condições de sobrecompra. Esse padrão é característico de momentos de forte impulso no mercado, nos quais a volatilidade elevada está associada a mudanças expressivas na percepção dos investidores sobre o ativo.

6. CONCLUSÃO

A análise exploratória e o pré-processamento dos dados foram fundamentais para compreender o comportamento das séries temporais dos preços das ações da B3. Através do estudo de Skewness e Kurtosis, os ativos foram divididos em três grupos distintos, permitindo uma caracterização detalhada de sua estabilidade, volatilidade e adequação a diferentes perfis de investidores. O Grupo 1 representou ativos mais previsíveis e estáveis, ideais para investidores conservadores. O Grupo 2 apresentou ativos com volatilidade intermediária, oferecendo um equilíbrio entre risco e retorno. Já o Grupo 3 agrupou ativos

altamente voláteis, indicados para investidores agressivos que buscam altos retornos, apesar do risco significativo.

A análise combinada de boxplots e violin plots revelou padrões importantes na distribuição dos preços das ações ao longo do tempo. O boxplot destacou quartis, dispersão e a presença de outliers, enquanto o violin plot permitiu a visualização da densidade dos preços, revelando concentrações de valores e possíveis distribuições multimodais. Essa abordagem possibilitou uma visão mais ampla da volatilidade e dos regimes de mercado, sendo essencial para identificar períodos de instabilidade nos preços.

A segmentação do dataset em períodos antes e depois de 2019 demonstrou o impacto da pandemia de COVID-19 nos preços dos ativos financeiros. Notou-se um aumento significativo na volatilidade de ações como VALE3 e PETR3, evidenciando que eventos macroeconômicos afetam diretamente a dinâmica do mercado financeiro. Antes de 2019, as distribuições eram mais concentradas e estáveis, enquanto após 2019, houve maior dispersão e mudanças nas medianas, refletindo as incertezas do período pós-pandemia.

A matriz de correlação entre os ativos revelou relações importantes entre diferentes setores da economia. Empresas do mesmo segmento, como BRAP3 e VALE3, apresentaram alta correlação, evidenciando que compartilham fatores macroeconômicos comuns. No setor financeiro, BBAS3 e ITUB4 também mostraram forte correlação, refletindo a influência de variáveis econômicas similares. Em contrapartida, setores mais independentes, como construção civil (DIRR3) e energia (PETR3), demonstraram correlações mais fracas, sugerindo que podem ser usados para diversificação de portfólios.

A análise da correlação para dados normalizados permitiu uma melhor comparação das tendências ao longo do tempo, destacando a semelhança de movimentos entre ativos fortemente correlacionados. Essa etapa do estudo reforça a importância de utilizar métricas estatísticas não apenas para avaliar o comportamento dos preços, mas também para auxiliar na construção de estratégias de diversificação de investimentos.

Por fim, a exploração e o pré-processamento dos dados provaram ser etapas cruciais para garantir a qualidade das informações utilizadas na modelagem preditiva. A análise detalhada das séries temporais revelou padrões valiosos e permitiu uma segmentação eficiente dos ativos, fornecendo insights que podem ser aplicados tanto em estratégias de investimento quanto no desenvolvimento de modelos de previsão mais robustos. Esse estudo

reforça que uma abordagem exploratória bem conduzida é essencial para qualquer análise quantitativa no mercado financeiro, permitindo uma melhor tomada de decisões e mitigação de riscos.

7. REFERÊNCIAS

1. BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. Time series analysis: forecasting and control. 5. ed. Hoboken: John Wiley & Sons, 2015.
2. CHOLLET, F. Deep learning with Python. Shelter Island: Manning Publications, 2021.
3. GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. Cambridge: MIT Press, 2016.
4. HAMILTON, J. D. Time series analysis. Princeton: Princeton University Press, 1994.
5. HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3. ed. Amsterdam: Elsevier, 2011.
6. HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning. 2. ed. New York: Springer, 2009.
7. HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: principles and practice. 2. ed. Melbourne: OTexts, 2018.
8. SHUMWAY, R. H.; STOFFER, D. S. Time series analysis and its applications. 4. ed. New York: Springer, 2017.
9. TUKEY, J. W. Exploratory data analysis. Reading: Addison-Wesley, 1977.
10. ENGLE, R. F. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, Princeton, v. 50, n. 4, p. 987-1007, 1982.
11. LIU, F. T.; TING, K. M.; ZHOU, Z. H. Isolation Forest. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM), 8., 2008. Anais [...]. Piscataway: IEEE, 2008.
12. OLIVEIRA, Bruno. Boxplot: como interpretar?. **Statplace**, 2019. Disponível em: [Boxplot: Você sabe como interpretar esse tipo de gráfico?](#). Acesso em: 31 jan. 2025.
13. HUNTER, J. D. *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007. DOI: 10.1109/MCSE.2007.55.
14. ASSIS, C. A. S. de; CARRANO, E. G.; PEREIRA, A. C. M. Predição de Tendências em Séries Financeiras Utilizando Metaclassificadores. **Economia Aplicada**, v. 24, n. 1, p. 29-78, 2020. DOI: <http://dx.doi.org/10.11606/1980-5330/ea148159>.
15. VARGAS, Gabriel M.; SILVESTRE, Leonardo J.; RIGO JR., Luís O.; ROCHA, Helder R. O. B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis. *IEEE Latin America Transactions*, v. 20, n. 7, p. 1067–1074, jul. 2022. Disponível em: <https://latamt.ieee9.org/index.php/transactions/article/view/6236>. Acesso em: 31/01/2025.
16. NASCIMENTO, S. N.; SANTOS, F. G.; FERREIRA, K.H.A. Previsão de Preços de Ações Utilizando Inteligência Artificial. Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Piauí, 2022. Disponível em: <https://sol.sbc.org.br/index.php/bwaif/article/download/20485/20313/#:~:text=O%20modelo%20de%20rede%20neural,%2C%20usando%20o%20software%20WEKA..> Acesso em: 31/01/2025.