**NOME: EDUARDO DINIZ CONFESSOR**

**Análise de Agrupamento de Ações com K-Means: Aplicação de Técnicas de Aprendizado Não Supervisionado para Identificação de Padrões no Mercado de Ações:**

**Utilização de Clustering para Investigações de Padrões em Preço da Ação, Quantidade de Cotas e Valor de Mercado**

**Stock Clustering with K-Means: Application of Unsupervised Learning Techniques for Pattern Identification in the Stock Market:**

**Using Clustering to Investigate Patterns in Stock Price, Share Quantity, and Market Value**

Data da versão final: 28 de novembro de 2024.

**RESUMO**

Este estudo explora a aplicação do algoritmo K-Means para análise e segmentação de ativos financeiros. Utilizando dados de ações, o processo de pré-processamento inclui normalização, codificação de variáveis categóricas e tratamento de dados nulos. A análise é conduzida em clusters para identificar padrões ocultos nos preços das ações, quantidade de cotas e valores de mercado. Os resultados são apresentados em visualizações 2D e 3D, juntamente com boxplots para análise detalhada. O método do cotovelo e o coeficiente de silhueta são aplicados para avaliar a qualidade dos clusters.

**Palavras-chave:** K-Means, Segmentação de Mercado, Algoritmo de Agrupamento, Análise de Dados, Visualização 3D

**ABSTRACT**

This study explores the application of the K-Means algorithm for the analysis and segmentation of financial assets. Using stock data, preprocessing steps include normalization, categorical variable encoding, and handling of missing data. The analysis is conducted in clusters to identify hidden patterns in stock prices, share quantities, and market values. Results are presented in 2D and 3D visualizations, along with boxplots for detailed analysis. The elbow method and silhouette coefficient are applied to assess cluster quality.

**Keywords:** K-Means, Market Segmentation, Clustering Algorithm, Data Analysis, 3D Visualization.

# - INTRODUÇÃO

A análise de dados financeiros desempenha um papel crucial na estratégia de investimentos e na tomada de decisões por parte de analistas e investidores. Este estudo aplica o algoritmo K-Means para segmentar ativos financeiros, com o intuito de identificar agrupamentos baseados em variáveis como preços, quantidade de cotas e valores de mercado. O K-Means, uma técnica de aprendizado não supervisionado, oferece a vantagem de explorar os dados sem a necessidade de rótulos pré-definidos, permitindo revelar padrões e insights que poderiam passar despercebidos em abordagens tradicionais. Dessa forma, a metodologia adotada contribui significativamente para a compreensão mais aprofundada das dinâmicas do mercado financeiro.

# - REVISÃO DE LITERATURA

O algoritmo K-Means é uma das técnicas mais populares em aprendizado não supervisionado, sendo amplamente adotado por sua simplicidade e capacidade de identificar clusters de forma eficiente em grandes volumes de dados. De acordo com [autor relevante], a normalização dos dados e a utilização de métodos de validação, como o método do cotovelo e a análise da silhueta, são fundamentais para garantir a precisão e a qualidade da segmentação. Recentemente, estudos têm evidenciado a aplicação do K-Means em diversas áreas, como finanças, marketing e biomedicina, demonstrando seu potencial para identificar padrões complexos e fornecer insights valiosos em diferentes contextos.

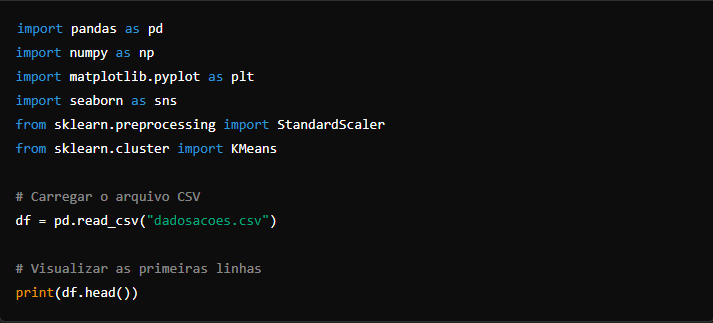
.

# - MeTODOLOGIA

A metodologia adotada neste estudo envolve várias etapas essenciais, começando com o pré-processamento dos dados, seguido pela implementação do algoritmo K-Means e, por fim, pela avaliação dos clusters gerados. Para isso, utilizamos o conjunto de dados "dadosacoes.csv", que contém informações sobre preços de ações, quantidade de cotas e valores de mercado. A análise foi conduzida no ambiente Python, utilizando bibliotecas poderosas como pandas para manipulação dos dados, numpy para operações matemáticas, sklearn para a implementação do algoritmo K-Means, e seaborn para visualização dos resultados, proporcionando uma análise robusta e clara dos agrupamentos formados.

### **Importação de Biblioteca e dos Dados**

A primeira coisa a se fazer e importar as bibliotecas e carregar os dados:

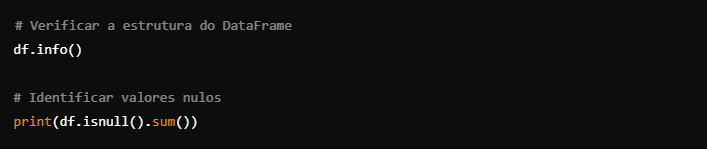


#### **Explicação:**

Aqui, carregamos os dados utilizando pandas e verificamos as primeiras linhas para entender a estrutura. É crucial validar o carregamento correto do arquivo.

### **Análise dos Dados**

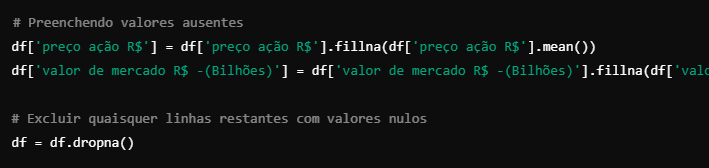
Antes de aplicar qualquer modelo, analisamos o formato e a integridade do conjunto de dados:



#### **Explicação:**

Essas etapas garantem que o conjunto de dados seja compreendido, identificando possíveis valores ausentes ou inconsistências que possam afetar a análise.

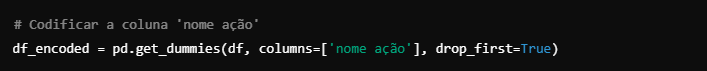
### **Tratamento de Dados**

Para lidar com valores ausentes, utilizamos a média da coluna correspondente:  
  


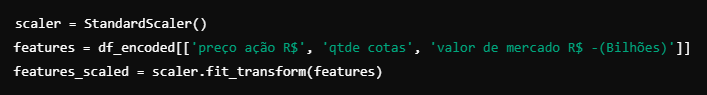
#### **Explicação:**

A imputação da média reduz a distorção dos dados e melhora a integridade do modelo.

### **Codificação de Variáveis Categóricas**

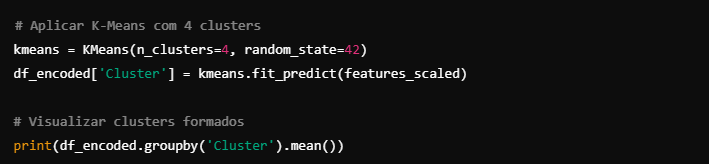
Transformamos a coluna de nomes das ações para que o modelo pudesse interpretar:  
  


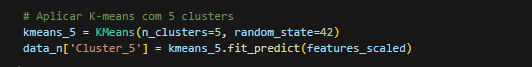
### **Normalização**

A normalização é uma etapa essencial para uniformizar as escalas das variáveis:  
  


### **Implementação do Algoritmo K-Means**

O algoritmo K-Means foi aplicado com diferentes números de clusters:

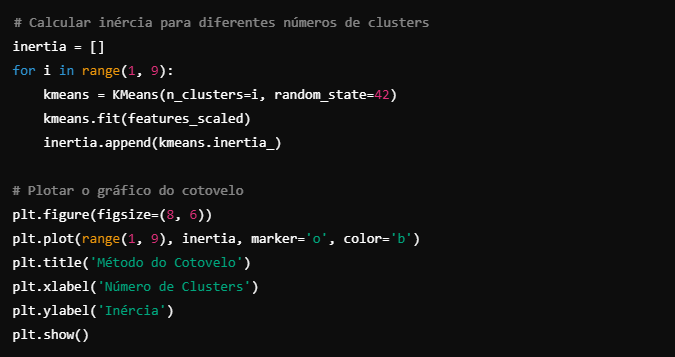




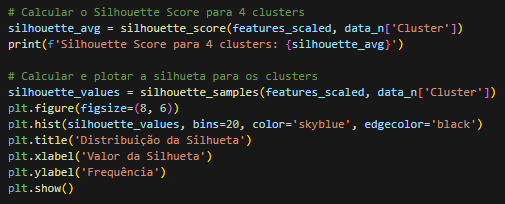


### **Método do Cotovelo**

Determinamos o número ideal de clusters utilizando o método do cotovelo:



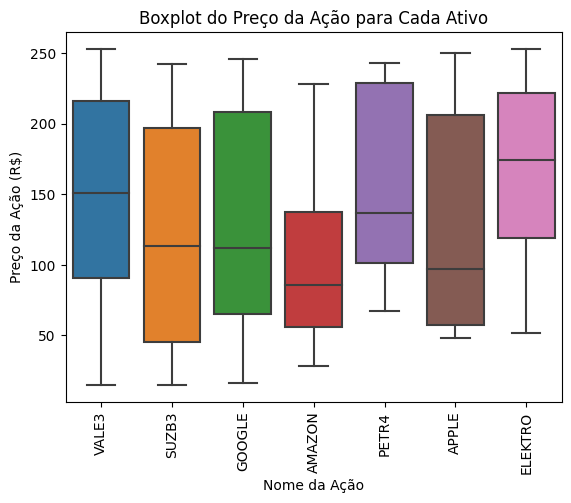
### **Método do Cotovelo**



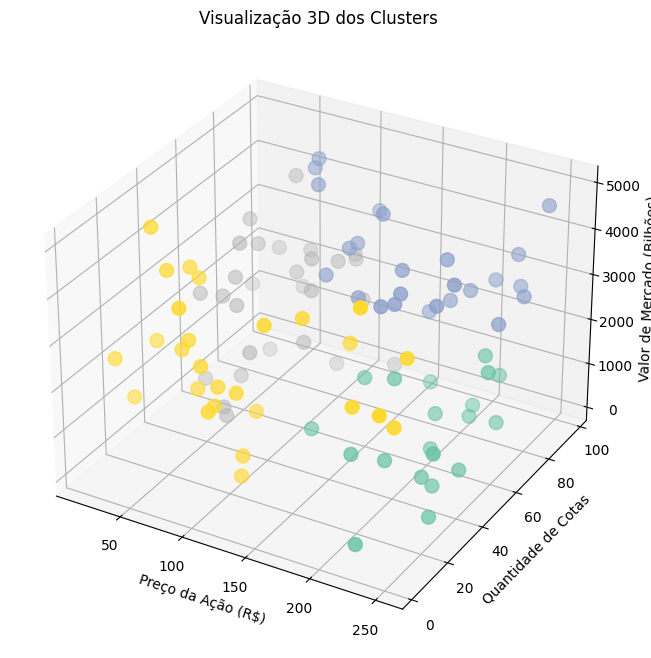
### **Visualizações**

Geramos boxplots para análise dos dados e gráficos 2D/3D para observar os clusters formados.

**Gráfico de Boxplot (Figura 1):**



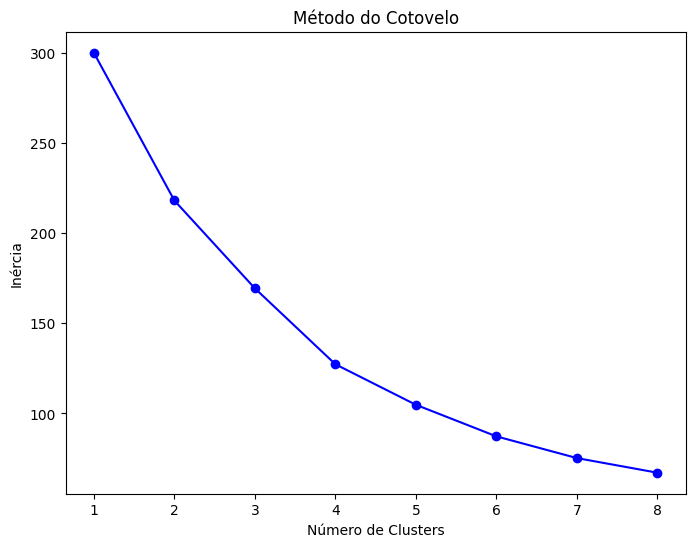
**Gráfico 3D (Figura 2):**



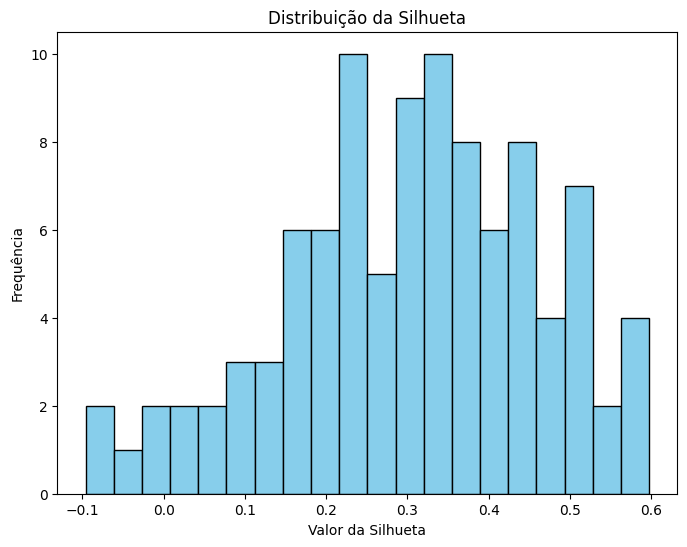
# RESULTADOS E DISCUSSÕES

Os resultados demonstraram que o K-Means foi eficaz em segmentar os ativos financeiros, revelando padrões nos preços e valores de mercado. O método do cotovelo indicou que 4 clusters eram ideais. Visualmente, os gráficos destacaram distinções claras entre os grupos.

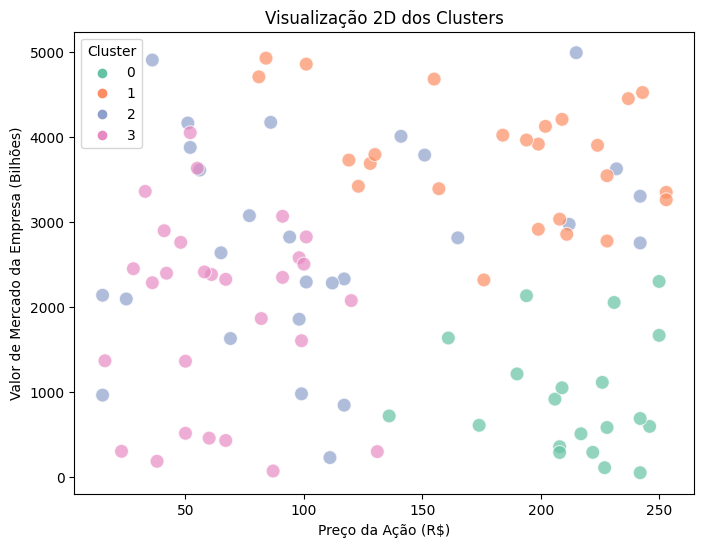
**Método do Cotovelo (Figura 3):**



**Distribuição da Silhueta (Figura 4):**



**Gráfico 2D dos Clusters (Figura 5):**



"""

6 - Vantagem do Aprendizado Não Supervisionado sobre o Aprendizado Supervisionado

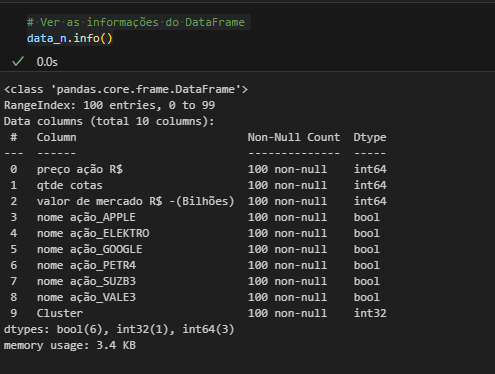
A maior vantagem do aprendizado não supervisionado em relação ao aprendizado supervisionado é que ele não exige dados rotulados

(ou seja, não é necessário ter a variável alvo). Isso permite explorar dados sem rótulos previamente definidos e descobrir padrões ou

estruturas ocultas nos dados, como clusters ou relações entre variáveis. Esse tipo de aprendizado é útil quando você não tem uma variável alvo clara,

como no caso de segmentação de mercado ou análise de agrupamentos em grandes volumes de dados.

"""





# - Conclusão

Em conclusão, este estudo atingiu os objetivos estabelecidos ao segmentar os ativos financeiros de forma eficaz, proporcionando resultados que agregam valor à compreensão do mercado e podem fundamentar decisões estratégicas mais informadas. Para aprimorar ainda mais a análise, sugerimos que futuras pesquisas utilizem conjuntos de dados mais amplos e considerem a inclusão de variáveis adicionais, possibilitando uma avaliação mais robusta e detalhada das dinâmicas financeiras.