TÍTULO DO ARTIGO: Análise Exploratória de Investimentos com Algoritmo K-means

TÍTULO DO ARTIGO EM INGLÊS: Exploratory Analysis of Investment using the K-means Algorithm

Nome do aluno: Leonardo Rodrigues de Oliveira
Nome do orientador: Daniel Filipe Vieira

Data da versão final: 02 de Dezembro de 2024.

RESUMO

Este artigo explora a aplicação de técnicas de aprendizado não supervisionado para agrupar ações de investimentos com características semelhantes, utilizando o algoritmo K-means. O conjunto de dados inclui informações sobre o nome da ação, preço, quantidade de cotas e valor de mercado da empresa. A análise passa pela importação e pré-processamento dos dados, seguida pela aplicação do K-means com diferentes números de clusters, análise visual dos resultados e exploração de características dos dados como preço e valor de mercado. A metodologia empregada oferece uma visão mais profunda sobre o comportamento e agrupamento de ativos, contribuindo para a tomada de decisões estratégicas no mercado de investimentos.

Palavras-chave: Aprendizado não supervisionado, K-means, agrupamento de ações, análise exploratória, Big Data, investimentos.

ABSTRACT

Keywords: x Unsupervised learning, K-means, stock clustering, exploratory analysis, Big Data, investments.

1 INTRODUÇÃO

O aprendizado de máquina, particularmente o aprendizado não supervisionado, tem ganhado destaque na análise de dados complexos, como aqueles relacionados ao mercado financeiro. A análise de dados de ações de investimentos permite identificar padrões, agrupamentos e comportamentos similares entre ativos financeiros, que podem orientar decisões de investimento mais informadas. Neste contexto, o algoritmo K-means é uma técnica amplamente utilizada para segmentar dados em grupos, sem a necessidade de rótulos ou supervisão explícita. Este estudo aplica o K-means para agrupar ações de uma companhia de investimentos, com base em características como preço da ação, quantidade de cotas e valor de mercado da empresa.

2 REVISÃO DE LITERATURA

O aprendizado não supervisionado é uma abordagem poderosa em que o modelo é treinado com dados sem rótulos, com o objetivo de descobrir estruturas e padrões ocultos. O algoritmo K-means é um dos métodos mais populares para agrupamento, dividindo dados em clusters baseados na proximidade das características. Estudos anteriores demonstraram sua eficácia na análise de grandes volumes de dados financeiros, como na segmentação de ações e na análise de risco de investimentos. Além disso, técnicas como a análise de silhueta e o gráfico do cotovelo são frequentemente usadas para avaliar a qualidade dos clusters formados, oferecendo uma métrica visual para ajudar na escolha do número adequado de clusters.

3 METODOLOGIA

A metodologia adotada envolve a análise de um conjunto de dados financeiros, com informações sobre as ações de uma companhia de investimentos. O processo de análise segue as seguintes etapas:

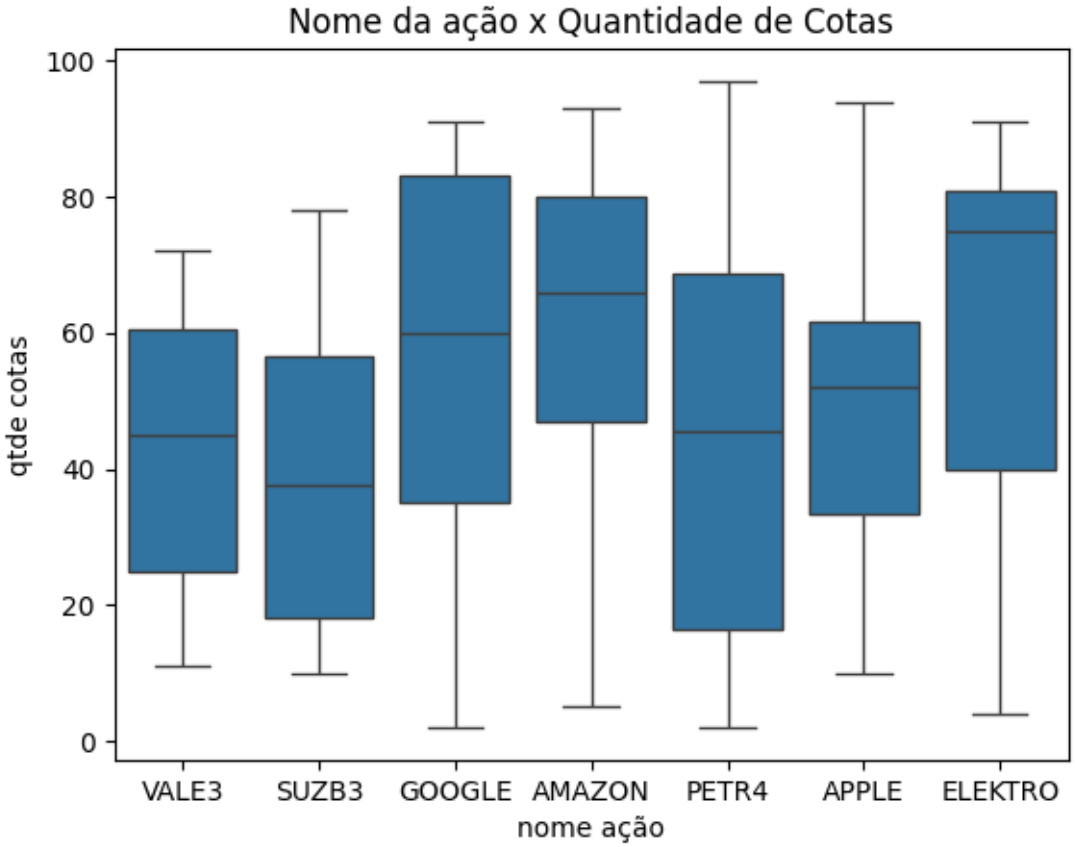
1. **Preparação e Importação dos Dados:** Os dados incluem o nome da ação, preço da ação, quantidade de cotas e valor de mercado da empresa.
2. **Pré-processamento:** Converte-se variáveis categóricas utilizando a função `pd.get_dummies` do pandas, tratando também valores ausentes, se necessário.
3. **Aplicação do K-means:** O algoritmo é aplicado para formar clusters com diferentes números de grupos (n=4, n=5, n=8). Após a aplicação, são realizados testes visuais utilizando gráficos de dispersão 2D e 3D.
4. **Exploração e Visualização:** Gráficos como boxplots são gerados para observar a distribuição de variáveis como o preço da ação e o valor de mercado. Também são utilizados os métodos `df.info()` e `df.describe()` para entender melhor o conjunto de dados.
5. **Avaliação de Qualidade dos Clusters:** O gráfico do cotovelo e o gráfico de silhueta são usados para determinar o número ideal de clusters e avaliar a coesão interna dos agrupamentos.

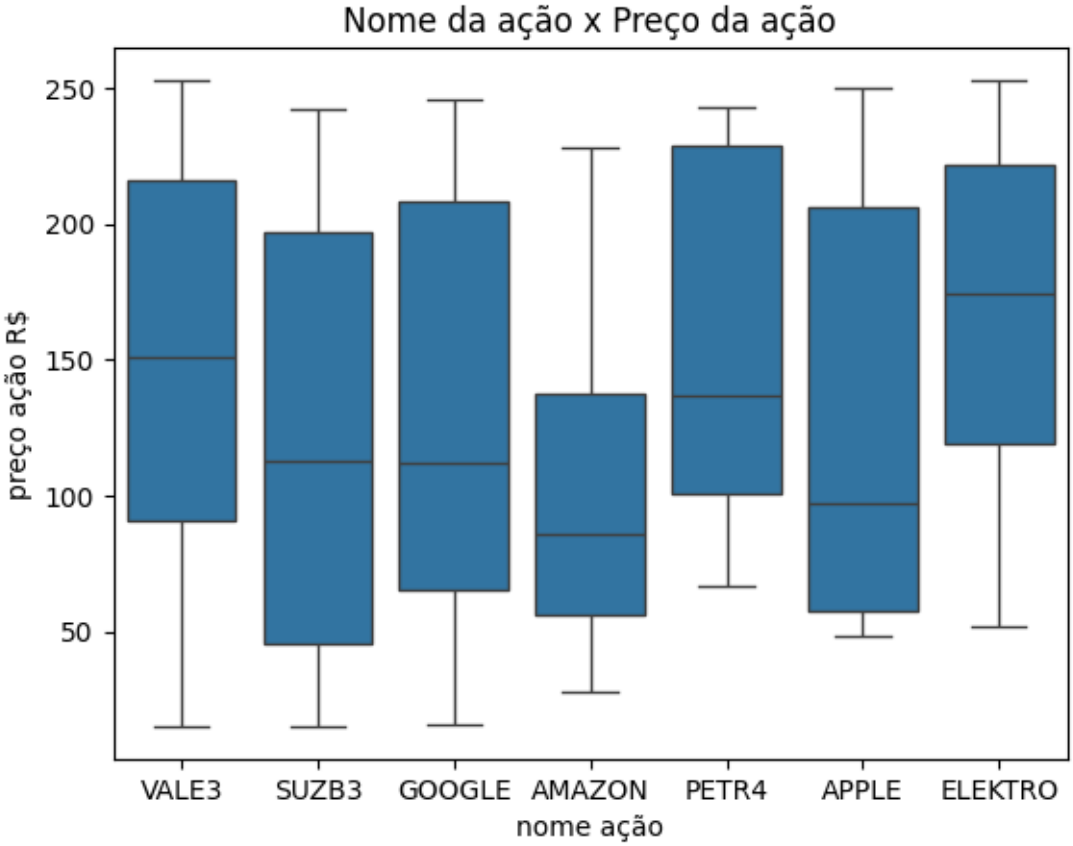
4 RESULTADOS E DISCUSSÕES

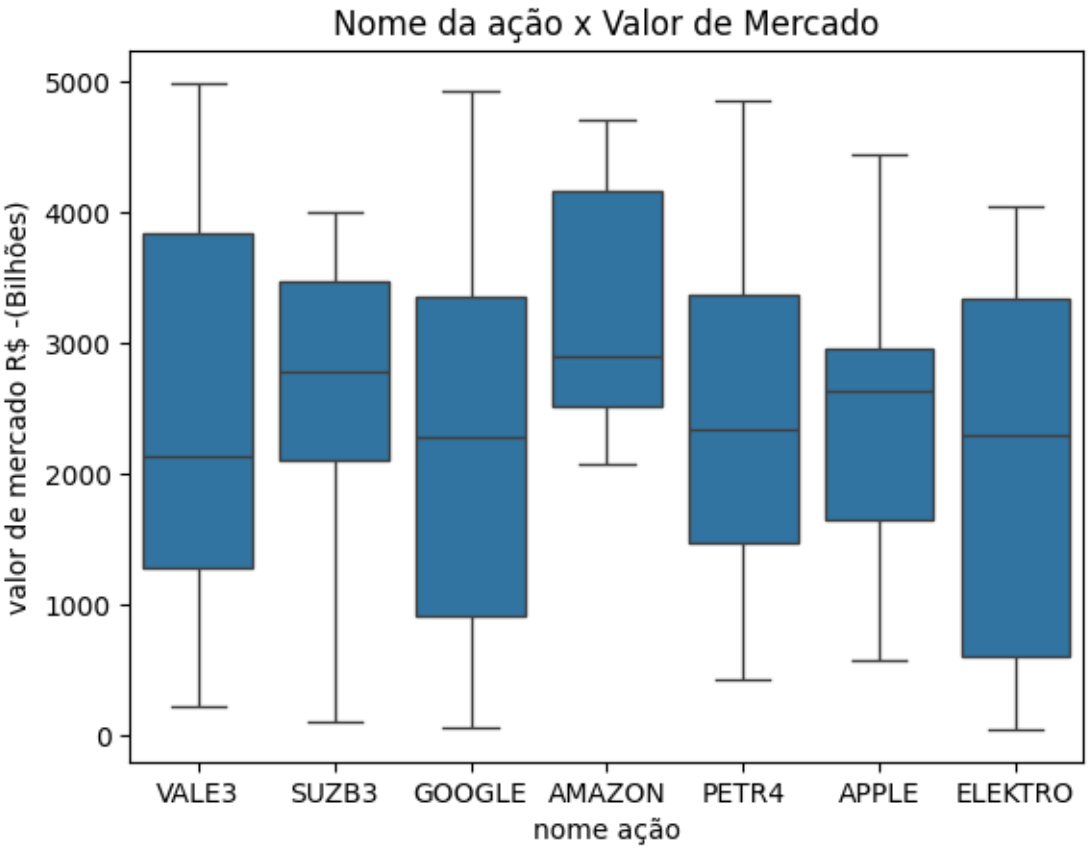
Os resultados obtidos após a aplicação do algoritmo K-means mostraram a formação de diferentes grupos de ações, dependendo do número de clusters definidos. Ao variar o número de clusters (4, 5 e 8), observou-se que, com o aumento de clusters, os grupos tendem a se tornar mais específicos, mas com menos coesão. A visualização dos clusters em 2D e 3D permitiu uma análise mais clara dos agrupamentos formados e ajudou a entender como as características das ações influenciam no agrupamento. O gráfico do cotovelo indicou que 4 clusters poderiam ser uma escolha razoável, enquanto o gráfico de silhueta forneceu uma indicação adicional sobre a qualidade do agrupamento.

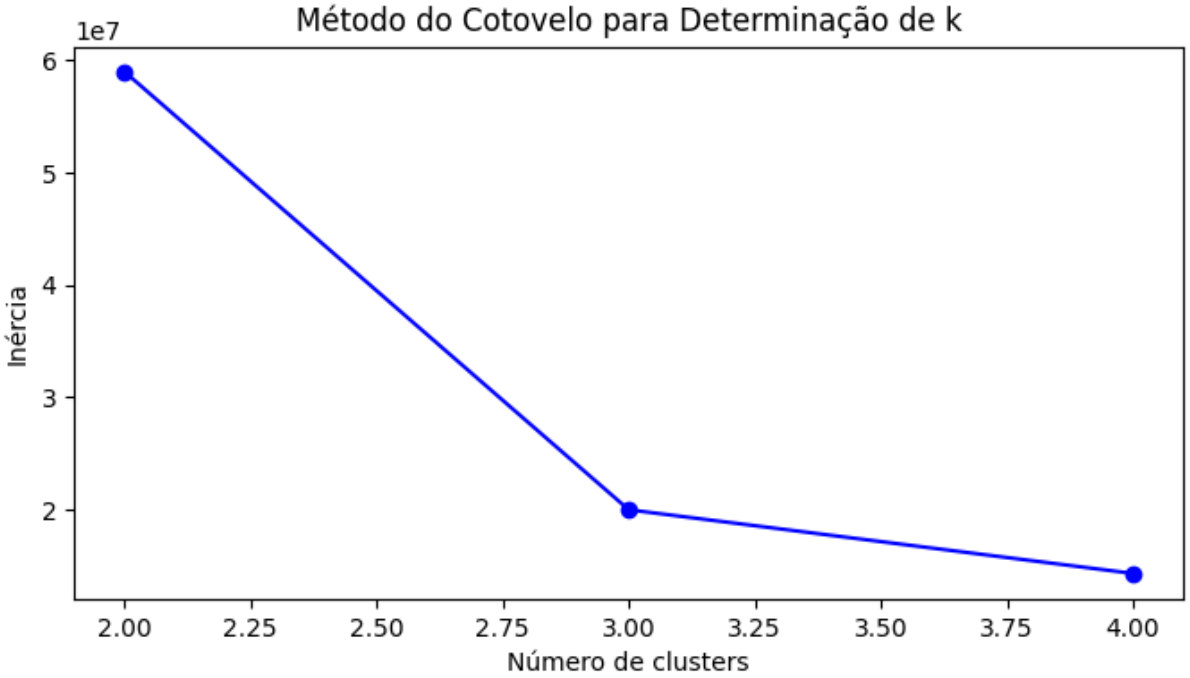
Além disso, a análise de variáveis, como o preço da ação e o valor de mercado, revelou a

existência de outliers em algumas ações, o que pode ser crucial para decisões de investimento. O uso de técnicas de visualização, como o boxplot, foi essencial para identificar essas anomalias.




Nome da ação x Preço da ação
preço ação R$
250
200
150
100
50
VALE3
SUZB3
GOOGLE
AMAZON
PETR4
APPLE
ELEKTRO
nome ação


Nome da ação x Valor de Mercado
valor de mercado R$ -(Bilhões)
5000
4000
3000
2000
1000
0
VALE3
SUZB3
GOOGLE
AMAZON
PETR4
APPLE
ELEKTRO
nome ação


Método do Cotovelo para Determinação de k
1e7
6
5
4
3
2
Inércia
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
Número de clusters

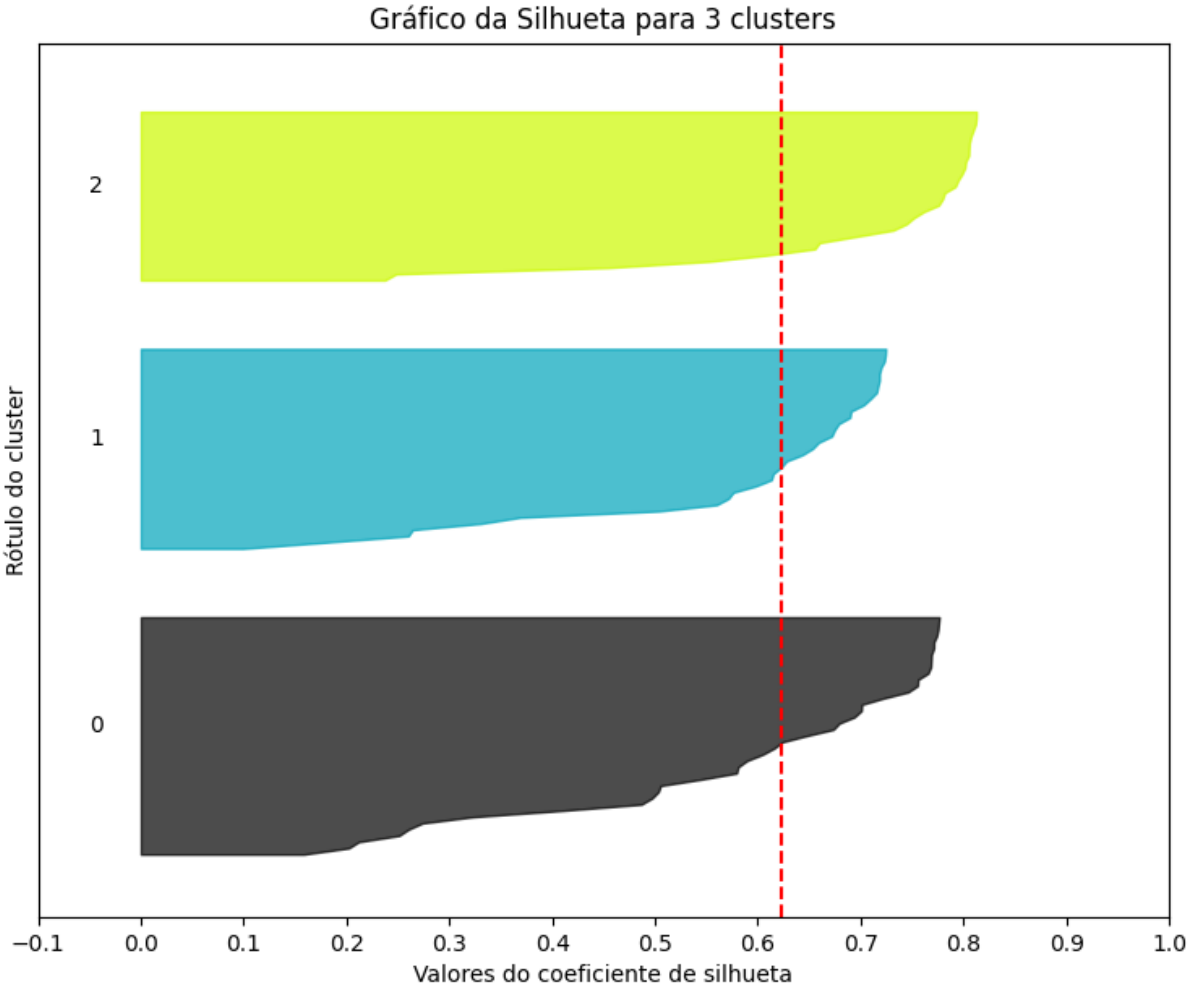
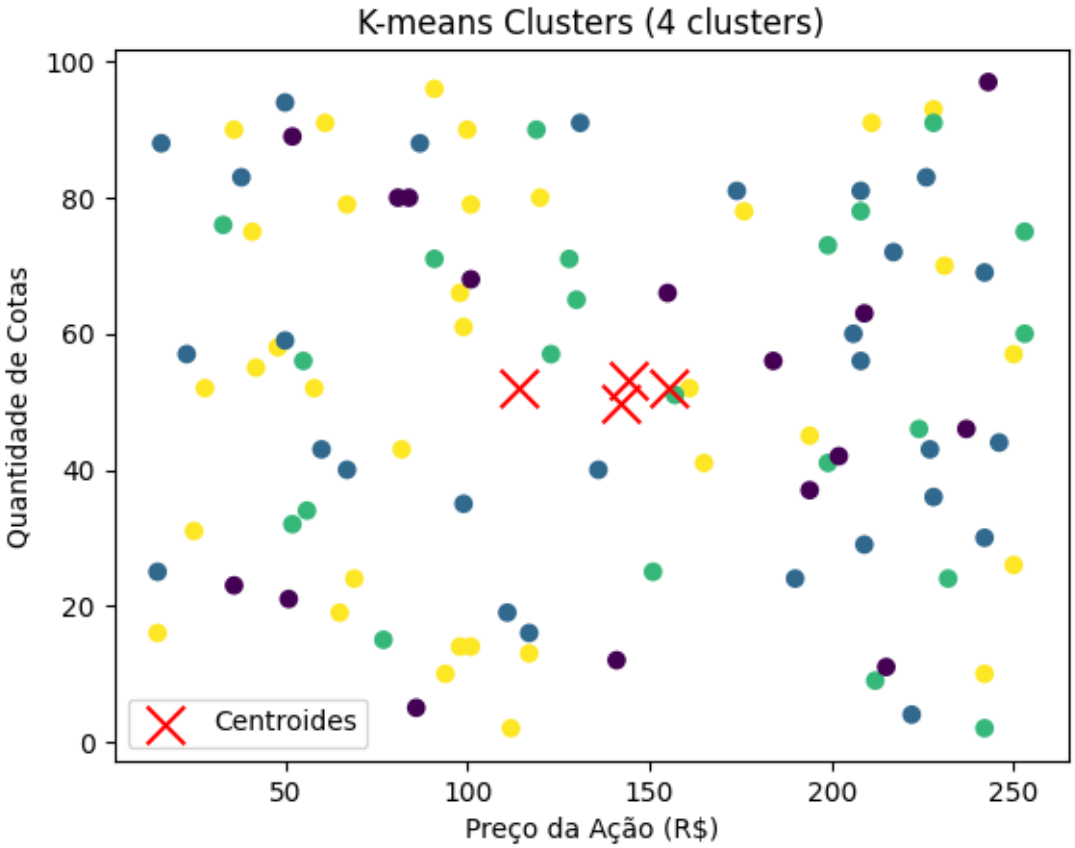
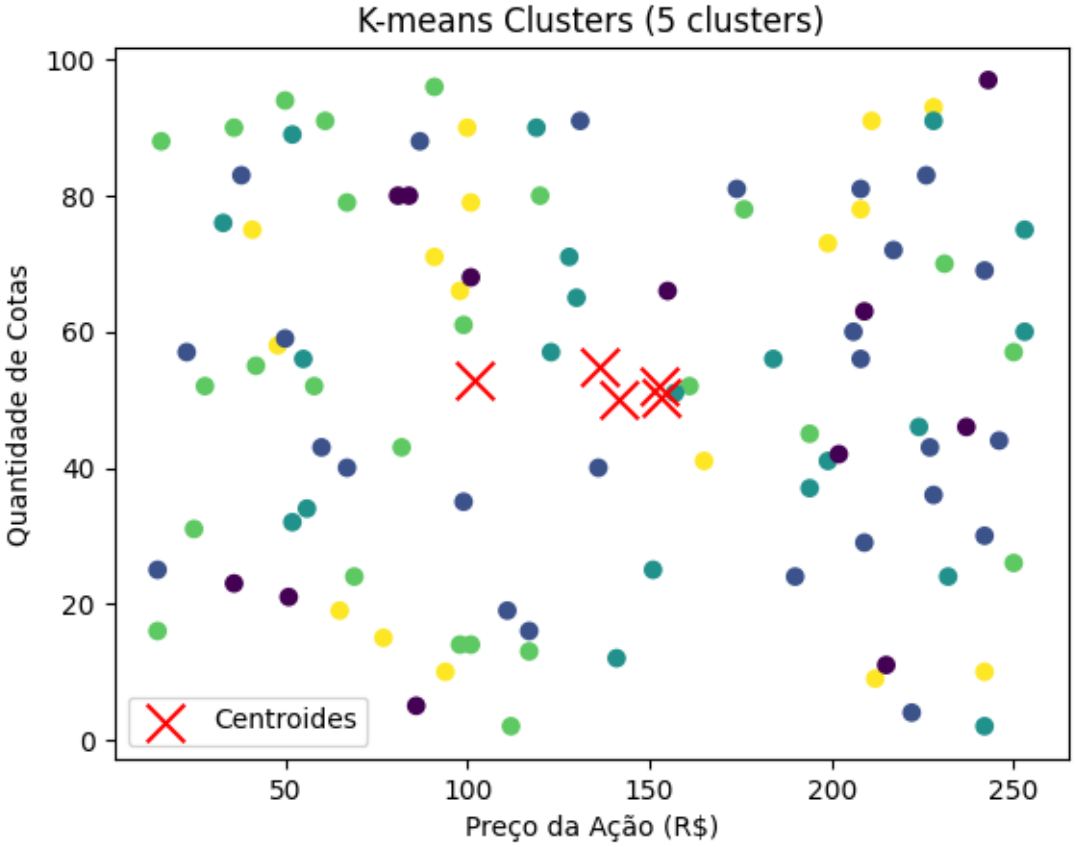
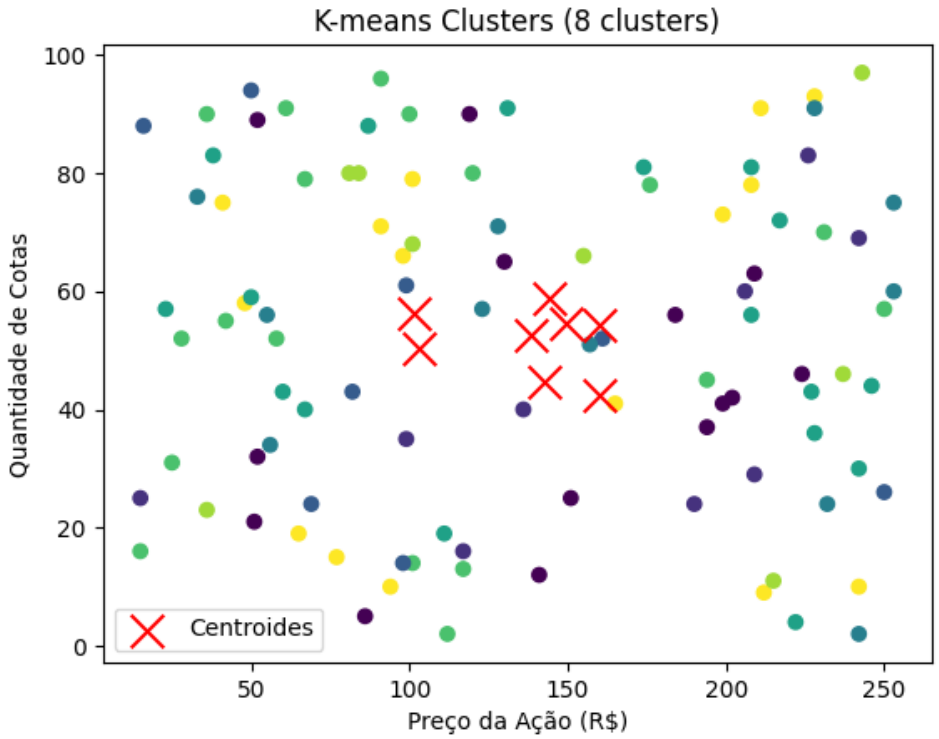
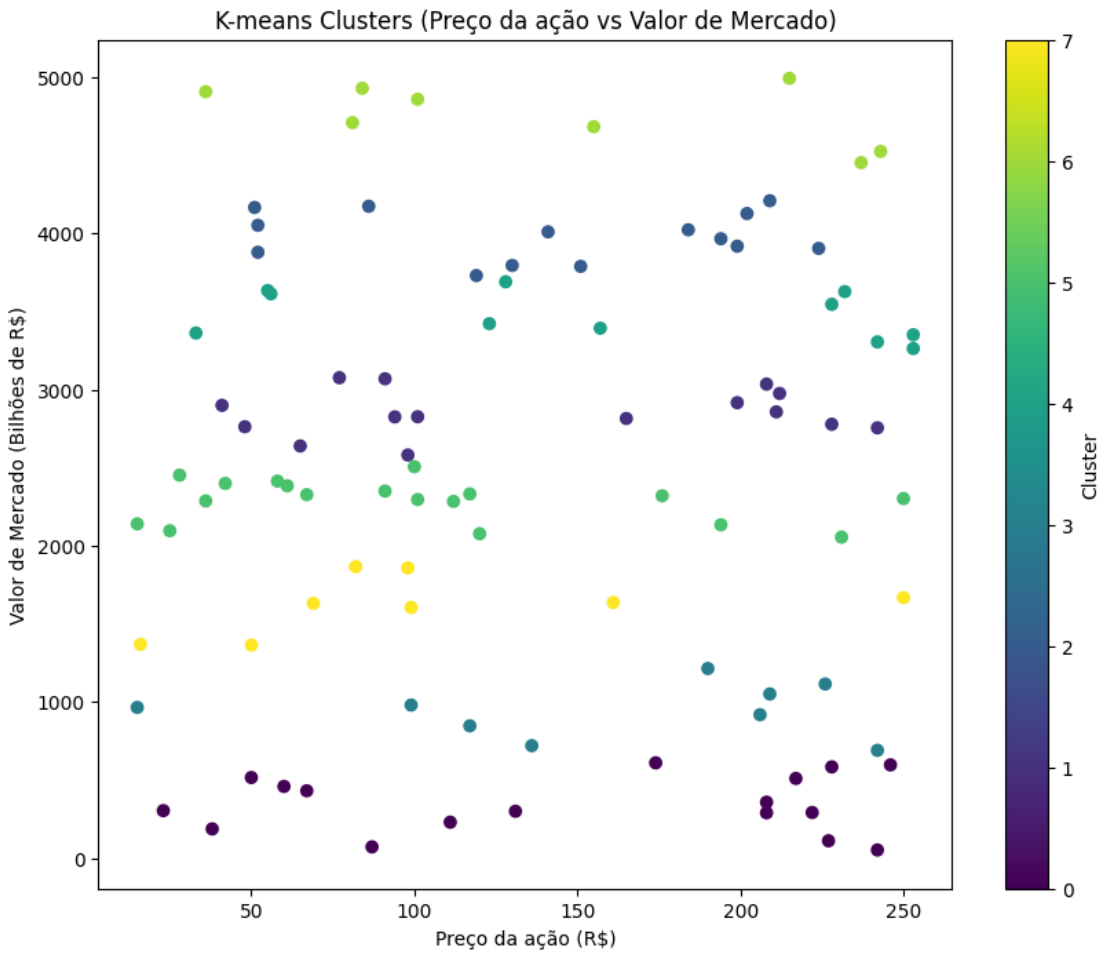


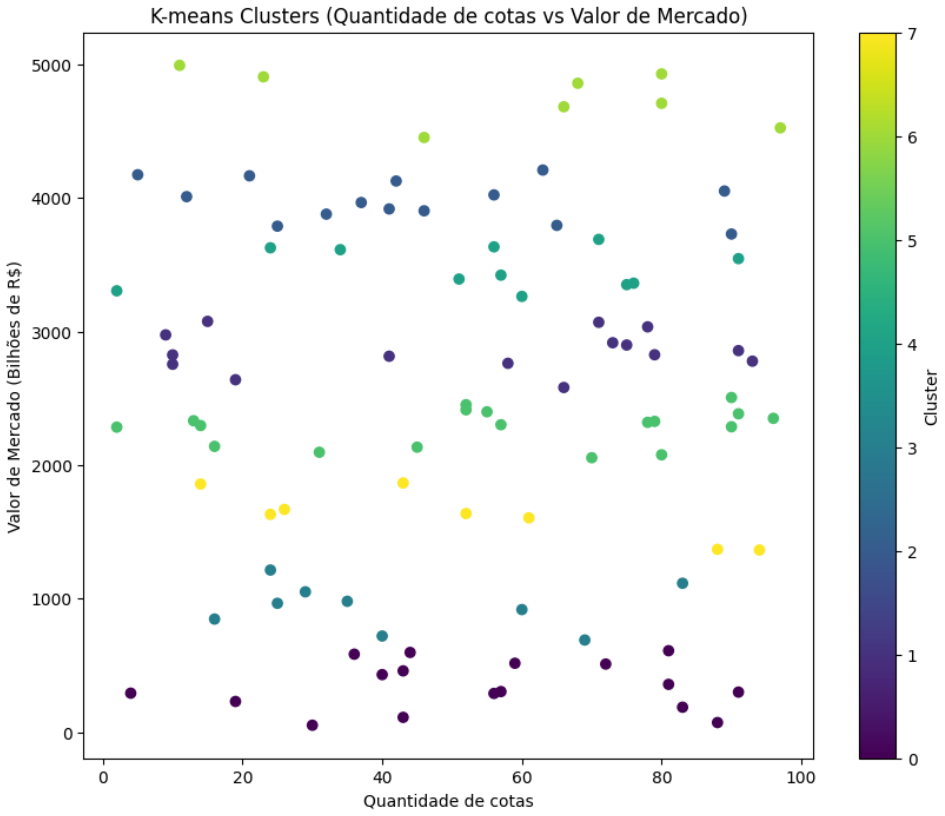
Gráfico da Silhueta para 3 clusters
2
1
0
Rótulo do cluster
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Valores do coeficiente de silhueta

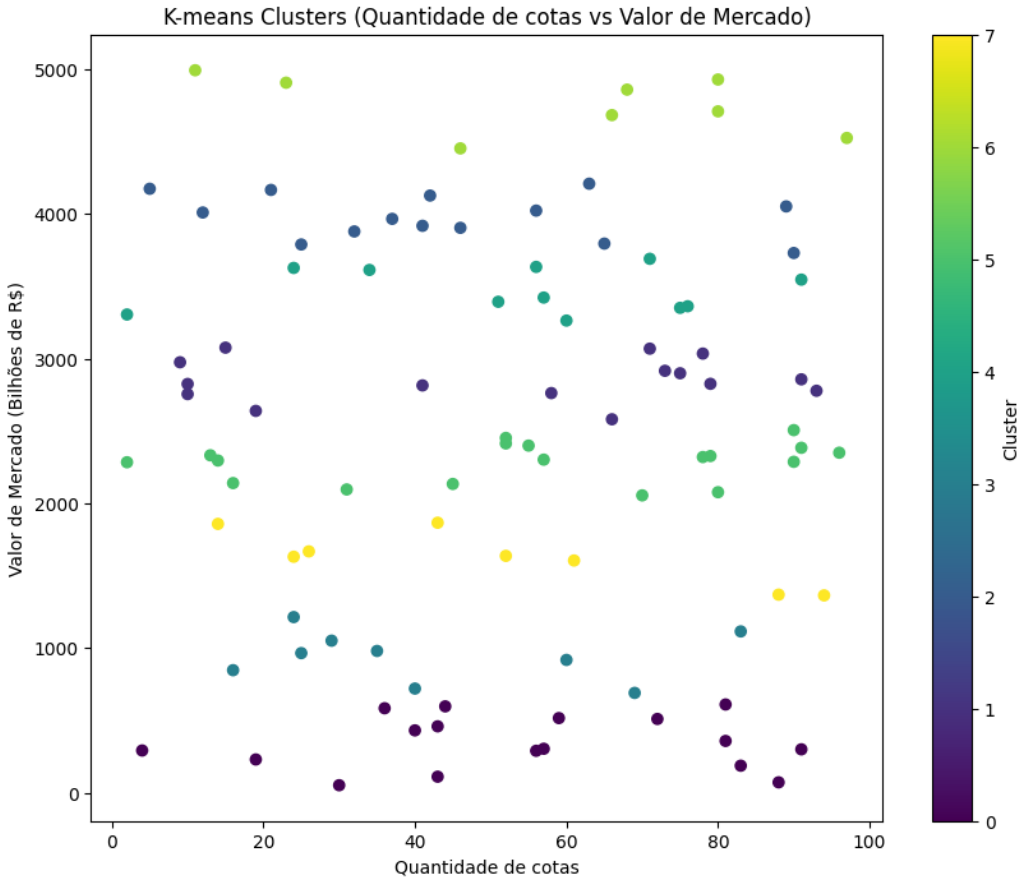

K-means Clusters (4 clusters)
Quantidade de Cotas
Preço da Ação (R$)
Centroides
0
20
40
60
80
100
50
100
150
200
250

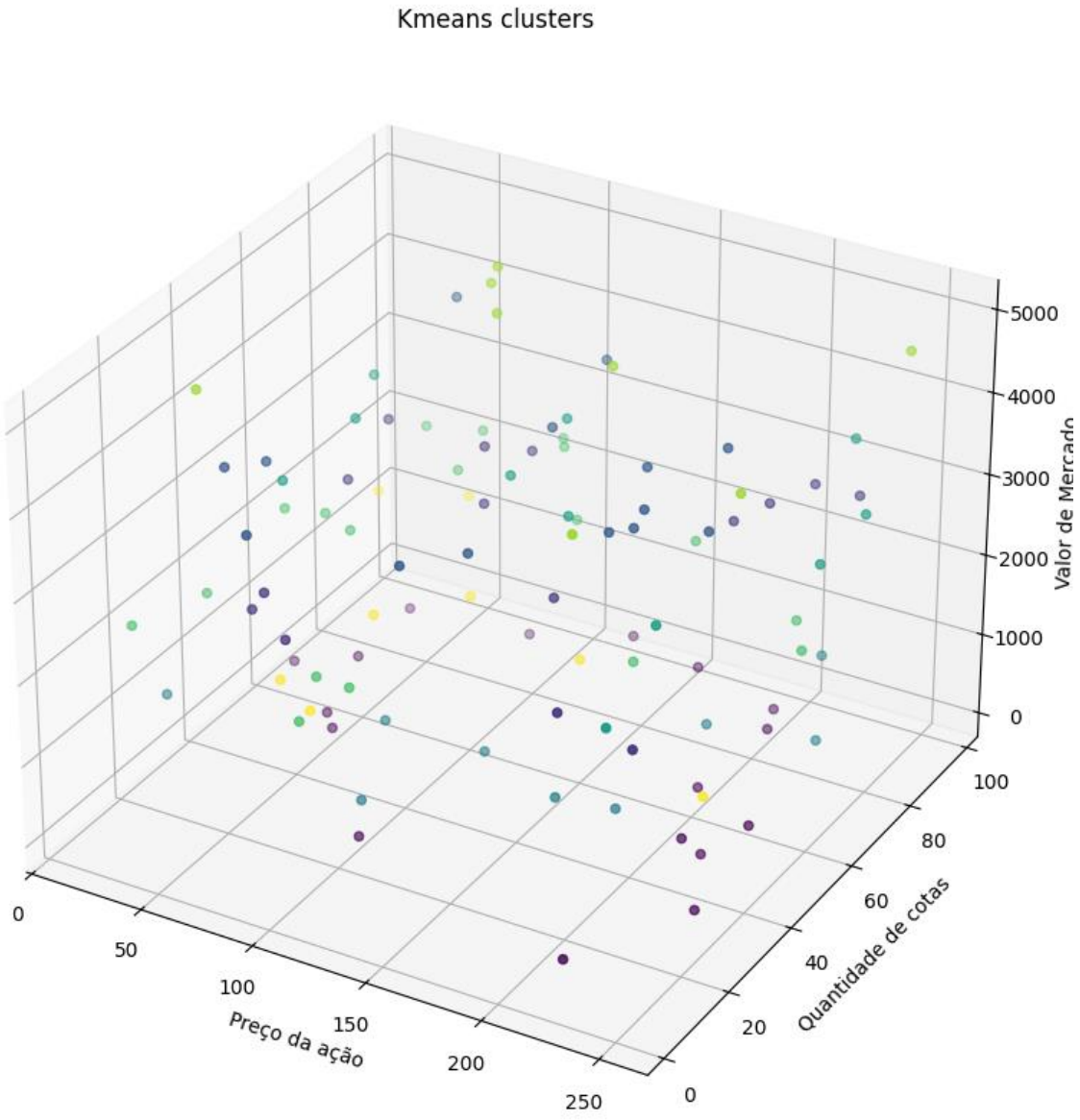

K-means Clusters (5 clusters)
Quantidade de Cotas
Preço da Ação (R$)
Centroides
0
20
40
60
80
100
50
100
150
200
250


K-means Clusters (8 clusters)
Quantidade de Cotas
Preço da Ação (R$)
Centroides
0
20
40
60
80
100
50
100
150
200
250


K-means Clusters (Preço da ação vs Valor de Mercado)
Valor de Mercado (Bilhões de R$)
Preço da ação (R$)
Cluster
0
1000
2000
3000
4000
5000
50
100
150
200
250
0
1
2
3
4
5
6
7


K-means Clusters (Quantidade de cotas vs Valor de Mercado)
Valor de Mercado (Bilhões de R$)
Quantidade de cotas
Cluster
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7


K-means Clusters (Quantidade de cotas vs Valor de Mercado)
Valor de Mercado (Bilhões de R$)
Quantidade de cotas
Cluster
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7



5 CONCLUSÃO

A aplicação do algoritmo K-means para a análise de ações de investimentos permitiu agrupar as ações em diferentes segmentos, baseados em suas características econômicas e financeiras. A utilização de técnicas de visualização e a exploração dos dados mostraram-se cruciais para a avaliação do comportamento das ações e para a tomada de decisões mais estratégicas. Embora o número de clusters tenha um impacto significativo na análise, a escolha do número ideal pode ser aprimorada com base nas métricas de qualidade, como o gráfico de silhueta e o gráfico do cotovelo.

O aprendizado não supervisionado, como o K-means, oferece uma vantagem substancial sobre o aprendizado supervisionado, principalmente em cenários em que os dados não estão rotulados e os padrões precisam ser descobertos de forma autônoma. Isso permite uma flexibilidade maior na análise e exploração de dados complexos e volumosos, como os

encontrados no mercado financeiro.

REFERÊNCIAS

- Jain, A. K. (2010). "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters, 31*(8), 651-666.
- Xu, R., & Wunsch, D. (2009). *Clustering*. Wiley-IEEE Press.
- Bihani, P., & Patil, A. (2018). "Stock market analysis using machine learning algorithms." *International Journal of Computer Science and Information Technologies*, 9(2), 134-138.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.