RELATÓRIO – Mineração de Dados

Nome: Lucas Luis Stasiak 101188 - **Turma:** 2023
Data: Julho de 2025

Dataset Utilizado

- **Nome:** USA House Sales Data
- **Origem:** Kaggle - USA House Sales
- **Tamanho:** Mais de 21 mil registros e 21 colunas

Este dataset contém dados de vendas de casas nos EUA, incluindo informações como número de quartos, banheiros, área útil, ano de construção, entre outros.

Técnica Aplicada

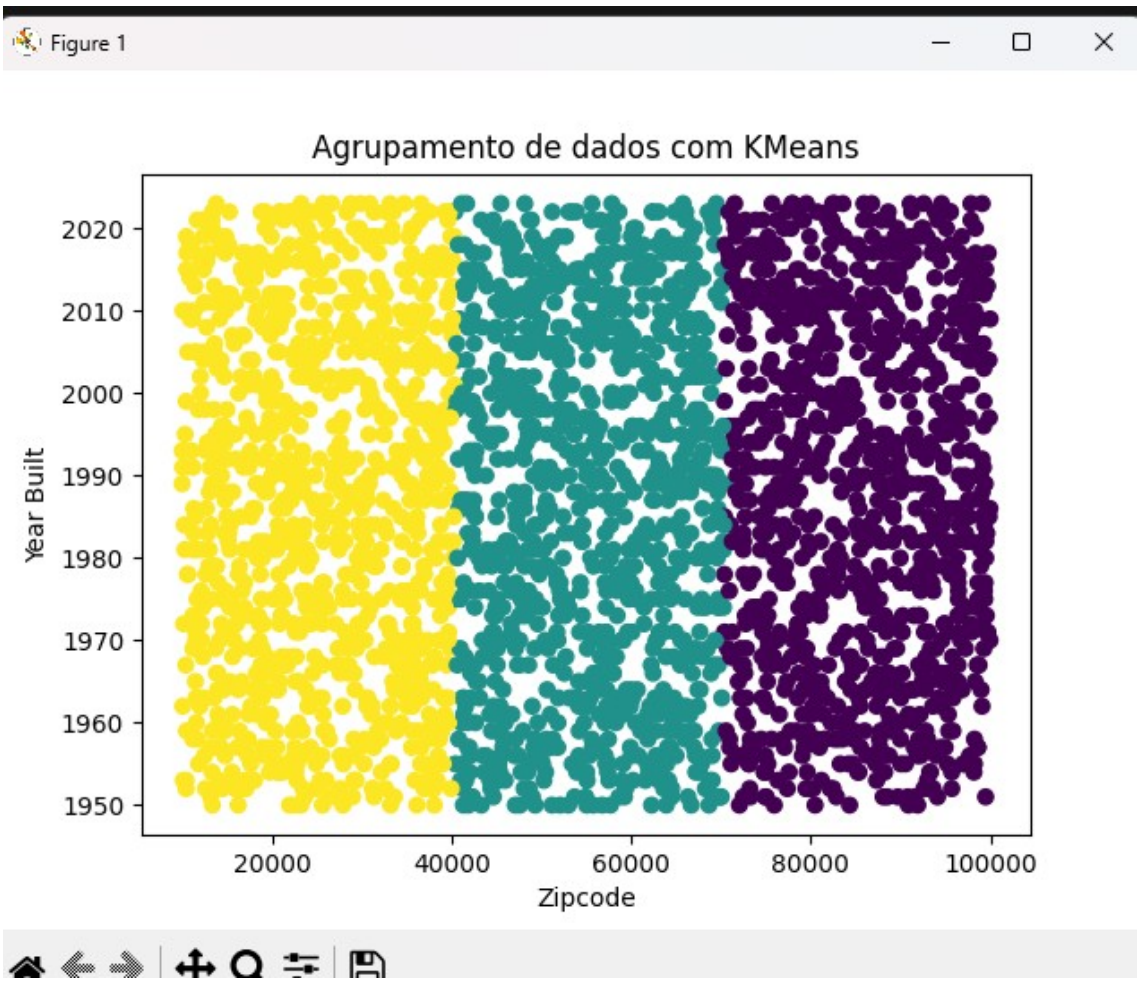
Foi utilizada a técnica de **agrupamento (clusterização)** com o algoritmo **KMeans** do `scikit-learn`, em Python.

Etapas do processo:

1. O dataset foi carregado com o `pandas`.
2. Apenas as colunas **numéricas** foram mantidas.
3. Linhas com dados ausentes foram removidas.
4. Foi aplicada a clusterização com **3 grupos (clusters)**.
5. Os resultados foram visualizados em um **gráfico de dispersão**, com cada cor representando um grupo diferente.

Resultado

A seguir está o gráfico gerado com a aplicação do KMeans. Ele representa as casas agrupadas por similaridade com base em duas variáveis numéricas principais do dataset:



Cada ponto representa uma casa, e a cor representa a qual grupo ela pertence.

Conclusão

O algoritmo KMeans foi capaz de **dividir as casas em 3 grupos distintos**, de acordo com características numéricas como preço, número de quartos, área útil, etc.
Esse tipo de análise é útil para identificar **padrões de mercado**, faixas de preço similares ou tipos de imóveis que compartilham características.

A clusterização é uma técnica não supervisionada eficaz para descobrir agrupamentos naturais em conjuntos de dados como este, sem a necessidade de rótulos pré-definidos.

Software Utilizado

- Python 3
- Bibliotecas: `pandas`, `scikit-learn`, `matplotlib`
- Ambiente: Visual Studio Code

Link repositório: https://github.com/Lucas-Stasiak/Trabalho-Minera-o-de-dados