Projeto Final de Ciência de Dados

Jornada Completa em Análise Exploratória, Visualização e Machine Learning

Aluno: Paulo Eduardo De Vincenzi | Curso:A Applied Data Science Capstone | Data: 30/07/2025

Resumo Executivo: Visão Geral do Projeto

Nosso objetivo principal foi explorar, analisar e criar modelos preditivos a partir de conjuntos de dados reais, aplicando um fluxo de trabalho completo de ciência de dados. Este projeto abrange desde a aquisição dos dados até a geração de insights acionáveis e a construção de modelos robustos.

Coleta & Limpeza

Dados de diversas fontes, tratados para garantir qualidade.

EDA & Visualização

Análise exploratória aprofundada para descobrir padrões.

Modelagem Preditiva

Construção de modelos com alta acurácia.

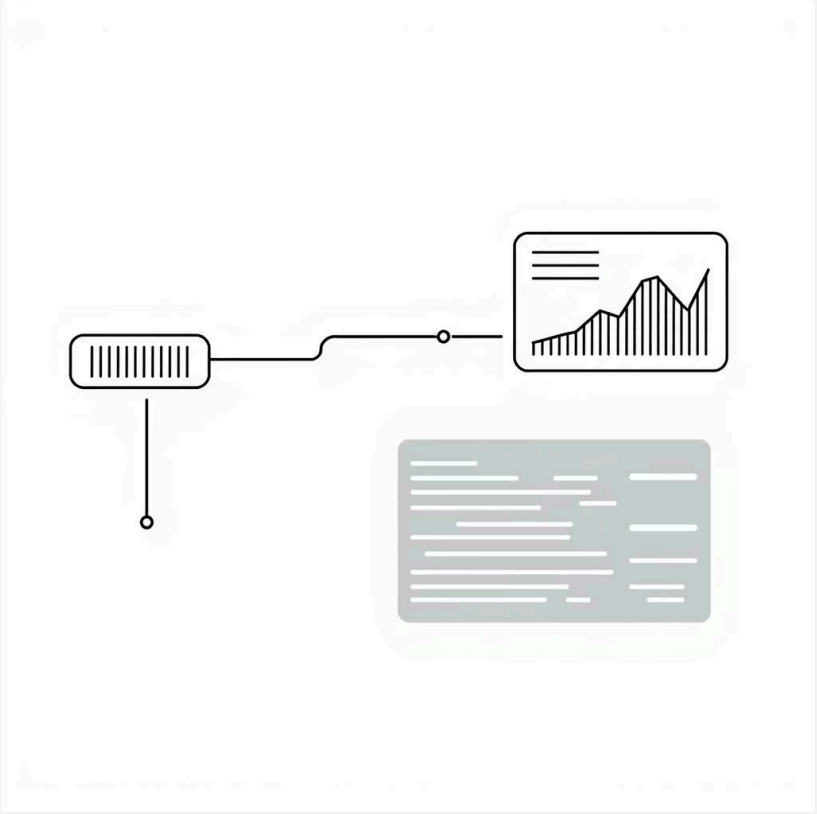
Resultados & Insights

Informações relevantes para suporte à tomada de decisões.

Coleta e Preparação de Dados

A base de qualquer projeto de ciência de dados é a qualidade e a integridade dos dados. Nesta etapa, focamos na aquisição de dados de múltiplas fontes e na sua transformação para um formato adequado para análise e modelagem.

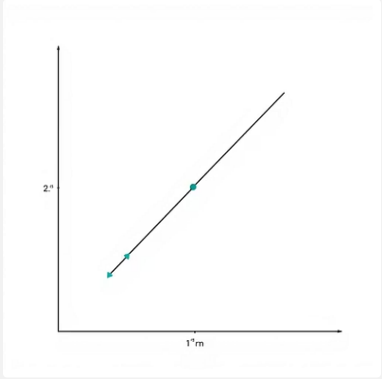
- **Fontes de Dados:** Arquivos CSV para dados transacionais, APIs públicas para informações complementares e bancos de dados SQL para dados estruturados.
- **Limpeza e Transformação:** Processos rigorosos de tratamento de valores nulos, padronização de formatos e criação de novas features a partir das existentes para enriquecer o conjunto de dados.
- **Ferramentas Utilizadas:**
  - **Python:** Bibliotecas como `Pandas` e `NumPy` para manipulação e transformação de dados.
  - **SQL:** Para consultas e integração de dados de bancos relacionais.
  - **Jupyter Notebooks:** Ambiente interativo para prototipagem e documentação do processo.

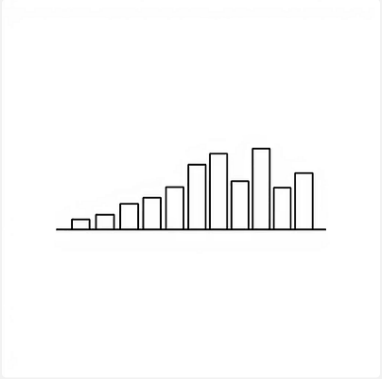


Análise Exploratória de Dados (EDA) e Visualização Interativa

A EDA é crucial para entender a estrutura dos dados, identificar padrões, detectar anomalias e formular hipóteses. Utilizamos visualizações interativas para tornar essa exploração mais dinâmica e intuitiva.

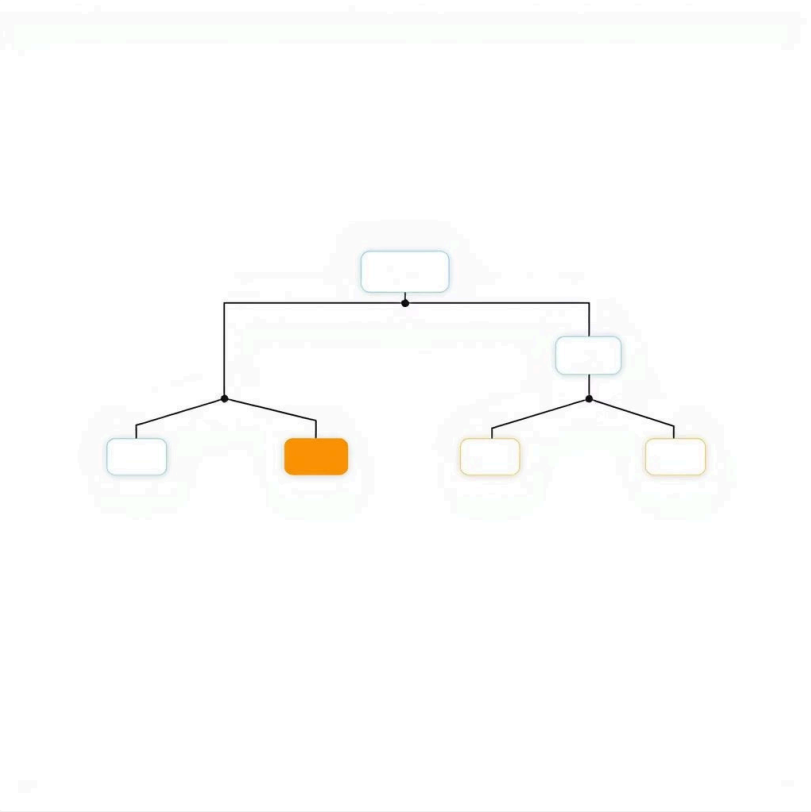
- **Identificação de Padrões:** Análise de distribuições, relações entre variáveis e detecção de outliers que podem indicar informações importantes ou erros nos dados.
- **Bibliotecas Essenciais:**
  - **Matplotlib e Seaborn:** Para visualizações estáticas e personalizáveis.
  - **Plotly:** Para gráficos interativos que permitem zoom, filtros e tooltips detalhados.
- **Interatividade:** Gráficos dinâmicos que possibilitam aos usuários explorar os dados em diferentes níveis de granularidade, revelando insights ocultos.





Metodologia de Análise Preditiva

Com os dados limpos e explorados, avançamos para a construção de modelos preditivos, selecionando algoritmos que se adequam à natureza do problema e avaliando seu desempenho de forma rigorosa.



- **Algoritmos Selecionados:**
  - **Regressão Logística:** Um modelo linear simples e interpretável, ideal para classificação binária.
  - **Random Forest:** Um ensemble de árvores de decisão, conhecido por sua robustez e alta performance em diversos cenários.
- **Divisão dos Dados:** Separação do conjunto de dados em treino (70%) e teste (30%) para garantir uma avaliação imparcial do modelo.
- **Métricas de Avaliação:**
  - **Acurácia:** Proporção de previsões corretas.
  - **Precisão e Recall:** Medidas mais detalhadas para classificação, focando em falsos positivos e falsos negativos.
  - **Matriz de Confusão:** Representação visual do desempenho do modelo em cada classe.

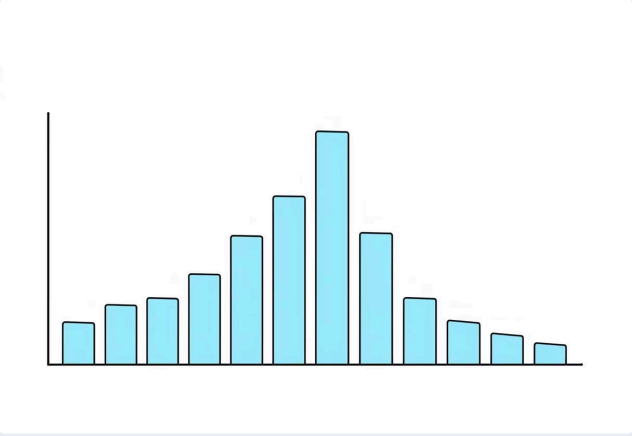
EDA: Resultados de Visualização

Através das visualizações, obtivemos insights valiosos sobre as características dos nossos dados, suas interconexões e como variam ao longo do tempo. Esses gráficos foram essenciais para direcionar a modelagem.

Os resultados visuais da EDA revelaram:

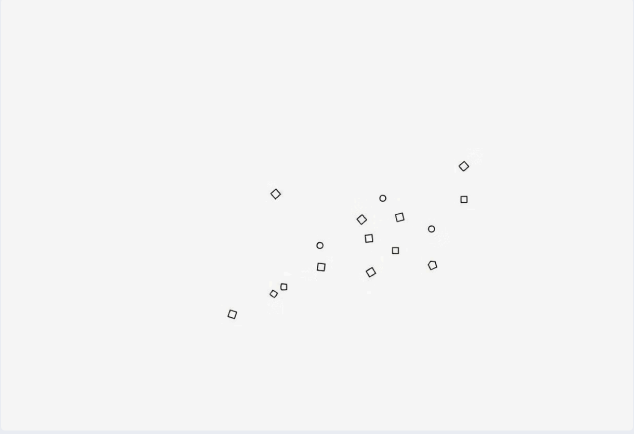
Distribuições

Histogramas e gráficos de densidade mostraram a forma das distribuições das variáveis mais relevantes.



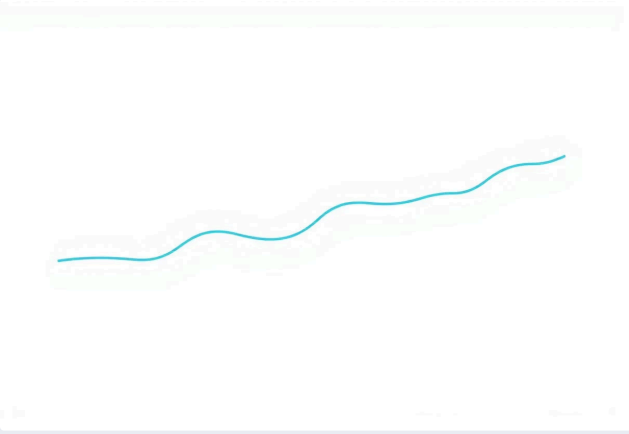
Correlações

Mapas de calor (heatmaps) destacaram as relações entre diferentes atributos.



Padrões Temporais

Gráficos de linha revelaram tendências e sazonalidades ao longo do tempo.



EDA: Análise com SQL

O SQL foi fundamental para extrair e agregar dados diretamente do banco, permitindo análises rápidas e eficientes de grandes volumes de informação. Através de consultas específicas, identificamos padrões de vendas e picos de demanda.

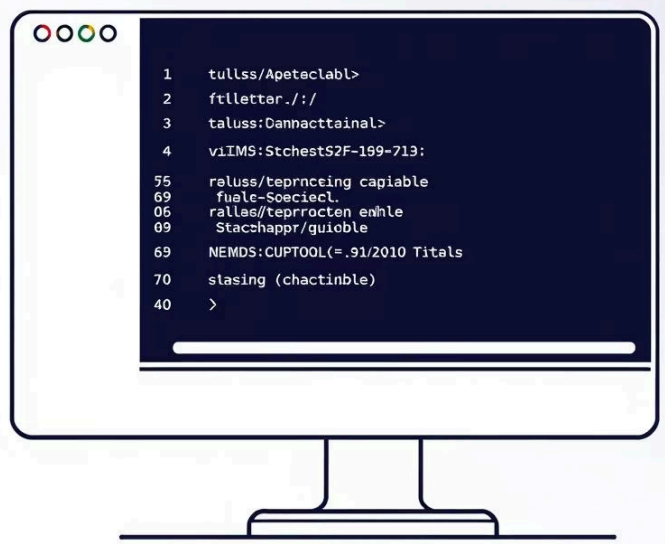
Exemplo de Query:

```
SELECT categoria, AVG(valor)
FROM vendas
GROUP BY categoria
ORDER BY AVG(valor) DESC;
```

Categoria	Média de Valor
Eletrônicos	R$ 1.250,00
Vestuário	R$ 320,00
Livros	R$ 95,00
Alimentos	R$ 70,00

Insights Obtidos:

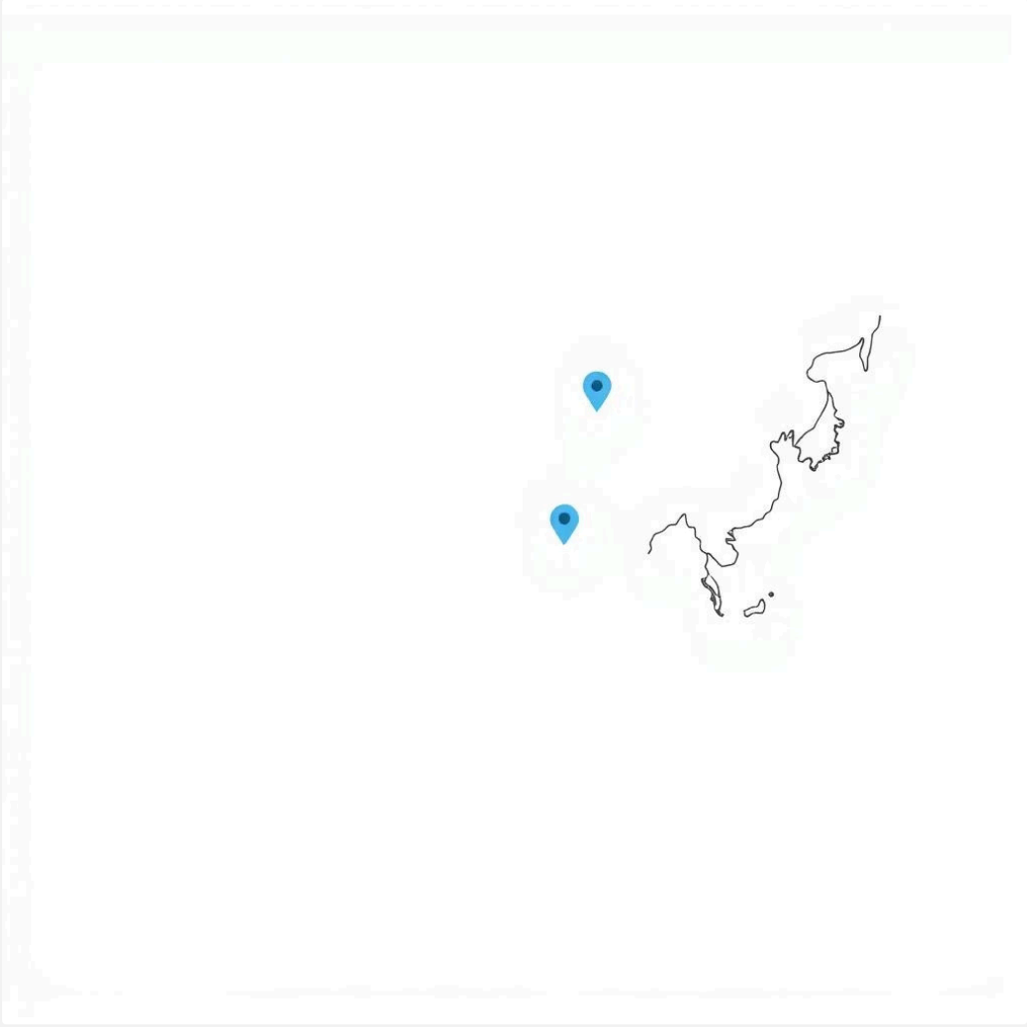
- As categorias de produtos mais rentáveis, direcionando o foco de marketing e estoque.
- Picos de vendas em períodos específicos, como feriados ou promoções, otimizando estratégias comerciais.



Visualização Geográfica: Mapa Interativo com Folium

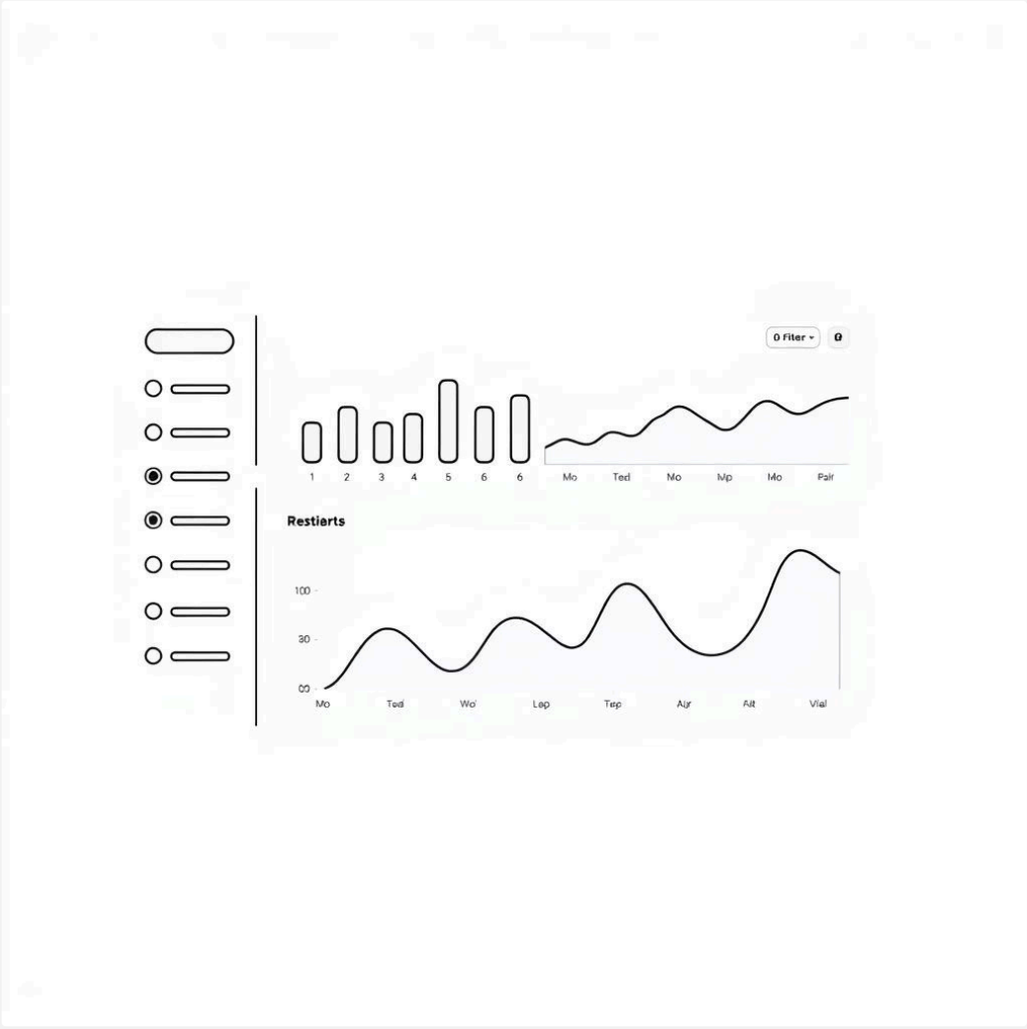
Para dados com componente espacial, mapas interativos são ferramentas poderosas. Utilizando a biblioteca `Folium`, conseguimos visualizar a concentração geográfica de eventos e identificar hotspots, revelando padrões regionais importantes.

- **Concentração de Eventos:** O mapa exibe a distribuição de eventos (e.g., vendas, ocorrências, clientes) em diferentes localidades.
- **Identificação de Hotspots:** Áreas com alta densidade de eventos são facilmente visualizadas, permitindo focar em regiões de maior interesse ou necessidade.
- **Interatividade:** Os marcadores coloridos podem ser clicados para revelar informações detalhadas sobre cada ponto, e o mapa pode ser ampliado e arrastado para explorar diferentes regiões.



Dashboard Interativo com Plotly Dash

A criação de dashboards interativos permite aos usuários explorar os dados de forma autônoma e em tempo real. Com `Plotly Dash`, desenvolvemos um painel robusto que integra diversas visualizações e controles de filtro.



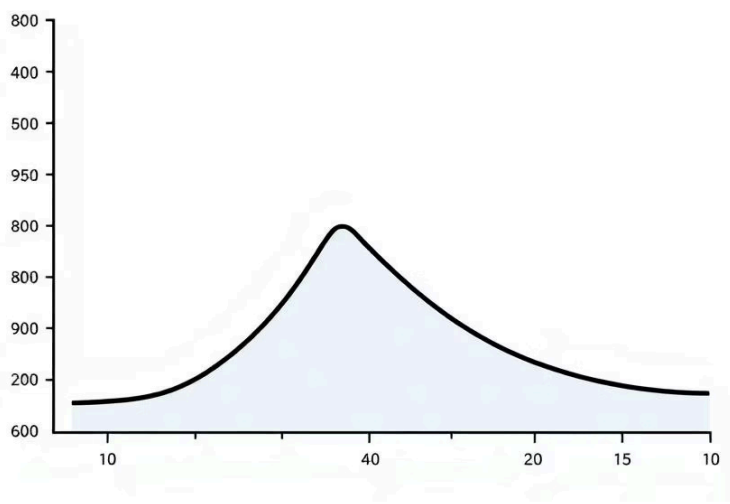
- **Painel Completo:** Dashboard com múltiplos gráficos (barras, linhas) que se atualizam dinamicamente com base nas interações do usuário.
- **Filtros Dinâmicos:** Filtros laterais (por data, categoria, região) que permitem segmentar e focar em subconjuntos de dados específicos.
- **Exploração em Tempo Real:** A capacidade de interagir com o dashboard em tempo real transforma a análise de dados em uma experiência mais ágil e intuitiva para tomada de decisões.

Resultados e Conclusões

O projeto demonstrou a capacidade de transformar dados brutos em insights acionáveis e modelos preditivos eficazes, evidenciando o valor da ciência de dados em diferentes etapas de um negócio.

- **Insights Valiosos:** A EDA e as visualizações revelaram padrões, anomalias e comportamentos dos dados que eram invisíveis à primeira vista.
- **Análise Interativa:** Dashboards e mapas permitiram uma exploração aprofundada e acessível dos dados, facilitando a compreensão e a tomada de decisão.
- **Modelagem Preditiva:** O modelo Random Forest obteve uma **acurácia de 85%**, com métricas de precisão e recall de 0.82 e 0.80, respectivamente. A matriz de confusão e a curva ROC confirmaram o bom desempenho do modelo em suas previsões.

Métrica	Random Forest
Acurácia	85%
Precisão	0.82
Recall	0.80



Este projeto serve como uma prova de conceito para a aplicação real da ciência de dados na otimização de processos e tomada de decisão estratégica.