Relatório Detalhado do Processo: Criação e Refinamento do Script de Análise de Dados

Data: 04 de Junho de 2025

1. Introdução

Este relatório documenta as etapas realizadas para atender à solicitação de criação de um script Python para análise de dados e Machine Learning, incluindo a organização do projeto e os ajustes para compatibilidade com ambientes de notebook como Jupyter.

2. Solicitação Inicial

O usuário solicitou um script Python modular para realizar um fluxo completo de análise de dados, incluindo:

- **Carregamento de Dados:** A partir de um arquivo `dados_exemplo.csv`.
- **Análise Exploratória de Dados (EDA):** Exibição de informações básicas, estatísticas descritivas, contagem de ausentes/únicos e visualizações (histograma, barras, dispersão).
- **Pré-processamento:** Tratamento de valores ausentes e codificação de variáveis categóricas.
- **Modelagem:** Treinamento de um modelo simples (Regressão Linear) e avaliação (MAE, R^2).
- **Requisitos de Bibliotecas:** Uso exclusivo de Pandas, NumPy, Matplotlib, Seaborn e Scikit-learn.
- **Formato:** Código modular em funções, comentado e adequado para execução em células de Jupyter Notebook.

3. Etapas de Desenvolvimento

3.1. Geração do Dataset de Exemplo

- Para garantir a funcionalidade do script, foi criado um script Python (`criar_csv.py`) para gerar um arquivo `dados_exemplo.csv` sintético.

- Este dataset contém colunas numéricas e categóricas, incluindo valores ausentes intencionais para testar as etapas de pré-processamento.
- O arquivo CSV foi salvo em `/home/ubuntu/dados_exemplo.csv`.

3.2. Criação do Script Principal (`analise_ml.py` - Versão 1)

- Foi desenvolvido o script `analise_ml.py` contendo funções modulares para cada etapa solicitada (carregamento, EDA, visualização, pré-processamento, modelagem, avaliação).
- O script foi estruturado com comentários explicativos e marcações (`# %% [code]`, `# %% [markdown]`) para facilitar a importação e execução em ambientes de notebook.
- Inicialmente, a execução do fluxo completo foi encapsulada em um bloco `if __name__ == '__main__':` para permitir a execução como um script independente.
- **Entregável:** Script `analise_ml.py` e `dados_exemplo.csv` foram enviados ao usuário.

3.3. Reorganização da Estrutura do Projeto

- O usuário forneceu uma imagem com uma estrutura de diretórios padrão para projetos de ciência de dados.
- **Ações Realizadas:**
  - Criação da estrutura de pastas: `analise-de-dados/`, `data/raw/`, `data/processed/`, `notebooks/`, `src/`, `models/`, `reports/`.
  - Movimentação dos arquivos: `dados_exemplo.csv` para `data/raw/` e `analise_ml.py` para `src/`.
  - Criação de arquivos auxiliares:
    - `.gitignore`: Para ignorar arquivos comuns (ambiente virtual, caches, etc.).
    - `README.md`: Com descrição do projeto, estrutura e instruções de uso.
    - `requirements.txt`: Listando as dependências (pandas, numpy, matplotlib, seaborn, scikit-learn).
    - `src/__init__.py`: Arquivo vazio para marcar `src` como um pacote Python.
- **Entregável:** Um arquivo zip (`analise_de_dados_projeto.zip`) contendo toda a estrutura organizada foi enviado ao usuário.

3.4. Ajuste para Execução Célula a Célula no Notebook (Versão 2)

- O usuário reportou que o script não funcionava ao executar as células individualmente no notebook, devido ao bloco `if __name__ == '__main__':`.
- **Ações Realizadas:**
  - O script `src/analise_ml.py` foi refatorado.
  - O bloco `if __name__ == '__main__':` foi removido.
  - As chamadas de função para executar o fluxo foram movidas para células `# %% [code]` separadas no final do script (Seção 6), permitindo a execução sequencial.
- **Entregável:** Um novo arquivo zip (`analise_de_dados_projeto_v2.zip`) com o script atualizado foi enviado.

3.5. Esclarecimento sobre Erros `NameError` (Versão 3 - Atual)

- O usuário apresentou capturas de tela mostrando erros `NameError: name '...' is not defined`.
- **Causa Identificada:** Esses erros ocorrem em notebooks quando as células que definem funções ou variáveis não são executadas antes das células que as utilizam.
- **Ações Realizadas:**
  - O script `src/analise_ml.py` foi novamente atualizado para incluir um comentário explicativo proeminente na Seção 6.
  - Este comentário instrui o usuário a **executar todas as células anteriores (Seções 0 a 5) que contêm as definições de funções antes de executar as células da Seção 6**.
  - Adicionou-se também uma verificação básica do diretório de trabalho para ajustar o caminho do CSV caso o script seja executado da raiz do projeto ou da pasta `src`.
- **Entregável Atual:** O projeto atualizado (nesta entrega) contém o script com os comentários didáticos adicionais.

4. Conclusão

O projeto foi desenvolvido e refinado iterativamente com base no feedback do usuário. A versão atual inclui:

- Um script Python modular (`src/analise_ml.py`) para análise de dados e ML.
- Um dataset de exemplo (`data/raw/dados_exemplo.csv`).
- Uma estrutura de projeto organizada.
- Arquivos auxiliares (`.gitignore`, `README.md`, `requirements.txt`).

- Formatação e comentários adequados para execução célula a célula em ambientes de notebook, com instruções claras para evitar erros comuns como `NameError`.

Espera-se que esta versão final atenda completamente aos requisitos e facilite o uso e a adaptação pelo usuário.