**Projeto para estudos dos indicadores da educação no Brasil. Dados referentes ao ano de 2021 e coletados no Kaggle pelo seguinte link:**

[**https://www.kaggle.com/datasets/jgabrielsb/2021-educational-data-about-brazil**](https://www.kaggle.com/datasets/jgabrielsb/2021-educational-data-about-brazil)

**Dados mesclado de: Pesquisas Estatísticas e Indicadores Educacionais**

**1. Estrutura do Projeto de Engenharia de Dados**

O projeto será dividido em quatro grandes etapas:

1. **ETL (Extração, Transformação e Carga)**
   * Extração: Ler e organizar os dados do CSV.
   * Transformação: Limpeza, normalização e criação de novas colunas úteis.
   * Carga: Salvar os dados transformados em um banco de dados (PostgreSQL ou outro).
2. **Modelagem e Integração com o Power BI**
   * Criar um modelo de dados otimizado.
   * Conectar o banco de dados ao Power BI.
3. **Criação do Painel de Indicadores**
   * Construção de gráficos e tabelas dinâmicas para análise.
   * Adicionar filtros interativos.
4. **Documentação e Otimização**
   * Criar documentação do pipeline.
   * Avaliar melhorias e desempenho.

**2. Detalhamento do ETL**

**1️⃣ Extração**

* **Tecnologias:** Python (pandas), SQLAlchemy para integração com bancos de dados.
* **Objetivo:** Carregar os dados do CSV para um DataFrame e verificar a qualidade dos dados.

**2️⃣ Transformação**

* **Passos:**
  + Remover colunas desnecessárias.
  + Lidar com valores nulos e inconsistências.
  + Criar colunas calculadas, como médias por região.
  + Normalizar os dados para facilitar análises.

**3️⃣Carga**

* **Banco de dados recomendado:** PostgreSQL (pela robustez e compatibilidade com Power BI).
* **Estrutura da Tabela:**
  + id (chave primária)
  + unidade\_geografica
  + ano (caso tenhamos séries temporais)
  + taxa\_aprovacao\_fundamental
  + taxa\_reprovacao\_fundamental
  + taxa\_abandono\_fundamental
  + taxa\_distorcao\_fundamental
  + taxa\_aprovacao\_medio
  + taxa\_reprovacao\_medio
  + taxa\_abandono\_medio
  + taxa\_distorcao\_medio
  + ... (outros indicadores importantes)

Agora, vamos começar a implementação do ETL. O primeiro passo é a **extração dos dados e uma análise inicial**. Vou fazer isso agora!

**📌 Análise Inicial dos Dados**

O CSV contém **33 registros** e **63 colunas**, com dados sobre educação por unidade geográfica. Aqui estão algumas observações iniciais:

✅ **Principais Indicadores:**

* Taxas de aprovação, reprovação, abandono e distorção idade-série (fundamental e médio).
* Remuneração média e carga horária dos professores (com valores nulos).
* Notas de matemática e português no ensino fundamental.
* Segmentação por diferentes níveis de ensino.

⚠️ **Problemas Identificados:**

* **Valores Nulos:** Algumas colunas (exemplo: remuneração média, INSE) têm valores ausentes.
* **Estrutura de Nomes:** Alguns nomes de colunas podem ser ajustados para facilitar análises.
* **Colunas Desnecessárias:** Algumas podem não ser úteis no Power BI e podem ser descartadas.

**🛠️ Próximos Passos – Transformação dos Dados**

Agora, vamos iniciar a **Transformação** no ETL:

1. **Corrigir nomes de colunas** para um formato mais padronizado.
2. **Tratar valores nulos** (preencher, remover ou substituir).
3. **Selecionar colunas relevantes** para análise no Power BI.
4. **Criar novas colunas** úteis, como médias por região.

Vou começar organizando os nomes das colunas! ​​

**✅ Transformação - Renomeação das Colunas**

Os nomes das colunas foram padronizados para facilitar a manipulação e integração com o Power BI.

📌 **Exemplos de ajustes:**

* "FUND Apr" → taxa\_aprovacao\_fundamental
* "MED Aban" → taxa\_abandono\_medio
* "Remuneração Média" → remuneracao\_media
* "FUND Nota Matemática" → nota\_matematica\_fundamental

**🛠️ Próximos Passos - Limpeza e Normalização**

Agora, vamos tratar valores nulos e normalizar os dados:

1. **Substituir valores nulos**: Definir estratégias para colunas com dados ausentes.
2. **Converter tipos de dados**: Certificar que as colunas numéricas estão no formato correto.
3. **Remover colunas desnecessárias**: Selecionar apenas as colunas essenciais.

Vou realizar essa limpeza agora! ​​

It seems like I can’t do more advanced data analysis right now. Please try again later.

Mas posso te guiar nos próximos passos! Para tratar os valores nulos no seu código, siga esta lógica:

1. **Preencher valores nulos na remuneração média** com a média dos valores disponíveis:

python

CopiarEditar

df['remuneracao\_media'].fillna(df['remuneracao\_media'].mean(), inplace=True)

1. **Preencher valores nulos em outras colunas numéricas com 0** (se fizer sentido para a análise):

python

CopiarEditar

numeric\_cols = df.select\_dtypes(include=['float64', 'int64']).columns

df[numeric\_cols] = df[numeric\_cols].fillna(0)

1. **Verificar se ainda há valores nulos** após o tratamento:

python

CopiarEditar

print(df.isnull().sum())

Tente rodar esses comandos no seu ambiente e me avise se precisar de ajustes! 🚀 ​

Outras ideias de projetos:

**1. Painel de Indicadores Educacionais**

* Criar um painel no Power BI mostrando métricas como aprovação, reprovação, abandono e distorção idade-série por região.
* Incluir gráficos de tendência e comparações entre estados/regiões.
* Criar mapas interativos para visualizar os dados geograficamente.

**2. Análise de Risco Educacional**

* Criar um modelo para identificar regiões com maior risco de evasão escolar.
* Aplicar técnicas de machine learning para prever abandono com base nas variáveis.

**3. Comparação entre Regiões e Estados**

* Criar um ranking de regiões e estados com base no desempenho educacional.
* Destacar regiões que melhoraram ou pioraram ao longo do tempo.

**4. Clusterização de Perfis Educacionais**

* Utilizar técnicas de agrupamento (K-Means, DBSCAN) para categorizar regiões com perfis educacionais semelhantes.
* Criar um dashboard mostrando os clusters e suas características.

**5. Análise Temporal da Educação**

* Se houver dados de diferentes anos, criar uma análise histórica para ver evolução da educação no Brasil.
* Incluir previsões para os próximos anos com modelos estatísticos.

**6. Correlação entre Indicadores**

* Analisar a relação entre abandono, reprovação e distorção idade-série.
* Investigar se existe um padrão entre as regiões mais afetadas.