**Plano do Projeto: Integração de Dados BeeCorp (CRISP-DM)**

**Fase 1: Entendimento do Negócio e Entendimento dos Dados**

**Objetivo:** Compreender o problema de negócio e a natureza dos dados para planejar o trabalho.

1. **Reunião de Kickoff (com a Empresa):** Participe do encontro com a BeeCorp para alinhar as expectativas e fazer perguntas sobre como os dados são usados, quais são as dores principais e como eles esperam usar a base unificada.
2. **Análise Exploratória dos Dados (AED):**
   * Carregue cada um dos arquivos .csv (Recrutamento\_Selecao, Profissional\_Ponta, Operacoes, Departamento\_Pessoal, Base\_Referencia) em um DataFrame do Pandas.
   * Para cada DataFrame, execute os seguintes comandos:
     + df.info(): Para ver os tipos de dados e a quantidade de valores nulos.
     + df.describe(): Para obter um resumo estatístico das colunas numéricas.
     + df.head() e df.tail(): Para visualizar as primeiras e últimas linhas.
     + Verifique a contagem de valores únicos nas colunas de cliente (df['Nome do Cliente'].nunique()). A variação entre as bases é a principal inconsistência que você precisa resolver.
3. **Planejamento de Limpeza e Unificação:** Com a AED em mãos, liste as inconsistências encontradas e discuta com seu grupo as estratégias para corrigi-las.

**Fase 2: Preparação dos Dados**

**Objetivo:** Limpar, padronizar e unificar as bases de dados.

1. **Criação do Ambiente de Projeto (com Python venv e Git)**
   * **Configuração Inicial (feito por apenas um membro):**
     + Crie um diretório para o projeto e inicialize o Git:

Bash

mkdir beecorp-projeto

cd beecorp-projeto

git init

* + - Crie e ative o ambiente virtual:

Bash

python -m venv venv

# No Windows:

.\venv\Scripts\activate

# No macOS/Linux:

source venv/bin/activate

* + - Crie o arquivo .gitignore para evitar que arquivos desnecessários sejam enviados ao repositório. Adicione as seguintes linhas:

Snippet de código

# Ignorar o ambiente virtual

venv/

* + - Instale as bibliotecas necessárias e crie o arquivo de dependências:

Bash

pip install pandas scikit-learn fuzzywuzzy python-levenshtein

pip freeze > requirements.txt

* + - Faça o primeiro commit e envie para um novo repositório no GitHub:

Bash

git add .

git commit -m "Configuração inicial do projeto"

git remote add origin <URL do seu repositório no GitHub>

git push -u origin main

* + **Configuração para o resto do grupo:**
    - Cada membro deve clonar o repositório:

Bash

git clone <URL do repositório>

cd beecorp-projeto

* + - Ative o ambiente virtual e instale as dependências:

Bash

# Ative o venv (mesmos comandos acima)

pip install -r requirements.txt

1. **Desenvolvimento do Código de Limpeza e Unificação:**
   * Crie um script Python (src/unificacao\_dados.py) para executar as seguintes tarefas:
     + **Carregamento:** Ler todos os arquivos .csv.
     + **Limpeza (Data Cleaning):**
       - Padronize os nomes dos clientes (ex: tudo para maiúsculas, remover caracteres especiais e abreviações).
       - Trate os valores nulos e remova duplicatas.
     + **Integração e Clusterização:**
       - Utilize o arquivo Base\_Referencia.csv como a **fonte principal** de verdade. Ele contém a "Razão Social", "Unidade", "Grupo Econômico" e um ID único, que é o objetivo final do projeto.
       - Use o ID da Base\_Referencia como a chave primária para unificar todos os outros arquivos.
       - Para clientes que não têm um ID correspondente, use a **clusterização (fuzzywuzzy)** para encontrar a correspondência mais provável na Base\_Referencia e atribuir o ID correto.
     + **Salvamento:** Salve a base de dados unificada final.

**Fase 3: Modelagem e Validação**

**Objetivo:** Desenvolver e validar a solução.

1. **Criação dos Relatórios:** Com a base de dados unificada, crie um script ou utilize uma ferramenta como Power BI para gerar um "esqueleto de relatórios" que a empresa possa usar para ver o DRE por cliente.
2. **Validação:**
   * Apresente a base unificada e os relatórios para a empresa nos encontros semanais.
   * Pergunte se os dados fazem sentido, se a unificação está correta e se a solução atende às necessidades deles.
   * Use o feedback para fazer ajustes no seu código.

**Fase 4: Implantação e Documentação**

**Objetivo:** Finalizar o projeto e entregar os artefatos.

1. **Documentação Final:** Elabore o relatório seguindo o modelo fornecido.
   * Inclua uma seção detalhando a metodologia CRISP-DM e como cada fase foi aplicada.
   * Descreva cada passo do seu código, as decisões tomadas para a limpeza e unificação, e os resultados alcançados.
2. **Entrega:** Submeta o relatório final e a base de dados unificada. A presença na banca avaliadora é crucial.

Seguindo este guia, vocês terão um fluxo de trabalho estruturado e profissional para a execução do projeto.