ADD - Analise de Dados

Aula 3 - Data Cleaning and Standardisation

Miguel Araujo - 03231034 - miguel.asilva@sptech.school

Aeris Bizaroli Rasmussen - 03231000 - aeris.rasmussen@sptech.school

1. Utilizando a base de dados fornecida como modelo aplicar as técnicas ministradas na aula para limpeza, padronização e enriquecimento de dados

- Dados Texto não devem ter acentuações
- Dados devem estar no formato exigido pelo SGBD.
- Dados Texto devem estar padronizados (maiúscula)
- Ler o arquivo CSV e gerar um arquivo XLS com dados tratados
- Detalhe as técnicas utilizadas para o tratamento de dados

Foi utilizado a biblioteca pandas para manipulação dos arquivos csv e xlsx, unicodedata para tratamentos de caracteres especiais e acentuações, chardet para identificar e tratar o enconding do arquivo e datatime para colocar na saída do processamento e ter controle de execução.

As técnicas utilizadas foram de:

- Técnicas de Data Cleaning:
  - Encoding Detection & Correction: Uso do chardet para identificar o encoding correto (ISO-8859-1 em vez de UTF-8);
  - Duplicate Removal: Remoção de valores duplicados;
  - Missing Data Imputation (constant value imputation): Tratamento de valores ausentes (Missing Values).
- Técnicas de Data Standardisation:
  - Accent Normalization: Remoção de acentuação;
  - Case Standardization: Padronização de maiúsculas/minúsculas;
  - Whitespace Normalization: Remoção de espaços extras;
  - Data Type & Format Standardization: Correção de formato de saída (CSV → xlsx).
- Técnicas de enriquecimento
  - Quality Reporting: Geração de Relatório de Qualidade (Contagem de registros ao final do script);
  - Data Lineage & Versioning: Timestamp no nome do arquivo

2. Identificar as Métricas e KPIs relevantes ao assunto tratado pelo arquivo Entregável

A métricas criadas foram criadas com o objetivo de atender uma persona que atua como Gestor Eleitoral (TSE/TRE).

Objetivo: Monitorar a integridade e eficiência do processo eleitoral.

Métricas relevantes:

QT_SECOES_NAO_INSTALADAS → Quantidade de seções eleitorais não instaladas.

QT_TOTAL_SECOES → Quantidade de seções eleitorais disponíveis.

QT_COMPARECIMENTO → Quantidade de eleitores que compareceram para votar.

KPIs:

% de seções não instaladas = (QT_SECOES_NAO_INSTALADAS / QT_TOTAL_SECOES) x 100

Taxa de abstenção = (QT_ABSTENCOES / QT_APTOS) x 100, cruzar com a taxa de comparecimento = (QT_COMPARECIMENTO / QT_APTOS)×100