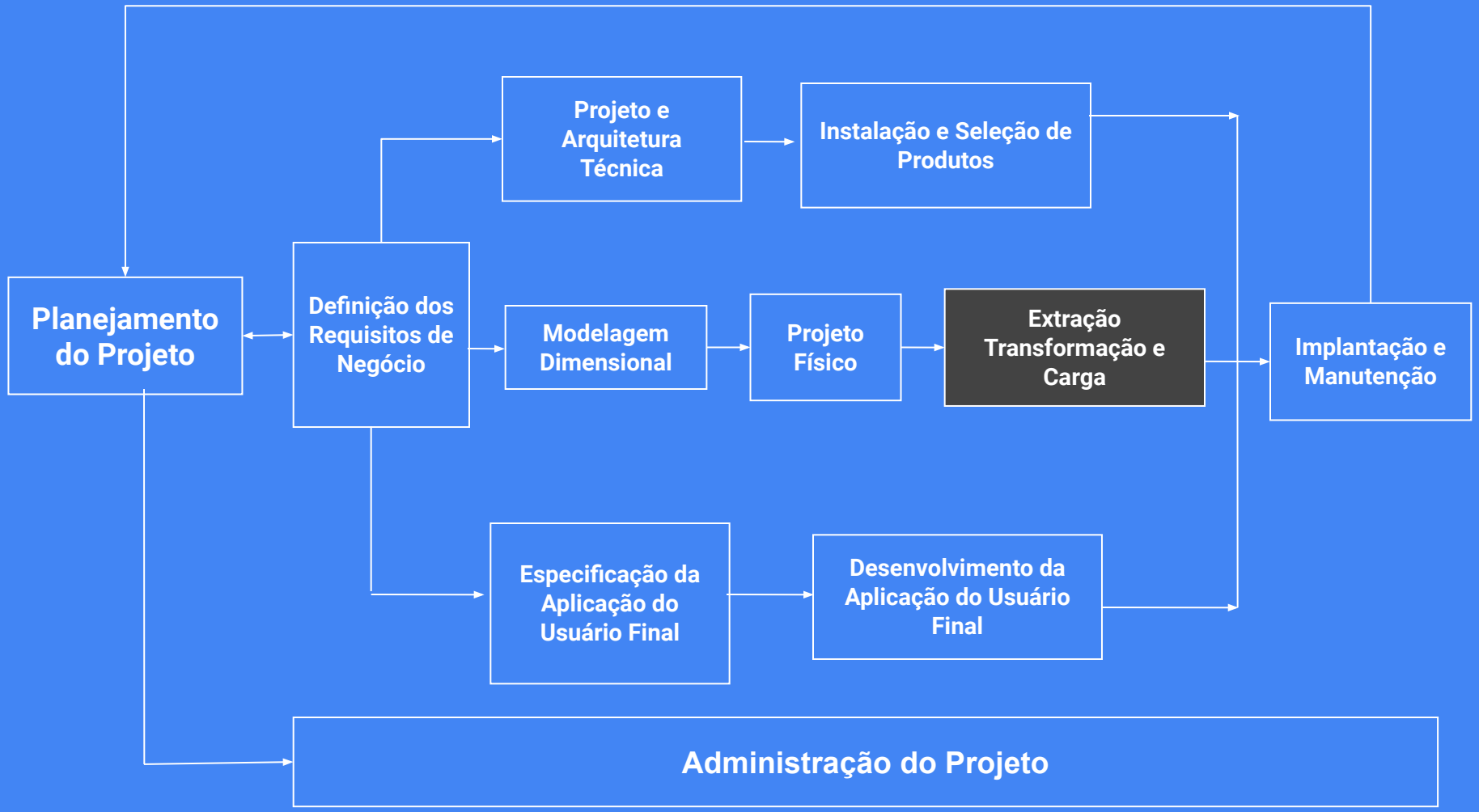
INE 5643
Data Warehouse
Aula 11 - Back Room - Análise e Transformação

Prof. Mateus Grellert
Prof. Renato Fileto
Créditos: Prof. Tite Todesco **(slides originais, adaptados pelos professores atuais)**

Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

Próxima Aula - Ciclo de Projeto DW



Introdução

- Na aula passada, vimos como extrair os dados de fontes e transformá-los em dimensões do nosso DWH

- Uma etapa importante que pode ser realizada após a extração consiste em fazer uma **exploração preliminar** dos dados
  - Entender um pouco sobre os dados
  - Detectar inconsistências
  - Analisar dados faltantes

Processo Geral de Exploração de Dados

Análise e Limpeza

- estatísticas sumarizadas
- detecção de outliers
- valores anômalos
- valores faltantes
- limpeza de dados
- normalização de fontes distintas
- deduplicação

Visualização

- boxplots
- mapas de calor
- histogramas e distribuições
- correlação

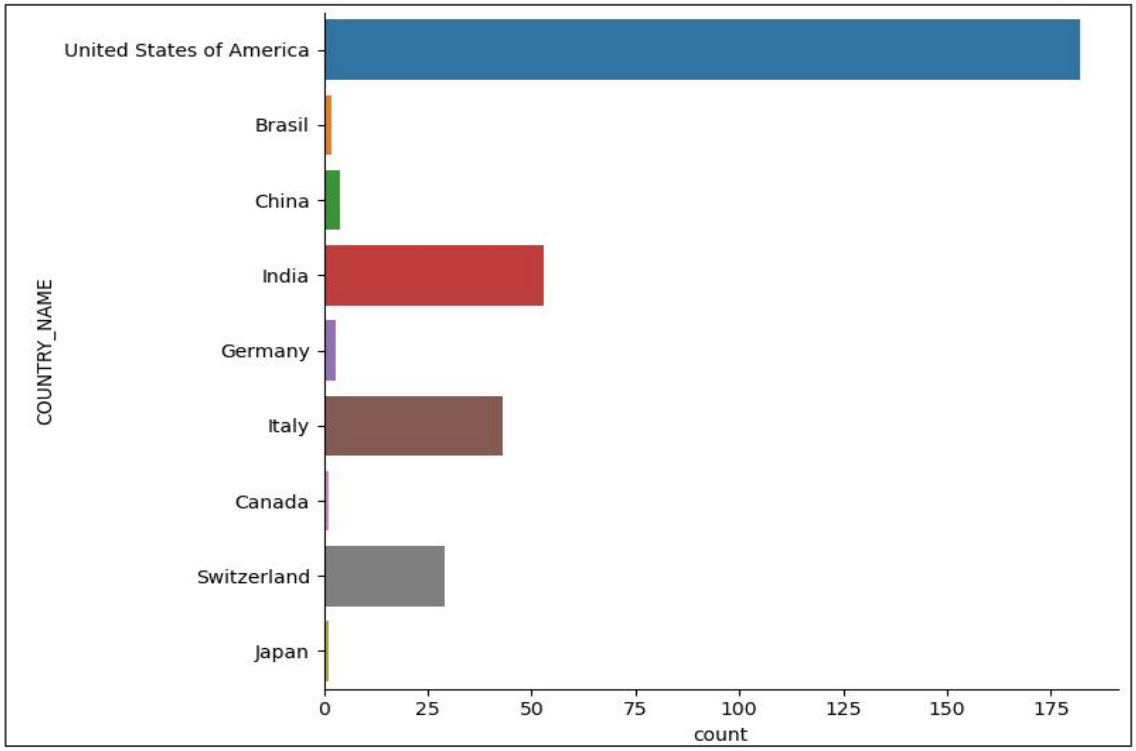
Transformação

- normalização global
- remoção de dados irrelevantes
- engenharia de dados

Analisando Dados

- Algumas medidas comuns para dados **numéricos**:
  - Média
  - desvio padrão
  - Mínimo
  - Máximo
- Algumas medidas comuns para dados **categóricos**:
  - Número de valores distintos
  - Maior, menor tamanho para strings
  - Primeira e última data
  - Frequência de cada valor

Analisando Dados Categóricos



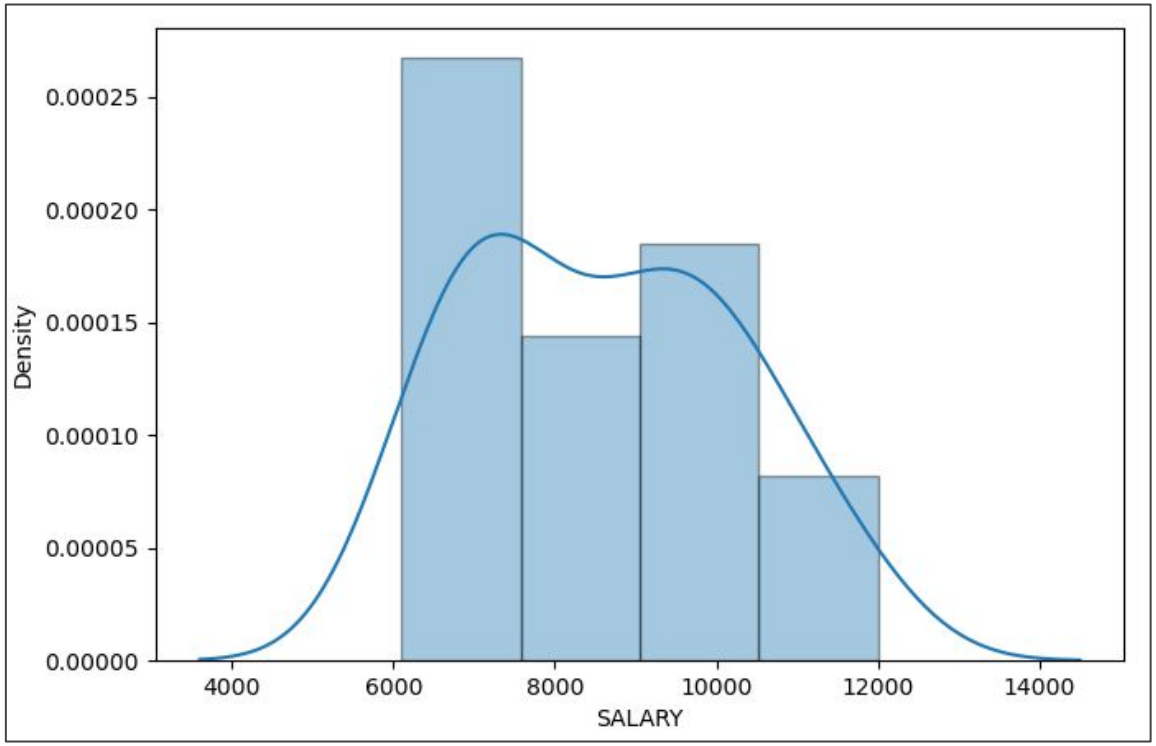
Exemplo para a tabela **Clientes** do CSV que usamos em aula.

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# abrindo o CSV como dataframe
df = pd.read_csv('customers_export.csv', sep = ',')

# plot de dados categoricos do tipo contagem (count)
sns.catplot(y='COUNTRY_NAME', kind = 'count', data = df)
# mostra na tela
plt.show()
```

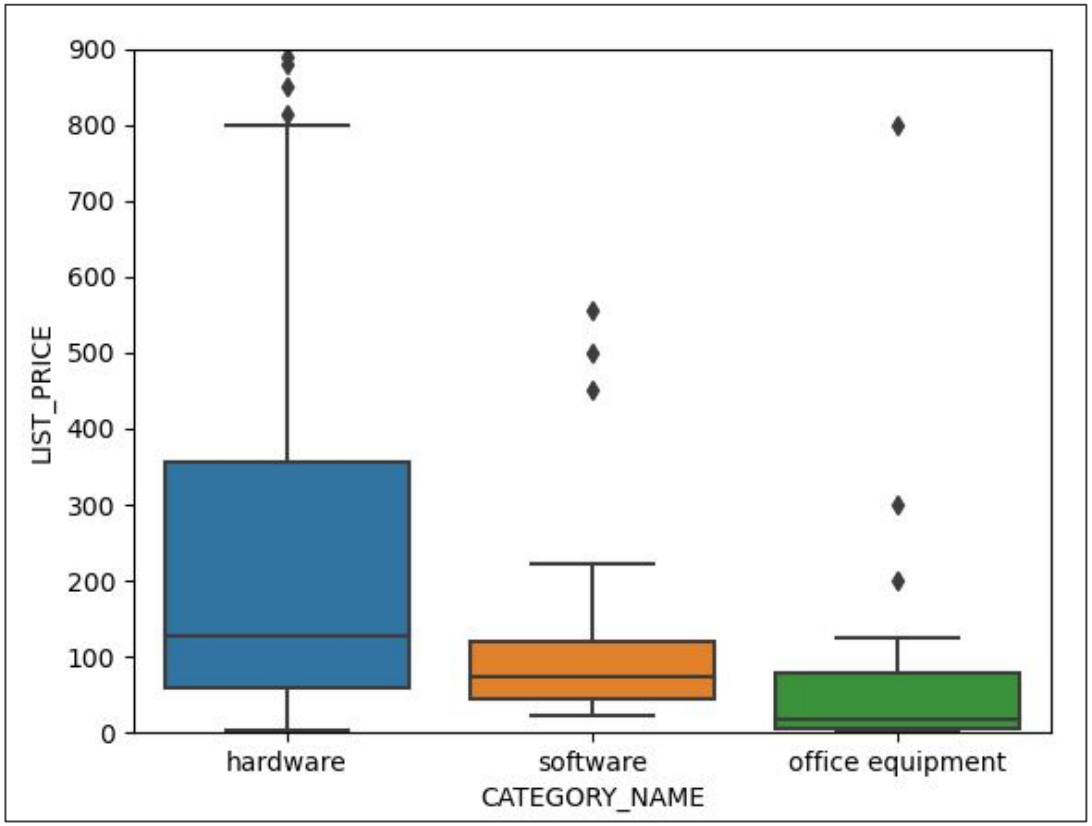
Analisando Dados Numéricos



Exemplo para a tabela **Representantes** do CSV que usamos em

```
22  df = pd.read_csv('salesrep_export.csv', sep = ',')
23
24  # distplot mostra a distribuicao com histograma de dados numericos
25  sns.distplot(df['SALARY'],hist_kws={'ec':'k'})
26  # mostra na tela
27  plt.show()
```

Analisando Dados Numéricos

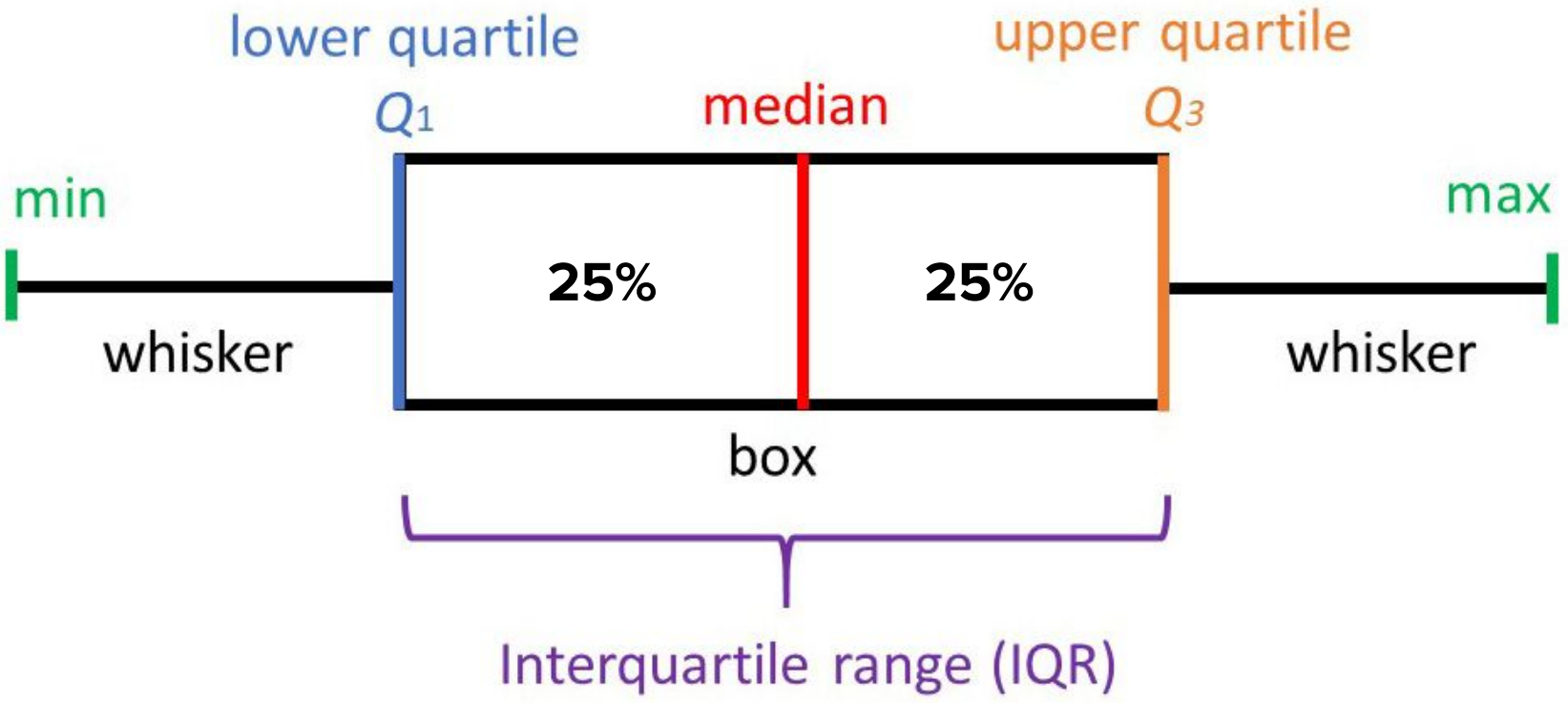


Exemplo de boxplot para a tabela **Produtos** do CSV que usamos

```
14  df = pd.read_csv('products_export.csv', sep = ',')
15
16  # boxplot permite que conhecer a distruiçao de valores numericos
17  sns.boxplot(x = 'CATEGORY_NAME', y = 'LIST_PRICE', data = df)
18  # mostra na tela
19  plt.ylim((0,900))
20  plt.show()
```

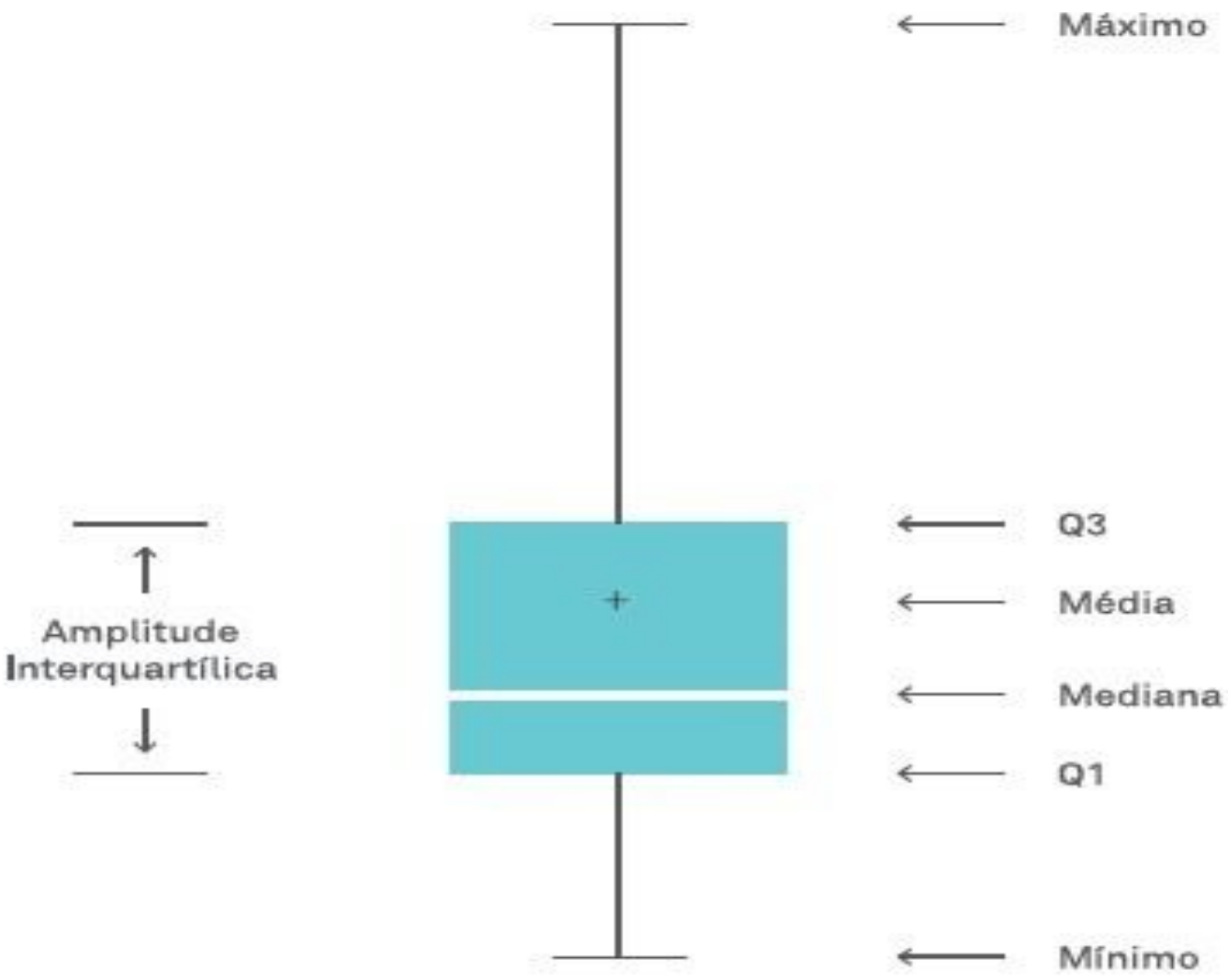
Boxplots

Boxplots apresentam várias informações sobre um dado e são muito utilizados na etapa de análise.



Fonte: https://www.simplypsychology.org/boxplots.html

BoxPlot


Máximo
Q3
Média
Mediana
Q1
Mínimo
Amplitude
Interquartílica

Consolidação dos Dados

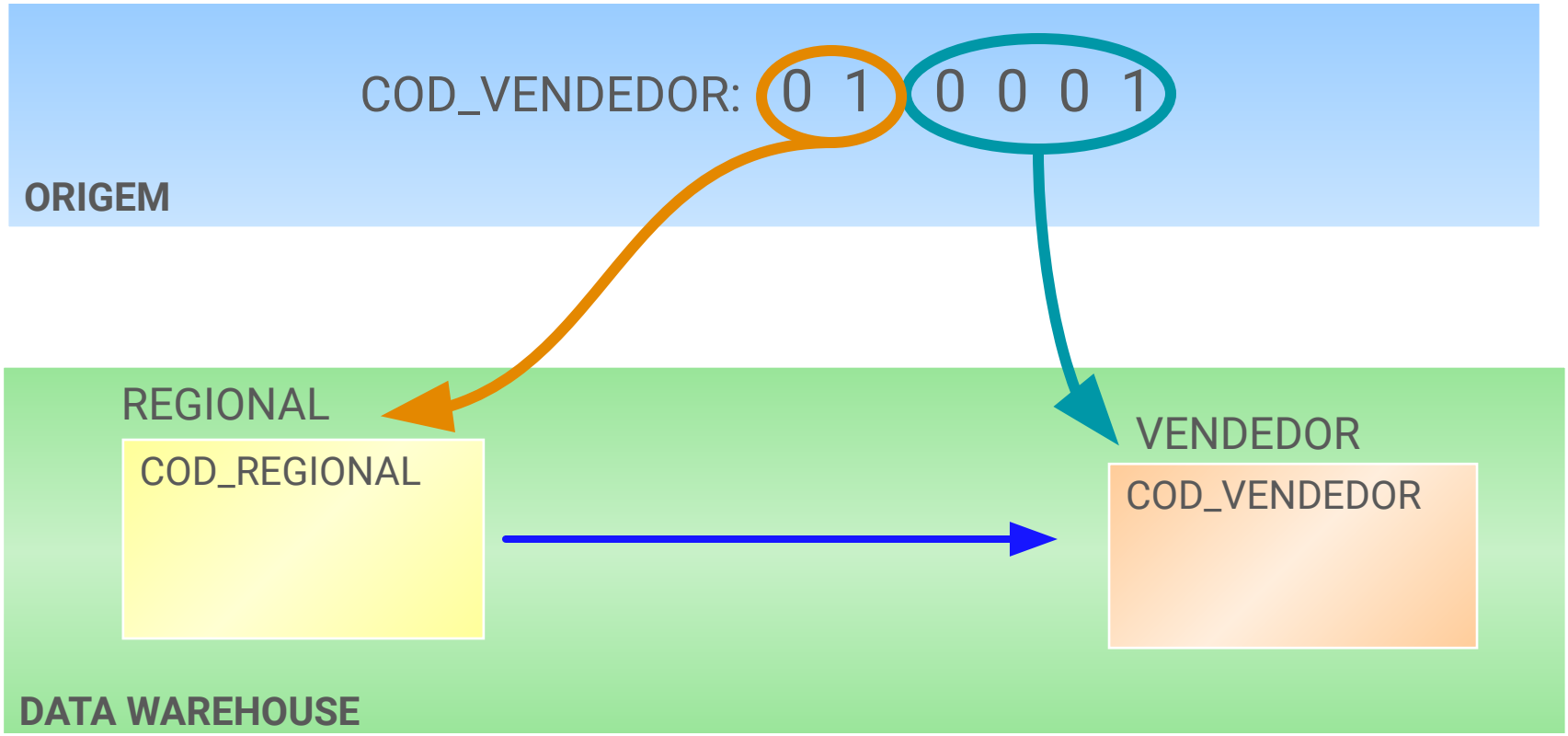
É O PROCESSO DE ANALISAR E COMBINAR DADOS DAS DIVERSAS FONTES EM UMA ESTRUTURA ÚNICA E INTEGRADA

PASSOS

- ANALISAR O **DOMÍNIO** DOS DADOS
- DETERMINAR **CHAVES PRIMÁRIAS** E SECUNDÁRIAS
- RECONCILIAR **SINÔNIMOS**, HOMÔNIMOS E ANÁLOGOS
- ENTENDER AS **REGRAS DE NEGÓCIO** E NUANCES DE SIGNIFICADO

Consolidação dos Dados

EXEMPLO



Conversão dos Dados

É O PROCESSO DE ANALISAR O CONTEÚDO DO DADO, ESPECIFICANDO COMO TRATÁ-LO PARA ESTAR DE ACORDO COM A ESTRUTURA INTEGRADA DO DATA WAREHOUSE

PASSOS

- MAPEAR OS VALORES DO ATRIBUTO NA FONTE E SUA RELAÇÃO COM O ALVO
- ESPECIFICAR VALORES DEFAULT
- ESPECIFICAR REGRAS DE CONVERSÃO

Conversão dos Dados

EXEMPLO

M = MASCULINO

F = FEMININO

FONTE A

1 = MASCULINO

2 = FEMININO

FONTE B

M = MASCULINO

F = FEMININO

b = NÃO IDENTIFICADO

DATA WAREHOUSE

Limpeza dos Dados

É O PROCESSO DE **CORREÇÃO** DOS DADOS IRREGULARES

PASSOS

- ESTABELECER UM CONJUNTO DE VALORES COMO **REFERÊNCIA**
- **AUDITAR** OS DADOS DE ENTRADA CONSIDERANDO ESTA REFERÊNCIA
- PESQUISAR CADA ELEMENTO DE ENTRADA QUE NÃO ESTÁ DE ACORDO COM A REFERÊNCIA
- COM BASE NESTA PESQUISA, REJEITAR ESTE ELEMENTO OU USÁ-LO PARA ATUALIZAR OS VALORES DE REFERÊNCIA

Limpeza dos Dados

EXEMPLOS

- VALORES DUMMY
  - CEP: 04014-012
  - DATA DE NASCIMENTO: 11-11-1111
  - SALÁRIO: R$ 99.999.999,99 (CLIENTE É UM EMPREGADO)
- AUSÊNCIA DE DADO
- CAMPOS COM VÁRIOS PROPÓSITOS
- DADOS CRIPTOGRAFADOS
- DADOS CONTRADITÓRIOS
- USO INADEQUADO DAS LINHAS DE ENDEREÇO
- VIOLAÇÃO DE REGRAS DE NEGÓCIO
- IDENTIFICADORES SEM UNICIDADE

Dados Faltantes

- Seguidamente não temos dados em algumas entradas dos nossos registros
  - Dados perdidos no processo
  - Dados preenchidos por humanos
  - ...
- Nas análises de BI, as linhas com dados faltantes **não são necessariamente um problema**
- Em **Data Mining**, a coisa muda de figura
  - Valores faltantes representam menos entradas utilizadas na geração de modelos
- Uma solução possível consiste em **imputar valores faltantes** utilizando alguma heurística

Geração de Dados (Data Engineering)

- Essa etapa da Transformação consiste gerar novos atributos que podem enriquecer as análises na fase OLAP
- A engenharia de dados usa os dados provenientes das fontes como entrada
- Esses dados podem ser tanto **atributos** de dimensões, como **fatos**
- Exemplos:
  - Flags dia_util, feriado, hora_comercial para a dimensão Tempo
  - Valores PIB_municipio, IPCA_regiao
  - Medidas específicas de domínio

Geração de Dados - Tempo

É possível derivar de um campo data (do calendário):

- Ano, semestre, trimestre, bimestre, mês, dia,...
- Dias da semana
- Dias úteis
- Fins de semana
- Datas especiais - feriados, eventos importantes, etc
- Estação do Ano
    - Essas duas últimas dependem de um referência ao país/cidade

É possível derivar de um campo hora (do dia):

- Hora comercial
- am/pm
- Turno do dia
- ...

Geração de Dados - Espaço

Costuma ser possível derivar de uma indicação de local ou de coordenadas geográficas:

- Continente, país, região, cidade, ...
- Hemisfério
- Idioma
- Índices socioeconômicos: IDH, IPCA, PIB
  - Discretizados em faixas/níveis
- Habitantes

Geração de Dados - Pessoa

- Faixa etária
- Sexo
- Cor
- Identidade de Gênero
- Profissão
- Renda
- Valores específicos de domínio

Sempre que trabalhamos com dados pessoais, é preciso ter muito cuidado com questões de **privacidade**.

O Brasil possui uma legislação específica para esses casos - a Lei Geral de Proteção de Dados (**LGPD**).

Outras Transformações

- **Conversões simples**: conversão de tipos de dados, atribuição de caixa alta e baixa a rótulos, etc;
- **Utilização de chaves artificiais**: através de uma *sequence* ou verificação da última chave gerada em tabela de mapeamento entre chaves naturais (operacionais) e chaves artificiais;
- **Junção de atributos de várias fontes**: junção de dados de variadas origens através de códigos ou critérios difusos;
- **Validação de relacionamentos um-para-um**: ordenação de valores e comparação entre duas fontes de dados para prevenção de repetição de valores.

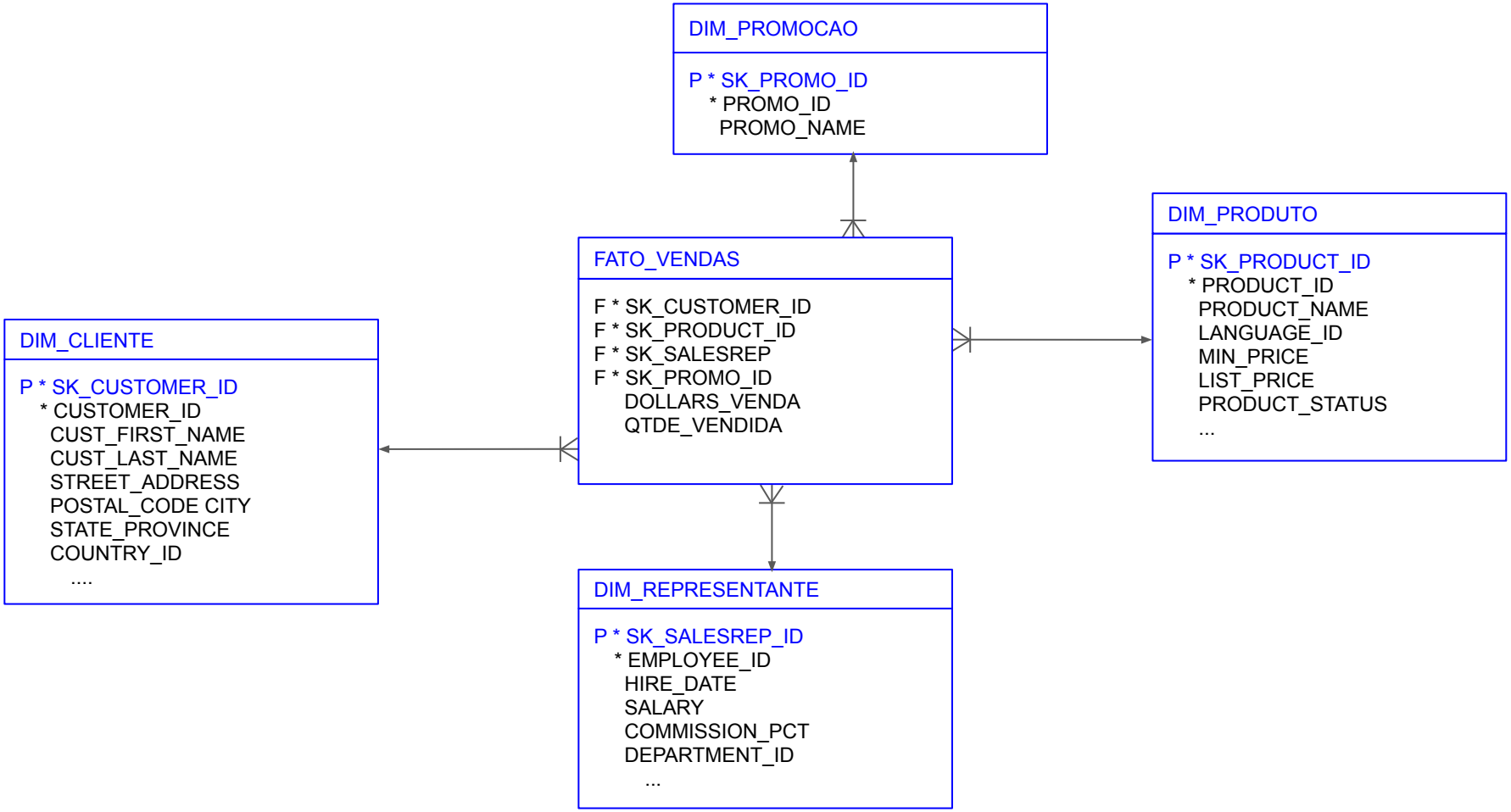
Bibliotecas Python

As seguintes bibliotecas possuem diversos métodos e estruturas de dados úteis para essa parte do projeto:

- **Matplotlib** (plots básicos)
- **Numpy** (cálculo eficiente sobre arrays)
- **Pandas** (DataFrames e métodos para análise de dados)
- **Seaborn** (plots avançados)
- **BeautifulSoup** (para scraping)
- **NLTK** ou **SpaCy**, por exemplo (para processamento de linguagem natural)

Um Exemplo de Transformação com PDI

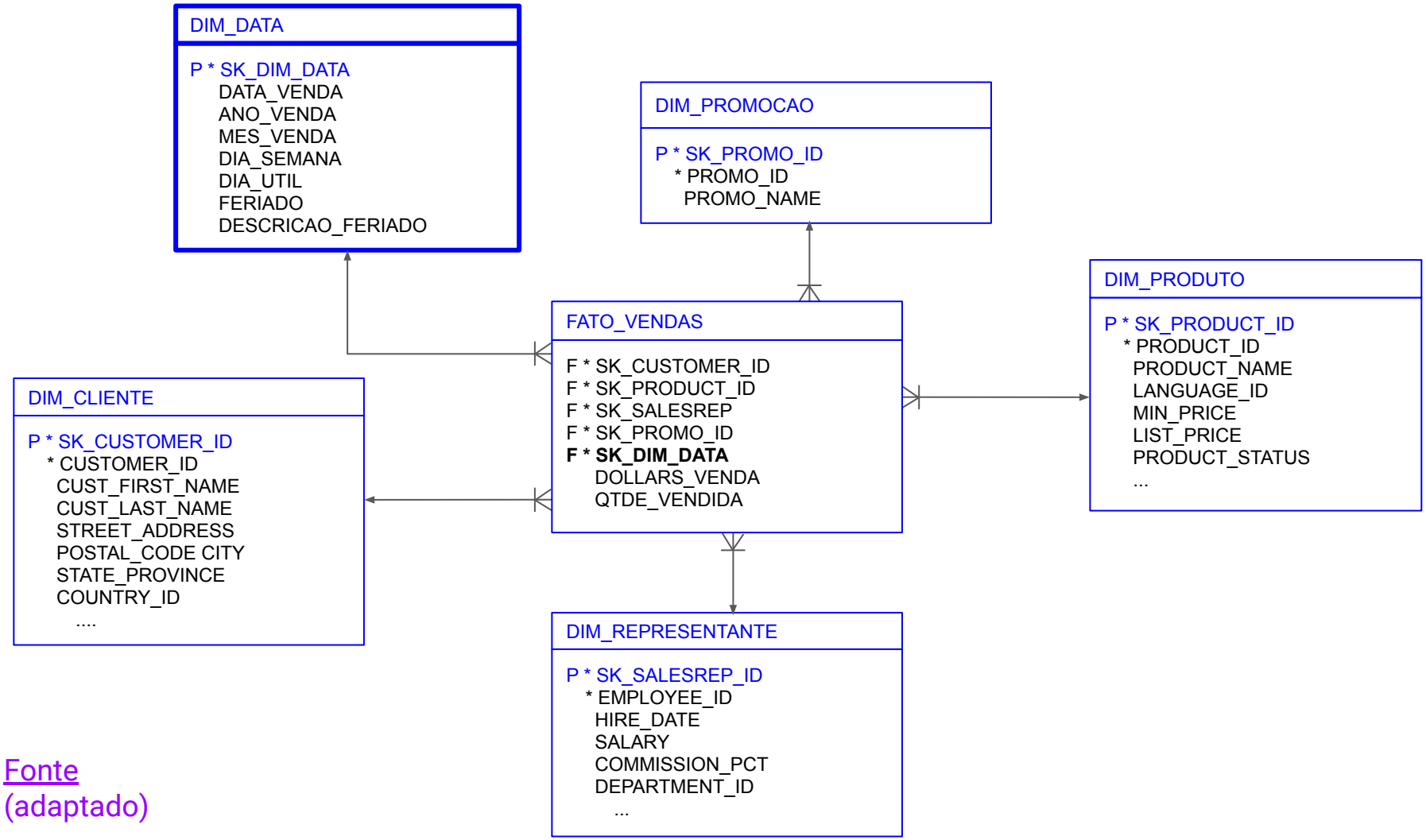
- Essa é versão do DW/DM da aula passada



Fonte: https://holowczak.com/building-etl-transformations-in-pentaho-data-integration-kettle/ (adaptado)

Um Exemplo de Transformação com PDI

- Vamos criar uma nova dimensão **Data** no Pentaho



Fonte
(adaptado)

Por hoje é só! 💪

Próxima aula:

- Carga na Tabela Fato
- Prática no PDI