**21.103 Sistemas de Gestão de Bases de Dados**

**Atividade Formativa Pre-processing**

Considere a seguinte sequência de estados no pré-processamento dos dados:

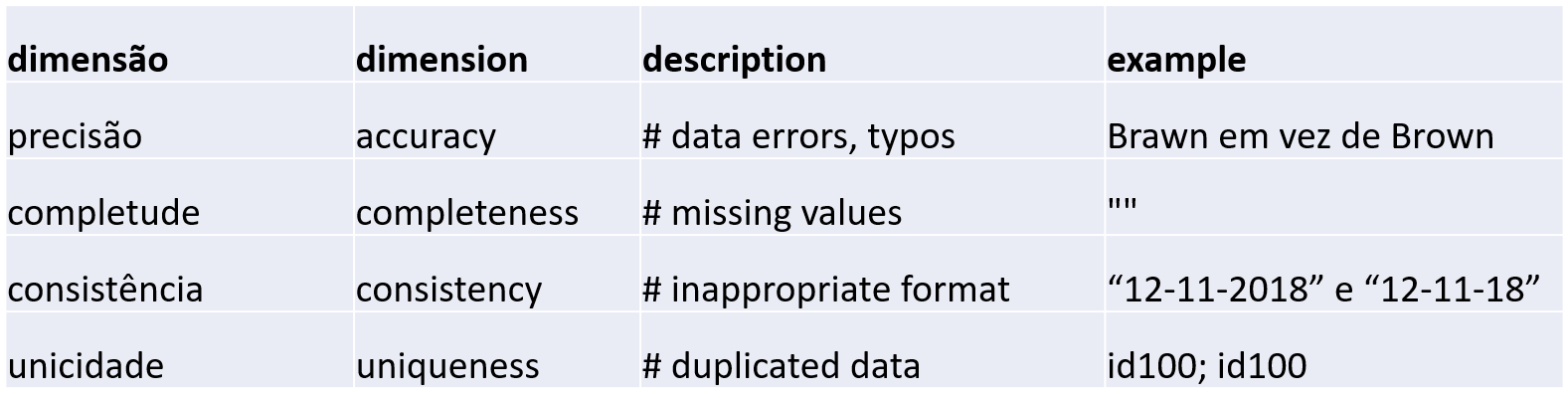
Data Collection -> Data Cleaning -> Data transformation -> Data Reduction



Ponto prévio)

O pré-processamento dos dados tem vários sinónimos como ETL (extract, transformation and loading), Feature Engineering e mais recentemente Data Engineering.

Em particular na Qualidade Dados ou Limpeza Dados temos várias dimensões:



Exemplos da falta de Qualidade Dados, no ano 2000:

- dados de preços errados poderão custar $2.5 biliões/ano;

- cerca de 14% é gasto faturação inadequada no sistema de saúde da US Attorney General;

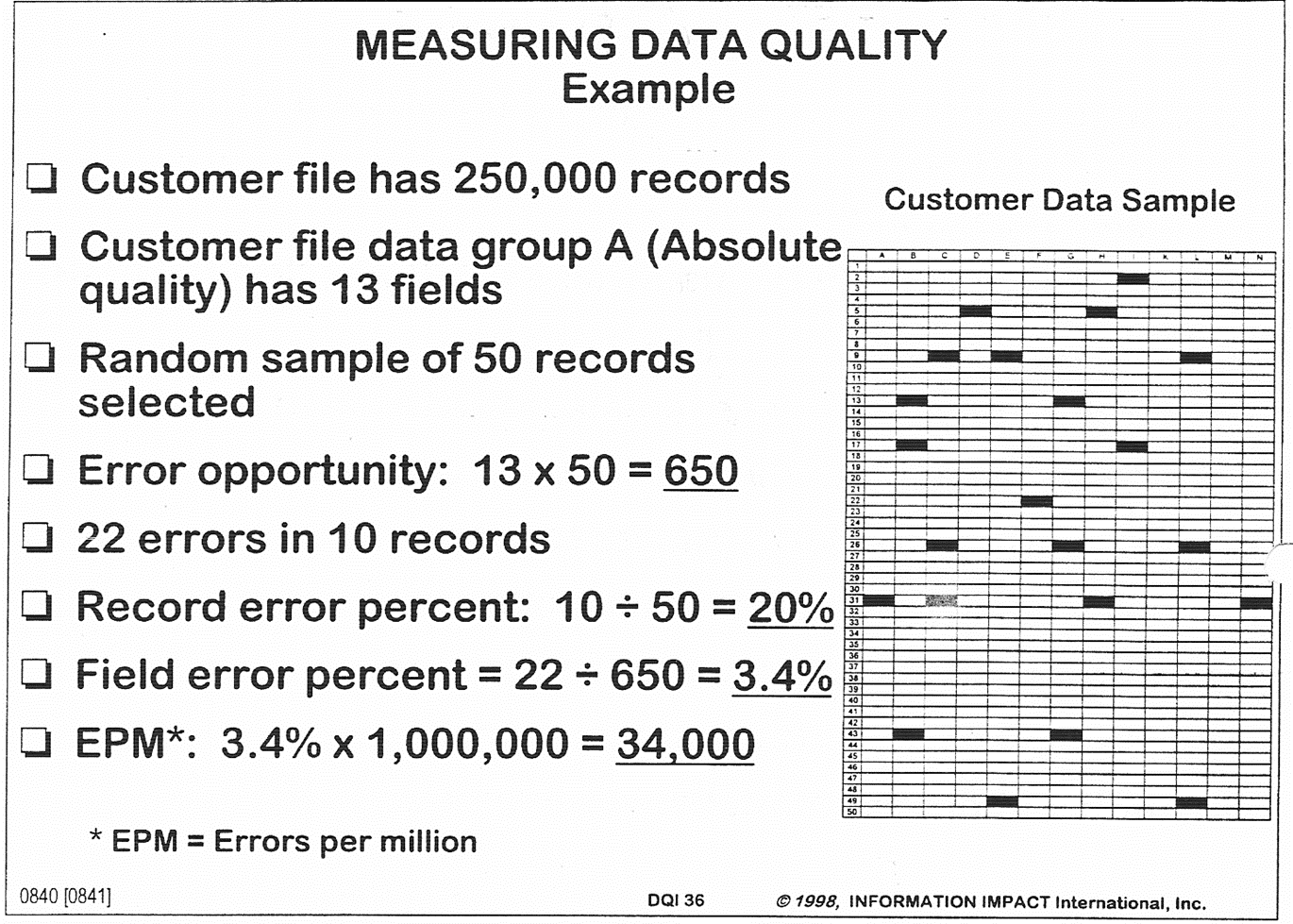
- datas mal formatadas produzem erros na faturação, uma empresa europeia descobriu que não faturava 4% das suas encomendas;

- o problema de qualidade de dados no ano 2000 fez com que 50 das maiores empresas tenham gasto $ 1.5 triliões;

- Barbra Streisand retirou o seu investimento porque se enganaram no seu nome (Barbara);

- 96.000 cheques de reembolso do IRS não foram entregues por endereço “errado”;

- uma escola enviou convites a crianças de 5 anos para apresentação da escola, um dos convites foi para uma senhora de 105 anos.



1) Relativamente à Limpeza dos dados exemplifique os 'missing data' e formas de os ultrapassar.

Existem duas formas de ultrapassar o “missing data” (NA): imputação de dados ou remoção da coluna.

- imputação dos dados usando a média (numéricos), moda (não numéricos) ou mediana;

- se uma coluna tem um percentagem de dados em falta superior a um determinado limite, pode ser removida.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **age** | **sex** | **region** | **income** | **married** | **children** |
| 48 | FEMALE | INNER\_CITY | 17,546.00 | NO | 1 |
| 40 | MALE | TOWN | 30,085.10 | YES | NA |
| 51 | FEMALE | INNER\_CITY | 16,575.40 | YES | NA |
| 23 | FEMALE | TOWN | NA | YES | 3 |
| 57 | FEMALE | RURAL | 50,576.30 | YES | 0 |
| 57 | FEMALE | TOWN | 37,869.60 | YES | NA |
| 22 | MALE | NA | 18,877.07 | NO | NA |
| 58 | MALE | TOWN | 24,946.60 | YES | 0 |
| 37 | FEMALE | SUBURBAN | 25,304.30 | YES | NA |
| 54 | MALE | TOWN | 24,212.10 | NA | NA |
| 66 | FEMALE | TOWN | 59,803.90 | YES | NA |
| 52 | FEMALE | INNER\_CITY | 26,658.80 | NO | 0 |
| 44 | FEMALE | TOWN | 15,735.80 | YES | NA |
| 66 | FEMALE | TOWN | 55,204.70 | YES | 1 |

Para “income” usar a média; para “region” usar a moda; remover coluna “children”.

2) Relativamente à Limpeza dos dados exemplifique os 'data errors' e formas de os ultrapassar.



Existem erros na “region”. Os nomes devem estar associados a dicionários que previnam as chamadas ‘gralhas’ (‘typos’).

3) Relativamente à Limpeza dos dados exemplifique os 'inappropriate formats' e formas de os ultrapassar.

|  |  |
| --- | --- |
| **id\_customer** | **last vist** |
| 7829 | 14/02/2022 |
| 3773 | 09/04/2022 |
| 7270 | 22/04/2022 |
| 4226 | "2022-07-12" |
| 3558 | 08/02/2022 |
| 7140 | 22/04/2022 |
| 5974 | "19/10/22" |
| 5235 | 25/03/2022 |
| 5119 | 01/02/2022 |

Todos os formatos de datas devem ser iguais no exemplo.

Para o efeito devem existir algoritmos que verifiquem a sintaxe do atributo.

4) Relativamente à Limpeza dos dados exemplifique os 'duplicated data' e formas de os ultrapassar.

|  |  |  |
| --- | --- | --- |
| **id\_cliente** | **nome\_cliente** | **morada** |
| I100 | João Silva | rua … |
| I100 | João Silva | rua … |
| … | … | … |

I100 está duplicado e corresponde ao mesmo cliente. Um dos registos deve ser removido.