Análise, construção e visualização de dados

Prof. Dr. Álvaro Campos Ferreira

alvaro.ferreira@idp.edu.br

Monitoria de Pensamento Computacional

Obtenção e Organização dos Dados

Tratamento e limpeza de dados

Qualidade de dados

Dimensões da qualidade de dados:

- Completude ou integridade
- Conformidade
- Validade
- Acurácia e precisão

Qualidade de dados

Problemas	Soluções
Erros de dados	Utilizar entrada de dados automatizada, formulários *web* para entrada de dados individuais, com checagem de integridade de dados, menus de localização e botões de opção
Dados duplicados	Redesenhar o modelo de dados e normalizar o banco de dados relacional
Dados comprometidos	Implementar uma abordagem de defesa profunda à segurança de dados
Dados faltando	Tornar campos obrigatórios nos formulários de entrada de dados

Fonte: Adaptado de Turban *et al.* (2013).

Qualidade de dados

Soluções para quando já não se pode corrigir a coleta:

- Descartar o registro ruim
- Atribuir um valor de sentinela
- Atribuir o valor médio ou mais frequente
- Calcular um valor substituto
- Atribuir um valor baseado em valores vizinhos

idp

Valores faltantes

Valores faltantes são representados por **nan** em DataFrames e não são a mesma coisa que valores nulos.

- Faltas (completamente) aleatórias
- Faltas não aleatórias

Outliers

Outliers são valores encontrados fora de um intervalo razoável com relação às outras observações. É um ponto diferente dos demais, não necessariamente um erro. Sua identificação é importante e pode possuir informações interessantes sobre o sistema.

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA