

Definindo e Identificando Valores Ausentes

Valores ausentes representam falta de informação e não falta de dados, como o termo pode sugerir.

Valores ausentes representam um problema e devem ser tratados.

Em qual das tabelas abaixo podemos identificar valores ausentes?

Idade	Peso (Kg)	Altura (Cm)
23	90	178
47		179



Idade	Peso (Kg)	Altura (Cm)
23	90	178
47	?	179



Idade	Peso (Kg)	Altura (Cm)
23	90	178
47	0	179



Como Decidir Qual a Melhor Estratégia de Tratamento de Valores Ausentes?

Não existe estratégia ideal para tratar valores ausentes.

Devemos analisar sempre cada cenário, cada problema e cada conjunto de dados e tentar gerar o menor impacto possível na limpeza dos dados.

Mas podemos facilmente identificar quais estratégias não devem ser usadas ou devem ser evitadas:

- Não preencha valores ausentes com valores aleatórios ou sem critério.
- Não use a média para preencher valores ausentes de uma variável quantitativa se esta não apresentar uma distribuição normal (moda e mediana seriam opções nesse cenário).
- Não invente categorias para tratar variáveis qualitativas (use termos como “desconhecida” ou “outros”, por exemplo).
- Não faça suposições. Tenha certeza.



Técnicas Para Tratamento de Valores Ausentes

Existem diversas técnicas para tratar valores ausentes:

- Listwise or Case Deletion
- Pairwise Deletion
- Imputation
- Multiple Imputation
- Regression Imputation
- Last Observation Carried Forward/Backward (Forward/Backward Fill)
- Maximum Likelihood
- Expectation-Maximization
- Sensitivity Analysis
- KNN (K Nearest Neighbors)



O Que é Imputação?

Imputação é uma técnica usada para substituir os valores ausentes por algum valor substituto para reter a maioria dos dados/informações do conjunto de dados.

É uma boa prática identificar e substituir os valores ausentes para cada coluna em seus dados antes de realizar o trabalho de análise. Isso é chamado de imputação de valores ausentes ou apenas imputação.

Essa técnica é usada porque a remoção dos dados do conjunto de dados pode não ser viável e pode levar a uma grande redução no tamanho do conjunto de dados, o que não só levanta preocupações quanto ao enviesamento do conjunto de dados, mas também pode levar a uma análise incorreta.



O Que é Imputação?

Dataset com valores
ausentes representados
em vermelho



Dataset com valores
ausentes tratados
representados em verde

