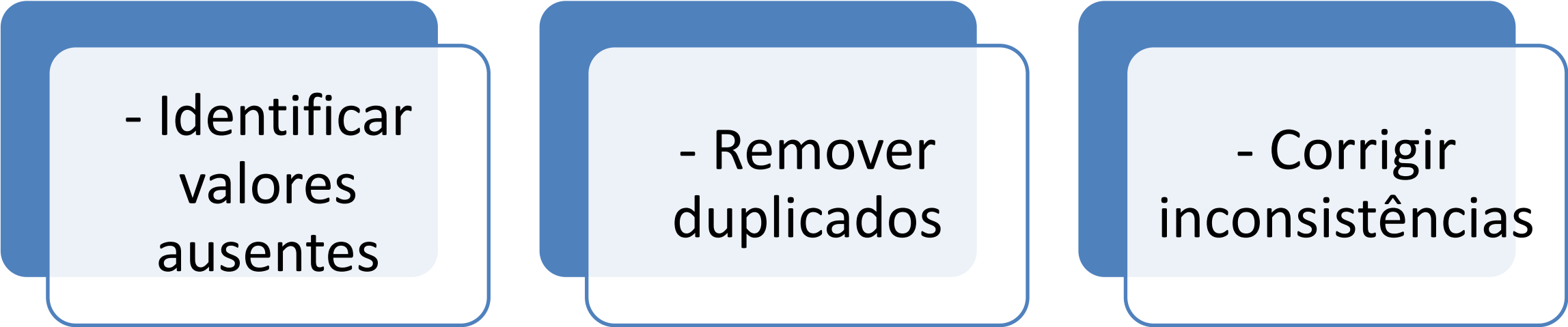


Aula 6 - Pandas: Limpeza de Dados

Tratamento de dados ausentes e duplicados

Objetivos



- Identificar
valores
ausentes

- Remover
duplicados

- Corrigir
inconsistências

O que são valores ausentes?

Valores ausentes (ou *missing values*) ocorrem quando há lacunas nos dados, representadas por **NaN** (*Not a Number*) no Pandas.

Esses valores podem surgir devido a falhas na coleta de dados, sensores defeituosos ou erros humanos.

Identificar valores ausentes

Use o método `.isnull()` ou `.notnull()` para verificar valores ausentes.

Utilize `.sum()` para contar o número de valores ausentes por coluna.

Estratégias para lidar com valores ausentes

1. Remover linhas ou colunas com valores ausentes:

python

```
1 df_limpo = df.dropna() # Remove linhas com NaN
```

2. Preencher valores ausentes:

python

```
1 df['Temperatura'].fillna(df['Temperatura'].mean(), inplace=True) # Preenche com a média
```

Remover Duplicados



Duplicatas podem distorcer análises estatísticas e modelos preditivos, pois dão mais peso a certos dados do que seria apropriado.



Método **.duplicated()** identifica linhas duplicadas.



Método **.drop_duplicates()** remove linhas duplicadas



Corrigir Inconsistências

Inconsistências ocorrem quando os dados não seguem um padrão uniforme, como variações em maiúsculas/minúsculas, espaços extras ou formatação incorreta.

Exemplos de correção de inconsistências

1. Padronizar texto (maiúsculas/minúsculas):

python

```
1 df['Coluna'] = df['Coluna'].str.upper() # Converte tudo para maiúsculas
2 df['Coluna'] = df['Coluna'].str.lower() # Converte tudo para minúsculas
```

2. Remover espaços extras:

python

```
1 df['Coluna'] = df['Coluna'].str.strip() # Remove espaços antes/depois
```

3. Corrigir valores inconsistentes:

python

```
1 df['Coluna'] = df['Coluna'].replace({'valor_errado': 'valor_correto'})
```


Aplicação

Sensores em estufas podem gerar leituras com falhas

Exercício

Tratar dados climáticos com valores ausentes e duplicados