**Para saber mais: formato wide e long**

Os dados em uma tabela podem estar organizados em formatos diferentes de acordo com o modo em que foram preenchidos ou da necessidade da análise que será feita. Os termos **wide** e **long** são usados para descrever a organização dos dados em um banco de dados relacional.

No formato **wide**, também conhecido como *tabelas de pivô*, as tabelas têm uma única linha de cabeçalho contendo diversas colunas, geralmente mais colunas do que linhas, e são úteis para armazenar informações que possuem muitos atributos, como questionários e pesquisas. As colunas podem conter valores ausentes para as células em que os dados não se aplicam.

Exemplo de uma tabela em formato wide:

| **Nível 1 - Setor** | **2017** | **2018** | **2019** | **2020** | **2021** |
| --- | --- | --- | --- | --- | --- |
| Processos industriais | 897.6 | 758.0 | 902.0 | 804.0 | 870.0 |
| Processos industriais | 130.0 | 168.0 | 98.0 | 125.0 | 134.0 |
| Mudança de Uso da Terra e Floresta | 378.03 | 388.89 | 452.0 | 935.0 | 1047.0 |

No formato **long**, também conhecido como *formato normalizado* ou *estreito*, as tabelas armazenam cada observação em uma linha separada. Ao invés de ter uma única linha de cabeçalho contendo todas as colunas, as tabelas no formato long geralmente têm duas colunas principais de informação, uma para os nomes das variáveis e outra para os valores das variáveis. São úteis para armazenar informações com muitas observações, como séries temporais ou dados de experimentos científicos.

Exemplo de uma tabela em formato long:

| **Nível 1 - Setor** | **Ano** | **Emissão** |
| --- | --- | --- |
| Processos industriais | 2017 | 897.6 |
| Processos industriais | 2017 | 130.0 |
| Mudança de Uso da Terra e Floresta | 2017 | 378.03 |
| Processos industriais | 2018 | 758.0 |
| Processos industriais | 2018 | 168.0 |
| Mudança de Uso da Terra e Floresta | 2018 | 388.89 |
| Processos industriais | 2019 | 902.0 |
| Processos industriais | 2019 | 98.0 |
| Mudança de Uso da Terra e Floresta | 2019 | 452.0 |
| Processos industriais | 2020 | 804.0 |
| Processos industriais | 2020 | 125.0 |
| Mudança de Uso da Terra e Floresta | 2020 | 935.0 |
| Processos industriais | 2021 | 870.0 |
| Processos industriais | 2021 | 134.0 |
| Mudança de Uso da Terra e Floresta | 2021 | 1047.0 |

Na biblioteca pandas, é possível fazer uma mudança do formato wide para o formato long com o uso do método [melt()](https://pandas.pydata.org/docs/reference/api/pandas.melt.html" \t "_blank). A sintaxe para utilizar o método melt() é a seguinte:

dataframe.melt(id\_vars = [...], value\_vars = [...], var\_name = '..', value\_name = '..')

O parâmetro id\_vars é utilizado para identificar os nomes das colunas que serão mantidas no momento da transformação. Já o parâmetro value\_vars é utilizado para identificar as colunas que serão transformadas nas colunas principais do formato long, uma contendo as categorias com os nomes das colunas e outra contendo os valores. O parâmetro var\_name é usado para nomear a coluna contendo as categorias e o parâmetro value\_name para nomear a coluna com os valores numéricos.

**Para saber mais: como funciona o groupby**

s grupos são gerados a partir das categorias de uma ou mais colunas da tabela com base na informação que desejamos agrupar. Nesta etapa, podemos visualizar os grupos formados com o atributo groups e visualizar os dados de um grupo específico a partir do método get\_group().

Já a etapa "aplicar" se refere ao momento de utilizar uma função matemática nos dados de cada grupo separadamente. Se o desejo for encontrar o valor total de cada grupo, então a soma será aplicada a cada grupo; se quisermos o valor médio, a função média será aplicada a cada grupo, e assim por diante. Existem diversas funções que podem ser aplicadas e cada uma dependerá do objetivo de análise dos dados:

* mean() : calcula a média dos grupos
* sum() : calcula a soma dos valores do grupo
* size() : calcula tamanhos de grupo
* count() : calcula a contagem do grupo
* std() : desvio padrão de grupos
* var() : calcula a variação de grupos
* sem() : erro padrão da média dos grupos
* describe() : gera estatísticas descritivas
* first() : calcula o primeiro dos valores do grupo
* last() : calcula o último dos valores do grupo
* nth() : pega o n-ésimo valor do grupo
* min() : calcula o mínimo de valores do grupo
* max() : calcula o máximo de valores do grupo

Por fim, a etapa "combinar" consiste em juntar todas as informações de agregação de cada grupo em uma única tabela, contendo o resumo das informações. A partir dela e com o resultado final, é possível extrair insights a partir dos dados e fazer comparações dos resultados entre cada grupo.