Faxina de dados

Integração

curso-r

Integração

Unidade amostral

Unidade amostral é o que define as observações de uma base de dados. Por exemplo, uma base pode ter como unidade amostral processos, ou como unidade amostral decisões, já que é possível ter mais de uma decisão por processo.

Cabe à pessoa gestora dos dados decidir como esses dados devem ser disponibilizados para análise.

Usualmente, a unidade amostral precisa de um **número identificador único** para permitir a integração com outras bases de dados.

Join

Join é a tarefa de juntar bases de dados pelas **linhas** que compartilham **variáveis** com **valores** em comum.

Exemplos:

- Juntar base de municípios com outras bases
- Juntar base de empresas com base da Receita
- Juntar bases internas da empresa

Tipos de join

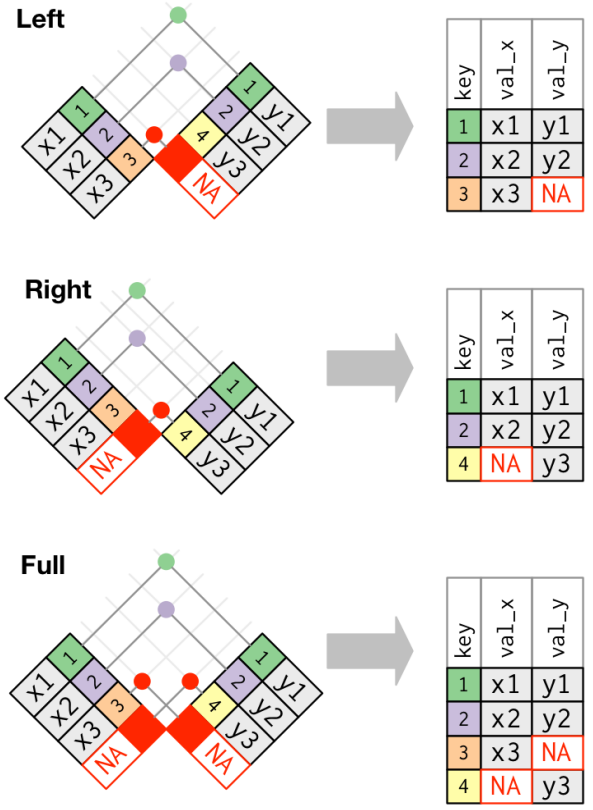

Imagem retirada do livro R4DS.

Partial matching

"Mogi-Mirim" pode estar escrito na forma "Mogi Mirim".

Quando fazemos o join diretamente com os dados assim, os programas são incapazes de reconhecer que estamos trabalhando com a mesma comarca.

Nesse caso, temos duas alternativas:

- Adicionar rotinas de arrumação de dados: caixa alta/baixa, acentos, traços, padronização do número de caracteres, entre outros.
- Utilizar mecanismos de partial matching, como Fuzzy Join (por distância de Levenshtein, Jaccard etc), ou de forma mais sofisticada com Record Linkage

```
stringdist::stringdist("Mogi-Mirim", "Mogi-Mirim")
stringdist::stringdist("Mogi-Mirim", "Mogi Mirim")
```

```
[1] 0
[1] 1
```

Exercício

Temos duas bases de processos judiciais e queremos juntá-las.

Uma vem da base interna da empresa e outra vem dos tribunais (dados públicos). Ambas têm como unidade amostral o processo, identificadas pelo número CNJ dos processos.

Ao fazer o join das bases, no entanto, obervamos 0 matchings. O que fazer?

	base1		base2
1	01405062220208260003	1	0378496-60.2020.8.26.0004
2	037849620208260004	2	0140506-22.2020.8.26.0003
3	06369773820208260002	3	0426277-81.2020.8.26.0003
4	016756916202082600005/001	4	0167569-16.2020.8.26.0005
5	4262778120208260003	5	0636977.38.2020.8.26.0002