

CIÊNCIA DE DADOS (BIG DATA)

ANÁLISE ESTATÍSTICA

Professor curador: Mário Olímpio de Menezes



Mackenzie



TRILHA 3

PARTE A – ANÁLISE EXPLORATÓRIA DE DADOS

PARTE A – ANÁLISE EXPLORATÓRIA DE DADOS

TIDY DATA

ORGANIZANDO O CONJUNTO DE DADOS

TIDY DATA

Um conjunto de dados é chamado *tidy* quando:

- Cada coluna representa uma variável.
- Cada linha representa uma observação.

Conjuntos de dados *tidy* são fáceis de modelar, visualizar, manipular.

Necessários para **análise exploratória de dados**.

TIDY DATA

Variáveis

Número Registro	Atributo 1	Atributo 2	Atributo 3	Atributo 4
1	Sim	Solteiro	125.00	Casa
2	Não	Casado	150.000,00	Apartamento

Observações

Valor

EXPERIMENTO – CRIAÇÃO DE TABELA DE DADOS

Estamos observando um grupo de 10 pessoas:

- 5 do sexo masculino.
- 5 do sexo feminino.

Criar uma tabela que sumarieze este grupo com relação a:

- Sexo.
- Estado de gravidez (Grávida ou Não Grávida).

EXPERIMENTO – CRIAÇÃO DE TABELA DE DADOS

- Quantas e quais são as variáveis neste experimento?
- Quantas observações temos?
- Como organizaríamos esta tabela?

EXPERIMENTO – CRIAÇÃO DE TABELA DE DADOS

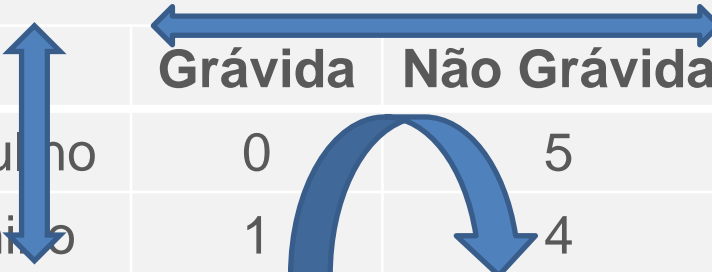
Uma maneira muito provável para a construção da tabela seria esta:

Sexo	Grávida	Não Grávida
Masculino	0	5
Feminino	1	4

TIDY DATA

Existem três variáveis neste conjunto de dados.

Quais são elas?



Sexo	Grávida	Não Grávida
Masculino	0	5
Feminino	1	4

CAUSAS COMUNS DE BAGUNÇA NOS DADOS

- Cabeçalhos das colunas são valores, não nomes de variáveis.
- Células são nomes de variáveis, não valores.
- Dados são espalhados em múltiplos arquivos.
 - Às vezes, isso ocorre porque estamos querendo juntar várias bases de dados.

DADOS NÃO *TIDY* – DESARRUMADOS

```
> head(raw)
```

	religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
1	Agnostic	27	34	60	81	76	137
2	Atheist	12	27	37	52	35	70
3	Buddhist	27	21	30	34	33	58
4	Catholic	418	617	732	670	638	1116
5	Don't know/refused	15	14	15	11	10	35
6	Evangelical Prot	575	869	1064	982	881	1486

ACERTANDO CONJUNTO DE DADOS *BAGUNÇADO*

Pacote

tidyr

do R

```
> library(tidyr)
```

```
gather(dados, key = "Coluna chave",  
       value = "Coluna valor",  
       <col_inic:col_fin>  
)
```

ARRUMANDO DADOS *NÃO TIDY*

```
> gather(naotidy, "Estado", "n", 2:3)
```

sexo	Estado	n
Masculino	Gravida	0
Feminino	Gravida	1
Masculino	NaoGravida	5
Feminino	NaoGravida	4

ARRUMANDO DADOS *NÃO TIDY*

```
> tidy <- gather(raw, key = "income", value = "n",  
  2:11)
```

```
> head(tidy)
```

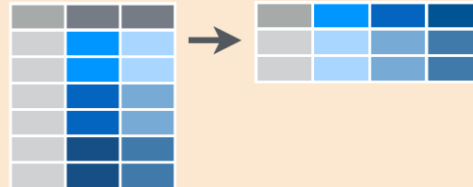
	religion	income	n
1	Agnostic	<\$10k	27
2	Atheist	<\$10k	12
3	Buddhist	<\$10k	27
4	Catholic	<\$10k	418
5	Don't know/refused	<\$10k	15
6	Evangelical Prot	<\$10k	575

ACERTANDO DADOS BAGUNÇADOS



`tidyr::gather(cases, "year", "n", 2:4)`

Junta colunas em linhas



`tidyr::spread(pollution, size, amount)`

Separa linhas em colunas



`tidyr::separate(storms, date, c("y", "m", "d"))`

Separa um coluna em várias



`tidyr::unite(data, col, ..., sep)`

Unifica várias colunas em uma

