

## Problemas básicos ao fundir bancos

### Transcrição

Obtivemos diversas informações interessantes a partir dos bancos de dados.

Aprendemos como descobrir quais vídeos são os mais assistidos, quais são os cursos que recebem o maior número de matrículas, entre outros.

Vamos tentar fazer uma análise mais avançada. Tentaremos descobrir se existe alguma **correlação** entre a popularidade dos cursos e o tempo que eles levam para ser concluídos.

É uma informação interessante que podemos levar para a empresa, para orientá-la na busca qualitativa, economizando tempo, dinheiro e esforço.

Uma vez que ela não terá que assistir a todos os vídeos e cursos para encontrar *insights* e elementos qualitativos comuns.

Por meio da análise que estamos fazendo, direcionamos a empresa agrupando os vídeos que possuem maior probabilidade de conter elementos comuns que os tornem populares.

O problema é que faltam informações referentes à popularidade, quantas matrículas um curso recebeu e o tempo médio de duração de cada curso em um mesmo banco de dados.

Precisamos juntar as informações em um único lugar.

Na verdade, não temos nem mesmo o tempo médio de duração de cada curso. Precisamos criar um objeto no RStudio que nos forneça essa informação, juntando as médias por curso.

Calcularemos a média de todas as matrículas de um curso e colocaremos no vetor, no objeto.

Sendo assim, criaremos um objeto com o nome de `sumario_estatistico`. Nele, resumizaremos uma estatística, ou seja, colocaremos a média de duração dos cursos calculada:

```
sumario_estatistico <- aggregate(duracao$dias, list(duracao$curso), mean, na.rm = T)
```

A função `aggregate` é responsável por agregar. Entre parênteses, especificamos o banco de dados ( `duracao` ) e a variável ( `dias` , referente ao tempo que os alunos levaram para concluir os cursos) que será trabalhada.

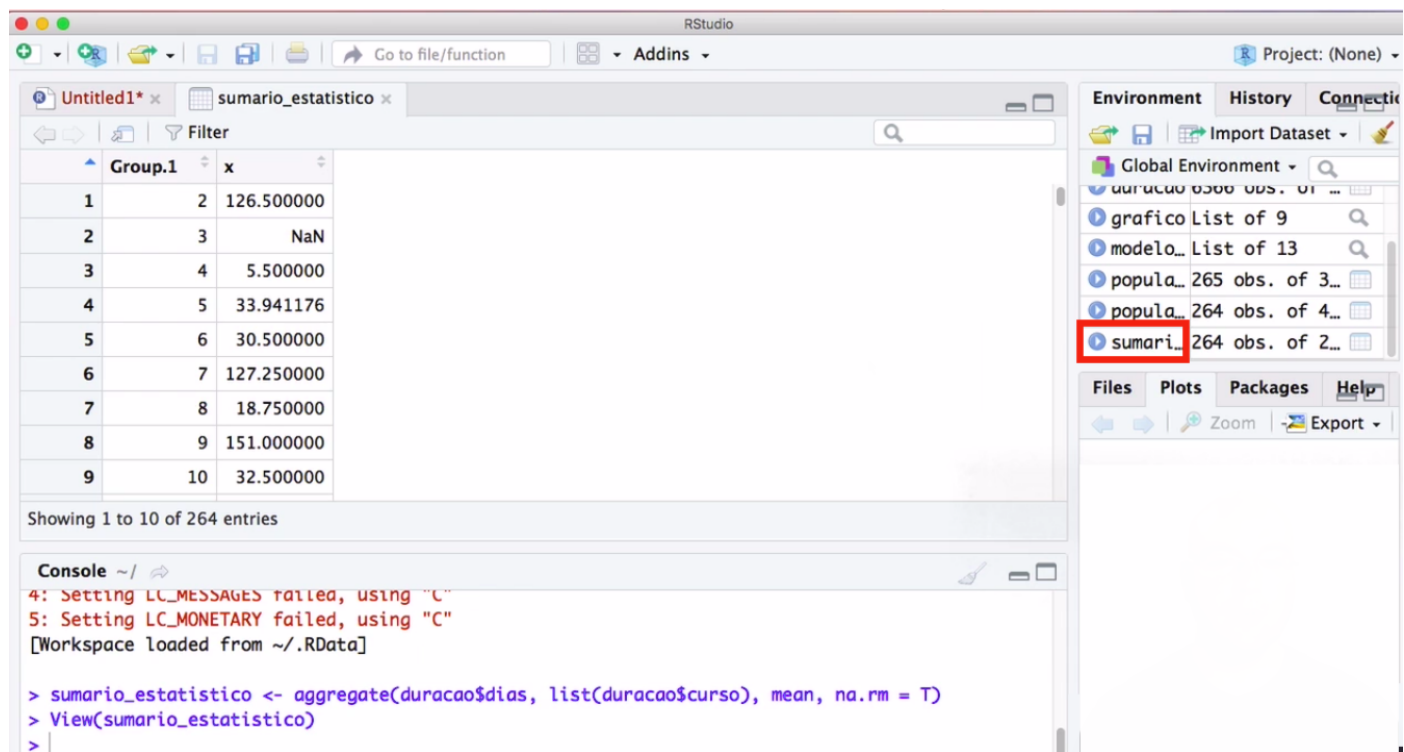
Após a vírgula ( , ), o parâmetro — função dentro de outra função — `list` refere-se à listagem dos cursos. Portanto, especificamos entre parênteses o mesmo banco de dados e consideramos a variável em questão ( `curso` ).

A medida que estamos buscando é a média, por isso acrescentamos `mean`.

Por fim, informamos ao programa que os dados indisponíveis ( `NA` ) devem ser ignorados por meio de `na.rm = T`. Lembrando que `T` equivale a `TRUE`, de verdadeiro.

Ao executar o comando, no Console, verificamos que o sumário foi criado.

Na janela superior direita, clicamos em `sumario_estatistico` e na janela superior esquerda, visualizaremos uma espécie de banco de dados com duas colunas: `Group.1` que informa um número por curso e `x` referente à média de duração de um curso em específico, não de todos.



The screenshot shows the RStudio interface. The main window displays a data frame named `sumario_estatistico` with two columns: `Group.1` and `x`. The data is as follows:

Group.1	x
1	2
2	126.500000
3	NaN
4	5.500000
5	33.941176
6	30.500000
7	127.250000
8	18.750000
9	151.000000
10	32.500000

The console shows the following commands:

```
> sumario_estatistico <- aggregate(duracao$dias, list(duracao$curso), mean, na.rm = T)
> View(sumario_estatistico)
>
```

Na primeira linha da planilha, temos que o curso 2 possui uma média de 126.5 dias para ser concluído.

O curso 3 possui dados indisponíveis `NaN`. O curso 4 tem uma média de 5.5 dias para ser concluído, e assim seguem os dados da planilha.

Agora, precisamos juntar essas duas informações em um mesmo banco de dados.

Agregaremos a duração média à popularidade, ou número de matrículas, de cada curso.

Criaremos um novo objeto `popularidade_e_duracao` para indicar o conteúdo do objeto no nome dele.

Atribuiremos (`<-`) a ele o resultado de um comando que agrega (`merge`) dois bancos de dados. Dentro de `merge`, passaremos os dois objetos ou bancos que queremos reunir.

Queremos juntar `sumario_estatistico` e `popularidade`. Portanto, após vírgula (`,`), faremos a indexação, que é a utilização de uma variável comum aos dois bancos de dados, pela qual o novo banco será construído.

Deve ser uma **variável comum** aos dois bancos de dados, com o mesmo nome.

A sintaxe do comando de indexação é `by = 'curso'`. `by` expressa "por" em inglês e `'curso'`, que está entre aspas simples (`' '`), é uma variável em comum a `popularidade` e `sumario_estatistico`.

No R Script, o comando ficará da seguinte forma:

```
popularidade_e_duracao <- merge(sumario_estatistico, popularidade, by = 'curso')
```

Ao executar, no Console, o retorno será um erro:

```
> popularidade_e_duracao <- merge(sumario_estatistico, popularidade, by = 'curso')  
Error in fix.by(by.x, x) : 'by' must specify a uniquely valid column
```

Ele diz que 'curso' não foi reconhecido como um variável comum aos dois bancos de dados.

Sabemos que a informação da variável `curso` está lá, mas talvez esteja com um nome diferente.

A sintaxe deve ser perfeita.

Para que o comando funcione, deve estar com o nome `curso` em ambos os bancos. Precisamos verificar se o nome deve ser corrigido em um ou mais de um deles.