02

Utilizando a função merge

Transcrição

Na janela superior direita do RStudio, se localizam os objetos que criamos. Ao clicarmos em `sumario_estatistico`, verificamos que não há uma coluna correspondente a `curso` na planilha. Se clicarmos em `popularidade`, acima de `sumario_estatistico` na visualização dessa planilha, ocorre o mesmo. Há uma coluna `course_id`, mas nenhuma possui o nome `curso`, utilizado como referência para juntar os bancos de dados.

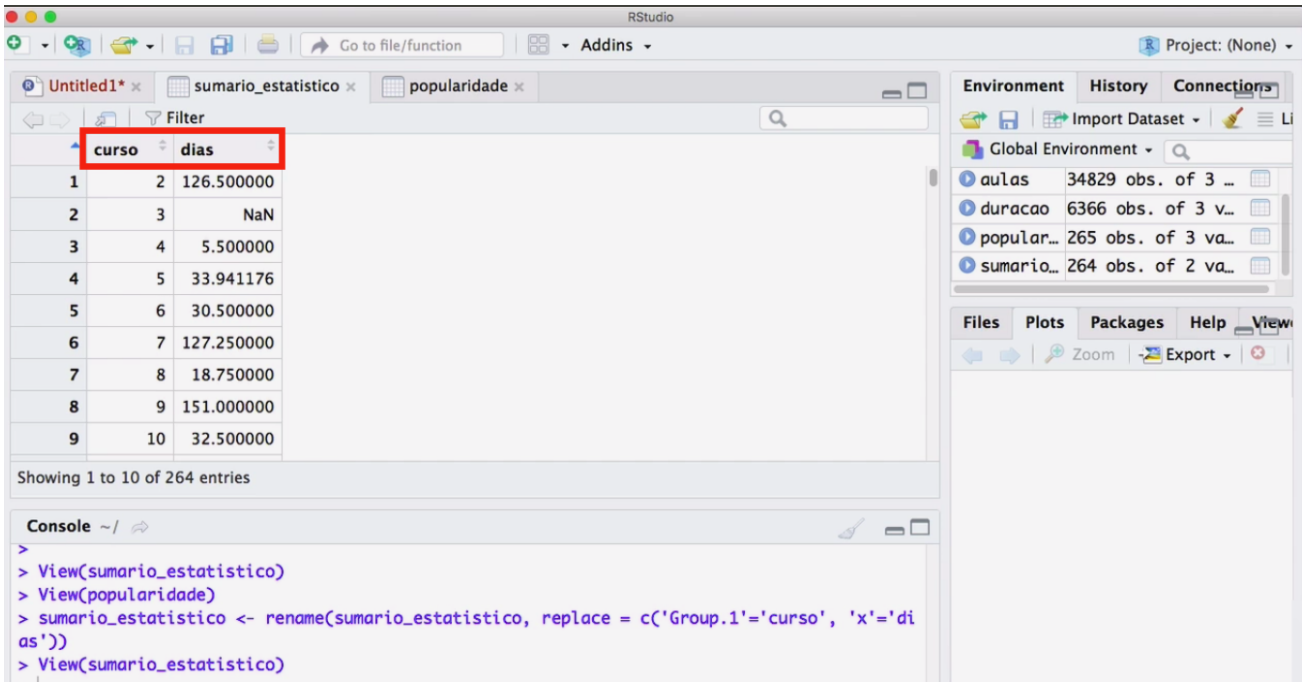
Precisamos alterar os nomes para que ambos fiquem com o mesmo. Faremos isso por meio da variável `rename`, usada anteriormente, para mudar o nome das colunas desses bancos de dados, um por um. Primeiro, mudaremos `sumario_estatistico`, atribuindo (`<-`) a ele a função `rename` e, entre parênteses, colocaremos o banco de dados que será alterado. Após a vírgula, digitaremos `replace` para substituir um nome por outro, seguido por sinal de igual (`=`). Criaremos o vetor (`c`) com os nomes e, entre aspas simples (`'`), colocaremos primeiro o nome que queremos substituir e, depois do sinal de igual (`=`), o novo nome. Após vírgula, trocaremos também o nome da coluna 'x' para 'dias':

```
sumario_estatistico <- rename(sumario_estatistico, replace = c(' Group.1'='curso', 'x'='dias'))
```

Ao executarmos o código, verificamos que está tudo correto no Console:

```
> sumario_estatistico <- rename(sumario_estatistico, replace = c(' Group.1'='curso', 'x'='dias'
```

Se clicarmos novamente em `sumario_estatistico`, na visualização, as colunas estarão renomeadas.



Utilizaremos o mesmo procedimento para alterar o banco de dados `popularidade`, atribuindo (`<-`) `rename` a ele e, entre parênteses, especificaremos o banco de dados. Em seguida, poderemos colocar `replace`, **ou não**, pois o RStudio sabe que aquele parâmetro, naquela posição e após vírgula, representa o comando.

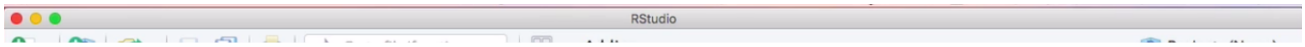
Utilizaremos essa sintaxe alternativa, omitindo `replace` e criando o vetor diretamente. Entre parênteses, especificaremos que `course_id` será `curso` e `freq` será `popularidade`, deixando os nomes das variáveis em português, tornando-os mais intuitivos:

```
popularidade <- rename(popularidade, c('course_id'='curso', 'freq'='popularidade'))
```

Ao ser executado, no Console conferiremos que aparentemente ele está correto.

```
> popularidade <- rename(popularidade, c('course_id'='curso', 'freq'='popularidade'))
```

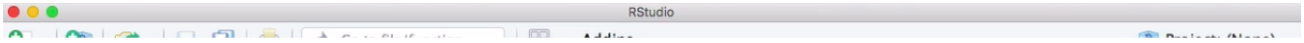
Verificaremos isto clicando no objeto `popularidade`. Na planilha, na janela superior esquerda, encontraremos as colunas `curso` e `popularidade`.



Agora, é possível fundir esses bancos de dados por meio do indicador `curso`. Posicionaremos o cursor na função `merge`, que tentamos executar anteriormente sem sucesso, e veremos se agora será criado um novo objeto, juntando os dois bancos de dados. Ao executarmos o código, teremos:

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

Há um novo banco de dados `popularidade_e_duracao` na janela superior direita. Ao clicarmos nele, à esquerda, visualizaremos uma planilha indexada por `curso`, variável comum a ambos.



Os mesmos cursos de cada banco de dados estão na primeira coluna, e `dias` e `popularidade`, que estavam separadas, agora estão em um mesmo banco. Isto é, fundimos os dois bancos, então temos na mesma planilha `curso`, a média de `dias` que os alunos levam para concluir um `curso` em específico, e também a `popularidade` deles. Está tudo pronto para analisarmos e tentarmos descobrir se há correlação entre essas duas variáveis.