

Chave de cruzamento

Transcrição

[00:00] O que vimos é que precisamos juntar informações numa mesma base, só que originalmente elas estão em bases diferentes. Como fazer isso? Falando alguns termos mais técnicos, isso normalmente é chamado de cruzamento, juntar informações de bases diferentes.

[00:32] Por exemplo, temos duas listas de contatos, uma com telefones e outra com e-mails. O que queremos na verdade é juntar tudo isso numa base só. Primeiro temos que pensar em como relacionar e-mail com telefone para ser uma mesma observação da tabela. Primeiro preciso descobrir a chave de cruzamento.

[01:08] Nas nossas listas de contato, posso usar o nome das pessoas que conheço para relacionar. Posso dizer que o nome é a chave de cruzamento nesse caso. Eu busco os correspondentes. No fim, vou ter uma base que tem nomes, telefones e e-mails.

[02:26] Agora já sabemos quantos jogos cada cliente alugou e precisamos colocar essa informação na nossa base, para ter todas as informações numa base só. Podemos fazer isso usando mescla, inclusive dentro de um data step. Quando usamos um data step para fazer essa união, dizemos que estamos fazendo um merge, que é a mescla entre bases.

[03:18] O padrão no começo é igual, escrevemos data, cad_cli_jogos. Nós vamos usar a opção merge, porque não queremos definir qual a base que vamos usar como entrada. Nós vamos juntar duas bases. Passamos as bases que queremos juntar.

[04:33] O próximo passo é passar para o SAS copiar a chave de cruzamento que nós vamos usar. Sabemos que o CPF é uma chave de identificação única de cada brasileiro. Vamos cruzar a informação da base de operações com o CPF na base de clientes. Inclusive vemos que nas duas bases ela tem o mesmo nome. Isso facilita o merge, porque precisamos ter chaves com o mesmo nome, usando o by.

[06:30] Ele deu um erro, porque a variável CPF é definida tanto quanto caractere quanto numérico. Esquecemos de analisar esse ponto. O CPF na base de operações é numérico, CPF raiz do cliente. E na base de clientes é caractere, CPF completo.

[07:17] Como vamos usar o CPF como chave de cruzamento, precisamos prestar atenção nessa particularidade que esse número tem antes de usar. Sabemos que o CPF é uma sequência de nove números e mais dois dígitos. Dizemos que esses dígitos são verificadores, porque no fundo não fazem parte realmente do número gerado do CPF. Eles são gerados a partir de um cálculo feito desses nove números anteriores. O número real do CPF são esses nove. Os dois últimos são colocamos como medida de segurança.

[08:35] Esse CPF com todos esses onze dígitos é normalmente chamado de CPF completo. Mas quando vamos utilizar esses CPFs nas bases, simplificamos só para a parte que interessa, que são os primeiros nove dígitos. Quando tiramos os verificadores, ficamos com os nove, chamamos de CPF raiz. Essa é a diferença. Precisamos transformar de forma que conversem entre si.

[09:35] Primeiro, vamos converter o CPF da base de clientes em CPF raiz. Iremos fazer isso de uma forma que já vimos. Vamos criar uma base temporária, chamar de cad_cli CPF raiz, que vai ser a mesma coisa que nossa base da Alura, mas vamos criar uma nova variável chamada CPF raiz que recebe os primeiros nove dígitos. Precisamos pegar os nove dígitos, mais os pontos separadores, dando onze caracteres desse string de CPF.

[10:52] Ainda precisamos transformar isso em número. Vamos usar o input, que vimos anteriormente desse texto, e precisamos falar para o SAS como ele tem que ler corretamente, passando em um formato que converse. O formato do SAS que usa números e pontos separadores é o COMMAX. Também precisamos falar o tamanho, que é onze, no formato colocamos ponto, e especificamos que não tem nenhuma casa decimal.

[12:36] Tudo certo. Já podemos cruzar a partir dessa chave. Não estamos mais usando a base de clientes. Mas aqui vemos que as chaves estão com nomes diferentes. Precisamos cruzar a partir do CPF raiz. Mas nessa base que tínhamos feito ele se chama CPF. É fácil consertar. Vimos que podemos renomear uma variável a partir da chamada dela, quando chamamos a base.

[13:32] Ele diz que conseguiu encontrar as variáveis, mas elas não estão devidamente ordenadas.