06

Removendo duplicações

Transcrição

[00:00] Lembram quando usamos o UNIVARIATE para analisar por diversas categorias, que usamos by e faixa de idade, e precisava estar ordenado? Quando usamos o próprio sort para ordenar uma base? Aqui é parecido. Estamos ordenando por CPF raiz. Vamos ordenar a base de clientes. Antes, tínhamos visto que podemos ordenar usando um PROCSORT. O começo do processo é o que estamos acostumados. Falamos qual é a base de entrada. Aqui vamos fazer algo diferente. Não queremos sobrescrever essa base de CPF raiz. Dizemos isso definindo outra base de saída, com out igual a. Não queremos sobrescrever a base porque precisamos aproveitar e garantir que essa base de cadastro de clientes não tem nenhum CPF repetido.

[01:35] Quando vamos cruzar bases, inclusive chamamos de chave única, vimos isso na primeira parte do curso, é importante não ter duplicações para conseguir fazer a paridade de informações corretamente. Imagine que você anotou em um papel que você está devendo vinte reais por João, mas você tem três amigos chamados João. Para quem você vai pagar? Fica difícil. É a mesma coisa. Se tenho informações de tabela iguais, como faço para cruzar?

[02:11] É importante garantir que não tem duplicação. Em poucos casos, exceções, pode ser que queiramos fazer o cruzamento com duplicação, mas aqui vamos aproveitar para garantir que não tem na nossa base de CPF.

[02:36] Falamos isso para o SAS passando um novo parâmetro. No SAS, escrevemos nodup. O próximo parâmetro é a variável que vamos ordenar. Selecionando e executando, nossa base está ordenada pelo CPF. Olhando na log, ela mostra se encontrou duplicações. Se tivéssemos encontrado alguma, teríamos um erro bem grande, porque CPF é uma chave única em todo território nacional. Nossa sorte é que não tivemos nenhuma observação duplicada.

[04:00] Essa opção nodup tira a duplicação se ela encontrar uma observação inteira, uma linha inteira que está igual. Ou seja, se olharmos o cadastro de cliente, ele só tiraria uma observação se tivesse duas linhas inteiras, com todas as variáveis iguais. Queremos ser mais restritivos. Queremos tirar a duplicação olhando só a chave, independente das outras variáveis. Então, colocamos nodupkey. É uma opção que usamos para retirar duplicações a partir da chave.

[05:05] Uma coisa interessante, olhando a base de cad_cli_cpf_sort, usando o PROCCONTENTS, existe uma informação a mais no conteúdo. No final, temos uma nova tabela com informações da ordenação. O SAS armazena a informação de que a base foi ordenada.

[06:17] Antes de usar um sorte para ordenar a base, vamos usar o CONTENTS para avaliar a nossa base. Ela já está ordenada também. Tem essa informação que o SAS armazenou de que ela foi ordenada. Às vezes bases ordenadas podem acabar não tendo essa informação de ordenação porque ela só aparece se a base for ordenada pelo SAS. Se ela for ordenada em outro ambiente, em outro sistema, e foi importada para o SAS, ele não sabe se ela está ordenada ou não, porque não foi ele mesmo quem ordenou. Mas como criamos essa base no SAS ele já traz essa informação para nós. Inclusive, ela não tem duplicação.

[07:24] Nós não ordenamos essa base em nenhum momento. Simplesmente criamos usando SQL. Isso mostra que o SQL também pode ordenar bases. Inclusive, quando usamos a opção de group by, ela já automaticamente ordena por essa chave de agrupamento.

[08:00] Como falamos para o SQL que queremos que ele ordene a base? Com order by. Nós já sabemos que se agrupar automaticamente ordena por essa chave de agrupamento. Vamos aqui mostrar que se eu colocar outra chave na ordenação, por definição sabemos que não vai ter duplicação, porque essa é a própria definição do processo. Agrupar é

justamente pegar todas as repetições que temos dessa informação na base e colocar tudo numa linha só. Não ter duplicação já é inerente do processo.

[09:20] Outro detalhe interessante também que podemos ver é o agrupe por. Podemos escrever o nome da variável, mas na hora de selecionar as variáveis que temos na base, temos o CPF, e uma outra, poderíamos trocar para um. Estaríamos falando para agrupar pela primeira variável, que é CPF.

[10:32] Não tivemos nenhum erro, estamos com a base de cadastro dos clientes.