02

Classificando a Idade

Transcrição

[00:00] Nós fizemos um gráfico com histograma da nossa variável de idade para ficar mais fácil de visualizar. Diversos exemplos da vida real vemos esse tipo de organização da informação, de separar por faixas de valores, ao invés de deixar bruto. Por exemplo, se vemos uma pesquisa, um dado estatístico, normalmente isso também acontece, vemos pessoas que são jovens, na faixa dos 18 aos 25 anos, ou pessoas que são idosas, que tem mais de 60 anos. Vemos essa separação, por exemplo, da idade, em faixas, como vimos no histograma. O pessoal da Alura Play também precisa dessa informação.

[00:50] Aqui nós só fizemos um gráfico. O próprio SAS organizou as faixas para visualizarmos, mas se quisermos fazer análises mais para a frente sobre idade, seria muito interessante ter uma variável que fizesse essa classificação para nós dentro da base.

[01:08] O que nós vamos fazer agora é criar uma nova variável que classifica nossa idade em faixas. Vemos nesse gráfico de histograma que o SAS construiu para nós sete faixas. Mas talvez isso seja demais. Menos seria melhor, porque nas idades maiores, tem pouca gente. Precisamos dar uma volumetria maior.

[01:46] Mas também seria interessante se fizéssemos com números ímpares de faixas para termos uma faixa central. Se tivermos um número ímpar, vamos ter uma categoria intermediária, um grupo no meio, um número de grupos para cima, um número para baixo dessas idades, conseguimos dividir de forma mais simétrica. Que tal dividir em cinco? É menos que sete, mas um número ímpar também.

[02:18] Vamos também fazer uma classificação bastante simples. Não é baseado em algoritmos difíceis, critérios complicados. Vamos simplesmente dividir em cinco faixas de mesma quantidade de clientes. No caso, nossa base de clientes tem cem pessoas. Dividiríamos em cinco faixas de vinte pessoas, mais ou menos.

[02:45] Criei um novo código. Já vou inclusive dar um título para ele, que vai ser classificação da variável de idade. Poderíamos até tentar fazer isso na mão, mas classificar, fazer esse ranqueamento de uma variável é um processo comum de acontecer no mercado. Inclusive, existe um procedimento do SAS que faz exatamente isso. Esse procedimento de ranqueamento, fazer essa divisão de variável em subgrupos de volumetria igual, é feito por um procedimento de ranking. Nós vamos usar ele, o PROCRANK.

[03:50] Nós passamos o primeiro parâmetro, no caso é a base de cadastro de cliente, que está na Alura. O segundo parâmetro que fazemos é que não vamos substituir essa base da Alura. Vamos primeiro construir outra base temporária, para ver se está tudo funcionando corretamente, e o próximo parâmetro é exatamente a base que vai sair desse nosso processo. Eu vou simplesmente chamar de base_ranks, que vai ser uma base com ranks.

[04:28] O próximo parâmetro que temos que passar é groups. Quantos grupos quero separar essa minha variável. Já tínhamos visto que queremos cinco. Agora finalmente colocamos o ponto e vírgula para encerrar a linha de comando. Mas ainda temos alguns outros parâmetros para passar.

[04:52] Temos que falar qual a variável que queremos classificar, ranquear. Passamos isso através do parâmetro VAR, de variável. A variável que queremos classificar é idade.

[05:08] O próximo parâmetro que temos que passar é o nome que queremos para essa variável. Passamos isso falando os ranks, onde estarão os ranks. Vamos chamar de faixa_idade.

[05:35] Não precisamos passar mais nenhum parâmetro, fechamos com run e executamos o código. Ele soltou uma base nova para nós, parecida com a nossa base de cliente mesmo, e tem a variável de faixa de idade.

[05:57] Vamos voltar para o nosso código e fazer uma análise nessa base que acabamos de criar, usando um PROCFREQ, para fazermos uma lista dessa nossa variável de faixa de idade para ver o que tem nela. A base que vamos passar de entrada é o ranks, queremos fazer uma tabela da variável faixa idade. Vamos executar e temos a variável faixa idade, que vai de zero a quatro, porque ele simplesmente faz uma espécie de ordenação, chama cada um desses grupos de forma ordinal mesmo, indo de zero até o número de grupos menos um. Estamos vendo quantas pessoas tem em cada uma dessas faixas. No caso, o zero são as menores idades e o quatro as maiores. Ele ordena nesse aspecto também.

[07:12] Estamos vendo que se fôssemos organizar e separar por faixas de mesma quantidade de pessoas, cada uma teria que ter vinte pessoas. Nossa base tem cem, dividido por cinco, cada uma teria que ter vinte. Mas não é exatamente o que aconteceu no nosso caso. Temos faixas que vão entre dezessete e vinte e duas pessoas. Não é tão ruim, mas acontece porque o PROCRANK não separa indivíduos que tem o mesmo valor da nossa variável que estamos categorizando. Por exemplo, precisaríamos fazer uma faixa com vinte pessoas, mas temos vinte e uma pessoas com dezoito anos. Ele não vai colocar vinte pessoas com dezoito anos numa faixa e depois uma outra faixa com uma pessoa de dezoito anos, mais dezenove pessoas com dezenove anos. Ele faz o melhor possível para deixar os mesmos valores de idade dentro da mesma faixa. No caso, é isso que vemos.