

Ausência de informação

Transcrição

[00:00] Temos o nosso primeiro programa, já comentamos e fizemos tudo que tínhamos que fazer, mas existe ainda um pequeno detalhe na base de cadastro de produtos que seria importante checar, que é a variável de data. Isso é importante porque a variável de lançamento é construída em cima dela. Se ela tiver problemas, a variável de lançamento também vai ter.

[00:42] Vamos criar um programa novo para fazer essas análises. Posso clicar no menu lateral, em novo, e na primeira opção. Eu não quero mais mexer no programa anterior, posso fechar e focar no que eu acabei de criar. Já vou começar dando um título. Vai ser análise sobre a variável de data da base de cadastro produto.

[01:50] Vamos usar o PROC FREQ da minha base Alura cadastro produto v2. Depois, quero a tabela de frequências da minha variável data. Posso executar. Tem várias datas diferentes nos meus jogos. Chegando no final, tem algo estranho. Na coluna de frequência acumulada, ele me diz que a base tem 497 linhas. Mas eu sabia que ela tinha 500. E também tem um novo rodapé falando que a frequência de missings é três.

[03:18] Eu sei que somados eles dão as 500 linhas que eu esperava encontrar. Lembra da nossa linha de contatos? Imagine que tem uma pessoa que você conhece que você sabe o telefone, mas não o e-mail, ou o contrário. Você ainda colocaria na lista de contatos, mas a informação fica em branco. Missing é justamente isso. Algo que está faltando na base.

[04:35] Como posso separar na base só as observações que estão faltando para eu verificar? Usando o comando de data, crio uma base temporária, a partir da base da Alura. Quero tirar alguma coisa dessa base, filtrar observações, para que só outras passem. Quero filtrar somente onde minha variável data é missing. Em inglês, digo where, que é o comando para filtrar, e depois a função. Minha variável data is missing.

[06:27] Encontrei três observações justamente em que a variável não está preenchida. Observem que nesses lugares tem um ponto. Como a variável data é numérica, o ponto significa um missing. Se eu quero escolher um missing, posso usar ao invés de um valor onde minha variável é igual a um ponto.

[07:33] Andando mais para a frente na base, quero dar uma olhada na variável de frag lançamento. Ficou bem estranho. Por que ela está preenchida e a data não? Isso acontece porque se lembrarmos de como criamos a variável de lançamento, ela é um condicional. Ela olhava para o valor da data, e se fosse menor ou maior que outro valor, era preenchida. Isso quer dizer que ao encontrar o missing, o SAS considera algo. Ele não dá erro. Ele consiga a ausência de informação como informação.

[08:50] O valor que o SAS dá para uma variável missing é bem interessante. Todas as flags de lançamento são zero. Minha condição é se a data é menor do que 20166. Sempre que minha data for missing, ele considera que é menor. É justamente esse o valor que o SAS dá para o missing. Imagine que de todos os valores que você tem, você tem o menor possível, o mínimo. Esse é o missing e é isso que pode dar problema na base.

[09:55] Como olhar as demais cópias desses três jogos para ver as datas preenchidas nela e tentar consertar a variável de data?