06

# Distribuição de frequências quantitativas - Classes personalizadas

### Transcrição

[0:00] Legal, gente, voltando aqui e continuando com o mesmo assunto, distribuição de frequências.

[0:04] Só que agora a gente vai trabalhar com variáveis quantitativas, que são as variáveis que não são naturalmente categorizadas.

[0:10] Então o primeiro passo, logicamente, é criar uma forma de categorizar elas para depois construir a distribuição de frequências.

[0:19] É isso que a gente vai fazer agora.

[0:20] Nesse vídeo eu vou mostrar pra você como criar classes personalizados.

[0:23] Eu peguei um exemplo aqui que é bem famoso, que é aquela categorização em classes de renda A, B, C, D, E e que classe e renda você pertence, de acordo com o rendimento que você recebe.

[0:36] Então a classe A seria rendas altas e classe E rendas mais baixas.

[0:40] Eu peguei essa classificação aqui de um trabalho que eu vi na internet, ele está dividido em salários mínimos onde a classe A está acima de 20 salários mínimos, a classe B entre 10 e 20, a C 4 e 10, D 2 e 4, E até 2 até salários mínimos.

[0:56] De 0 a 2 salários mínimos... Eu não lembro se eu falei isso para vocês, mas eu acho que eu tinha dito que a nossa PNAD que é de onde veio o nosso dataset é do ano de 2015, então no ano de 2015 o salário mínimo era de 788 reais.

[1:10] Então eu fiz essa transformação aqui para gente, para facilitar o nosso trabalho e já está aqui as classes A, B, C, D e E, em reais daquela época, 2015.

[1:20] Então é isso que a gente tem que construir.

[1:21] O primeiro passo para fazer essa construção é descobrir quais são os valares mínimos e máximos da nossa variável.

[1:29] Para fazer isso nós vamos estar usando os dados de renda, eu vou vir com dados ponto renda, que é a variável de renda... O mínimo eu descubro assim, ponto min, eu venho com o valor mínimo, que é 0, dados ponto renda ponto máx., eu venho com um valor máximo que é 200 mil reais.

[1:50] Isso aqui já acende uma luz para a gente pensar, "será que esse valor está certo?", "será que é um outline?", a gente anota que a gente vai ver isso mais para frente.

[1:58] Continuando, eu vou usar uma funcionalidade do pandas que é o cut, esse cut precisa de alguns parâmetros e pra isso eu vou construir isso previamente.

[2:07] O primeiro parâmetro é o que? São os limites das classes.

[2:11] Eu vou colocar dentro dessa variável classes que vai ser uma lista do Python com os limites das classe que eu quero construir.

[2:21] O primeiro limite é o limite mínimo, que eu já identifiquei sendo 0. O segundo a gente tira aqui da nossa construção aqui em cima, o próximo depois do zero 1576, correto? 1576.

[2:34] O próximo depois de 1576 tá aqui, é daqui a 3152. 3152, e assim por diante.

[2:43] O próximo é 7880, e o próximo é 15760, 15760.

[2:52] E agora a gente chegou em um ponto onde a gente vai ter que ir até o limite superior que é quem? 200 mil, acima de 15760 coloquei 200 mil.

[2:56] Criado essa variável classe eu quero não mostrar de uma forma feinha, quero mostrar de uma forma que são as classes A, B, C, D e E, então vou criar um label, labels...

[3:19] E vou dar justamente a mesma coisa uma lista no Python com os labels.

[3:29] Como eu comecei com a classe menor eu tenho que começar com o label menor.

[3:33] Então eu vou começar do E,D, C e assim por diante.

[3:37] A classe E, classe D, classe C, aqui B e aqui A, bem simples, já fizemos o primeiro passo, criamos as classes.

[3:47] Agora vamos entender como é que a função, o método cut funciona no Pandas, vamos lá.

[3:54] Pra utilizar esse cara já deixei aqui o ajudinha dele, eu vou vir chamar o Pandas PD ponto cut, e vou passar pra ele a variável que eu estou trabalhando no parâmetro X, para dados ponto renda que eu quero fazer as classes da variável renda.

[4:14] Um outro parâmetro, vou pular uma linha aqui pra ficar mais claro, se não a gente vai esticar muito, o outro parâmetro que eu quero é passar pra ele quais as classes, quais os limites das classes, pra ele construir as classes para mim.

[4:25] Foi aquela variável que a gente criou em cima, classes.

[4:28] Eu vou passar para parâmetro bins do cut, classes... Como a gente criou um label e eu quero mostrar isso num label, eu não quero mostrar com números estranhos, eu vou passar pra ele também esse parâmetro dentro do parâmetro labels.

[4:49] Vai no labels que eu criei, parâmetro labels... construído.

[4:46] Uma última coisa que eu quero mostrar para vocês que é bem interessante é, que, por default esse método não inclui nas classes, a classe inferior que é a classe menor e o zero, e não inclui... Então eu tenho que dizer pro cut que eu quero que incluía esse zero, porque eu tenho pessoas com essa informação, com renda zero, eu quero que eles apareçam na contagem.

[5:16] Para fazer isso eu tenho um parâmetro include lowest que eu falo que tem que ser igual a true.

[5:26] Rodando isso aqui a gente vai criar uma series, como você pode ver com um índice igual ao do meu dataset, meu dataframe, onde o que ele está dizendo aqui é que é a variável do registro zero, a renda desse registro. ela se encaixa na classe E, do registro 1 é a mesma coisa, do 2 é a mesma coisa, do 3 ele já está dizendo que se encaixa na classe C.

[5:54] Vamos fazer uma coisa aqui interessante, vou criar aqui uma célula acima e vou botar meus dados... Só os cinco primeiros, só pra gente ver como está funcionando.

[6:13] Tá vendo aqui? Aqui os índices são iguais, zero, um, dois, três eles seguem esse cara, aqui estão mostrando até o quatro, vou mostrar para vocês, a renda aqui é 800, 1150, 880, aqui no nosso amigo três já é 3500.

[6:28] Esses caras aqui você pode ver que estão realmente até 1576, estão na classe E, ele classificou aqui como classe E.

[6:39] Esse cara aqui com 3500 ele está em que classe? 3500 tá na classe C. De 3152 até 7880, ele está aqui nessa classe C, o cut classificou o cara como classe C, e assim que ele fez, sucessivamente.

[6:54] Então vamos deletar nossa célula aqui para não confundir a nossa cabeça e vamos continuar.

[7:02] Tendo isso aqui o que eu faço? Como a gente já fez nos exemplos anteriores, basicamente a mesma coisa, a gente vai passar para aquele método value counts só que vou chamar de uma forma diferente, antes eu tava chamando dados ponto a ponto variável ponto value counts, ele fazia a contagem da própria variável, aqui não.

[7:18] Aqui eu vou recortar isso aqui para não fazer uma confusão recortar o que está construído.

[7:24] Eu vou chamar de outra forma, vou chamar pelo próprio pandas, pd ponto value counts, e vou passar para ele agora isso que a gente acabou de criar aqui, desculpem...

[7:43] Passar pra baixo o nome de atuação aqui pra ficar mais claro e vou rodar de novo.

[7:48] Ela criou aqui, o value counts, aqui tá o cara, e ele já fez a contagem pra gente do jeito que a gente queria.

[7:55] O que é que isso aqui? É justamente aquilo que a gente tinha feito antes, aquilo que eu estava chamando de frequência para depois juntar um percentual.

[8:05] Vamos fazer isso aqui de novo, vamos criar essa mesma tabelinha, bonitinho, pra gente mostrar pro nosso chefe.

[8:13] Está feito aqui, o que eu tenho que fazer agora? Aquela coluna de percentual, vamos fazer aqui, rápido.

[8:21] Vou chamar, percentual... Vou colocar aqui embaixo para ver, qual o parâmetro que eu preciso passar para fazer o percentual dentro da função value counts? Normalize, a gente já conhece isso, igual a true.

[8:37] Ele vai rodar aqui, vai fazer para mim, tá aqui, é isso aqui que eu quero.

[8:41] O próximo passo é exatamente igual a este aqui.

[9:02] Vou copiar esse aqui e é justamente isso que vamos botar aqui embaixo, para não termos que digitar tudo de novo.

[9:11] Vamos lá, só que eu vou mudar aqui o nome. vou chamar isso aqui não mas qualitativas, a gente está fazendo agora com quantitativas.

[9:22] Eu vou botar aqui mais uma coisa, como que a gente está chamando aqui em cima? Personalizados. Quantitativos personalizados. Um nome de variável enorme, mas tudo bem, a gente entende.

[9:38] Vou pular aqui pra não fugir do que a gente tem.

[9:44] Frequência, já está certinho, que frequência vem daqui, de frequência aqui em cima e o percentual tá aqui, é exatamente isso que eu quero.

[9:52] Só que eu esqueci de mostrar o nosso resultado. Beleza.

[10:00]Uma coisa que eu posso fazer aqui, porque fica estranho começar do E ir até o A, porque ele está ordenando por isso aqui, olha que louco.[10:10] Então a gente pode mudar isso também, vamos fazer isso, aqui nos permite que a gente faça tudo o que a gente quiser.

[10:15] Ponto. Vou fazer um sort pelo index pelo, sort index, vai ser assim? Então tá bom.

[10:27] Ascending igual a false e agora a gente vai fazer o default.

[10:25] Não quero default, pronto ,resolvido.

[10:39] Essa distribuição de frequência é a que eu quero, é a que eu vou analisar depois, é a que a gente vai tirar conclusões depois, primeiro estou construindo as coisas para depois a gente começar a fazer uma autoanálise.

[10:48] Perceba que já pouquíssimas pessoas colocam assim e ela vai fazendo... Aumentando violentamente, a classe E é gigantesca.

[20:58] Essas coisas a gente vai ter que perceber, depois a gente vai ver foi isso em um gráfico, como é que vai ficar.

[11:03] E a gente vai começar a fala de assimetria.

[11:06] Então, no próximo vídeo a gente vai continuar com esse mesmo assuntinho que a gente tá tratando agora, mas com outra forma de criar essa categorização de variáveis quantitativas.

[11:15] Legal? Até o próximo vídeo.