08

Distribuição de frequências quantitativas - Classes de amplitude fixa

Transcrição

[0:00] Legal, pessoal. O último macete aqui na forma de construir distribuições de frequência.

[0:03] O que eu quero mostrar para vocês agora, como criar essas tabelas, quando a gente não tem classes previamente organizadas.

[0:09] Como a gente tinha na última aula, a classe A, B, C, D e E.

[0:13] Aqui não. Aqui eu vou mostrar para vocês uma regra que otimiza a escolha da quantidade de classes, que a gente vai ter nessa tabela de distribuição de frequências.

[0:22] Que é essa Regra de Sturges aqui, que tem essa formulazinha simpática, que considera somente o número de observações que a gente tem da variável. Beleza?

[0:30] Então, vamos lá. A primeira coisa que a gente tem que fazer, importar o NumPy, que é essa biblioteca de análise matemática. Ela tem diversas fórmulas matemáticas, ela tem alguns macetes para você trabalhar com álgebra linear, matrizes.

[0:44] Então, a gente vai usar ela aqui, unicamente, para utilizar a função log na base 10, que está aqui na nossa fórmula da regra. Legal?

[0:52] Então, vamos importar o NumPy. Import numpy as np, que é o padrão que o pessoal usa aí, dá esse apelidinho. No panda a gente deu pd, aqui a gente está dando o np para o NumPy.

[1:05] Então, vamos lá. Essa fórmula é bem simples, a única coisa que ela precisa é como eu já falei, número de observações, que é esse n aqui no final.

[1:12] Então vamos achar o n da nossa distribuição, do nosso data set.

[1:17] A gente já viu isso, eu acho, lá em cima. Então, n vai ser igual, criando a variável aqui. Como é que a gente faz para achar isso num DataFrame?

[1:24] Passo dados ponto shape e eu quero pegar, aqui no shape, eu vou rodar aqui para você ver, com n.

[1:31] Desculpa. Botar aqui o n. Ele vai vir com número de registro, de observações, e o número de variáveis que tem.

[1:39] Então, só quero número de registro, vou pegar só o primeiro elemento ali.

[1:43] Então, n é 76840. Agora para descobrir o número de classes é só aplicar essa fórmula, que está aqui em cima.

[1:51] K é igual a um mais, eu vou abrir e fechar um parêntese aqui, 10 dividido por três vezes, e agora eu vou chamar o NumPy, np ponto, a função do log na base 10 é bem simples, é log10, abre e fecha parêntese e passo o n que a gente acabou de calcular aqui, n.

[2:17] Shift Enter, aqui, vamos ver qual o K que a gente criou aqui.

[2:21] Então, ele está me dizendo aqui que uma forma ótima de eu visualizar isso, baseado nessa quantidade de registros que eu tenho, é com 17 classes, aproximadamente. Um pouco mais do que isso.

[2:32] Então, a gente, como não tem como fazer 17,28 classes, a gente vai arredondar esse número.

[2:38] Primeira coisa que eu vou fazer, K ponto round e eu quero zero casas decimais.

[2:52] Eu vou mostrar esse K aqui, só que ele vai fazer o quê? Ele vai me passar um 17 ponto zero, então não quero esse ponto zero aqui. Eu quero um número inteiro, então eu vou englobar tudo isso aqui na função int do Python, que ele vai pegar só a parte inteira do número. Ou seja, K igual a 17.

[3:09] Então, guarda essa formulazinha aqui para você criar o seu K, sempre que você precisar dele.

[3:15] O segundo passo é o quê? O que a gente já fez nas outras aulas.

[3:18] Vamos lá, pd ponto value counts. Pular aqui, a gente vai fazer com cut, como a gente tinha feito antes.

[3:30] Então, como é que faz? Pd ponto cut. Vou de novo aqui, pular e vou fazer uma endentação para ficar bem organizado, para a gente entender depois.

[3:39] Ele pede o parâmetro x, que é o quê? Eu vou fazer do mesmo jeito que eu fiz antes, vou fazer com a variável de renda.

[3:44] Então, dados ponto renda. Esse é o primeiro parâmetro. Qual é o outro? São as classes.

[3:52] Lembra o que a gente passou no bins, só aqui, ao invés de a gente passar as classes, a gente vai passar o número de classes que a gente quer e vai fazer isso igualzinho.

[3:59] Vai fazer 17 classes de mesma amplitude, mesmo tamanho. Por isso que eu coloquei aqui em cima, se eu não me engano, classes de amplitude fixa. As amplitudes das classes são todas iguais, mesmo tamanho.

[4:12] Mais uma coisinha que está faltando aqui. Lembra que é do include lowest igual a true, que é para quê? Quero que ele inclua o limite mais baixo.

[4:27] Nesse caso aqui talvez não precisasse, mas é bom a gente deixar aqui para a gente não esquecer desse cara sempre.

[4:33] Tudo bem, ele criou aqui, está vendo? Eu vou deixar para vocês de exercício, se você quiser fazer, lógico, os labels. Eu não fiz os labels aqui, porque eu queria mostrar para vocês, justamente, as classes que ele mesmo criou para mim.

[4:46] Aqui estão as 17 classes. Você vai reparar e falar assim: "puxa, está uma coisa esquisita aqui. Parece que essa classe tinha que estar em último lugar, olha o 200 mil aqui".

[4:55] Por quê? Porque essa funcionalidade aqui, value counts, ele faz um sort, mas o sort que ele faz é pelos números que estão aqui do lado.

[5:05] Ou seja, você pode reparar aqui que está organizado do maior para o menor.

[5:11] Eu não quero isso. Quero ver as classes, na ordem das classes.

[5:14] O que temos que fazer aqui, tem um parâmetro para value counts que é o sort.

[5:23] Por padrão ele vem como true, o que eu quero é false.

[5:28] Rodou lá. Tudo organizadinho do jeito que a gente queria.

[5:33] Essa é a forma de criar as classes. Para organizar do jeito que a gente fez antes, frequência é igual a isso, já vou copiar aqui, Shift Enter.

[5:46] Ele criou, não mostra porque a gente não quis mostrar.

[5:51] Aqui, percentual. Lembra? Aqui, dentro do value counts, a gente passa o normalize igual a true. Ele criou aquele percentual também. Vamos mostrar aqui.

[6:12] Percentual. Está aqui aquele percentual criado, certinho.

[6:18] O que a gente tem que fazer? O que a gente já fez nas outras aulas, que é isso aqui. Vou até copiar aqui para a gente não ter que digitar isso tudo. Podem copiar aí também.

[6:30] Vou fazer aqui. Só que aqui eu não vou chamar de dist freq quantitativas personalizadas, eu vou botar aqui amplitude fixa.

[6:45] Vai ficar um nome variável enorme, mas você vai entender bem do que se trata, você que está acompanhando.

[6:53] Shift Enter, está lá a nossa tabelinha, já organizadinha, toda bonitinha. Você querendo fazer aqueles macetes que a gente fez nas últimas aulas, manda ver.

[7:04] Colocar um título aqui, criar um label também. Beleza?

[7:09] Então, é isso. Agora, o que a gente vai ver no próximo vídeo é justamente como visualizar isso graficamente, visualizar todas essas distribuições de frequência que a gente fez, de maneira gráfica, com os histogramas. Beleza?

[7:21] No próximo vídeo a gente vê isso. Abraço.