

Mediana

Transcrição

[0:00] Legal pessoal, continuando aqui com medida de tendência central, vamos falar aqui da mediana.

[0:04] O que é mediana e como a gente obtém a medida: a mediana é o valor que divide a minha série exatamente ao meio.

[0:12] Tenho, por exemplo: vamos dizer que eu tenha uma série à idade de dez pessoas.

[0:17] Como é que vou obter a mediana?

[0:18] Vou organizar isso de forma crescente do menor para o maior, do mais novo para o mais velho; vou pegar o valor que parte exatamente essa série ao meio, ou seja, o valor que vai deixar os cinco mais novos para um lado e os cinco mais velhos para o outro.

[0:33] Legal, pescou?

[0:34] É isso que a gente vai fazer agora.

[0:36] Como eu disse agora a pouco: deixei um roteiro; antes de falar disto, deixei um roteiro de como calcular e obter a mediana com essa formulazinha que são bem simples de entender.

[0:45] Primeiro passo: é ordenar os dados, depois agente identifica o numero de observações.

[0:51] Porque quando a gente vai separar, cortar exatamente no meio; ser par ou impar vai interferir um pouco nisso, e aqui tem estas formulas que estão explicando.

[1:00] Então vou fazer um exemplo: eu vou fazer o impar e vou deixar um exercício para você fazer o par, e depois na próxima aula você verá a solução dele, beleza?

[1:09] Então vamos lá. Vamos ordenar: venho aqui para baixo: [1, 3, 4, 6 e 9.] Como esse dado é impar, eu faço essa conta aqui: n é um nome de registro, que no caso aqui, é [5], aqui seria: $[5+1=6 \div 2= 3]$.

[1:34] Isso aqui está querendo dizer que o meu elemento mediano está na posição [3].

[1:39] Venho aqui e conto: [1, 2] e o terceiro aqui, é o elemento mediano; neste caso aqui, o [4] é a minha mediana, legal?

[1:47] Vamos fazer isso aqui, passo a passo com o nosso DF, que a gente já conhece; aquele dataset, que a gente criou lá, em cima para fazer de exemplo.

[1:55] Eu vou fazer primeiro aqui, vou chamar de nota, do fulano de novo: notas _ fulano = me deixa eu copiar isto aqui, que eu vou repetir isso aqui, muitas vezes, vai = a DF. Fulano.

[2:13] Vou mostrar aqui, para vocês; vou fazer passo a passo para gente entender como isso funciona: estão aqui, as notas do fulano.

[2:19] Primeiro passo: organizar o dataset, name: Fulano: dtype: int64. Então vamos lá: nota _ fulano = nota_ fulano. sort _ values ; isso aqui é um método do pandas. short values, vamos colocar aqui em baixo, para a gente verificar esse cara sorteado. Olha, ele já fez o short das notas do fulano.

[2:44] Uma coisa que eu vou mostrar vocês aqui, que é necessário só para gente entender a posição, logicamente, que isso aqui, você não vai ter que fazer toda vez que for calcular a mediana, é só para gente entender.

[2:54] Então vamos lá, novamente: `notas_fulano = notas _ fulano. Reset _ index` . Eu vou usar outro método do pandas : que é o `reset index`, você já vai entender o que este cara faz: repare, ele tirou o nosso índice.

[3:14] Antes, era os nomes das matérias, e ele transformou este índice em uma variável da minha Data Firewall antes era uma sample, agora é uma Data Firewall.

[3:26] Ele pegou o que era índice e transformou em uma das variáveis e me colocou o índice numérico, é isso que quero para mostrar você; somente para isso.

[3:34] A posição do elemento mediano, aqui a gente tem quantos elementos?

[3:38] Vai de [0 a 6], porque em computação , geralmente a gente começa as contagens no [0], então aqui, ele está dizendo que tem [7], ou seja, é um ímpar. O quê a gente tem que fazer quando o cara é ímpar?

[3:49] Vamos lá em cima.

[3:50] Aqui, olha , para eu determinar qual é o elemento mediano eu tenho que fazer esta continha aqui: $n+1/2$.

[3:56] Então vamos lá fazer esta conta: a gente já obteve o `n` aqui. Quer dizer: vamos obter o `n`. `1 n= notas_ fulano. shape [0], ok? 2 n - 7`.

[4:14] Então para ver o elemento `md` como está escrito lá em cima: o mediano, eu vou fazer aquela continha lá, que é: `1 elemento _mediano - md - n+1/2`. Vou fazer aqui ó: `2 elemento _md`; o elemento mediano é o 4.

[4:40] Está na posição 4, agora, vou prestar atenção aqui, posição 4, é o quê? Não é esse cara aqui, né? A gente vai contar: é [1, é 2, é 3, é 4]; esse cara, aqui, é o meu elemento mediano.

[4:50] Como eu já estou com uma tabela pequeninha, já automaticamente eu já sei qual é a mediana: é 8.

[4:55] Mas se a gente tiver com uma tabela enorme, a gente não ia saber, a gente tem que fazer a continha, então vamos fazer.

[4:59] Quer dizer: neste caso a gente não precisa fazer a conta, precisa só pegar o elemento mediano, no caso, você tem que prestar atenção se não, você vai ter que fazer uma conta a mais, que é a media aritmética entre dois valores.

[5:10] Então vamos lá: `nota _ fulano.loc` , que é uma forma da gente selecionar um item dentro de um Data Firewall ou de uma sample , eu vou passar esse cara, aqui, que é o elemento mediano.

[5:26] lembra? Se eu passar isso aqui, o elemento mediano; por acaso ele vai falar que é o 8, porque tem um 8, logo aqui em baixo, mas aí, ele pegou esse cara aqui, pegou errado.

[5:35] Como ele começou a contar no 0, então nesse caso, aqui, nós temos que fazer o menos 1. `Notas _ fulano. loc. [elemento_med-1]`.

[5:41] Repare que ele vai pegar: aqui, geografia 8, geografia 8. Beleza?

[5:47] É isso aqui: para a gente calcular esse mesmo carinha, usando o nosso amigo pandas, o macete das pandas, sem ter que fazer esta conta toda, ele pode colocar aqui: `notas _ fulano`, a gente já fez, já está com ele pronto.

[6:04] `Notas _ fulano. media..` Ok?

[6:09] Ele vai trazer o 8, aqui, que é a mediana, legal? Aqui em baixo tem o desenhinho do par, repare que tem uma modificação aqui.

[6:20] Eu deixei aqui para você. Vamos fazer com as notas do beltrano? E eu já fiz aqui uma amostra Porque a gente tem sete observações, correto?

[6:31] Eu fiz uma mostra aqui, eu coloquei seis aqui.

[6:33] Este sample aqui é para gente fazer uma amostra aleatória dentro de uma Data Firewall ou de uma sample, aqui a gente tem um Data Firewall, na verdade a gente tem um sample, eu fiz uma seleção pelo Beltrano.

[6:43] Eu quero pegar somente seis elementos ali dentro, eu quero um valor par, correto? Para gente poder fazer essa brincadeira aqui.

[6:48] Esse random_state, é como se fosse um CID, um gerador de números aleatórios, eu coloquei um número aqui, 101, você repita esse número, que você vai ter o mesmo resultado que estou tendo aqui.

[6:58] É uma geração de número aleatório, só que travado o CID, você tem o mesmo experimento, a mesma seleção outra vez, legal?

[7:08] Eu fiz aqui, esse cara, você trabalha com ele aqui, faça essa mesma coisa que eu fiz aqui em cima, lembrando: que as notas _ beltrano. Median. dele é 6.25, é isso que você vai ter que botar em cima, legal?

[7:18] Fazendo isso, para fazer a análise do nosso projeto, obter aquela análise descritiva dos nossos dados, a gente vai fazer desse jeito aqui, que são dados: Renda. Median. Vou fazer com a renda aqui, depois você manda vê aí, faz para os outros, porque o chefe é seu, você que tem que mostrar para ele, legal?[7:40] É uma outra forma de obter, a gente vai estudar um pouco melhor isso, quando a gente estiver falando de medidas separatrizes, é da seguinte forma, aqui: dados. Renda. Quantile., aí você repete para outro, quantile, sem passar absolutamente nada, ele vai pegar, porque o default dele aqui, se eu não me engano; deixa-me eu colocar aqui uma ? no final, rodar aqui, está vendo o default dele aqui: q=0.5, 0.5 no q, ali, significa: que você esta querendo partir exatamente ao meio. Pegar a median, ok?

[8:14] Vamos lá, fazer de novo para deixar prontinho para você estudar. Ok? É isso aí, com quantile, você consegue também obter o median e outros valores que a gente vai ver quando estiver falando de medidas separatrizes. Beleza? O próximo vídeo a gente vai falar da moda. Vejo você lá. Abraço.