02

Desvio médio absoluto

Transcrição

[0:00] Legal, vamos falar de medidas de dispersão agora.

[0:03] Nos vídeos anteriores a gente falou de medidas tendência central, medidas para separatrizes, e a gente viu que essas estatísticas descritivas somalizam de forma bem importante o conjunto de dados que a gente está analisando, mas às vezes elas não são suficientes para caracterizar, para distinguir bem o conjunto de dados que são diferentes.

[0:21] Principalmente quando eles têm uma variação muito forte, muito significativa... Você deve estar lembrando do nosso dataframe de exemplo, porque tem as notas daqueles três alunos, que é o DF, que a gente criou lá em cima.

[0:34] Aqui se a gente pegar a média desse pessoal, a nota média desses caras, deixa eu ver aqui, não estou escrevendo, df.mean, a gente vai reparar que dois alunos têm a mesma nota média, fulano e sicrano.

[0:48] Como temos poucas observações desses dados, no olhômetro a gente consegue fazer uma análise prévia, e a gente vê que sicrano tem notas menos dispersas, ele é mais constante, tira notas bem próximas em todas as matérias, ele tá ali em 7, 5, 8, 7 e por aí vai.

[1:05] Já fulano a gente repara que parece que ele leva mais a sério algumas matérias e outras que ele não gosta, por exemplo, não leva muito a sério.

[1:14] Por exemplo, inglês ele tirou 4, em histórias tirou 6, mas nas outras ele teve notas boas, o que mostra que esse cara já não é tão constante como sicrano que leva a sério todas as matérias.

[1:26] Se você, por exemplo, calcular a mediana aqui você vai reparar que eles também tem a mesma mediana, ou seja, essas estatísticas que a gente estudou de tendência central elas não caracterizam perfeitamente, elas não conseguem identificar esse tipo de problema que a gente está vendo no olho.

[1:43] Mas imagine um conjunto de dados enorme, a gente não vai conseguir no olho identificar isso, a gente precisa de uma estatística que nos forneça essa informação de dispersão, por isso a gente está estudando agora as medidas de dispersão.

[1:57] Vamos começar falando do desvio médio absoluto, que tem essa formulazinha simpática aqui, que é uma média também, um divido, porém está aqui um somatório desses desvios, que é o Xi, que é o valor de cada nota, no caso do nosso data frame, menos a média geral.

[2:14] Por exemplo, tem aqui beltrano, tiro a média dessas notas e faço essa conta desvio, 10 menos a média, 2 menos a média e somo tudo isso.

[2:22] Só que aqui, reparem, tem essas duas barras aqui que indicam que eu só vou pegar os valores absolutos, ou seja, os valores sem sinal, os valores positivos, ou seja, se, por exemplo, a média aqui for maior que o valor, ou seja, se isso aqui, por exemplo, for 10, o X barra for 10, e Xi for 1, essa conta aqui vai dar menos oito, concorda? Eu não quero menos 8, eu vou pegar só o 8, eu vou tirar o menos, que eu já explico para você porquê dessa situação.

[2:49] Porque fazer isso, eu já vou mostrar para você.

[2:53] Então vamos lá, vamos reproduzir aqui para fulano essa estatística, ver como ela funciona e depois a gente vê como calcula de maneira fácil, utilizando os recursos do Pandas.

[3:03] Então vamos lá, vamos chamar aqui de notas. Notas fulano vai ser igual a DF, na verdade, eu quero 2 colchetes X, mas eu vou mostrar para você porque eu quero 2 colchetes X.

[3:20] Porque eu quero criar um dataframe, desculpa, eu não mostrei.

[3:21] E aqui isso se eu passo um colchetes isso aqui é uma series, se eu passo 2 ele vai criar pra mim um data frame.

[3:36] Por que data frame? Eu vou colocar minhas variáveis aqui nesse cara pra justamente a gente entender melhor como funciona.

[3:42] Qual o próximo passo? Eu já tenho esse Xi aqui e eu preciso da média, então vamos lá.

[3:49] Vamos botar aqui nota média fulano.

[3:49] Vai ser o quê? Notas fulano ponto mean, eu venho aqui, copio e colo a nota média embaixo pra mostrar pra vocês a nota média do cara.

[4:22] Só que aqui ele me mostra uma series de novo, então para resolver os pormenores, só o valor eu passo aqui o zerinho, o índice zero , ele vai pegar dentro daquela series a nota média ,que é 7.71, que a gente já tinha visto no começo desse vídeo.

[4:35] Guardei isso aqui, o que eu vou fazer agora? Eu vou criar aqueles desvios dentro do meu data frame notas fulano.

[4:45] Então vamos lá, notas fulano para criar uma nova variável, que vai ser o desvio.

[4:52] Como é que o cálculo o desvio? Lembra lá na nossa nota, vai ser notas fulano, eu passo aqui o nome dele, fulano.

[5:06] Fulano menos esse cara aqui, a nota média de fulano. Copio esse cara aqui e colo aqui pra mostrar isso e aqui tão os desvios.

[5:23] Reparem que eles têm valores negativos.

[5:29] Ou seja nesse caso aqui a média é maior que esse valor, como a gente tinha confirmado, é maior e esses valores são negativos, mas eu não quero isso aqui, por que eu não quero isso aqui? Lembra que eu disse para você que ia explicar pra você porque valores absolutos? Se eu pego esse cara aqui...

[5:47] Pego aqui um desvio e somo esse camarada aqui, que é justamente o que a nossa formula lá em cima faz, o que eu vou ter? Eu, na verdade, vou ter zero.

[6:03] A soma dos desvios em relação à média é zero.

[6:07] Aqui por conta de arredondamento de casas decimais ele me deu um valor muito pequeno, muito próximo de zero, repare que esse E menos 16 quer dizer que aqui 16 é zero vírgula dezesseis zeros, oito, oito, oito.

[6:21] Ou seja, é um valor bem pequeno, bem próximo de zero, que por conta das casas decimais aqui eles não mostra o zero absoluto.

[6:29] É por isto que eu... Porque se eu fizesse essa conta aqui, zero dividido por n, fica 0, isso não faz sentido nenhum.

[6:38] E aí eu pego só os valores positivos para fazer essa soma, vamos lá, vamos fazer essa soma com valores positivos.

[6:46] O que eu vou fazer aqui é justamente isso aqui, desvio médio... Criar um novo onde vou passar aquelas barrinhas aqui, o nome da variável que eu estou mudando.

[6:58] Entra duas barras, desvio igual ao desvio médio ponto abs, que é pegar um valor absoluto... Que eu esqueci, novamente, eu tenho essa mania de não mostrar aquilo que eu estou fazendo... Vamos lá... Tá aqui.

[7:20] Ele está aqui agora, mostrei esse valor, podemos aqui cortar aqui e mostrar tudo porque ele criou uma nova variável e aqui os valores estão positivos, é o mesmo cara daqui só que desconsiderando os sinais.

[7:33] Agora sim eu posso fazer aquela soma que falei.

[7:38] Então, uma coisa que eu queria mostrar antes, eu vou até tirar essas células aqui, selecionando elas aqui...

[8:04] O que é que eu estou calculando aqui com esse cara? Eu criei esse graficozinho, não se estresse com essas coisas aqui porque isso aqui é só uma demonstração.

[8:13] Você pode tentar entender, aqui eu criei um gráfico onde eu passei uma reta aqui, que a média representa a média dos nossos dados, aqui 7.71.

[8:23] Aqui essas bolinhas seriam os dados que são observados e esse tracejado representa o desvio.

[8:30] No caso do desvio médio esses valores aqui de baixo eles vão pular para cima porque eu estou pegando sem o sinal, eles vão pular para cima e eu vou calcular a média desses valores aqui, lógico que, positivos e eu vou calcular essa média.

[8:45] E esse aqui é o cálculo dessa medida de dispersão, à medida de variação dos dados, veja que que no caso do fulano é uma coisa que tem uma variação interessante.

[8:56] Só coloquei isso aqui para você tentar entender que média é essa que a gente está calculando.

[9:02] Então vamos lá, calculando nossa média agora eu vou pegar simplesmente isso aqui, notas fulano desvio médio, copio isso aqui e vou calcular aqui com uma média ponto mean.

[9:19] É justamente isso que ele está fazendo aqui em cima na nossa forma formulazinha, ele soma todos esses caras e divide por N, pura e simplesmente por uma média desses desvios. Tá aqui.

[9:32] Depois de toda essa conta que a gente fez, a gente chega à conclusão de que o desvio médio absoluto, eu vou botar aqui, médio absoluto. Desvio médio absoluto.

[9:51] Eu vou copiar e colar embaixo se não vou esquecer de mostrar de novo... É notas fulano, e aí eu pego só as notas do fulano mesmo e faço isso aqui, ponto mad, essa é a função para calcular esse cara.

[10:12] Vou mostrar aqui embaixo, ele tem que dar exatamente igual a essa média que a gente calculou aqui em cima de forma mais intensa, a gente tem que fazer a conta na mão, aqui a gente não precisou, e aqui obviamente você vai usar isso no seu dia a dia de estatístico, de cientista de dados.

[10:29] Vai usar essa função mad que é justamente quem calcula o desvio médio absoluto com o Pandas.

[10:28] Aplica essa forma aqui no conjunto de dados.

[10:42] No próximo vídeo a gente já vai falar de medidas de dispersão, que são as mais famosas, que é a variância de padrão, um segue o outro, mas elas são coisas distintas.

[10:53] A gente vai entender o porquê dessa distinção. Veja você no próximo o vídeo. Abraço.