02

Média

Transcrição

[0:00] Legal galera, vamos começar um novo tópico de nosso curso de estatística, vamos começar a falar de medidas de tendência central, e nesse grupo de assuntos aqui, a gente vai começar a falar sobre média aritmética, todo mundo deve ter ouvido falar, vamos falar também da mediana e da moda, e no final vamos falar de uma relação entre as três medidas, o que a gente via começar a falar é estatística descritiva, e você lembra que nosso assunto, nosso projeto envolve uma análise descritiva de conjunto de dados, então vamos utilizar muito isso aqui, e as outras estatísticas que a gente vai ver nas próximas sessões. Fiz aqui um dataframe para ajudar no entendimento de como funciona as estatísticas, o cálculo dessas medidas, beleza? Um dataframe pequeno, para gente entender e aplicar no dataset maior.

[0:47] Vamos lá, clico aqui, shift enter, vai rodar, e isso aqui é um dataframe que fiz com a nota de três alunos em algumas matérias, e com o boletim desses caras.

[0:59] Qual a motivação de começar a mexer nessas coisas, a gente viu nas aulas anteriores, o que a gente vem fazendo é uma sumarização dos dados, para gente tentar entender melhor um conjunto de dados enorme, tem conjunto com mais de 70 mil observações. Aqui eu vou tentar sumarizar mais ainda, com uma medida pegar uma informação importante do conjunto de dados.

[1:30] vamos começar falando da média, o que é a média? A média não deixa de ser um centro de massa da distribuição de uma variável, o ponto de equilíbrio, o ponto que equilibra variável, e por isso, por ser esse ponto de equilíbrio, é muito sensível aos extremos, é muito importante observar, porque vamos falar bastante sobre isso, e algumas variáveis, a média não é a medida mais indicada para representar a variável, justamente por causa dos extremos, você lembra, na nossa variável renda, tem muita gente ganhando pouco, e pouquíssima gente ganhando muito, isso dá um desequilíbrio, faz com que a média não represente muito bem o conjunto de dados. Vamos começar entendendo como calculamos a média. Vamos fazer manualmente primeiro, os outros faremos de outra forma.

[2:23] Vamos copiar as notas de fulano, e vamos colar dentro de um parêntese, para aplicar a formulazinha que deixei, todas as medidas deixo numa fórmula para você entender como é calculado, a média nada mais é Do que o somatório de todos os valores que estamos estudando, tu fala da nota de um fulano, a soma de todas as notas do fulano dividido pelo número de notas, de matérias.

[2:48] Que são aqui 7, então vamos trocar a virgula pelo símbolo de mais, porque temos que somar, um somatório, vou fazer para todos eles. Rapidamente.

[3:02] Tirar o mouse aqui da frente. E, para seguir nossa fórmula, temos que somar esses caras e dividir pelo número de itens que tem, número de matérias, temos 7.

[3:25] Somou tudo, dividiu por 7, temos a nota média do fulano, 7.71. Mas logicamente, não vamos fazer a média desse jeito toda vez, imagine um conjunto de dataset com 70 mil itens, somar 70 mil e depois dividir, não dá.

[3:45] O pandas tem um macete bem simples, disponibiliza uma função, vamos lá. Vou começar com o DF, por que DF? São as iniciais de Dataframe.

[3:55] Criei o DF, vou fazer a nota do fulano, mas de forma mais simples, DF quero fulano, e quero a média. Min. Só isso. Calculando a média, igual de cima.

[4:15] Voltando ao nosso dataset, vamos calcular a média pro nosso dataset, como fazemos isso? Da mesma forma, quero a renda média, .min.

[4:28] Pulou aqui, 2 mil reais a renda média. Lembra que tem as influencias que vamos ver nas próximas aulas, a gente vai ver que a renda média não é bem assim, a média está muito menor do que isso. Mas uma coisa interessante que eu queria mostrar para vocês aqui também, a gente tem, vou inserir uma célula aqui, para mostrar nossos dados.

[4:56] .head para mostrar só os 5 primeiros. Lembra que a gente no começo do curso estávamos falando dos tipos de variável que temos, e alguns tipos de variável, não conseguimos calcular a média, por motivo óbvio, as variáveis qualitativas nominais, categóricas, sexo, a gente tem a cor, você não consegue calcular a média, como que você vai calcular a cor média, o sexo médio, não tem como. O que podemos fazer é usar as variáveis para ajudar a gente nas análises, por exemplo, não posso calcular a média delas aqui mas posso usar ela como um by, calcular a renda média por sexo.

[5:37] Como vou fazer isso? Vou apagar esse, vou chamar de novo os dados, o pandas tem uma funcionalidade que se chama group by, e passo por qual variável quero agrupar meus dados, estávamos falando de sexo então vamos fazer com sexo, .sex, vou fazer uma coisa errada para você ver como isso funciona. .Min.

[6:05] Olha lá, ele calcula a média de todas as variáveis do meu dataset, do meu dataframe, para cada tipo de sexo, o 0 sabemos que é o masculino, e o 1 sabemos que é feminino, aqui, o F médio não faz sentido, então temos que escolher a variável que queremos calcular a média, podemos colocar aqui fora, antes do Min, pegamos a renda, e ele vai vir com a Renda aqui.

[6:42] É basicamente isso que eu queria mostrar sobre a média, você tem o ferramental para obter uma média, então no próximo vídeo, vamos partir para outra medida de tendência central que é a mediana, tá bom?

[6:53] Então no próximo vídeo a gente se vê, abraço.