04

Boxplot

Transcrição

[0:00] Ok, falando ainda sobre medidas separatrizes, vamos mostrar agora, o Boxplot.

[0:06] Eu deixei para falar do Boxplot agora, justamente porque ele é construído a partir de medidas separatrizes.

[0:09] A gente vai o usar os quartis pra construir o Boxplot, como você pode ver aqui em baixo.

[0:14] O Boxplot é uma representação gráfica bastante importante porque ele mostra muita coisa da nossa variável, da área de distribuição que a gente está estudando, como por exemplo: informações de posição; informações de dispersão; informações de assimetria; informações de dados discrepantes, como a gente vai ver aqui.

[0:30] Já inicialmente, eu posso mostrar pra vocês aqui esse desenhinho, e como ele é construído; quais as estatísticas que eu uso.

[0:36] Lembra do primeiro e o terceiro quartil em mediana?

[0:38] Elas estão aqui. O Q1 representa, essa parte aqui, representa o primeiro quartil, 25 por cento fica a baixo desse valor, 75 por cento para cima.

[0:47] O Q3 aqui, é o terceiro quartil, 75 por 25, aqui, temos a mediana; esse traço aqui no meio está representando a mediana.

[0:55] Essa diferença entre esses dois aqui, é o IIQ, que é o índice interquartil: intervalo entre o quartil ok?

[1:03] E esses extremos aqui, são construídos da seguinte forma: Q1 - 1,5xIIQ, esse carinha aqui; esse cara aqui, é uma informação de dispersão dos dados; já começa por aqui, aqui também, ok?

[1:18] E aqui, esse limite inferior: está bem abaixo desse limite aqui.

[1:23] E acima desse limite aqui, a gente tem candidatos a dados discrepantes; não quer dizer que eles são realmente outliers como está escrito aqui: perdoem-me, mas são candidatos a outliers.

[1:35] Então a gente como pesquisadora a gente vai ter que olhar com mais calma essas informações e decidir se realmente são ou não, ok?

[1:43] Vamos ver como a gente constrói alguns Boxplot, e vamos começar fazer algumas analisezinhas.

[1:49] Utilizando ainda o nosso amigo siban, eu vou fazer aquele mesmo esqueminha, a configuração já está pronta aqui, eu sou vou criar o Boxplot.

[1:59] Então vamos lá, vamos chamar o siban, eu estou assumido que você já rodou todo o código; ele está importado lá em cima; sns, que é o siban; eu passo aqui: Boxplot e aqui dentro eu vou passar inicialmente três parâmetros.

[2:18] Eu vou começar com x, o eixo, e vou fazer primeiro com a altura que seria o cara mais simétrico que a gente tem para você visualizar o Boxplot.

[2:31] Depois eu passo pra ele o data, e vou dizer de onde vem os meus dados, eu passei só a altura aqui, não diz de onde vem

[2:38] Então eu passo esse parâmetro data, e digo: para ele usar o data frawley dados, e pego a variável altura dentro da data frawley dados. para Boxplot ficar igualzinho esse aqui; se não ele vai desenhar ao contrario: em pé; eu vou passar outro parâmetro aqui: que é o orient-h, para ele ficar na posição horizontal .

[3:02] aqui tem o desenho do nosso Boxplot feito pelo siban, repare aqui, como eu disse: essas bolinhas seriam o dados candidatos a dados discrepantes, então a gente tem que tomar bastante cuidado quando for analisar esses dados aqui, principalmente com uma variável dessa, bastante assimétrica; vejam como ela é bem assimétrica mesmo, bem simétrica que é o dado da altura.

[3:22] Você vê que está bem dividido bem no meinho aqui, parece uma coisa, realmente você vê pelo Boxplot que é um dado simétrico, ok?

[3:31] Outra construção interessante desse tipo de figura, aqui, é que você pode comparar; fazer como se fosse uma Bayer: colocar outra variável nessa brincadeira aqui, para gente fazer uma comparação.

[3:43] Vou copiar aqui, o cara de cima, que eu vou fazer ainda com a altura.

[3:48] Você pode que os títulos estão de altura, metros, ok?

[3:51] Só que aqui agora, vou colocar uma variável extra no eixo y: Y= só para gente brincar aqui, vamos colocar o sexo, a gente lembra que o 0 é homem, se não me engano; e 1 é mulher.

[4:07] As outras configurações ficam do mesmo jeito. Este tipo de analise é bem interessante, a gente vai ver quando a gente fizer com a renda; quando a gente fizer com outro tipo de variável você vai ver que tem umas diferenças e a gente já pode criar algumas.

[4:20] Como está variável, foi montada por mim, ela é totalmente simétrica, perfeitinha.

[4:25] Aqui, a gente não vê quase diferença nenhuma. Se fosse uma variável perfeita, de altura, por exemplo: aqui a gente teria uma diferença bem significativa entre homens e mulheres concordam?

[4:35] Mas como essa variável é um fake, a gente não tira conclusão nenhuma dela.

[4:39] Então vamos lá, vamos continuar e vamos brincar com a renda, então; vamos lá, fazer essa mesma coisa: pode copiar e colar aqui.

[4:45] Só vamos mudar aqui, os parâmetros renda, o resto fica do mesmo jeito, Ok?

[4:54] Olha que violência, aquele cara de duzentos mil.

[4:58] Você nem vê o Boxplot aqui, de tão absurdo que é essa assimetria; fica evidente aqui a simetria à direita: O cara puxando muito para cá.

[5:08] Então só pra gente visualizar, vou fazer uma brincadeira diferente dentro de dados.

[5:13] Eu passo pra ele aqui uma query, vai fazer uma query pra gente selecionar só um grupo de pessoas.

[5:20] Por exemplo: vamos passar aqui, uma string dentro dessa query.

[5:26] Estamos trabalhando com a renda , eu quero só que ganha menos de 10 mil ,somente para a gente conseguir visualizar aqui, nosso Boxplot.

[5:36] Aqui já melhorou um pouco nossa visualização. Ainda continua simétrico, tá, você pode ver aqui o comportamento assimétrico desse cara, mas a gente já consegue visualizar ele melhor, tá legal?

[5:47] Você pode usar este artifício aqui, de query para você fazer algumas analises, umas visualizações de forma mais adequada, mais simples para você; você consegue mostrar e visualizar tudo melhor.

[5:59] Vou fazer a mesma coisa aqui, para esse cara, eu copiei tudo?

[6:02] Não é isso que eu queria.

[6:04] Vou copiar só isso aqui e vamos visualizar com essa mesma query aqui.

[6:09] Agora, jogando o Y aqui , e o sexo, vamos ver se a gente tem um diferencial aqui.

[6:20] Olha lá, olhe aqui: o tipo de analise interessante que a gente já pode começa a fazer.

[6:24] Lembra -se que esse cara aqui, são homens e aqui, são mulheres, São todos os chefes de família, são pessoas de referência na pesquisa; aqueles caras que responderam o questionário.

[6:32] Geralmente são os chefes da família.

[6:34] Olha como os homens tem uma renda superior as da mulheres.

[6:40] Existe essa desigualdade e a gente consegue visualizar isso nos dados aqui.

[6:44] Repara: a mediana é maior para o homem, a mediana mulher é menor.

[6:48] As variáveis: primeiro quartil, segundo e terceiro quartis estão aqui, também. Todos esses valores bem, a gente vê aqui, que tem uma diferença, já pode fazer uma pequena analise aqui, mostrando que a Wendy indica uma diferença no nível de renda por sexo, ok?

[7:05] Continuando: Vamos fazer agora, com anos de estudo; vamos assumir que esses anos de estudo são realmente anos e não uma classe, ok?

[7:13] para fingir de estudo, para não atrapalhar o nosso estudo aqui.

[7:20] Então vamos lá, vamos fazer aqui: anos de estudos, ok?

[7:36] Olha lá, aqui. Para geral os anos de estudo: a gente vê que tem uma leve puxada para o lado de cá, uma leve simetria a esquerda, a mediana bem puxada para o lado de cá, a gente tinha verificado isso anteriormente.

[7:52] Vamos fazer este mesmo cara com sexo, para a gente ver se tem uma diferença, e agora, se o homem tem uma vantagem ou desvantagem.

[8:01] A gente já ouviu falar sobre isso: que as mulheres têm as qualidades maiores que as dos homens; e ganha menos; uma coisa estranha, mas acontece.

[8:08] Então vamos lá ver. fazendo com o sexo e a gente comprova isso mesmo.

[8:14] Aqui, você vê que o nível, a quantidade de estudo das mulheres que são chefe de família é superior a dos homens, você vê que a mediana é maior e já estão a 12, a dos homens estão entre 8 e 10, talvez 9. Interessante; você vê que essa diferença: para cima na educação e para baixo , no nível de renda.

[8:45] Percebeu como é interessante esse tipo de análise com o Boxplot?

[8:49] Usando medidas separatrizes a gente consegue visualizar essas coisas.

[8:52] aqui eu deixei uma figurinha muito semelhante aquela que a gente já tinha visto lá da simetria, onde a gente comparava as estatísticas descritivas: moda, mediana e media.

[9:04] Aqui a gente consegue também fazer uma visualização no Boxplot e ter uma ideia se é assimétrica direita, a esquerda ou se é totalmente assimétrica.

[9:13] A gente verificou isso aqui em cima, olhem aqui: assimétrica a direita; aqui perfeitamente assimétrica, a gente construiu essa variável.

[9:21] E aqui a gente tem uma leve assimetria a esquerda: no caso as mulheres bem fortes bem parecido com isso aqui estão vendo?

[9:32] Aqui seria: a altura e aqui será a renda, legal?

[9:35] Pessoal, é isso que eu queria mostrar para vocês; vai para a ultima sessão falar de medidas de dispersão: definir padrão, variância, por aí, vai, beleza?

[9:45] Vejo você lá. Abraço.