05

Tratamento de pontos discrepantes

Transcrição

[00:00] Vamos dar continuidade no nosso curso de análise de dados, usando o RStudio, agora eu vou falar para vocês como que a gente faz o tratamento de outliers, pontos discrepantes na base de dados. Então, outliers... da gente conhecido, geralmente por ser... dá um problema muito grave, se a gente não fizer esse tratamento na base de dados.

[00:23] Então, pontos discrepantes, são aqueles pontos que alteram o nosso valor da média, alteram a conclusão da análise, quando a gente está fazendo essa análise de dados dentro do R ou utilizando qualquer outro tipo de programa, então conhecido também como outliers. Então, primeiramente aqui, a gente vai utilizar para demonstrar isso, vamos verificar como é que está distribuído o número de filhos.

[00:54] A gente está usando ainda, só lembrando, a gente continua usando a nossa base de funcionários, então a gente vai utilizar... vamos investigar como é que está a distribuição do número de filhos desses funcionários na base de dados. Então, vamos usar os recursos que a gente já viu nas aulas anteriores.

[01:12] Então, a primeira coisa que a gente vai utilizar aqui, uma visualização gráfica, eu vou abrir um chunk aqui, então vamos fazer uma... a visualização gráfica da variável, número de filhos. Vamos entender como é que o número de filhos está distribuído na gráfica. Então, a gente vai usar aqui qplot, a base que a gente está utilizando, é a func_t, funcionário já tratada, a gente já fez alguns tratamentos aqui.

[01:56] E setando a variável filhos. Vou executar. Então, quando a gente executa aqui, a gente tem a distribuição no eixo X, o número de filhos e funcionários e a gente percebe que tem uma grande concentração aqui no número de filhos mais baixos, aqui não dá para ver a escala aqui, mas a gente vê que tem um ponto aqui que está marcado acima de 40 filhos, isso aqui deve ser algum erro na base, ninguém vai ter mais que 40 filhos, é impossível.

[02:27] Então pode ser um erro de digitação, alguma coisa, então esses dados que geralmente atrapalham as conclusões, tomar sempre cuidado, dar uma olhada, por isso que é importante fazer análise descritiva, cuidado para não acontecer esse tipo de coisa, porque se a gente pegar e olhar... eu vou colocar aqui, vamos olhar a estatística descritiva, então usar a função stat.desc.

[02:51] De novo, você seta a base, func_t e a variável filhos, vou executar. Então, quando a gente faz a análise aqui, se a gente fosse tomar alguma decisão sem olhar com um pouco mais de critério, o que que a gente estaria falando? A gente está falando que o número de filhos médios desse grupo de funcionários é 4.16, ou seja, um número elevado de filhos até... fala: “Poxa, na base de dados, então o número médio é 4 filhos”.

[03:23] A mediana dois filhos. Então a mediana, ela já faz essa correção. Então, por isso que muitas vezes é importante até utilizar a mediana, do que a média. A média... a mediana não é muito alterada por pontos discrepantes, já a média é, você vê que a média está dando 4.19, possivelmente por causa desse ponto aqui, que está elevando para cima.

[03:48] Uma outra visualização gráfica que a gente poderia estar utilizando aqui, que a gente já... a gente pode utilizar, que a gente já viu, que é bastante importante, é o gráfico de boxplot. Então, a gente utiliza o boxplot aqui também na nossa base e número de filhos. Quando a gente fizer o boxplot, a gente vai perceber isso.

[04:12] O boxplot já nem dá para ver muito, por quê? Porque realmente, está parecendo esse número de filhos aqui com mais de 50 filhos, possivelmente, com certeza é um erro de digitação, um erro da base. O que que a gente vai ter que fazer? A gente vai ter que excluir esse ponto da nossa análise, porque senão a gente vai tomar conclusões errôneas.

[04:34] Então, vamos excluir esse ponto e vamos ver como é que fica. Para excluir esse ponto, a gente utiliza uma função que a gente... eu mostrei aqui para vocês, que é o filter, então eu vou excluir ou melhor até, para lembrar aqui, eu vou filtrar a base de dados. Como é que eu vou fazer esse filtro?

[05:05] A gente já sabe que esse ponto discrepante aqui, ele está acima de 50, a gente poderia classificar em 50 mesmo ou então, a gente pode ser um pouco mais razoável, o que que seria um número de filhos naturalmente? Isso, a gente pode fazer um corte para filtrar esse ponto aqui, pode colocar acima de 10 filhos, por exemplo, também já é uma coisa bem absurda, mas ok.

[05:30] Aí, você vê que entre 10 e 50, aqui não tem ninguém, provavelmente todo mundo vai estar abaixo de 50. Então, vamos fazer isso, vamos pegar e filtrar a nossa base aqui, colocando um corte, para que se a gente tiver um número de filhos acima disso, a gente não vai utilizar na análise. Então, o que que a gente usa?

[05:47] A gente usa a... a gente vai usar o filter, então eu vou criar uma nova base aqui, func_t2, aonde... Opa, deixa eu colocar o chunk aqui, não pus, tudo dentro aqui, só passar para dentro aqui, senão depois a gente executa, não vai funcionar. Aqui dentro, ok. Então, essa base nova vai ser o quê? O que que a gente vai fazer?

[06:16] A gente vai usar a função filter, a gente filtrar na nossa base, que a gente está utilizando, que é a base atual, func_t e a gente coloca a condição, filhos, vou colocar aqui, menor, igual a 10, 10 filhos já é um número bem grande, mas como a gente já viu que não tem ninguém aqui, então, a gente poderia colocar até menor, mas ok. Vou colocar aqui, executei.

[06:48] Então, criei essa base aqui de funcionário dois agora e agora a gente vai fazer a nossa análise aqui em cima do funcionário dois, vamos ver como é que fica, a primeira coisa que a gente vai fazer já, vamos repetir a análise da estatística descritiva aqui nessa base filtrada e setando a variável filhos, vamos ver como é que ficou.

[07:12] Olha, agora já ficou bem mais razoável, você vê que a mediana não foi alterada, a mediana que a gente tinha aqui, mesmo com o outlier, mesmo com esse ponto discrepante era dois, então olha, a mediana, ela não é alterada, ela não é influenciada por valores muito grandes ou muitos baixos na nossa base. A média era 4.19, agora ficou 1.65, bem mais razoável.

[07:36] Então, a gente está falando que na nossa base de dados aqui, o número de filhos médio é de 1.65, ficou bem mais razoável. Vamos repetir aqui, colocar de novo... escrever o nosso boxplot, vamos olhar como é que fica o boxplot agora, dessa variável, apontando para func_t2 agora, que é a nossa base filtrada, número de filhos. Vamos ver como é que ficou.

[08:07] Opa, não rodou. Vamos ver de novo porque não rodou. Vamos ver de novo. Ah, está aqui, rodou. Então, olha, veja bem, aqui é o boxplot. Você vê que a linha escura aqui é exatamente a nossa mediana, agora sim, agora a gente consegue visualizar já. Então assim, tem um ponto discrepante aqui, com uma família com cinco filhos, mas ok, cinco filhos é uma coisa razoável, até mesmo porque a gente viu aqui que não tem ninguém acima de cinco filhos.

[08:41] O mínimo aqui com zero filhos e o máximo com esses cinco filhos, então já ficou bem mais razoável. Aqui, sim, a gente poderia tirar essa conclusão, que olha, você tem família com zero filhos, máximo cinco filhos, o número médio de filhos é 1.65, a mediana dois e um desvio padrão de 1.2 filhos por família de funcionários.

[09:06] Então, muito importante fazer esse tratamento realmente dos pontos discrepantes, para que não haja nenhuma dúvida ou a gente tome conclusões errôneas na hora em que a gente está fazendo essa análise de qualquer campo, de qualquer variável, aqui dentro da nossa base de dados. Era isso, então, que eu tinha para mostrar para vocês aqui de pontos discrepantes, espero que vocês tenham gostado.