04

Visualização Gráfica - Boxplot

Transcrição

[00:00] Vamos continuar o nosso curso de análise de dados, utilizando o RStudio. Então, vamos voltar aqui, hoje eu vou falar para vocês uma utilização de um gráfico chamado boxplot, que ele auxilia na análise de dados, sempre com conjunto de dados, a responder algumas perguntas de imediato, sem que e a gente precisa fazer muita transformação, muita análise.

[00:27] Então, a gente vai usar essa visualização gráfica pelo boxplot, só lembrando aqui, os pacotes que a gente está utilizando, que a gente já... mostrei para vocês como é que instalava, mas a gente precisa chamar esses pacotes, que é o tidyverse, magrittr, dplyr.

[00:41] Se der problema em algum desses pacotes, na hora que vocês estiverem utilizando, reinstalem de novo, é só dar um install packages e chamar o pacote. Executar a biblioteca, já está executada. Lembrando, a base que a gente está utilizando aqui, é uma base de funcionários, ainda vamos tratar essa base de funcionários.

[01:02] Essa base de funcionários, nós temos aqui os dados de alguns funcionários da empresa, referente à salário, idade, número de filhos, então é uma base de funcionários. Então, a primeira coisa que a gente vai fazer aqui é essa visualização gráfica boxplot.

[01:28] Eu vou abrir aqui um chunk novo e para utilizar o boxplot, a gente vai fazer da seguinte maneira. É só colocar o boxplot aqui, a variável que a gente quer e pela instrução. Então, boxplot, vou fazer a segunda pergunta, a análise que a gente quer fazer aqui de imediato, por exemplo.

[01:48] Então, vamos imaginar, será que o salário que existe nessa base, será que o salário que a gente tem na base aqui, ele se difere por diferentes graus de instrução. Então, a gente poderia perguntar, uma pergunta clássica aqui, será que o salário dos funcionários difere por diferentes graus de instrução?

[02:24] Então, como é que a gente responderia essa pergunta? Pelo boxplot, a gente vai conseguir fazer essa diferenciação já e responder de imediato. Então, a gente vai colocar aqui a nossa variável no nosso campo, só salário, eu coloco aqui o tio, que ele vai salário por, ele tem a ideia desse tio, salário por, instrução.

[02:58] E buscando a nossa base, func_t, que é funcionário tratado, onde a gente já fez alguns tratamentos, retirou duplicidade, a gente já fez esse tratamento anteriormente, vamos continuar usando essa base. Vou executar aqui. O que que ele faz? Então, ele monta esse gráfico de caixas aqui, o que que representa esses gráficos de caixas.

[03:22] No eixo Y, eu tenho salário, então tem o número e salários, cinco salários, 10, 15, 20 salários mínimo, aqui no eixo X, o grau de instrução, 1º grau, 2º grau e superior. Essa reta do meio aqui, ela representa a mediana da base, mínimo a linha inferior, máximo a linha superior, isso aqui é como se fosse uma distribuição de quartil, eu ordeno toda a base e separo, fatio a base.

[03:50] Então aqui eu tenho, a mediana representa 50% dos dados ordenados a esquerda, então é como se fosse uma balança, eu tenho 50% a esquerda e 50% a direita, todos ordenados. A linha de baixo aqui, eu tenho 25% do primeiro quartil e aqui 75% do primeiro quartil.

[04:08] Quanto mais aberta essa distribuição, os dados vão estar mais dispersos, quanto maior a caixa, mais disperso, quanto menor, menos disperso e aqui, a gente já conseguiu ver então uma visualização, que a gente vê nitidamente que

essas medianas aqui, se diferem em relação ao grau de instrução, quanto maior o grau de instrução, quem tem nível superior, maior o número de salários.

[04:30] Então, rapidamente a gente já vê essa diferença claramente aqui. Um outro boxplot que a gente poderia estar executando aqui, eu vou copiar o comando, eu vou trocar a variável instrução por região. Então, outra pergunta, será que salário difere... tem diferença por região? Então, também a gente poderia buscar essa pergunta.

[05:05] E assim, note que eu estou usando uma variável quantitativa que é salários, versos uma que é do tipo fator, são qualitativos, então eu posso cruzar essas duas variáveis, ele vai fazer a média mediana dos salários, da variável quantitativa e vai classificar pela minha variável de fator, a minha variável região.

[05:27] Eu vou executar essa linha aqui, então, veja bem, ele executou. O que que ele fez? Ele fez um outro gráfico aqui, que de fato, ele colocou salário na capital, salário no interior e outros. Veja que o salário mediano da capital, ele é até menor, a linha está abaixo de 10 salários mínimos e no interior está até acima.

[05:51] Porém, a dispersão aqui na capital é maior, ou seja, eu tenho salários até mais concentrados mais altos aqui na população, na capital e aqui no interior, eu tenho essa dispersão muito grande, em outros aqui, não tem nada, não apareceu nada. Se eu executar, vou executar os dois aqui ao mesmo tempo, ele vai fazer esses dois gráficos aqui.

[06:16] Veja bem, esses dois gráficos, eu poderia fazer uma outra pergunta aqui agora, será que existe diferença entre o grau de instrução em diferentes regiões? Então, eu estaria juntando essas duas informações e que é um outro tipo de gráfico que a gente pode usar o boxplot, como se fosse um grid agora.

[06:37] Eu vou fazer a seguinte pergunta. Então, a pergunta que eu poderia estar fazendo aqui, que eu vou fazer para a gente tentar responder, para a gente responder pelo boxplot, tentando juntar essas informações é a seguinte pergunta.

[06:54] Então, existe diferença aqui entre salário, versos grau de instrução, que a gente já viu que realmente tem, grau de instrução a gente viu no primeiro boxplot, que existe essa diferença, mas eu vou colocar só mais uma variável aqui, para a gente tentar observar isso junto: salário, grau de instrução e região.

[07:38] Vou colocar isso, existe a possibilidade de a gente colocar isso num mesmo gráfico, num boxplot. Então, a gente vai montar esse gráfico, que é como se fosse um grid. Veja bem, é como se tivesse três informações aqui, eu não vou montar um gráfico com três eixos, aí da para fazer gráficos tipo grid, que eu vou mostrar aqui.

[07:58] Então, para a gente fazer esse tipo de gráfico, a gente vai usar, a gente precisa usar uma outra função auxiliar aqui do R, que é o ggplot. O ggplot possibilita fazer esse tipo de gráfico. Então, ggplot, ele precisa... o primeiro comando aqui, o primeiro é data, a base que você está usando, a base que a gente está utilizando aqui é a func_t, funcionário tratado.

[08:32] Aí, eu coloco mais, agora, eu começo a colocar as opções do meu gráfico aqui. Aqui, eu tenho que definir qual que é o tipo de gráfico que eu vou usar, no ggplot, eu tenho vários tipos de gráficos diferentes aqui, ele abre, geométrico, barra, boxplot. O tipo de gráfico que a gente vai usar, ainda é o boxplot.

[08:56] Então, geométrico boxplot, opção de mapping, como se ele fizesse um mapa, um grid que ele vai estar fazendo, aonde eu estou colocando no eixo. Agora, eu vou ter que definir os meus eixos, X e Y, o aes, eu defino os meus eixos X e Y, onde no eixo X, eu vou colocar aquela informação da minha região.

[09:21] Então, vai aparecer região, vírgula, defino aqui o eixo Y, quem vai ser a informação do eixo Y, continua sendo salário, como nos dois boxplot anteriores que a gente acabou de fazer e aqui existe a possibilidade de eu colocar... aí que entra essa diferença.

[09:44] Eu vou colocar aqui, você define o color, você consegue definir o color aqui, como sendo uma outra variável, que eu vou colocar aqui a minha instrução. Então, eu vou colocar aqui essas variáveis e por fim, eu só preciso fazer mais uma outra condição aqui, colocar facet_grid.

[10:17] E aí, eu falo o que que eu quero, que é a minha região, é a variável região, por... que é o tio e aí um ponto. Então, ele vai cruzar essas regiões pelas demais variáveis aqui. Vou executar. Olha, veja bem o que que ele fez. Então, ele colocou aqui... Ah, o que que aconteceu?

[10:47] O gráfico não saiu colorido, ele separou salário, vamos ver porque que ele não saiu colorido, provavelmente o nosso comando aqui de colors não pegou, não está certo, mas a gente já vai arrumar. O eixo X tem salário... no eixo Y, salário, quer dizer. No eixo X a região e aí, um grid aqui, exatamente da região separado por capital, interior e sem classificação, que é o NA.

[11:16] Vamos arrumar aqui o colors, acredito que seja isso aqui, color, igual, vou rodar de novo. Ah, agora sim, então color, igual a instrução. Veja bem, o que que ele fez? Ele está definindo a... agora ficou colorido, porque eu estou falando para ele definir a cor por uma variável, isso no ggplot é possível.

[11:40] Então, eu tenho aqui a minha variável vermelha, a cor vermelha define o 1º grau de instrução, a cor verde o 2º grau e azul é o grau superior. Então, a nossa pergunta, qual que era? Existe diferença entre salário, grau de instrução e região?

[11:59] Olha, vamos ver, na capital, eu sei... eu já sabia anteriormente que existia diferença aqui, entre diferentes graus de instrução, mas de modo geral. Agora, a gente colocou, foi mais a fundo e olhou, falou: poxa, na capital, como é que tem? Olha, capital existe essa diferença nitidamente, então quem tem nível superior aqui recebe até... chega até 20 salários mínimos.

[12:24] Veja que existe plano superior... existe essa dispersão muito pequena, a caixa do boxplot está bem pequena, então existe essa pouca dispersão. Quando a gente olha para interior, essa dispersão já é maior.

[12:36] Existe também essa diferença de novo, de 1º, 2º grau, quem tem nível superior ganha mais, superior aos demais, mas, porém e com uma dispersão maior aonde 2º grau e superior ainda é uns limites aqui de primeiro quartil e terceiro quartil se encontram.

[12:53] E quando a gente vai para uma outra classificação aqui, que é o sem classificação na região, está como NA, são um pouco mais misturados. Então, veja bem, esse gráfico, a gente conseguiu responder muitas coisa, a gente conseguiu juntar muita coisa e responder muita coisa.

[13:09] Esse tipo de gráfico é bastante interessante para a gente... em visual também, aonde a gente consegue responder perguntas desse tipo, quando eu quero cruzar duas ou três variáveis, quando eu faço o boxplot sozinho, eu tenho essa informação, porém eu não consigo fazer essa junção, esse grid.

[12:32] Então, aqui eu poderia estar cruzando também outras variáveis, como por exemplo o salário versos região, número de filhos, estado civil e assim por diante, depois até tentem fazer, mudar essas outras informações cruzando com estado civil, cruzando com número de filhos.

[13:55] Esse boxplot é bastante interessante, realmente, para a gente ter uma visualização rápida da base de dados, cruzando informações de variáveis quantitativas, versos informações qualitativas de classificação, a gente consegue realmente ter um mapa bastante interessante.

[14:14] Então era isso que eu tinha para mostrar sobre o gráfico de boxplot, espero que vocês tenham gostado.