

Gráfico de pontos II

Transcrição

[0:00] Nós geramos dois gráficos: um para ciência humanas, que é esse gráfico aqui na aba inferior direita e outro para matemática, correto?

[0:11] Esses gráficos poderiam ser úteis porque transmitem a informação principal, ou seja, a relação da idade com a média de cada respectiva matéria.

[0:23] Porém, tem uma desvantagem que seria quando comparar a média de uma nota por idade de uma forma mais fácil.

[0:31] Por exemplo, eu tenho aqui os dois gráficos separados, aqui eu só tenho de matemática e aqui eu só tenho de ciências humanas. Como eu poderia visualizar, comparar, as duas médias com os dois agrupamentos de idade de forma mais fácil? Correto?

[0:46] Isso facilitaria a nossa vida, a vida do analista de negócios de dados, que você que está construindo o gráfico e também o principal objetivo, facilitar o entendimento do nosso cliente, que é o cursinho, os analistas de negócio do cursinho ou até mesmo os professores.

[1:06] Após gerar os dados com a média de idade para ciências humanas e matemática, que estão nos conjuntos de dados separados, ou seja, esses dois grupos aqui, `notas_ciencias_humanas_idade` e `notas_matematica_idade`, ambos os conjuntos de dados têm em comum o campo `idade`, correto?

[1:31] Se executarmos aqui `ó`, `média_ciencias_humanas_idade` e a matemática também vai ter em comum a `idade`.

[1:41] E outro campo que tem a média de cada matéria, respectiva matéria, matemática e ciências humanas.

[1:50] Para gerar um gráfico com ambos os conjuntos de dados faremos um `merge` dos conjuntos de dados, desses 2 conjuntos de dados, transformando em 1 só.

[1:59] Então vamos chamar o novo conjunto de dados de `notas_ciencias_humanas_matematica_idade` e vamos fazer um `merge` `notas_ciencias_humanas_idade` e `notas_matematica_idade`.

[2:23] Vamos colocar parâmetro `all` com o `to`.

[2:31] Esse parâmetro `all` ele vai fazer o `merge` no conjunto por inteiro, então para não haver nenhum problema na hora de executar. Pronto. Fizemos um `merge`. No final, todos os dados vão estar em um mesmo objeto, que eu vou te mostrar agora.

[2:47] Porém, para elaborar o gráfico que desejamos, ou seja, obter informações das médias, notas de ciências humanas e matemática por idade, vamos ter que fazer uma outra transformação.

[2:58] Qual será essa outra transformação? Eu vou mostrar aqui já para você. Ó: `nota_ciencias_humanas_matematica_idade` vamos dar um `view`. Temos aqui o campo `idade`, média de ciências humanas e média de matemática.

[3:13] Então para cada idade vai ter uma média.

[3:20] E aquele parâmetro `all` é justamente esse daqui: a idade 98 tem uma média para uma média de ciências humanas, mas para a média de matemática não há uma média para essa idade. Então a própria função inseriu o `NA` por conta daquele parâmetro `all`.

[3:37] Se não tivesse esse parâmetro aqui, ele ia dar um erro de execução e não ia realizar o merge.

[3:43] Então vamos fazer agora as devidas modificações para gerar o gráfico que desejamos.

[3:49] A transformação que nós faremos agora nesse conjunto de dados aqui `notas_ciencias_humanas_matematica_idade` nós vamos converter a coluna em linha, ou seja, vamos converter essas duas colunas aqui em linhas, para cada idade, haverá um registro para média de ciências humanas e uma outra para matemática.

[4:07] Para fazer isso, você vai utilizar a função, o pacote `library reshape2`.

[4:22] Se você não tiver instalado esse pacote, `install.packages`, do primeiro curso, que nós também utilizamos, você vai instalar aqui `install.packages`, `reshape2`, executa esse comando aqui e após instalar, você executa esse comando `library` para habilitar o pacote.

[4:44] Então, como é que vai ficar aqui nossa... vamos chamar um novo conjunto de dados `notas_ciencias_humanas_matematica_idade`, que é o objeto que nós criamos anteriormente.

[4:59] E vamos chamar a função `melt` `notas_ciencias_humanas_matematica_idade`, perceba que eu estou usando o mesmo objeto, ou seja, estou subscrivendo o mesmo objeto, porque é uma modificação simples e se der um erro eu posso executar aqui o merge novamente, que é uma função também bem simples e que não vai gerar prejuízo nenhum.

[5:31] E também passar aqui o `id.vars`. Esse parâmetro `id.vars` indica qual a minha coluna dessas 3 aqui vai ser mantida no meu conjunto de dados e não vai ser transformada em linha. Vamos executar aqui para você ter uma visualização. Não é `true` e sim o nome da coluna: `idade`.

[5:58] Pronto, executei, vou abrir aqui.

[6:02] Nós temos aqui ó, agora todos os dados. Continuamos ainda com 3 colunas, porém a coluna de média de ciências humanas se transformou em linhas, vou ordenar aqui pela idade, ó, virou uma linha e média da nota de matemática virou outra linha.

[6:26] E assim sucessivamente. Então para cada idade agora eu tenho uma linha com a média para determinada matéria: ciências humanas e matemática. 19 anos, média de ciências humanas, 19 anos, média matemática, ok?

[6:42] Por que eu fiz isso? Nós fizemos isso para gerar esse gráfico bonito agora: `ggplot2`, `data` `notas_ciencias_humanas_matematica_idade`, vamos concatenar aqui, `geom_point`, `aes` vamos mapear aqui o eixo x, que é a coluna `idade`, o eixo y, que é com a nova coluna agora, `value`, vamos olhar aqui novamente, olha.

[7:22] O próprio R atribuiu o nome dessas colunas, tá? Não se preocupe, esses nomes de colunas não são atribuídos de forma manual e sim pela própria função. `Value` vai na coluna, no eixo y.

[7:43] E agora nós vamos definir as cores, com o `variable`, que também é a nova coluna do nosso conjunto de dados. Vamos executar aqui, vai dar um warning. Lembrando que warning não é erro.

[7:54] Esse warning aqui está dizendo que foram removidos aqueles dois valores NAs que eu mostrei anteriormente para você, que tinha em um conjunto de dados e não tinha em outro, e o próprio R insere o valor NA, e como não é possível plotar NA, a própria função remove.

[8:13] E aqui nós temos o nosso gráfico, vamos dar um zoom aqui.

[8:15] A legenda `variable`, em vermelho é a média de notas para ciências humanas e em azul a nota para matemática. Agora nós podemos visualizar o gráfico mais facilmente.

[8:26] Agora nós podemos fazer uma análise por idade da médias das duas matérias de forma mais fácil e junto.

[8:35] Por exemplo, nesse gráfico, podemos observar que há um padrão na média para notas de ciências humanas aqui em vermelho, como foi visto anteriormente no gráfico quando ele estava sozinho, porém há uma queda para ciências, desculpa, para a matéria de matemática a partir do momento em que o aluno vai ficando mais velho.

[8:57] Entre 20 e até 60 anos, mais ou menos, nós temos um padrão para a nota de ciências humanas, já em matemática é o contrário: entre 20 e 60 anos, a nota já vai caindo consideravelmente.

[9:14] A partir disso, ambas as matérias ficam bem dispersas, a partir de 60, 62 anos, quanto mais velho, mais disperso as notas ficam.

[9:26] Outra informação bem importante, além de matemática serem as notas menores de acordo com que as idades ficam maiores, ou seja, a pessoa fica mais velha, as médias das notas de matemática vão diminuindo e não mantêm um padrão.

[9:47] Com essas informações, os gestores do cursinho podem intensificar as aulas nas matérias principalmente de matemática, que eles podem observar que há um desempenho pior em relação à matéria de humanas, então podem dedicar mais os estudos dos alunos nessa matéria aqui, com a intenção de melhorar as médias, pelo menos do cursinho, e que podem influenciar as médias estadual e até mesmo nacional, inclusive de pessoas de mais idade, que aqui quanto mais velho vai ficando, mais a média vai decaindo.