01

Gráfico de bolhas I

Transcrição

[0:00] Ainda analisando a média das notas, o cursinho deseja obter informações sobre as médias de ciências humanas, matemática e redação apenas para os locais onde ela tem uma filial.

[0:14] O objetivo do cursinho é fazer uma comparação e tentar identificar se há alguma relação entre essas médias de ciências humanas, matemática e redação, assim também sendo possível identificar alguma deficiência ou sucesso nos estudos nas regiões onde ela tem filial.

[0:33] O primeiro passo nós vamos fazer um filtro dos valores NA, como vocês já sabem bem.

[0:38] Também vamos filtrar os estudantes que tenham acima de 17 anos e vamos selecionar os registros apenas nos locais onde o cursinho tem filial, ou seja, Ceará, Distrito Federal, Minas Gerais e Rio Grande do Sul.

[0:52] E por fim calcular a média para as matérias de ciências humanas, matemática e redação.

[0:59] Anteriormente, nós fizemos todas essas operações separadamente, agora nós vamos fazer tudo junto ,que você já está mais familiarizado com o código, com a biblioteca dplyr, vamos fazer tudo junto, para você ver também como é possível fazer diferentes operações em um mesmo comando.

[1:19] Vamos colocar aqui o Enem, que é o pacote, desculpa, que é o conjunto de dados que nós vamos trabalhar, a função filter que vocês já conhecem bem, a negação, que é exclamação. Is.na nota ciências humanas a condição e, ou and, is.na nota matemática e is.na nota redação.

[2:01] Vamos colocar o e aqui, vamos quebrar a linha para ficar mais visível. Um outro and, que agora vamos fazer o filtro is.na da coluna idade, porque isso pode interferir, vamos analisar com a idade também, onde também idade seja maior que 17 e por fim, e UF Prova, vamos quebrar a linha aqui para ficar melhor de visualizar, vamos utilizar uma nota condição agora.

[2:47] O in aqui é uma condição para verificar se aquele registro está contendo, está presente dentro de uma lista de opções, no nosso caso vai ser os locais onde o cursinho têm filiais, que é o Ceará, vírgula, Distrito Federal, Minas Gerais e Rio Grande do Sul.

[3:14] Fizemos aqui a parte do filtro, como você pode ver, fizemos várias condições em um mesmo filtro.

[3:19] Porém, vamos fazer já agora o cálculo das médias group_by, idade, UF Prova, vamos agrupar por essas duas colunas e fazer a média, summmarise, média nota matemática recebe mean da coluna nota matemática, vírgula, média nota ciçencias humanas mean nota ciências humanas, vírgula, média nota redação mean nota redação.

[4:17] Pronto, fizemos várias operações em um comando só, de uma vez e vamos salvar tudo isso em um novo objeto chamado notas matemática redação.

[4:32] Vamos alinhar aqui as operações para ficar formatado e vamos executar.

[4:46] Criamos um novo objeto, vamos dar um view aqui nesse objeto para você ver como é que ficou. Notas matemática idade.

[4:59] Como você pode ver aqui ó, opa, matemática, redação.

[5:09] Temos a coluna idade, UF Prova, a nota em matemática, nota de ciências humanas e média nota redação. Então temos o cálculo da média para cada idade e para cada estado.

[5:21] Então você pode ver que têm várias linhas com 18 anos, para os estados que desejamos, Ceará, Distrito Federal, Minas Gerais e Rio Grande do Sul.

[5:28] Agora nós temos o conjunto de dados com os dados que nós desejamos, vamos gerar o gráfico contendo esses registros.

[5:35] Primeiro vamos chamar aqui a função ggplot, data, vamos passar o conjunto de dados notas matemática redação que acabamos de criar, geom_point, função que utilizamos anteriormente para gerar os gráficos de pontos.

[5:59] O eixo x vai receber média nota ciências humanas, o eixo y vai receber média nota matemática.

[6:12] Vamos executar aqui e vamos fazer o passo a passo para você ver o que está acontecendo.

[6:16] Criamos um gráfico semelhante ao que a gente criou no começo do curso, que é um gráfico de pontos entre duas variáveis numéricas: a média matemática aqui no eixo y e ciências humanas no eixo x. Ok?

[6:32] Vamos primeiro agora inserir mais uma outra informação.

[6:38] Podemos usar o parâmetro color para diferenciar as cores das bolhas para UF Prova.

[6:47] Vamos copiar esse código aqui que vai ser o mesmo. Copiamos aqui e vamos inserir o parâmetro color, e vamos inserir UF Prova, o nome da coluna UF Prova. Vamos executar, opa, está errado aqui color.

[7:07] Vamos executar aqui esse código. Pronto.

[7:20] Agora nós temos informações da média nota matemática, e nota de ciências humanas por estado.

[7:24] A própria lib, a função ggplot preencheu as cores de acordo com os valores distintos da UF Prova, como fizemos anteriormente para cada das matérias de matemática e ciências da natureza que fizemos anteriormente.

[7:41] Agora só falta inserir a outra variável que é a média da redação. Lembrando que nós fizemos com 3 variáveis. O nosso objetivo é inserir 4 variáveis no mesmo gráfico.