02

Gráfico de Pontos

Transcrição

[0:00] Outra demanda, outra informação importante para o cursinho ficar sabendo é ver a relação da média das notas da matéria de ciências humanas e matemática com a idade dos candidatos que fizeram as provas.

[0:16] Essa análise é importante para o cursinho traçar estratégias de ensino para respectivas matérias. Por exemplo, se o método adotado está atendendo as necessidades, ou se deve ser alterado.

[0:26] E se o método atual atende a todas as faixas etárias e se há a necessidade de intensificar os estudos das matérias em ciências humanas ou matemática para aumentar as médias.

[0:42] Em análise de dados há um gráfico bem específico para verificar relação entre duas variáveis: é o scatterplot ou também conhecido como gráfico de pontos.

[0:52] Primeiro, vamos utilizar um filtro para criar um novo subconjunto de dados para construir esse gráfico de pontos que vamos fazer agora.

[1:01] Vamos então eliminar, vamos chamar nossa base de dados Enem e vamos eliminar todos os registros, você já sabe bem, NA, ou seja, registros com valores não definidos, da coluna nota ciências humanas, só que agora vamos utilizar uma condição que é o &, um and. em lógica de programação e NA na coluna idade e todos os registros com idade maior que 17 anos.

[1:38] Por que essa condição da idade? Porque as pessoas que fazem essa prova que têm menos de 17 anos geralmente são treinee, isso está de acordo com as leis, as normas do próprio ENEM.

[1:51] Então a nota válida é apenas a partir dos 17 anos que a nota começa a ser válida. Então vamos fazer esse filtro para melhorar a acurácia, melhorar as informações que nós vamos coletar.

[2:09] E vamos salvar isso tudo num objeto chamado notas_ciencias_humanas. Vamos executar.

[2:20] Como nosso objetivo aqui é analisar a média de cada matéria de ciências humanas e matemática por idade dos candidatos e candidatas, agora nós temos que calcular a média, com base nesse novo conjunto que nós acabamos de criar.

[2:34] Então vamos calcular a média notas_ciencias_humanas, vamos usar esse objeto, vamos aqui colocar fazer um group_by, como você já deve saber, porque nós queremos agrupar com campo idade, porque nós queremos a média da matéria ciências humanas com base na idade.

[3:01] Vamos vir aqui e vamos criar um novo conjunto de dados com a função summarise, vamos colocar aqui media_nota_ciencias_humanas, vai receber a função mean, que vai calcular a média na coluna nota_ciencias_humanas desse conjunto de dados aqui que nós estamos trabalhando.

[3:25] E vamos salvar tudo isso num conjunto chamado notas_ciencias_humanas_idade.

[3:38] O nome está um pouco grande, mas não se preocupe com isso porque é apenas uma forma de ficar mais intuitivo.

[3:44] Você futuramente pode escolher o próprio nome das variáveis, um nome menor, mais resumido, a sua escolha.

[3:54] Vamos executar esse código agora. Pronto. Vamos visualizar esse conjunto aqui no console.

[4:04] Nós temos duas colunas? Idade e media_nota_ciencias_humanas.

[4:11] Então para cada idade vai ter uma média para cada idade.

[4:18] A média somente das ciências humanas. Posteriormente nós vamos fazer de matemática.

[4:22] Pronto, nós já fizemos os filtros na primeira parte do código, calculamos a média por idade para a nota de ciências humanas, agora nós vamos fazer o que?

[4:32] Plotar o gráfico utilizando a função ggplot, passando data notas_ciencias_humanas_idade. O próprio R já recomenda para a gente o conjunto de dados e vamos utilizar a função geom_point.

[4:49] Essa função já vai gerar o gráfico de pontos para nós. Para variar um pouquinho vamos aqui mapear o eixo x que vai ser idade, eixo y, que é a media_nota_ciencias_humanas, que é desse novo conjunto de dados aqui que nós calculamos a média da nota por idade.

[5:23] E vamos executar. Pronto. Nós temos o gráfico aqui na aba direita, no inferior direito. No eixo x a idade e no eixo y, a média.

[5:36] Agora nós temos informações sobre a média de ciências humanas por idade, como nós podemos ver aqui no gráfico abaixo.

[5:42] É possível observar que a média mantém um padrão até aproximadamente 60 anos, que é mais ou menos aqui. Vem aqui essa linha até 60 anos, mais ou menos aqui, certo?

[5:55] E depois disso, depois dos 60 anos, a média fica bem dispersa, que é a partir dessa linha, a média vai tanto para baixo quanto para cima, ela fica bem dispersa os pontos.

[6:08] Outra observação interessante é o valor para idade que chega até 125 anos.

[6:16] Nós fizemos o filtro a partir de 17 anos, porém nós temos aqui uma idade de até 125 anos, o que pode ser um erro na base de dados sendo um outro tipo de análise a ser feita em outro momento, que seria para corrigir esses dados ou eliminar esses registros para melhor visualização e melhor coleta de informação dessas demandas.

[6:38] Mas até aqui nós geramos dados e informações os dados aqui nessas duas novas bases de dados, nota_ciencias_humanas e nota_ciencias_humanas_idade e o gráfico apenas para a média da matéria de ciências humanas por idade para ela.

[6:55] E o nosso objetivo também é visualizar e ter informações da matéria de matemática. Então para gerar os dados e as informações da visualização para a matéria de matemática, o processo é idêntico.

[7:06] Vamos aqui criar um novo objeto chamado notas_mt, que significa matemática, Enem, vamos fazer os filtros necessários, filter is, a negação, ou seja, registros que não sejam NA em nota matemática e não seja NA no campo idade e idade maior que 17 também.

[7:45] Vamos executar, criamos um novo conjunto de dados apenas com essas informações mais limpas, corretas.

[7:55] Vamos gerar outros objeto chamado notas_matematica_idade, notas_mt, que acabamos de criar, vamos calcular a média das notas, processo idêntico ao feito anteriormente. Idade.

[8:26] E por fim, o cálculo da média: vamos chamar a função summarise, vamos chamar o campo média_nota_matemática para receber o cálculo da média em nota_mt.

[8:43] Vamos executar, geramos os registros e agora vamos plotar. Ggplot, data notas_matematica_idade, vamos concatenar, geom_point, aes para mapear os valores do eixo x, que é a idade e y vai receber a media_nota_matematica. Vamos plotar e pronto. O gráfico aqui foi gerado do lado direito.

[9:24] Temos informações aqui que as notas de matemática, as médias das notas de matemática são pouco mais decrescentes, ou seja, elas têm uma queda maior que a nota de ciências humanas, vamos comparar aqui. Ó: a média segue um padrão acima de 500 e depois fica dispersa.

[9:48] Porém a de matemática já começa abaixo de 500, vai descendo, descendo, descendo e depois dispersa depois dos 60 anos.