01

Gráfico de linhas das médias das notas

Transcrição

[0:00] Bom, até aqui nós geramos 4 grandes gráficos, obtendo 4 grandes informações.

[0:09] O primeiro foi a identificação da escolha do idioma por sexo, entre espanhol e inglês entre os alunos do sexo masculino e feminino.

[0:19] Depois nós geramos um gráfico de situação escolar por estado, também uma informação bem útil para os cursinhos e até mesmo outros gestores, outras autoridades, para verificar a condição do aluno que está realizando a prova do Enem.

[0:34] Também geramos uma pirâmide de média de idade por estado, dividido entre o sexo masculino e feminino.

[0:41] E por fim, agora por último, geramos um scartterplot, um gráfico de pontos comparando a média das notas da matéria de ciências humanas e matemática por idade das pessoas que estão fazendo as provas.

[0:58] Agora, outra informação bem importante a ser obtida é de como está o desempenho com base nas notas em cada matéria ao longo dos anos para verificar se as notas estão melhorando ou piorando, se há uma grande diferença nas médias entre todas as matérias.

[1:17] Assim, o cursinho poderá ter uma visão geral de como está o desempenho dos alunos, se há uma necessidade de aumentar a intensidade dos estudos em determinada matéria e posteriormente fazer uma comparação com a média dos estados onde o cursinho inclusive tem filiais.

[1:35] Agora vamos criar um novo conjunto de dados com a média de cada matéria, como de costume e como eu já venho avisando para você desde o primeiro curso que ao se trabalhar com análise de dados, nós fazemos sempre muita manipulação de dados, criando novos conjuntos de dados e fazendo também a limpeza deles.

[1:54] Vamos aqui então chamar o nosso objeto principal, Enem, fazer um filtro, semelhante ao que já fizemos anteriormente, is.na, agora vamos fazer para todas as matérias, todas as colunas que possuem notas das matérias, ou seja, notas ciências humanas, notas ciências da natureza, linguagens e códigos, matemática e redação, ou seja, nós temos aqui 5 colunas para fazer as devidas alterações.

[2:25] Então vamos começar aqui com ciências humanas, fazer os respectivos filtros, que seria o que?

[2:35] O filtro é eliminar todas as linhas que têm valores NA. Vamos aqui fazer uma quebra, is.na notas ciências da natureza, você pode também copiar para agilizar nosso processo e substituir aqui, vamos fazer nota agora linguagens e códigos, quebrando a linha aqui, linguagens e códigos, nota matemática e por fim, nota redação.

[3:19] Pronto, nós temos aqui as 5 colunas que nós queremos eliminar.

[3:29] E agora, vamos colocar o concatenador aqui, vamos fazer um group_by, porque nós queremos calcular a média, group_by.

[3:41] Só que a nossa média agora não vai ser mais por idade e sim por ano, nós queremos comparar as médias durante os anos que nós temos aqui.

[3:49] Então vai ser a coluna ano. E por fim vamos calcular a média, summarise, para cada matéria.

[3:59] Como é feito isso? Vamos aqui chamar, nomear aqui média ch, que é para ciências humanas, você vai chamar a função mean e colocar a coluna nota ciências humanas, pronto.

[4:15] Colocar a vírgula e agora você vai repetir esse mesmo código para cada coluna.

[4:24] Substituindo aqui ch por cn, nota ciências da natureza, a outra agora é linguagens e códigos, vamos chamar de lc, nota linguagens e códigos, a próxima é mt, que é matemática, nota matemática e por fim, a nota de redação, nota redação.

[5:07] Vamos colocar aqui média red. Vamos eliminar aqui. Vamos executar.

[5:25] Há um erro aqui no nosso código, o e aqui que a gente não finalizou e também, vamos já salvar num objeto chamado média anos.

[5:38] Então vamos aqui recapitular o que estamos fazendo em cada parte desse código.

[5:43] A primeira parte aqui, nós estamos fazendo o filtro que vocês já conhecem bem que é eliminando registros com valores NA. Essa parte aqui do filtro.

[5:56] Depois, nós estamos fazendo um group_by, que é por ano.

[5:57] E, posteriormente, por fim, nós estamos fazendo a principal coisa que é calcular a média para cada matéria. Porque nós queremos comparar a média de cada matéria por ano.

[6:08] E agora nós podemos executar, pronto. Foi executado. Vamos visualizar esse novo objeto, média anos, aqui ó. Ano 2010, 11, 12, 13, 14, 15, 16 até 17 e cada coluna com a sua respectiva média.

[6:29] Média em ciências humanas, média em ciências da natureza, linguagens e códigos, matemática e redação. Nenhum valor NA, que é importante isso.

[6:42] E agora vamos fazer o principal que é gerar o gráfico de linhas.

[6:50] Para gerar o gráfico de linhas, cada gráfico nós vamos fazer aqui ggplot data, vamos chamar esse objeto que nós criamos, média anos e nós podemos fazer a seguinte forma: geom_line, essa função que gera o gráfico de linhas, mapear o eixo x, que vai receber os valores de ano, e o eixo y, que vai receber a média, vamos fazer um exemplo aqui, ciências da natureza.

[7:29] Vamos colocar o mais. Vamos copiar esse código aqui, porque nós vamos inserir quase exatamente a mesma coisa. Vamos trocar aqui o cn por ch e essa é a única modificação a ser feita. Pronto.

[7:42] Nós geramos o gráfico aqui a direita, média ch, ano e duas linhas.

[7:54] O gráfico ficou vem ruim de ser visualizado, não é mesmo? Não conseguimos saber o que cada linha representa, mas porém nós podemos fazer uma pequena alteração de cores.

[8:07] Vamos colocar aqui color green para ciências humanas e, embaixo, para ciências humanas, vamos colocar uma cor azul, blue. Lembrando que no R é tudo praticamente em inglês: as funções e os valores a serem utilizados.

[8:23] Nós geramos um novo gráfico. O gráfico melhorou um pouco, não foi? Mas ele ainda continua não intuitivo, seria necessário inserir uma legenda aqui a direita para cada linha que nós inserimos.

[8:35] Lembrando que a gente ainda tem mais 3 matérias para inserir, ou seja, teremos que inserir mais 3 linhas e as respectivas legendas.

[8:45] Isso seria um trabalho complexo e desnecessário, já que o pacote gplot2 faz isso tudo automaticamente para nós, basta termos os dados corretos.

[8:52] Então, vamos criar um novo conjunto de dados chamado média anos 2, utilizando a função melt, lembrando que essa função vai converter as colunas em linhas. Vamos passar os dados, que é média anos, escolher o id.vars, que é a coluna que nós desejamos manter, que é ano.

[9:13] Vamos executar. Vamos dar um view aqui nessa média anos 2. Pronto.

[9:22] Você pode perceber que não existe mais aquele tanto de colunas, apenas 3 colunas, uma com o ano, com a matéria e a sua respectiva média. Agora nós podemos gerar o gráfico com esse novo conjunto de dados.

[9:34] Vamos passar o data média anos 2, utilizando a função geom_line, mapeando o eixo x, que vai receber ano, o y vai receber value, que é o nome da coluna que a própria função melt criou e o color vai receber variable, que é a coluna também que a própria função criou.

[00:10:00] Vamos aqui salvar tudo dentro do objetivo chamado plot line notas, vamos salvar aqui, vamos executar, vamos agora procurar aqui executar o objeto.

[00:10:20] Pronto, o nosso gráfico já está feito com apenas 2 linhas, não é mesmo? Com cada matéria e suas respectivas cores. Isso tudo em apenas 2 linhas, tendo os dados corretos.

[00:10:33] O nosso gráfico está pronto em apenas 2 linhas aqui ó. Mesmo nós tendo que criar um novo conjunto de dados, com uma simples função, de forma mais prática, foi bem mais prático e fácil gerar o gráfico com esse código do que gerar uma linha para cada matéria e definindo manualmente também os valores de cada linha. Vamos dar o zoom no gráfico para analisar melhor.

[00:10:59] Pronto. O gráfico ficou bem mais legível.

[00:11:02] E analisando esse resultado, podemos extrair as seguintes informações: houve uma variação grande nas médias entre 2010, que é o primeiro ano com os nossos dados, e 2017, que é o último aqui no final, você pode ver aqui há uma grande variação entre as médias de todas as matérias.

[00:11:25] Com exceção de letras e códigos, que é a verde aqui, que é a segunda linha de baixo para cima, que é onde está o mouse, e ciências da natureza em vermelho, que é essa linha aqui maior, todo o restante teve uma queda ou manteve uma média, como você pode observar aqui ó, teve uma queda tanto aqui na roxa, que é a média de redação, você pode olhar aqui na legenda a roxa é a média de redação, esse azulzinho mais claro aqui é a média de matemática, então houveram quedas aqui nas médias.

[00:12:07] Todas começaram aqui bem próximos, com exceção aqui de ciências da natureza, que é esse amarelo aqui, ela subiu, então houve essa variação.

[00:12:20] Outra informação que podemos coletar é que as médias em 2010 eram bem distantes, o valor mínimo aqui está entre 480 e pouco e a máxima 570, se você olhar aqui na primeira linha à esquerda, o valor mínimo está entre 580 e 500 e o valor máximo está acima de 560.

[00:12:48] Mas em 2017, as médias estão bem próximas.

[00:12:54] Mesmo as médias mais altas elas ficaram bem próximas que foi entre os valores aqui 510 e 530. As vezes isso pode ser bom, ou às vezes isso pode ser ruim.

[00:13:06] Com base nessas informações, o cursinho pode focar em matérias como das ciências da natureza e letras e códigos, já que são as maiores médias. Letras e códigos tá aqui. Então o cursinho pode dar uma focada nessas matérias para aproximar as notas nos próximos anos, deixando as notas das matérias mais próximas, mais normalizadas, e os alunos tendo desempenho bom em todas as matérias.