01

Gráfico de Pirâmide

Transcrição

[0:00] Até aqui, nós já fizemos um gráfico de colunas, para idiomas da prova por sexo, que é esse gráfico aqui, nós também fizemos um gráfico para situação escolar, que foi esse último, mostrando já duas informações bem úteis para o cursinho.

[0:18] Agora nós vamos preparar os dados e fazer as visualizações para a seguinte solicitação do cliente: vamos identificar, visualizar a média de idade por sexo e UF da prova. Como o cursinho tem filiais em diferentes estados do Brasil, obtendo essas informações sobre as idades dos estudantes por estados será possível elaborar melhores metodologias de estudos para faixa etária para cada estado ou até mesmo melhores campanhas publicitárias, campanhas mais voltadas para alunos de determinadas idades.

[0:51] Primeiro, vamos fazer, vamos verificar a coluna idade, que vamos trabalhar agora. Idade.

[1:08] Pronto, bem no início, lá nas primeiras aulas, nós fizemos essa verificação das colunas numéricas, que aqui nós estamos apresentando, visualizando informações básicas de estatísticas descritivas para valor numérico, que é o valor mínimo, primeiro quartil, mediana, média, terceiro quartil e máximo. E esses valores NAs aqui que são registros que não têm valores definidos, ou seja, NAs dentro do conjunto de dados. Ou seja, nós temos 43.

[1:37] Vamos então eliminar esses registros porque eles podem interferir na nossa visualização, na elaboração do gráfico. Então vamos criar um novo conjunto de gráficos chamado idade UF, vamos chamar aqui o Enem, vamos filtrar, vamos pegar apenas os registros que não é NA na coluna idade.

[2:08] Aqui nós vamos pegar todos os registros que são NA. A negação no R é esse símbolo aqui de exclamação. Você vai colocar bem no início, ou seja, eu estou negando essa condição aqui, então eu quero pegar todos os registros que não são NA.

[2:24] Vamos executar, filtramos, agora nós vamos trocar aqui e verificar no conjunto de dados. Salvamos em um novo conjunto de dados o filtro. Pronto, já não tem mais aqui no final à direita os valores NAs. Então nosso conjunto de dados está agora em condições ideais para gerar os gráficos.

[2:47] Como nós queremos identificar, visualizar informações referentes a média de idade para cada UF, nós vamos ter que agora calcular a média desse campo, ou seja, da idade. Então vamos chamar aqui o conjunto, vamos trabalhar com um novo conjunto de dados que usamos anteriormente.

[3:06] Vamos chamar aqui o group_by e vamos agrupar pela coluna uf prova e pela coluna Sexo. Por quê? Eu quero fazer um agrupamento por UF Prova e Sexo para então calcular a média. Vamos colocar aqui summarise, vamos colocar o nome do campo aqui média igual a mean, que é a função do R que vai calcular a média da coluna idade.

[3:48] Vamos salvar isso dentro de um objeto chamado média idade sexo uf para o conjunto ficar bem intuitivo. Vamos alinhar aqui, vamos executar, pronto, já foi executado. Vamos visualizar esse conjunto de dados na aba. Pronto.

[4:11] Aqui nós temos, vocês podem observar, para cada estado o sexo e a média. Acre, feminino, 24. Acre, masculino, 23. Alagoas, feminino. Alagoas, masculino. E assim por diante.

[4:29] Porém, você pode observar aqui que temos duas linhas em branco. Ou seja, que são valores incorretos para a coluna estado. Então agora vamos eliminar essas linhas. Vamos salvar no mesmo objeto. Média idade.

[4:49] E vamos aplicar novamente um filtro, aplicando a seguinte condição: UF Prova diferente de vazio. Vamos colocar aqui, vamos executar. Vamos abrir novamente aqui. Pronto, os dois registros foram eliminados e nós temos agora um conjunto de dados apenas com a média de idade para cada estado dividido por sexo ainda por cima. 3 colunas: estado, sexo e média.

[5:21] Agora que nós temos um conjunto de dados com a média para cada estado e também por sexo, fizemos os tratamentos, os filtros, limpamos os dados, vamos gerar o gráfico com esses registros. Vamos aqui usar a função já conhecida: ggplot.

[5:38] Vamos colocar data o objeto que acabamos de criar: média_idade_sexo_uf. Vamos aqui usar geom_bar, mapear os eixos x e y. O x vai ser UF Prova, o y vai ser o valor da média. Vamos também colorir de acordo com valores do sexo. Pronto, mapeamos.

[6:17] Vamos utilizar aqui a função position, position_dodge, para deixar as barras uma do lado da outra e não sobrepostas. O stat identity. Identity porque nós estamos passando já o valor aqui para o eixo y, então a função geom_bar vai usar esses valores para plotar no eixo y.

[6:44] E por fim, vamos fazer a rotação aqui dessas barras para a horizontal. Vamos executar. Pronto. Vamos dar um zoom aqui.

[7:00] O gráfico da média por sexo e estado está pronto. Porém, se você apresentar esse gráfico para o cliente, ele não vai entender literalmente nada. Até para nós que estamos desenvolvendo está um pouco confuso.

[7:14] Por exemplo: quais as maiores médias? Quais as menores médias? Como está as médias entre os estados somente para o sexo feminino ou para o sexo masculino? Então essa visualização está poluída, está errada, e a intenção da análise de gráficos é justamente o contrário, justamente o oposto do que está apresentado nesse gráfico aqui.

[7:35] Nós temos que apresentar informações através de gráficos de forma fácil a ser visualizado e interpretativos. Então vamos alterar esse gráfico e fazer umas transformações para deixá-lo do jeito mais adequado e profissional para apresentar para o nosso cliente e até mesmo para nós coletarmos informações úteis a partir dele.

[7:59] Como acontece no gráfico de pizza que fizemos no primeiro curso, não há uma função específica para gerar um gráfico que nós vamos trabalhar agora que é o chamado gráfico de pirâmide. Você já vai ver visualmente como é esse gráfico. Então vamos ter que elaborar esse gráfico manualmente, utilizando ainda o pacote ggplot2. Então vamos arrumar aqui a função ggplot, passando data e o objeto que nós queremos.

[8:29] Mas agora, a primeira mudança, bem sutil, mas importante para futuras alterações nos gráficos, já vou mostrar para vocês. Em vez de mapearmos o eixo x e y, que fizemos aqui com aes, que é a função de mapeamento, em vez de fazermos isso dentro da função geom_bar, como fizemos anteriormente, nós vamos fazer dentro da função ggplot. Vamos colocar aqui aes, eixo x vai receber ainda UF Prova, certo?

[9:06] Mas vamos já fazer uma ordenação desses valores. Reorder, vamos ordenar de forma decrescente, ou seja, do menor para o maior, então vamos passar aqui o campo que desejamos média. Essa função nós já utilizamos no curso anterior, então vocês já devem conhecer bem. Essa é a primeira mudança aqui.

[9:31] A segunda mudança é do eixo y. Dentro do eixo y, nós vamos utilizar uma função chamada if else, você também já utilizou no primeiro curso. O que vai acontecer aqui? Para gerar o gráfico de pirâmide, nós precisamos ter valores positivos, ou seja, à direita de 0 e negativos, à esquerda de 0.

[9:54] Então nós vamos fazer a seguinte coisa: vamos fazer uma condição, toda vez que aparecer um valor para o sexo masculino, nós vamos negativar o valor da média. Caso contrário, a média vai continuar normal, está certo? Não se preocupe que a gente já vai ver como vai ficar esse resultado. E o fill, que você já sabe bem, definir cores distintas para os valores da coluna sexo, ou seja, masculino e feminino. OK. Nós já fizemos aqui a primeira inserção, vamos fechar aqui.

[00:10:37] Agora vamos utilizar a função geom_bar, que ela vai receber apenas o stat, stat identity. Identity lembrando que é porque nós estamos passando o valor de y. Vamos fazer uma rotação coord_flip. Vamos executar. Pronto, você pode ver aqui na aba à direita o gráfico de pirâmide está pronto. Vamos dar um zoom aqui para eu te explicar.

[00:11:15] Lembra que eu te disse que nós precisaríamos de valores negativos à esquerda de 0 e positivos à direita de 0? Justamente por isso aqui. Os valores negativos são para o sexo masculino, já os valores positivos para o sexo feminino.

[00:11:31] Como diz aqui nossa legenda, o vermelho feminino e o azul masculino. Então nós temos aqui a média para cada estado. Para cada estado nós temos uma média para o sexo masculino à esquerda e para o sexo feminino à direita.

[00:11:47] Porém, nós temos ainda que fazer outras modificações. A principal são esses valores aqui negativos. Porque não existe média de idade negativa, correto? Então vamos alterar esses valores.

[00:12:00] Para alterar esse valor é muito fácil. Antes, já vamos salvar esse gráfico aqui dentro de um objeto chamado plot_piram_idade. Vamos executar, pronto. Nós temos aqui plot_piram_idade, se executar, o gráfico vai ser plotado. O mesmo gráfico, não se preocupe.

[00:12:27] Para alterar esses valores negativos, nós vamos utilizar a função scale_y_continuous. Por que y? Essa função significa o que? Que você vai fazer alterações manuais nos valores contínuos, ou seja, numéricos, no eixo y, como indica aqui, como eu falei para você já desde o início, desde o outro curso, que as funções no r são bem intuitivas, porém no inglês.

[00:12:51] No y por que? Esse valor do x, atualmente x, ele é o y. Lembrando que essa função coord_flip ela faz a rotação das barras de verticais para horizontais, porém ela faz a rotação completa. Então o meu eixo x ficou no lugar do eixo y e o meu eixo y foi para o meu eixo x. Mas originalmente, ainda no objeto, cada um é um e é seu respectivo valor. Aqui no x também para funções e alterações, nós temos que utilizar funções relacionadas ao y e aqui no y nós temos que usar funções relacionadas ao eixo x. Vou só aqui mostrar para vocês.

[00:13:37] Vamos só alterar os labels. Os labels são os rótulos, que nós vamos alterar, que são esses números embaixo do eixo x e a função abs. A abs ela arredonda valores negativos para positivos de forma mais fácil. Pronto. Você já pode olhar aqui que os valores foram alterados e temos valores positivos à esquerda e valores positivos à direita porque são informações numéricas positivas que só temos a direita e a esquerda para ficar bem dividido entre o sexo masculino e o feminino.

[00:14:13] Então vamos salvar essa alteração dentro do objeto, executar e o gráfico já está com a primeira alteração.

[00:14:26] Agora com essas pequenas modificações nós conseguimos analisar esse gráfico melhor e extrair informações úteis. Vamos lá: primeiro, vamos, é possível identificar com esse gráfico que as médias são bem próximas entre elas, em todos os estados, e que variam entre 20 e 30 anos, você pode ver aqui. 20 para o masculino, que está acima de 20, aproximadamente para 30 e de 20 aproximando ali de 25, também 30.

[00:15:02] Entre 20 e 25 anos, tanto do sexo masculino como do sexo feminino. E olhando mais para cima, a média é um pouco menor, mas não está abaixo de 20 anos, então nós podemos concluir em todos os estados a média a partir de 20 anos. Com essa informação, o cursinho já pode preparar, por exemplo, campanhas publicitárias mais voltadas para pessoas a partir dos 20 anos, até os 25, 30 anos. Mas ainda, se olharmos aqui esse gráfico, é possível melhorar mais ainda, possibilitando extrair mais informações mais detalhadas e é o que faremos agora.