01

Gráfico variáveis sexo e tipo_lingua

Transcrição

[0:00] Após prepararmos os dados para gerar as informações solicitadas pelo cursinho, agora nós vamos gerar os gráficos com pacote gplot2, que você habilitou lá bem no início do curso e vamos gerar os gráficos e as análises desses dados.

[0:18] O primeiro gráfico que será elaborado é o tipo de língua, ou seja, o idioma que o aluno fez a prova de línguas, que é o espanhol ou o inglês.

[0:31] Vamos aqui ggplot, vamos passar os dados utilizando a função geom_bar, passando a função aes, lembrando que ela mapeia os valores que vão no eixo x e y. Vamos aqui utilizar a coluna que nós queremos, ou seja, tipo língua e lembrando que vamos usar aqui também o parâmetro stat que é para indicar se a própria função geom_bar vai fazer a contagem, a tabela de frequência e a contagem dos valores distintos em tipo língua ou se nós vamos passar esses valores manualmente.

[1:24] Como nós vamos pedir para a própria função fazer essa contagem, vamos executar. Pronto.

[1:32] Vamos observar aqui. Nesse gráfico é gerada informações do idioma da prova que tem espanhol no meio, o inglês aqui à direita e tem aquele ponto que nós identificamos bem lá no início.

[1:45] Esse valor é inválido. Se você olhar no dicionário de dados, vai verificar que não há nenhuma representação para esse valor aqui, então vamos tratar esse valor aqui agora.

[1:55] Lembrando a você que nós não fizemos essa limpeza antes porque poderia excluir informações úteis de outros campos.

[2:06] Então por isso que nós preferimos, escolhemos, preferimos deixar esse valor aqui que nós podemos limpar agora e que não vai interferir em outras análises de outros campos das outras colunas. Como é que vamos limpar isso daqui?

[2:17] Primeiro você vai utilizar aqui, vamos utilizar a função dplyr, que também importamos, vamos passar o conjunto de dados que nós vamos utilizar, ou seja Enem, esse símbolo aqui: %<%, lembrando que esse caractere aqui é a concatenação do pacote dplyr.

[2:40] Vamos utilizar a função filter, passando a coluna tipo língua, diferente, passando a condição diferente de ponto. Ou seja, eu quero todos os valores, todos os registros que sejam diferentes de ponto.

[2:59] E vamos selecionar apenas as colunas sexo e tipo. Select_ passando aqui dots coluna sexo e coluna tipo língua. Preste bem atenção nesses valores aqui ó, eles estão entre aspas e esse valor aqui da coluna não está entre aspas.

[3:18] Por que isso acontece? Porque essa função aqui já reconhece as colunas que estão dentro dessa base de dados do Enem. Aqui, nós estamos fazendo um select passando um array, então ela vai verificar, ela está recebendo um array de string para verificar a existência dessas colunas na base de dados.

[3:58] E vamos salvar toda essa nova base de dados em tp língua sexo. Vamos atribuir aqui, vamos fazer a concatenação para ficar limpo e vamos executar.

[4:14] Pronto. A execução já foi feita. O filtro já foi feito.

[4:17] Agora vamos utilizar novamente a função ggplot, porém passando como data essa nova base de dados tp língua sexo, vamos abrir aqui só para dar uma verificada. Temos apenas 2 campos sexo e o tipo língua, indicando o sexo e o idioma que

aquela pessoa fez a prova.

[4:48] Vamos agora utilizar o geom_bar. Lembrando passando o aes. O valor de x eu quero a coluna sexo, mas agora nós vamos fazer uma diferenciação de idiomas para cada sexo. Nós podemos fazer isso com a função fill tipo língua e novamente o stat do tipo count. Vamos executar esse código.

[5:26] Pronto. Agora nós temos um gráfico com as informações tipo língua por sexo. Temos aqui a legenda: tipo língua espanhol e inglês, eliminamos aquele ponto, dividido por feminino e masculino, que são os sexos das pessoas que fizeram a prova.

[5:43] Porém, há uma forma de melhorar essa visualização aqui, colocando as barras uma do lado da outra, inserindo o parâmetro position, ainda na função geom_bar. Vamos inserir aqui position, o próprio R já indica e vamos utilizar outra função chamada position_dodge.

[6:10] Vamos executar e você vai ver o resultado dessa função. Pronto. Temos aqui o gráfico de tipo língua por sexo, dividido aqui em feminino e masculino, espanhol e inglês. Espanhol é meio avermelhado e o inglês em azul. E o tipo língua.

[6:33] Quando a gente utiliza essa função position_dodge ela transforma aquela visualização anterior, com uma barra em cima da outra, coloca uma do lado da outra. A visualização ficando bem mais fácil e bem mais limpa.

[6:49] Agora que o nosso gráfico está mais limpo, mais fácil de interpretá-lo, qual é a conclusão que nós podemos tirar aqui? Qual é a análise que a gente pode fazer?

[6:58] Primeiro que o espanhol é a língua mais escolhida para realizar a prova de idiomas do Enem, tanto aqui no feminino, quanto no masculino. Além de as pessoas do sexo feminino escolhem majoritariamente o espanhol para fazer, olha só a diferença entre o espanhol e o inglês.

[7:18] Porém, já do sexo masculino, está mais equilibrado, a diferença não é tão grande. Então, e tanto do sexo masculino e feminino, estão próximo a escolha pra fazer a prova de inglês.

[7:35] Então, essas informações podem ser passadas para a escolinha, para o curso preparatório e se dedicar um pouco mais nos estudos do espanhol, tentando aumentar, melhorar o desempenho de seus alunos nas provas de línguas.