01

Normalizando coluna sexo

Transcrição

[0:00] Como nós analisamos anteriormente, há algumas colunas que não estão condizendo com a apresentação delas, por exemplo, essas colunas de nota que deveriam ser numéricas, porque é uma nota da matéria da prova e elas foram armazenadas, reconhecidas pelo próprio R como textual. Nessa aula, nós iremos fazer algumas transformações em algumas colunas.

[0:27] Essas operações são necessárias porque há colunas em um conjunto de dados que têm valores distintos, porém que representam a mesma informação, ou seja, se nós gerarmos gráficos com esses dados, nós iremos gerar gráficos com informações repetidas ou até mesmo confusas. Vou mostrar já um exemplo para você.

[0:47] Também há casos em que estão armazenados apenas valores numéricos, porém que representam valores categóricos, ou seja, valores textuais. Tudo isso é identificado ao olhar as documentações dos dados, que também estão disponíveis para você no material do curso.

[1:01] Primeiro, vamos verificar a coluna sexo, nessa coluna, deve armazenar valores referentes ao sexo da pessoa que fez a prova, ou seja, masculino ou feminino.

[1:19] Para verificar quais valores estão armazenados, você pode utilizar a função table Enem, passando a coluna sexo. Vamos executar, pronto.

[1:30] Nós temos o resultado resumido, ou seja uma contagem dos valores distintos dentro dessa coluna. Temos o valor 0, o valor 1, o valor F e o valor M. Ou seja, cada valor desse distinto representando o sexo masculino ou o feminino. O F e o M nós já sabemos que o F é o feminino e o M é o masculino, porém o 0 e o 1 nós não sabemos.

[1:53] Para saber o que representa o 0 e o 1, você pode fazer uma pesquisa no dicionário de dados, também disponível para você no material do curso. Lá, vai indicar que o valor 1 representa o valor feminino e o valor 0, representa o sexo masculino.

[2:13] Então, você tem diferentes valores que representam a mesma informação.

[2:18] Você tem que fazer essas transformações para os valores ficarem únicos e apresentarem as informações de acordo com o que foi solicitado. Por exemplo, se você gerar um gráfico do jeito que está esses valores aqui, você vai gerar um gráfico com os valores 0, 1, F e M. Primeiro problema: o seu cliente, o usuário, não vai saber o que significa 0 e 1 e segundo, você vai ter 2 informações repetidas: o 0 e o F, que representam a mesma informação, ou seja, o sexo feminino e o 1 e o M, para sexo masculino.

[2:57] Nós vamos começar normalizando o valor 1 para o feminino de acordo com o que está na documentação.

[3:08] Você vai utilizar a função gsub, o primeiro parâmetro é o valor que você quer encontrar nos registros, o segundo parâmetro é o valor que você deseja substituir, no caso "FEMININO" e o último parâmetro são os registros em questão que serão substituídos e consultados, por exemplo aqui, é o Enem sexo, a coluna que nós desejamos substituir.

[3:41] Você vai salvar tudo isso dentro da coluna sexo. Passando aqui, vamos executar, pronto, já executou. Se executarmos aqui novamente o table, agora nós temos 4 valores ainda, o 0, o F, o FEMININO e o M, ou seja, nós não temos o valor 1 mais.

[4:05] Agora nós vamos substituir o F, vamos copiar essa linha aqui, vamos normalizar tudo para "FEMININO". Você vai utilizar uma expressão regular dessa forma aqui: o tiozinho, o F e o cifrão.

[4:24] Esse valor representa que eu quero pegar qualquer valor que se inicia e finaliza com a letra F, ou seja, apenas registros que têm apenas a letra F e nada mais além disso.

[4:38] Vamos executar aqui novamente, pronto. Executou, vamos executar aqui, pronto. Agora eu tenho apenas 3 valores: o 0, o FEMININO e o M, porque eu normalizei.

[4:53] Agora vamos fazer para o sexo masculino, pode copiar aqui, só que agora vamos trocar o 1 pelo 0 e o FEMININO por MASCULINO. E embaixo do mesmo jeito, vamos trocar o F pelo M e vamos substituir aqui pelo MASCULINO.

[5:15] Vamos executar aqui essas duas linhas. Daí executando essas duas linhas, executou.

[5:25] Vamos executar o table novamente, vamos limpar nosso console aqui, vamos executar o table novamente e pronto. Agora nós temos apenas 2 valores distintos dentro da coluna sexo, que é o correto: FEMININO e MASCULINO.