02

Normalizando coluna tipo_lingua

Transcrição

[0:00] Agora que nós já finalizamos a normalização e a correção da coluna sexo, vamos verificar agora a coluna Enem tipo língua. Lembrando que essa coluna é para armazenar um valor que representa o idioma que o aluno escolheu para fazer a prova de idiomas, ou seja, inglês ou espanhol. Vamos executar aqui.

[0:28] Nós temos 3 valores distintos: um ponto, o 0 e o 1. Ao olharmos a documentação, como eu já disse para vocês que está disponível lá nos documentos do curso, nós podemos observar que o valor 0 representa o tipo de inglesa e o valor 1, o espanhol.

[0:48] Se olharmos lá em todas as documentações, o valor ponto não é encontrado, ou seja, podemos concluir que esse valor, o ponto, ele foi inserido de forma incorreta, é um valor errado na inserção dos dados. Deve ter tido algum problema na hora de inserir os registros e não observaram essa questão na hora de validar os dados.

[1:11] Então vamos nos atentar a esse valor 0 e 1. Vamos fazer a substituição lá utilizando a mesma função, gsub.

[1:25] Procurar o valor 0, vamos substituir para o valor INGLÊS, que é o que representa de acordo com as documentações, consultando na coluna tipo língua.

[1:40] E vamos salvar esses valores tudo na mesma coluna. Já vamos fazer também a substituição para o espanhol, que no caso agora é o 1, vamos inserir aqui ESPANHOL.

[1:58] Vamos executar essas duas linhas, demora um tempinho, agora vamos verificar novamente, pronto, executou, vamos verificar novamente, pronto. Nós temos o ponto, que nós não vamos eliminar, não vamos tratar esses dados agora, o valor ESPANHOL, e o valor INGLÊS. Os dados foram normalizados.

[2:21] Agora, vamos verificar a coluna UF Prova, vamos utilizar ainda a função table Enem e a coluna UF Prova. Se executarmos, nós não temos um número muito grande de dados, dá para olhar na mão, porém, dá para analisar de uma forma mais fácil pela quantidade de estados.

[2:54] Nós sabemos que no Brasil existem 27 estados. Então vamos utilizar uma função aqui chamada length, executando, pronto. Ele retornou 28. Ou seja, há 28 valores distintos nessa coluna UF Prova.

[3:09] Sabemos que essa coluna representa o UF onde a prova foi realizada, ou seja, o correto seria se tivesse apenas 27 valores. Podemos olhar aqui no início que tem um valor em branco, ou seja, na hora de inserir os dados na base de dados do Enem, houve um erro de inserção, algum outro problema que nós não sabemos e esse valor foi armazenado em branco.

[3:33] Não vamos tratar esse valor nesse momento porque se eliminarmos esses registros, vamos eliminar 1480 registros no geral e isso pode interferir em análises de outras colunas, porque de repente, quando formos analisar por exemplo a coluna sexo, esse valor em branco não vai interferir na coluna sexo, então não há necessidade de eliminarmos por enquanto.

[4:03] Então vamos deixar do jeito que está e vamos partir para a próxima coluna.