03

Normalizando coluna situacao_conclusao

Transcrição

[0:00] Bom, agora vamos verificar a coluna chamada situação conclusão, essa coluna aqui. Vamos colocar Enem, situação, conclusão. Essa coluna indica qual a situação da escolaridade da pessoa que está fazendo a prova do Enem. Vamos identificar aqui os valores armazenados nessa coluna. Vamos executar, vamos limpar aqui, vamos limpar o console e pronto.

[0:40] Nós podemos observar que essa coluna é representada por números que vão de 1 a 4, o que representa cada valor você pode encontrar nas documentações disponíveis para você no conjunto de dados.

[0:59] No dicionário de dados lá tem o que representa cada valor.

[1:03] Então vamos fazer a substituição de cada valor pela string, pelo valor real dessa situação do aluno. Vamos começar aqui, vamos fazer situação, nós vamos substituir os valores nessa coluna e vamos continuar usando ainda a função gsub. Vamos inserir aqui o 1, a gente está procurando esse padrão aqui, vamos substituir por CONCLUIDO, relembrando que esse valor eu sei que ele é concluído, ou seja, que o 1 representa o valor concluído, porque eu olhei nas documentações disponíveis sobre os valores dos dados.

[1:47] E vamos procurar lá na Enem, situação conclusão. Vamos executar aqui, executou. Vamos verificar novamente aqui e pronto. Fizemos a substituição do valor 1. Agora temos 2, 3, 4 e CONCLUIDO.

[2:07] Agora vamos fazer isso para o restante dos valores. Vamos aqui copiar e colar para agilizar nosso trabalho. Valor 2 que representa "CONCLUIRÁ NO ANO"; o valor 3, que representa "CONCLUIRÁ APÓS (ANO)"; e por fim o valor 4, que representa "NÃO CONCLUIDO", vamos abreviar aqui... e "NÃO CURSANDO".

[2:59] Vamos arrumar aqui a acentuação, concluído, concluirá e o não concluirá.

[3:15] Vamos aqui executar novamente todas as linhas seguidas, você pode executar uma atrás da outra que ao executar não vai finalizar ou você pode selecionar e executar de uma vez. Vamos verificar aqui novamente e pronto.

[3:32] Agora nós temos 4 valores distintos que não são mais números: CONCLUÍDO, CONCLUIRÁ APÓS UM ANO, CONCLUIRÁ NO ANO e NÃO CONCLUÍDO E NÃO CURSANDO. Tá ok?

[3:42] Agora, vamos verificar as colunas das notas. Inicialmente, quando a gente viu lá no str, quando executamos a função str, vimos que a coluna das notas tinham alguns valores um pouco estranhos.

[3:59] Então vamos fazer aqui, utilizando uma função diferente da table, o summary. Vamos executar Enem.

[4:10] Vamos primeiro verificar a nota de ciências humanas. Vamos executar. Essa função, o que ela faz? Ela traz um resumo numérico, ou seja, de estatísticas descritivas, alguns cálculos de estatística descritivos em colunas numéricas.

[4:28] Como a gente viu bem lá no início, quando executou a função str, essas colunas de notas estavam no formado de chr, ou seja, texto, então foi por isso que essa função retornou isso daqui: class character, mode chracter também. E o tamanho. Essa não é o nosso objetivo, correto?

[4:53] Nós precisamos agora transformar essas colunas em números, que é o valor correto delas, já que elas armazenam notas.

[5:02] Então, para isso, nós podemos utilizar uma função chamada as.numeric. Vamos aqui as.numeric e colocar um exemplo aqui para você, do número 4, porém ele não está em uma forma numérica e sim em uma forma textual, como eu estou usando aqui as aspas, e isso transforma o caractere em textual. Vamos executar. Pronto, ele converteu aqui embaixo. 4.

[5:35] Vamos inserir outro exemplo aqui. 2.5, executa, 2.5. Porém, ao tentar converter um caractere que não tem como converter para número, ou seja, letras, ele vai retornar o valor NA e dar um warning: NAs introduced by coercion.

[5:58] Por quê? Esse valor aqui não tem possibilidade de converter para número. Outro exemplo aqui também, caracteres especiais como o ponto. Vamos executar aqui, o mesmo valor: NA, warning message: NAs introduced by coercion. Ou seja, esse valor aqui, esse caractere aqui não foi possível transformar em número.

[6:22] Agora, nós vamos aplicar essa mudança a todos os valores dessa coluna.

[6:26] Como é que vamos fazer isso? Vamos aqui Enem, vamos utilizar a coluna notas ciências humanas e vamos fazer aqui, não vamos usar o gsub, vamos utilizar essa função que eu acabei de mostrar para você, as.numeric.

[6:49] E novamente, Enem nota ciências humanas. Vamos executar aqui, deu o warning. Por que deu esse warning?

[7:03] Provavelmente, quase 99% de certeza que algumas linhas dessa base de dados, nessa coluna notas ciências humanas, elas tinham algum valor que não foi possível converter, ou seja, alguma letra ou caractere especial, como nós vimos aqui anteriormente, que não é possível converter. Mas agora, utilizando novamente a função summary notas ciências humanas, vamos limpar o console aqui e vamos executar... Pronto.

[7:31] Como eu falei para você, o Summary, ele retorna alguns valores estatísticos descritivos daquela coluna, por exemplo, o valor mínimo, 0.0; o primeiro quartil 473; a mediana, 534; a média, 529; o terceiro quartil, 588; o valor máximo, 883.7 e a quantidade de linhas NAs, ou seja, quantidade de registros que não foram possível converter em números, que tem mais de 760 mil linhas. É um número considerável alto.

[8:10] Agora vamos fazer isso para todas as outras colunas. Vamos aqui agilizar o nosso trabalho copiando a função que já temos ali, correto? Nós vamos trabalhar na coluna, todas as colunas de nota, então vamos aqui.

[8:34] Notas ciências humanas a gente já fez, ciências da natureza, vamos aqui nota ciência da natureza, depois, nota linguagens e códigos, depois nota matemática e, por fim, nota redação. Vamos inserir aqui também, nota redação, redação... Nota de linguagens de códigos, nota matemática. Vamos aqui verificar, nós temos 1, 2, 3, 4, 5 colunas de notas.

[9:22] Então se a gente verificar aqui nós temos 1, 2, 3, 4 e 5 colunas de notas.

[9:29] Vamos executar todas essas linhas, você pode selecionar e executar todas de uma vez. Todas vão aparecer warning message, porque há números que não são corretos.

[9:45] Tendo a execução, agora vamos dar uma olhada novamente com a str Enem, vamos verificar aqui as colunas das notas. Olha aqui, todas as colunas de notas agora estão no tipo numérico. Nota redação, ciências da natureza, ciências humanas, linguagens e matemática. Pronto.

[00:10:15] Essas foram algumas modificações e transformações necessárias, agora vamos dar continuidade nas nossas análises criando gráficos para a escola que te contratou para fazer a análise das notas do Enem e agora vamos fazer os gráficos e análises desses dados.