4.2 -Tratando valores faltantes (parte 3)

00;00;00;12 - 00;00;27;27
Bom, agora vamos dar uma olhada na coluna de salário, filtrando os valores nulos da coluna. Mas antes, vamos organizar, vamos colocar aqui uma célula de texto, os nossos três jogos da velha e "tratando coluna salário". Perfeito. Vamos apagar aqui também essa célula para não ficar ali. Ótimo. Agora a gente vai filtrar os valores nulos de salário. A gente já sabe, né? Fácil.

00;00;28;10 - 00;00;59;05
Tabela dados... Da tabela dados da coluna salário, e o "isnull" aqui para a gente poder filtrar. Vamos ver aqui... Olha, são 557 linhas nulas. Nós temos a coluna de faixa salarial, então vamos ver o que essas pessoas que não colocaram salário marcaram de faixa salarial. Talvez a gente possa fazer igual fizemos para a idade. Vamos lá dar uma olhada.

00;01;00;04 - 00;01;33;27
A gente pode até copiar o que está aqui mesmo, de cima, pra ir já facilitando pra gente. E aí, depois do filtro, aí a gente pega aqui tudo o que está com salário nulo. A gente quer a coluna faixa salarial e o ponto value counts. Ok. Opa, escrevi errado a coluna: faixa sala... O que que eu escrevi aqui, hein gente? SA-LA-RI-AL.

00;01;34;22 - 00;02;01;09
Acho que agora vai. Agora foi. Executando essa célula aqui, temos um vazio. Vejam: as pessoas que deixaram a coluna salário nulas também deixaram a coluna de faixa salarial nula. O que restou foi pegar a média geral de salários. Porém, a gente ainda não viu se tem valores discrepantes nessa coluna, nós vamos aprender melhor sobre isso na próxima aula.

00;02;01;21 - 00;02;20;07
Então, aqui, ao invés de usar a média, nós vamos utilizar a mediana, porque ela é menos afetada por esses valores discrepantes. A gente até viu isso na aula de estatística. Vamos entender um pouquinho melhor sobre isso na próxima aula, mas salário é uma variável que tem muita chance de ter algumas pessoas que recebem muito mais ou muito menos do que as outras.

00;02;20;17 - 00;02;58;29
Então, é mais seguro se usarmos a mediana. Então vamos filtrar aqui os valores de salário nulos e substituir pela mediana. Nós já aprendemos como calcular a mediana, então vamos lá... Se a gente colocar: dados. Qual é a coluna? Salário ponto median e parênteses vazios. A gente tem aqui a mediana, que é 7625. Bom, podemos usar o "loc" e substituir por esse cálculo. Então a gente pode fazer assim: da tabela dados com a função loc, colchetes...

00;02;59;11 - 00;03;28;17
E aí a gente faz o filtro de tudo o que seja da tabela e da coluna salário, que seja nulo. Então, tabela dados, coluna salário, ponto is null. É só isso. A gente não precisa de um outro filtro aqui, a gente só está querendo pegar tudo que é salário nulo. Ok.

00;03;28;20 - 00;03;52;20
Se a gente rodar aqui, a gente vai ter a tabela em que todos os dados aqui as pessoas não colocaram salário. Então a coluna salário está vazia. Só que a gente quer localizar exatamente da coluna salário, a gente não quer como retorno a tabela. Então a gente coloca aqui: vírgula e a coluna salário. E aí, se a gente executar, aqui realmente a gente tem como retorno as linhas que estão nulas. E qual o valor que a gente quer atribuir aqui?

00;03;52;20 - 00;04;19;26
A gente quer atribuir a mediana. Igual. E aqui em cima a gente calculou a mediana e gente nem colocou, nem atribuiu algum valor, né? E aí é uma outra forma aqui que eu posso falar para vocês: a gente pode simplesmente copiar isso aqui e colocar aqui. Mas, veja bem, fica muita informação. É por isso que a gente gosta de colocar os cálculos em uma variável: mediana salário.

00;04;21;22 - 00;04;49;09
Ok. Agora sim, a gente pode pegar lá embaixo e colar. Executando, vai dar erro. E olha esse erro aqui: mediana salário não foi definida. Lembram lá nas primeiras aulas que a gente falou sobre a ordem de execução das células, a gente mudou aqui em cima, mas a gente não executou. Então, é óbvio que se a gente executasse a segunda primeiro, vai dar ruim.

00;04;49;09 - 00;05;21;04
Então a gente volta aqui, a gente atribuiu. Vamos executar essa célula. Ok, agora a gente tem mediana salário e aí sim a gente executa a de baixo. E agora sim, a gente atribuiu a mediana a todos os valores nulos da coluna salário. E pronto, não temos mais nenhum nulo em salário. E é isso nessa aula pessoal. Vimos as várias formas que podemos tratar os valores nulos nos nossos dados e aproveitamos para tratar os valores nulos das colunas de idade, gênero e salário. Nos vemos na próxima aula!