4.2 -Tratando valores faltantes (parte 2)

00;00;00;05 - 00;00;21;20
Vamos trabalhar aqui esses valores nulos da coluna de idade, que é uma coluna muito importante para a gente conseguir responder algumas perguntas com os nossos dados. Bom, para manter tudo organizado, o que é que a gente vai fazer? Célula de texto aqui embaixo, colocar aqui alguns jogos da velha para poder ficar igual. Vou até olhar quantos jogos da velha que eu coloquei aqui em cima.

00;00;22;04 - 00;01;00;13
3. Beleza, tá certinho. Opa, vamos apagar isso aqui porque quero só meu texto, ótimo! Agora aqui embaixo, clico duas vezes e “trabalhando coluna de idade”. Ok, beleza. Aí a gente já tem a nossa célula aqui. Bom, vamos ver a quantidade de valores nulos que a gente tem na coluna de idade.

00;01;00;13 - 00;01;26;21
Se a gente fizer assim: tabela dados, coluna idade, ponto isnull, a gente vai ter como resposta uma coluna com “false” pra quando a célula de idade está preenchida e “true” pra quando está com o valor faltante. Então esse false representa se a linha da coluna está preenchida e o true é quando eu não tenho o valor. Aqui na linha zero, a gente pode ver que ela está preenchida, já que o valor é false.

00;01;27;03 - 00;01;53;02
Na linha 469 a gente tem que o valor de true, então está faltando algum valor aqui. Ok? Bom, para fazer a contagem de quanto true e false, basta a gente fazer assim: a gente usar aquela função que a gente já aprendeu, que é o value counts. Depois da função isnull, a gente coloca ponto (mais uma função) value counts, parênteses e executa.

00;01;54;00 - 00;02;24;23
Então, aqui a gente pode dizer que a gente tem 4197 valores falses, ou seja, valores que estão preenchidos na coluna de idade, e 74 valores true, ou seja, valores que estão nulos. Ok? A gente poderia criar um modelo para preencher esses valores, mas vale a pena ter todo o trabalho de criação de modelo para preencher 74 valores? Eu acho que não. A melhor prática, nesse caso, é substituir os nulos pela média das idades.

00;02;24;23 - 00;02;52;21
Porém, vamos tentar ser mais específicos nessa média. Vamos olhar aqui as colunas da nossa tabela para ver se existe alguma outra informação que podemos utilizar. Como o que a gente vê mesmo quais são todas as colunas? Dados, que é a nossa tabela, columns. Executando aqui, a gente tem o nome de todas as colunas. A gente tem uma coluna aqui que chama “faixa idade”. Nela, as pessoas preencheram a faixa etária delas.

00;02;53;02 - 00;03;14;24
Será que a gente tem a faixa etária dessas pessoas que deixaram a coluna de idade nula? Tipo, elas não colocaram a idade, mas elas colocaram faixa etária delas. Vamos fazer um filtro e, em seguida, um value counts. Assim: dados (que é a nossa tabela), abre colchetes... Aí qual é a nossa coluna dentro da tabela de dados? É a idade...

00;03;16;01 - 00;03;52;14
Ponto isnull. O que a gente está fazendo aqui? A gente está fazendo primeiro um filtro de todos os valores nulos da coluna idade, ok? E aí a gente vai olhar qual é a quantidade de pessoas por faixa de idade ou faixa etária usando essa coluna de faixa idade. Então ok. Aqui, nesse filtro, eu filtro todas as pessoas que deixaram idade nula e aí eu quero, lá na coluna de faixa idade, fazer um value counts.

00;03;52;14 - 00;04;15;12
Assim eu vou ter a faixa etária das pessoas que não preencheram a idade. Vamos executar e vamos ver o que vai aparecer. Bom, primeiro a gente tem 74 valores nulos. Então, se a gente contar aqui 68 + 6 = 74. Então, todas as pessoas que deixaram a idade nula, elas preencheram a faixa idade. Então isso já é algo bom pra gente, a gente pode usar essa faixa de idade.

00;04;15;12 - 00;04;38;10
E 68 dessas pessoas que deixaram a idade nula, colocaram que são +50 e 6 colocaram que estão entre 17 e 21. Então podemos ser mais exatos na nossa substituição de valores faltantes pela média. Podemos fazer a média das idades entre 17 e 21 e substituir esses seis valores ausentes por essa média.

00;04;38;24 - 00;05;05;11
E depois pegamos a média das idades com a faixa +55 e substituímos os 68 valores para essa média. Que tal? Acho que fica mais assertivo. Para pegarmos a média da faixa de 17 a 21, vamos usar dois conceitos que já vimos nas aulas passadas: filtro e o cálculo da média. Vocês lembram como que faz? Olha, eu aconselho dar um pause e tentar aí um pouquinho ok? Vamos lá:

00;05;05;11 - 00;05;25;12
Vamos filtrar esses valores e fazer a média usando a função “min”. Então a gente coloca aqui qual é a nossa tabela, que é dados, da nossa tabela qual que é a coluna que a gente vai filtrar, que é a faixa de idade, e qual é a faixa que a gente quer? Essa de 17 a 21. Vou até copiar para não ter o perigo de errar: ctrl+c.

00;05;26;13 - 00;05;47;23
Aspas simples ou duplas, ctrl+v. Beleza. Aqui a gente tem o filtro, mas a gente quer é o quê? A média. Então a gente vai colocar: idade (que é a nossa coluna que a gente quer calcular a média) ponto min (que é a nossa função de média do Pandas), ok? E a gente pode executar. Ok.

00;05;48;01 - 00;06;21;04
A média é 20,20 anos. Vamos colocar esse valor de média em alguma variável. Vamos chamar de média 17, 21. Então aqui antes: média urdenline 17 urdenline 21, igual... Então essa variável está recebendo o valor da média. Executando... perfeito. Agora, nós precisamos filtrar exatamente as linhas em que a faixa etária seja 17 e 21 e a idade esteja vazia.

00;06;21;07 - 00;06;43;11
Ou seja, nós vamos usar dois filtros, mas aqui nós vamos usar um método "loc", de localização mesmo. Nós queremos localizar essas exatas células que estão nulas, ou seja, a linha e a coluna, e não apenas a coluna inteira que aparece quando fazemos um filtro simples. Então começamos assim: colocamos nossa tabela ponto loc e abrimos aqui o colchete.

00;06;44;26 - 00;07;12;11
Bom, como eu falei, queremos encontrar as linhas que estejam na faixa de 17 e 21 e que estejam com a idade nula. Então a gente vai fazer assim: dentro desses colchetes, a gente vai abrir um parêntese e aí a gente coloca o nosso filtro, que é da tabela de dados, a coluna de faixa idade e a gente quer que seja igual, ou seja, dois sinais de igual (pra comparação), à faixa etária de 17 a 21.

00;07;12;28 - 00;07;38;29
Ok, isso é uma das condições. A outra condição que a gente quer é que na coluna de idade esteja nulo. Ok, então a gente vai fazer. A gente vai usar o & que a gente já aprendeu em outras aulas e vamos colocar essa outra condição também entre parênteses que vai ser: dados (que é a nossa tabela), nossa coluna de idade entre colchetes e entre aspas simples ou duplas, ponto isnull.

00;07;39;02 - 00;08;07;21
Lembrando que essa função isnull vai filtrar aqui pra gente todos os valores nulos da coluna de idade. Ok. E aí nessa função loc, a gente faz uma vírgula e coloca qual que é a coluna que a gente está querendo preencher os valores nulos. No caso, a gente está querendo preencher os valores nulos da coluna de idade. Então vamos colocar assim...

00;08;07;21 - 00;08;38;28
Na verdade, gente, o loc vai localizar esses pontos em que os valores da coluna idade estão nulos. Coluna idade nulo e na faixa etária, de 17 a 21. Vamos executar aqui que a gente vai dar uma olhada. Aqui, olha, ele filtrou pra gente as exatas seis linhas em que idade está nula e na faixa de idade 17 a 21. Ok? Então vamos atribuir pra elas a média que já calculamos antes, que está guardada numa variável média 17 21.

00;08;39;11 - 00;09;11;27
Ok? Então a gente pode só fazer assim: igual (que é atribuição), copiar aqui a nossa variável média 17 21 e colocar aqui. E aí, quando a gente executar, vai estar atribuído a média do 17 21 a esses seis valores nulos aqui. Vamos executar e pronto. Ou seja, estamos substituindo esses valores. Se a gente olhar a faixa etária dos dados faltantes novamente, a gente vai ter isso aqui... Vou copiar pra poder ficar mais fácil.

00;09;12;02 - 00;09;39;23
Foi aqui que a gente viu que é o filtro dos valores nulos. A ideia é que esses seis aqui desapareçam, né? Que a gente substituiu. Eu vou colocar, rodar de novo. E olha, sumiram mesmo porque a gente já substituiu pelos valores da média 17 - 21. Ok. Bom, vamos lá calcular a média de idade dessas pessoas aqui que colocaram que a faixa de idade delas é +55.

00;09;39;23 - 00;10;09;13
Vamos copiar aqui, olha, essa parte de cima do média 17 21 para facilitar pra gente. Bom, vamos lá então calcular a média da idade dos +55. Vamos copiar aqui de cima mesmo, esse finalzinho aqui do média, e a gente só muda os valores para facilitar aqui pra gente. Selecionei, ctrl+c, seleciona essa célula de código, ctrl+v, e aí eu mudo a faixa de idade que agora é 55+. Beleza.

00;10;09;24 - 00;10;34;09
A ideia é que a gente fez um filtro de todo mundo, que é 55+, eu quero a média da coluna de idade e ok, vamos executar. E agora? Olha o resultado aqui, gente, o valor deu “nan”. Será por quê? Vamos retirar a parte da função min, para visualizarmos a coluna de idade para essa faixa de idade.

00;10;34;09 - 00;10;58;09
Eu vou tirar o min, ou seja, agora eu não estou calculando média. Eu só quero ver o que tem nessa coluna de idade, pras pessoas que responderam que são da faixa idade 55+. Ok, vou executar. Bom, e isso abre espaço para algumas questões, porque as pessoas que marcaram a faixa de idade +55 resolveram não colocar a idade delas. E as questões podem ser assim...

00;10;58;09 - 00;11;24;17
Será que é algum erro de preenchimento de formulário? Podemos até pensar que pessoas mais velhas não querem colocar a idade delas, mas é estranho que absolutamente nenhuma pessoa marcou qual é a idade delas. Tipo, a gente tem 68 pessoas, e nenhuma resolveu colocar idade. Bom, a gente supõe que pessoas +55 tenham mais tempo de carreira, então faz sentido que elas estejam em um nível de sênioridade.

00;11;25;02 - 00;11;49;04
Vamos dar uma olhada na coluna de nível para as pessoas que marcaram +55. Então aqui, por exemplo, a gente viu a coluna de idade. Vamos copiar isso aqui, ir para uma célula de código abaixo, colar e, ao invés de ver a coluna de idade, a gente vai dar uma olhada na coluna de nível. Lembrando, gente, que esqueceu qual é o nome da coluna, porque tem que ser exatamente igual, dados ponto columns e copia de lá.

00;11;49;16 - 00;12;18;29
Vamos executar aqui... A gente tem pleno, júnior, pleno, algumas pessoas não responderam e com essa quantidade de dados aqui, apareceu só um sênior. Então asssim, talvez esse preenchimento de faixa da idade +55 tenha sido mesmo um erro de formulário, né? Aqui é válido explorar a planilha, levantar mais hipóteses e discutir com seu time de trabalho também.

00;12;18;29 - 00;12;42;15
Ter uma troca de ideia e ver o que as outras pessoas acham que pode acabar levantando pontos interessantes. Muitas vezes a gente recebe os dados prontos, mas um outro time interno da empresa que coletou pode acabar ajudando a gente a verificar se foi algum erro de formulário. Mas no nosso caso, como não podemos supor que essas pessoas realmente sejam +55, vamos preencher esses valores nulos com a média geral das idades mesmo.

00;12;42;15 - 00;13;20;18
Calculando a média geral e atribuindo o valor para a variável de média geral. Então, a gente vai criar uma variável "média geral" aqui e vamos atribuir o valor da média mesmo, que é da tabela dados, que é da coluna idade ponto min, ok? E assim calcula a média de todo mundo da coluna de idade. Vou até copiar aqui a média geral, colocar uma linha abaixo para gente poder também saber qual é a idade.

00;13;21;25 - 00;14;02;07
A média é 31,15. Então, o que a gente quer agora? A gente quer localizar esses 68 valores nulos aqui, que a faixa de idade é +55. Localizando esses 68, a gente vai colocar o valor de média geral, ok? Então primeiro vamos usar o loc pra gente poder localizar. Vamos lá: dados, a gente coloca ponto loc, abre colchetes e aqui a gente vai fazer o primeiro filtro, que é o filtro de quem marcou que a faixa de idade é +55, ou 55+.

00;14;02;29 - 00;14;34;20
Então vamos: parênteses dados... Qual que é a nossa coluna que a gente quer? Faixa idade em que essa coluna seja igual a 55+. Ok, nosso primeiro filtro de quem marcou 55+. E a gente quer também todas as linhas em que a coluna idade seja nula. Então o nosso segundo filtro: abrindo parênteses, dados, coluna de idade, isnull, beleza?

00;14;36;03 - 00;15;08;06
Então a gente está fazendo dois filtros. A gente quer localizar onde todo mundo marcou +55 e a idade seja nula. Vamos rodar. Gente, deu erro: primeiro erro aqui que ele apontou é que não existe "isnuul". Exatamente, não existe. Coloquei um N aqui a mais e um L a menos. E faltam os parênteses, né? Ok, vamos rodar.

00;15;08;06 - 00;15;38;10
Aqui a gente localizou, só que a gente não quer a tabela inteira, a gente quer de uma coluna específica, a gente quer da coluna de idade, a gente quer achar todos os valores da coluna idade que estão vazios. Então, por isso a gente vai colocar um: vírgula idade. E aí, se a gente executar de novo, a gente tem aqui... Agora sim a gente não tem a tabela, a gente tem a coluna idade que está vazia, beleza? A função loc é pra isso, a gente localizou.

00;15;38;28 - 00;16;06;04
Só que a gente quer atribuir a média geral nesses dados e por isso a gente coloca um valor de igual aqui, um simbolozinho de igual, um só, de atribuição. E aí a gente coloca a média geral. Opa, esse negócio está aparecendo aqui. Pronto, aí sim a gente localizou e está atribuindo o valor. A gente executa... Opa, executei a de cima. Executa aqui e pronto, está atribuindo.

00;16;07;10 - 00;16;46;01
Bom, se a gente fizer aquele value counts de novo, a ideia é que não apareça nenhum valor nulo. Então vamos copiar lá de cima para ficar mais fácil e colocar aqui embaixo. Antes a gente tinha seis nulos na faixa etária 17 - 21. A gente já atribuiu o valor da média, sobrou 68 da faixa 55+, agora a ideia é que eu não apareça nada, hein! Orações. Não apareceu nada. Sucesso, galera, e é isso.