4.2 -Tratando valores faltantes (parte 1)

00;00;00;18 - 00;00;28;02
Olá pessoal! Vamos pra mais uma aula do curso. Bom, nas últimas aulas nós fizemos algumas análises utilizando funções estatísticas como “Numpy” e Pandas. Bem legal, mas vocês lembram que algumas colunas que nós analisamos existiam valores nulos? Vamos organizar, né? Primeiro, voltando, vamos organizar aqui o nosso notebook. A gente começou mais uma aula, a gente coloca aqui uma célula de texto. E nessa aula a gente vai falar sobre valores faltantes, então vou colocar aqui: valores faltantes.

00;00;28;02 - 00;01;02;08
É só pra gente deixar bem organizadinho. Bom, teve uma função que a gente aprendeu, que foi a “info”, que passava algumas informações sobre a nossa tabela. Vamos lá colocar: dados ponto info e o parênteses vazios. Quando a gente rodar, a gente tem algumas algumas informações aqui. E lá no final, eu tinha falado que tem uma coluna que faz a contagem da quantidade de valores não nulos, ou seja, tem alguns valores nulos em algumas colunas.

00;01;03;03 - 00;01;28;15
Bom, primeiro eu preciso dizer que não tem um único tipo de tratamento para valores nulos ou valores faltantes. Então a gente pode ver aqui quais as formas que podemos tratar esses dados e depois a gente pode discutir o que a gente pode fazer com nosso caso da nossa tabela. Beleza? Uma das coisas que podemos fazer com esses dados é deletar geral, deletar todas as linhas que têm valores faltantes.

00;01;28;15 - 00;02;00;21
Por exemplo, a pessoa deixou só a informação de idade em branco. Deletamos todas as respostas dela, como nível, raça, etc. Simples, rápido e prático. Mas tem desvantagens: se a nossa tabela tiver muitos valores nulos, provavelmente a gente vai perder muita informação deletando tudo e fica bem chato se a gente tiver poucos dados para analisar. Então, apesar de ser atrativa a ideia de deletar tudo, é sempre bom verificar se temos dados suficientes e se deletar os valores faltantes vai afetar muito a nossa análise.

00;02;01;07 - 00;02;33;29
Por exemplo, no nosso caso nossa planilha tem 4271 linhas. Se a gente deletar todos os valores faltantes... Por exemplo, aqui tem coluna que tem só 837 linhas. O restante seria tudo apagado, então ficaríamos com pouquíssimos dados. Uma outra coisa que a gente pode fazer é substituir os valores faltantes pela média ou pela mediana. O bom dessa prática é que a gente não vai perder nenhum dado, já que não deletamos nada e funciona bem com uma base de dados pequena ou mais homogênea.

00;02;34;13 - 00;02;56;22
Para valores categóricos, podemos fazer quase a mesma coisa usando a moda, ou seja, valor que se repete mais. Porém, um fator negativo dessa prática para ambos os casos é não levar em

consideração a covariância dos fatores. Ou seja, tem fatores que variam em conjunto. Por exemplo, quanto mais chove de 0 a 4 milímetros, aumenta as pessoas que vão ao cinema.

00;02;57;12 - 00;03;25;23
Porém, quando chove muito, a partir de cinco milímetros, diminui essa quantidade, já que as pessoas nem querem mais sair de casa. Só substituir por uma média pode ser que a relação entre essas duas variáveis, milímetros de chuva e ida ao cinema, não sejam correspondidas nos valores substituídos. Uma última possibilidade é criar um modelo de aprendizado de máquina para resolver o problema, pode dar ótimos resultados e vai levar em consideração a covariância dos dados.

00;03;26;09 - 00;03;46;22
A desvantagem é que é muito mais trabalhoso do que substituir valores pela média, por exemplo. Bom, sabendo de todas essas possibilidades, nós podemos voltar aqui para o nosso notebook com a análise da tabela e a primeira coisa que precisamos saber aqui é como saber quais colunas têm os valores nulos e por isso a gente tem essa função de info: ponto info.

00;03;46;24 - 00;04;19;01
A gente consegue olhar aqui: se a nossa tabela tem 4271 linhas, vai ter colunas que vão estar todas preenchidas, como essas aqui do início. Mas se não tem essa quantidade de linhas, essa contagem aqui, significa que tem linhas em branco. Tipo 4262 não é 4271, então tem linhas em branco. E quando eu falo em branco, significa que, por exemplo, a gente tinha um formulário de onde veio esses dados, então a pessoa, na hora de preencher esse formulário, ela deixou em branco.

00;04;19;11 - 00;04;39;05
Às vezes ela não quis colocar a idade dela. Então ficou lá o campo de idade vazio. E aí ela pode ter respondido todo o resto, mas o de idade ficou vazio. A coluna de salário também tem nulos. Vamos dar uma olhada aqui onde está a coluna... É a última coluna. São 3694 valores não nulos.

00;04;39;13 - 00;05;18;13
A coluna de gênero, vamos procurar aqui e vamos marcar. Já está marcado aqui. A gente vai pra cá... Só selecionei aqui mesmo para a gente poder conseguir fazer essa linha e saber quais são os valores. A coluna de gênero tem 4262 linhas não nulas, ou seja, 9 valores nulos, porque o total seria 4271 menos 4262. 9 nulos. Começando aqui pela coluna de gênero, a gente pode fazer um agrupamento, como a gente viu nas aulas anteriores, mas como a gente também está deixando tudo organizadinho, vamos criar aqui, adicionar uma célula de texto.

00;05;19;09 - 00;05;46;20
Vamos criar um textinho aqui menor, vamos colocar assim: trabalhando coluna de gênero, para ficar legal. Aí a gente executa aqui. Essa linha de cima, que ficou vazia de código, a gente pode apagar só pra ela não ficar ali flutuando no limbo. Ok, agora a gente pode fazer o nosso groupby: nossa tabela dados, ponto groupby, que é a nossa função, a gente abre parênteses, aí essa função ela recebe alguns parâmetros.

00;05;46;20 - 00;06;14;21
Primeiro, qual é a coluna: coluna de gênero, vírgula... A gente quer que apareçam os valores nulos, aqui a gente está querendo conversar sobre valores nulos. Então a gente vai usar o dropna igual a false... Lembrando que essa função, quando a gente não coloca ela, por default, por padrão, ela tira os valores nulos.

00;06;14;28 - 00;06;35;02
Mas como a gente quer que apareça, a gente vai colocar assim. Como a gente quer que faça uma contagem de valores únicos, a gente coloca lá qual que é a coluna que a gente vai usar para fazer essa contagem, que é a nossa coluna de id ponto nunique, que vai fazer a contagem de valores únicos. A gente vai executar.

00;06;36;19 - 00;07;02;25
E aqui a gente tem a quantidade de pessoas que responderam que são do gênero feminino, masculino, a quantidade de pessoas que preferiram não informar e a quantidade de nulos. Temos 9 valores nulos. Sabendo o contexto nesse caso e observando que tem opção em que as pessoas podiam marcar “prefiro não informar”, podemos preencher esses 9 valores nulos por essa opção de “prefiro não informar”. Para preencher todos os valores nulos de uma coluna por um único valor,nós podemos usar a função “fillna”.

00;07;03;06 - 00;07;31;08
Que é meio que um preenchimento, preencher os nulos. E a gente faz assim: a gente coloca qual é a tabela e qual é a coluna que a gente vai querer preencher os nulos. Aqui, no caso de gênero, a gente coloca ponto a função que é o “fillna”, parênteses e aí dentro dos parênteses, a gente vai colocar o parâmetro que é o que a gente quer colocar no lugar desses nulos.

00;07;31;16 - 00;07;59;05
A gente está querendo colocar essa frase, aqui: prefiro não informar. Vou até copiar para poder ficar exatamente igual. Eu coloquei aqui aspas, tanto faz aspas simples ou duplas. E dentro eu coloco: prefiro não informar. E aí a gente vai executar. Olha só: aqui, a saída quando a gente executa é a coluna gênero, com os nulos preenchidos. Só que a gente quer que isso fique direto na nossa coluna, né?

00;07;59;13 - 00;08;22;01
Então, a gente quer atribuir essa coluna à nossa coluna de gênero. A gente pode fazer assim: a gente copia esse início aqui, que é a nossa tabela e nossa coluna. A gente cola aqui no início e coloca um igual. O que a gente está fazendo aí? A gente está atribuindo o valor do fillna lá na coluna de gênero, ok?

00;08;22;11 - 00;08;46;25
E aí, quando a gente executa essa célula, a gente faz essa atribuição. E se a gente copiar agora, esse groupby aqui de cima: ctrl+c e ctrl+d nessa célula de código aqui embaixo e executar, a ideia é que não apareça nenhum valor nulo. Perfeito.

00;08;46;25 - 00;09;07;03
Então antes a gente tinha 9 nulos e 12 “prefiro não informar”. Agora a gente tem 21 prefiro não informar, porque a gente substituiu os nulos por prefiro não informar. Beleza? Assim, preenchemos todos os valores nulos de uma coluna por um valor. Em casos de colunas numéricas, existe o padrão de preencher os nulos por -1, desde que isso não atrapalhe as análises.