3.5 - Repetindo a primeira análise (parte 2)

00;00;00;17 - 00;00;25;23
Bom, continuando, a próxima análise que fizemos foi ver a quantidade de pessoas por gênero. Nesse caso, já não é um filtro, é um agrupamento. No módulo anterior nós fizemos isso utilizando a ferramenta Tabela Dinâmica para realizar esse agrupamento. Na Biblioteca Pandas, existe a função "groupby", que vai realizar esse agrupamento para nós. A gente usa a função assim... Mais uma linha de código...

00;00;26;12 - 00;00;54;26
Qual é a nossa tabela? Nossa tabela é a de dados ponto groupby, que é agrupar por. E a gente quer agrupar por que? Por gênero. Então a gente abre aqui os nossos parênteses, nossas aspas simples ou duplas, o nome da coluna, no caso "gênero". E aí, como a gente quer saber a quantidade de pessoas que responderam gênero feminino ou masculino, lá na nossa aula do módulo 2,

00;00;54;28 - 00;01;23;04
na tabela dinâmica a gente colocou a coluna de gênero e para fazer a conta dos valores únicos, a gente usou a coluna de id, que cada item dessa coluna é único e a gente fez a soma por isso. Então a gente também vai usar aqui essa coluna. Só que agora a gente não vai colocar dentro dos parênteses, a gente vai colocar fora e entre colchetes. A gente abre o colchete, nossas aspas duplas e a coluna de id, que é ID maiúsculo.

00;01;23;21 - 00;01;50;27
E aí a gente usa uma função que chama "nunique". Esse nunique vai fazer a soma dos valores únicos para a gente. Então ele vai basicamente somar a quantidade de valores únicos de id. E executando a célula, aqui a gente tem uma tabelinha, que é a quantidade de valores para cada item na coluna.

00;01;50;27 - 00;02;16;16
Agora a gente sabe, por exemplo, que a gente tem 3194 pessoas, que assinalaram que são do gênero masculino e 1055 que assinalaram que são do gênero feminino e 12 que preferiram não informar. Uma coisa interessante é que não apareceram os valores nulos, porque pode ter pessoas que deixaram essa coluna em branco, aqueles valores em branco mesmo.

00;02;16;16 - 00;02;51;26
Se a gente quiser que apareça a quantidade de nulos, podemos usar um parâmetro da função groupby, que se chama "dropna". Esse drop é de cortar os nulos, né? Eu vou copiar o que está aqui em cima só pra gente poder conseguir comparar. Então copiei o que está em cima, colei em uma célula de código embaixo. E aí dentro da função groupby, depois de gênero, eu coloco uma vírgula e coloco "dropna". E aí o dropna é um parâmetro que recebe um valor.

00;02;52;13 - 00;03;18;21
O valor que ele recebe é um valor de true ou false. Por padrão ele é true, ou seja, ele vai retirar os valores nulos, por isso a gente não enxerga aqui. E se a gente colocar ele igual a false, a gente está querendo dizer: olha, não corta os valores nulos não, eu quero ver a quantidade. Então, se a gente executar aqui agora, a gente consegue ver esse NaN, que é de valores nulos e a gente consegue ver que são nove valores nulos.

00;03;19;07 - 00;03;39;18
Ok? Bom, gente, aproveitando o que a gente fez aqui, esse groupby, vamos falar um pouquinho sobre como olhar e saber que é uma função, saber que é um filtro. Por exemplo, a gente colocou dados, que é a nossa tabela, ponto... Esse ponto indica que a gente está chamando já uma função "ponto groupby", ou seja, uma função.

00;03;39;18 - 00;04;04;16
E normalmente as funções têm esses parênteses, que é onde a gente vai colocar os parâmetros. Ah, Carol, teve essa função aqui, o nunique, que teve parênteses e não teve nenhum parâmetro. Porque a gente está usando o formato padrão dela, o default dela. Mas se a gente quisesse, a gente poderia pesquisar sobre essa função: colocar lá no Google: nunique Python, e aí vai ter a documentação dessa função e a gente vai poder colocar parâmetros, enfim...

00;04;04;16 - 00;04;28;26
Por isso que sempre que tiver uma dúvida sobre a função, pesquisa ela, porque às vezes vai ter parâmetros que você vai poder utilizar. Então tá: ponto, vem a função... E depois, quando a gente quer filtrar ou pegar uma coluna específica da nossa tabela, a gente usa os colchetes e aí o nome da tabela, entre aspas duplas ou simples.

00;04;29;10 - 00;04;50;23
Uma coisa legal aqui também é que a gente pode usar duas funções ao mesmo tempo. A gente está usando groupby e nunique juntos. Então a gente aplicou primeiro o groupby e depois a gente fez a contagem de números únicos. E está tudo certo. A gente pode juntar várias coisas ali, funcionando. Inclusive, a gente tem que saber interpretar o resultado.

00;04;50;23 - 00;05;14;05
Porque imagina que aqui a gente rodasse isso e na contagem de feminino, desse 10 mil e alguma coisa. Poxa, a gente olhou na última aula que a nossa tabela tem 4 mil linhas. Alguma coisa está errada com esse resultado. Então é importante a gente já ir construindo nosso notebook e já ir observando os valores para quando der algum resultado muito diferente a gente vai conseguir pensar: opa, tem alguma coisa errada nisso aí. Beleza?

00;05;14;29 - 00;05;38;17
Bom, voltando aqui, uma outra forma é utilizar a função do pandas chamada "value counts", que é contar valores para fazer essa contagem aqui de forma mais rápida. Essa função também pode receber o parâmetro de dropna, para visualizarmos a quantidade de valores nulos. Então a função fica assim: vamos criar mais uma célula de código e dados, que é a nossa tabela.

00;05;39;26 - 00;06;08;03
Qual é a nossa coluna? Gênero. E aqui no final a gente coloca ponto (para falar que é um função) value underline counts, parênteses... E aí dentro dos parênteses a gente coloca o parâmetro dropna igual a false. Se a gente não colocasse nada, se a gente não colocasse o dropna, se a gente só colocasse os parênteses vazios, por padrão, a função ia tirar os valores nulos.

00;06;08;03 - 00;06;34;29
Mas a gente quer ver, então a gente coloca esse dropna false. A gente executa essa célula. Opa, deu um erro, vamos ver aqui. Eu escrevi errado, né gente? Vamos ficar atentos também ao nome das funções: value_counts. Tem um "s" aqui no final. Aí a gente executa e dá certinho. Mesma coisa: quantidade de valores masculinos, femininos, prefiro não informar e a quantidade de nulos.

00;06;34;29 - 00;07;14;16
Utilizem essas funções para saber a quantidade de pessoas por raça, por exemplo, em outras outras colunas, também. Beleza? Inclusive nós podemos realizar filtros e utilizar o value counts ao mesmo tempo, vocês acreditam? Naquele filtro lá de pessoas acima dos 30, por exemplo, vamos supor que queremos saber a quantidade de pessoas por nível acima dos 30. Então nós queremos apenas respostas acima dos 30 anos. Ou seja: nosso filtro de idade, nossa tabela dados - colchete, dados - nossa coluna de idade maior que 30.

00;07;15;19 - 00;07;39;20
Então aqui a gente tem o nosso filtro de idade, mas a gente quer o quê? A gente quer saber a quantidade de pessoas por nível. Então, depois do nosso filtro, a gente vai colocar qual que é a coluna que a gente quer saber a quantidade, no caso: nível. Coluna, então a gente coloca colchetes e o nome da coluna: nível.

00;07;40;10 - 00;08;06;15
Tem que ser exatamente igual, gente, ficou na dúvida? Dados ponto columns para poder listar ali o nome das colunas e tal. E aqui depois a gente coloca o value counts, aqui o nosso parêntese, porque a gente não vai colocar nenhum parâmetro. Executa. E aqui a gente tem a nossa tabelinha, que fala que acima dos 30 anos de idade a gente tem aqui essa quantidade de pessoas por nível.

00;08;06;29 - 00;08;33;06
Ok? Então a gente fez primeiro um filtro e depois desse filtro a gente fez uma contagem de pessoas por nível, ok? A maior quantidade é de pessoas sêniors. Então são pessoas que começaram como júnior e foram evoluindo dentro suas carreiras. Será que a gente consegue olhar ali só pessoas do gênero feminino acima dos 30 e igual. Nesse caso, a gente continua realizando filtros, mas acrescentamos o filtro de gênero.

00;08;33;07 - 00;09;04;05
A gente já aprendeu a fazer dois filtros ao mesmo tempo, né? Eu vou copiar aqui para poder facilitar. Vou criar mais uma célula de código, colar aqui embaixo. Beleza, eu já tenho o filtro de maior que 30 anos, agora eu quero o filtro de gênero. Então eu coloco aqui dentro o & e aí eu faço o meu filtro de gênero: igual igual a feminino.

00;09;05;04 - 00;09;28;13
Tô esquecendo alguma coisa? Tô, né, gente? Tô esquecendo os parênteses. Vamos colocar aqui, selecionar tudo aqui. Parênteses, que já vem tudo rápido e prático... Selecionar do lado de cá... Vamos ver o que a gente está fazendo aqui: a gente está fazendo dois filtros. A gente está querendo ver todo mundo que é maior que 30 anos que marcou que é do gênero feminino. E aí a gente quer saber a contagem por nível.

00;09;28;15 - 00;09;57;19
Beleza? Vamos executar pra gente poder ver o que vai acontecer. Veja bem, temos mais nível pleno que júnior. Podemos talvez concluir ou começar a pensar que as pessoas do gênero feminino encontram mais dificuldade de alcançar um nível de senioridade, né? Olha aí, algumas questões começando a surgir na nossa análise. Bom, outra coisa legal que fizemos lá no Módulo 2 foi usar a tabela dinâmica para saber a quantidade de mulheres e homens gestores.

00;09;58;15 - 00;10;28;06
O pandas tem uma função de tabela dinâmica chamada Pivot Table, que é tabela dinâmica em inglês. Para usar essa função, primeiro precisamos colocar o "pd", de pandas - que é aquele apelidinho que a gente deu pro pandas - então pivot table... Assim: vamos criar mais uma célula de código, pivot table.... Assim... Naquela ideia de avisar de qual biblioteca que é a função que a gente vai usar, então a gente está usando o pivot table de qual biblioteca?

00;10;28;06 - 00;10;54;09
Do Pandas. Dentro dos parênteses, precisamos colocar os parâmetros dessa função. Primeiro, colocamos qual é a tabela, que no caso é dados. Então: tabela de dados. Utilizamos a vírgula para separar os parâmetros e, em seguida, o parâmetro values, que é de valores. Então: vírgula values igual... Igual ao quê? Ao id, porque a gente está somando pela coluna de id.

00;10;54;19 - 00;11;22;09
Coluna, então a gente coloca colchetes, aspas e id. O próximo parâmetro é o de index, que é o mesmo que linhas. Lá no módulo 2 nós colocamos o gênero como linha quando a gente fez a nossa tabela dinâmica, e aqui também. Então a gente vai colocar aqui depois uma vírgula, index e aí o igual. E aí a gente coloca nossa coluna, que é de gênero.

00;11;23;15 - 00;11;48;09
Ok, e o próximo parâmetro da função é o Columns, que é de coluna. Qual coluna que a gente está querendo? A coluna de gestor. Então a gente vai colocar aqui a coluna de gestor. Então a gente coloca aqui: vírgula columns - coluna em inglês - igual a, abre colchetes, gestor. Aí você pensa: nossa, Carol, agora eu esqueci qual que é o nome da coluna de gestor.

00;11;48;25 - 00;12;16;03
E aí só pra poder ensinar também como apaga uma célula, vou colocar uma célula de código embaixo, vou colocar: dados ponto columns, só pra gente poder saber como escreve essa coluna de gestor, vou executar, procurar aqui... Cadê gestor? Olha só, tem um sinal de interrogação no final, então isso tem que ser exatamente igual, gente. A gente volta na célula de cima, coloca as aspas e coloca lá o nome da coluna.

00;12;16;15 - 00;12;35;14
Só que a gente quer apagar essa célula que a gente executou embaixo porque a gente não precisa mais dela, a gente só queria saber qual era o nome da coluna. Aqui no canto, aqui nessas opções de seta pra cima e para baixo, tem uma lixeirinha no final. Aí ela diz “excluir célula”. A gente clicou, excluiu a célula, seguimos a nossa vida, ok?

00;12;36;04 - 00;12;58;09
Pra finalizar a função, a gente tem um último parâmetro que é chamado “aggfunc”, que é “aggegation function”, que basicamente é onde a gente informa qual a função que a gente vai realizar nesse pivot table: se vamos somar, se a gente quer é a média, etc. Lá no Excel, essa parte ficava em “valores”, onde a gente colocava o “count”, de contar valores.

00;12;58;21 - 00;13;27;29
Aqui nós também vamos usar o count, mas com essa função de aggegation function. Então aqui, depois do gestor, a gente coloca vírgula, aggfunc, igual, e aí aqui dentro a gente coloca o count, beleza? E aí a gente tem a nossa função. Bom, nós colocamos ali a função que a gente vai aplicar, que é o count, ou seja, que a gente irá contar a quantidade de id único que tem por gênero na coluna de gestor.

00;13;28;19 - 00;13;54;27
Ok? Lembrando que os parâmetros serão sempre em uma ordem específica. Caso tenha dúvidas, a gente pode pesquisar pela função lá no Google e procurar a documentação dela. E a gente vai executar e ver o resultado. Olha, deu um erro aqui falando de sintaxe. Toda vez que é sintaxe, pode procurar que a gente escreveu alguma coisa errada. Vamos dar uma olhada aqui o que a gente fez de errado nessa função.

00;13;55;14 - 00;14;30;10
Bom, aqui a gente colocou os dados, o value... Olha só: em index, eu coloquei a coluna, mas eu não fechei o colchete. Vai dar erro mesmo. Vamos fechar aqui o nosso colchete e vamos executar de novo. Deu mais um erro. Vamos ver o que mais... Aqui, olha, eu tinha fechado o colchete lá na frente, o que não faz sentido, era para poder fechar aqui. Então a gente vai tirar esse colchete que está sobrando e agora a gente vai executar de novo e vai funcionar. Agora funcionou.

00;14;30;10 - 00;14;57;17
Ok. Prontinho, temos nas linhas os gêneros e nas colunas o false para quem não é gestor e o true pra quem é gestor. Assim, conseguimos ver que 132 pessoas do gênero feminino são gestoras e 578 pessoas do gênero masculino são gestores. Bem mais masculino do que feminino. A gente pode pensar algumas coisas aqui, como, por exemplo, como a maior quantidade de pessoas que responderam

00;14;57;17 - 00;15;29;06
o formulário é do gênero masculino, então seria natural que fosse a maior quantidade em todas as categorias. Porém, sabendo como nossa sociedade é, podemos pensar também que pessoas do gênero masculino conseguem chegar a cargos maiores com mais facilidade do que pessoas do gênero feminino, né? Um lembrete pessoal: se ficar na dúvida na utilização de funções do pandas, vá lá no cookbook e vê alguma aplicação ou vem aqui no Google e digita a função que está na dúvida: pandas pivot table e clica aqui na primeira opção.

00;15;29;06 - 00;15;57;11
Quer ver? Vamos abrir uma nova aba: Pandas pivot table. Clique aqui na primeira opção, que já é a documentação. E aí já tem pandas pivot table. E dentro dos parênteses, a gente tem todos os parâmetros que essa função aceita. Não necessariamente a gente precisa usar tudo. A gente está usando várias funções aqui, inclusive, com os parênteses vazios, que a gente está deixando no modo padrão, default.

00;15;58;22 - 00;16;20;26
Aqui em baixo a gente tem os parâmetros e as inscrições. E mais embaixo aqui ainda, a gente tem algumas aplicações, alguns exemplos. Aqui do lado esquerdo a gente tem várias funções do pandas e sempre é possível vir aqui e ver como a função é, tentar aplicar e tal. É isso por hoje, gente. Até a próxima.