

Distribuição de frequências qualitativas - Método 2

Transcrição

[0:00] Muito bem, continuando com nossa construção de tabela de distribuição de frequência, vou mostrar pra vocês o método do pandas agora pra fazer cruzamentos entre variáveis e construir essas tabelas de distribuição de frequência com o cruzamento dessas variáveis.

[0:14] Aqui deixei um link pra ajuda do crosstab, que é o método que a gente vai usar do Pandas, você entra, se tiver querendo saber mais informações, dá uma olhada aqui como utilizar isso, estou deixando esses links, não sei se falei anteriormente, pra todos os métodos que estou utilizando, pra ajudar você aqui no caso do Value Count que usamos na aula anterior, tem uma ajudinha, está tudo documentado bonitinho.

[0:37] Então, eu deixei uma célula onde estou criando um dicionário para transformar nosso dado numérico em uma string, por exemplo, 0 significa masculino, 1 significa feminino, criei um dicionariozinho aqui que vou utilizar daqui a pouco, aqui mesma coisa pra cor, isso a gente consegue ver, você deve lembrar, variáveis utilizadas, aqui a gente tem, sexo, 0 masculino, 1 feminino, e a cor, mesma coisa que eu coloquei embaixo, deixei construído pra não precisar digitar agora. Shift enter, rodamos, vamos começar a fazer a tabela de frequência, frequência vai ser igual, chamo pandas, Pd.crosstab e passo pra ele a variável que quero na linha, dado sexo, na linha sexo, e depois vou pular uma linha, vou passar a variável que quero nas colunas, dados, cor, se não me engano, vou passar aqui, isso é o suficiente pra construir a tabela, a tabela está aqui, você vê que nas linhas tenho as informações, sexo, e nas colunas, de cor.

[2:02] Aqui você vê a contagem e cruzamento entre elas, vamos melhorar isso aqui pra poder visualizar melhor. Quero fazer o seguinte, vou passar frequência novamente, Rename, vou renomear vou transformar esse 0 em masculino e feminino, e mesma coisa nas colunas.

[2:23] Rename, passo pra ele o parâmetro index, e o que tem nas linhas, sexo, então construí aqui em cima, Sexo, e passo pra ele esse dicionário, Python pode passar, vai dar tudo certo. Vírgula, posso fazer isso, in place equal to, que ele vai modificar o próprio dataframe que está tratando.

[2:48] se eu não fizesse isso, ia te que fazer isso, frequência igual a isso aqui e tiraria esse cara daqui, eliminaria e ficaria assim, estou fazendo um pouco diferente, fazendo a modificação do próprio dataframe, vou rodar pra você visualizar, aqui, temos informação masculino e feminino, fazer a mesma coisa mas com pequena modificação.

[3:16] No caso não vai ser mais index, vai ser columns, e também não vai ter sexo aqui, vai ser cor.

[3:26] Aqui tenho uma tabelinha de frequência onde tenho cruzamento de duas variáveis, sexo e cor, estou dizendo nessa contagem que tenho 22 mil, 194 pessoas do sexo masculino da cor branca, 117 pessoas do sexo feminino da cor amarela, com isso você consegue fazer uma quantidade boa de análise e tentando sempre entender como está funcionando a distribuição dos seus dados. A mesma coisa posso fazer pra criar tabela de percentuais, percentual vou mudar os nomes pra manter armazenado aí, em um dataframe, continua do mesmo jeito aqui, e aqui dentro do crosstab, normalize=true, sem o Y.

[4:25] Muito bem, fez aqui, posso multiplicar por 100, pra ficar daquele jeito bonitinho, como fizemos na aula anterior. 28% das pessoas do sexo masculino e da cor branca, em todo o universo que estamos trabalhando. Aqui temos 14% feminino, da cor parda, e vamos entendendo como nossos dados estão organizados. Uma coisa que vou plantar, uma sementinha, que vou pedir pra vocês no nosso notebook de exercício, não falei dele, porque estava deixando pro final, mas esse nosso projeto é um projeto de análise descritiva e vou deixar o notebook no final vazio pra você fazer isso, com um roteirinho já previamente organizado, pra você fazer essa análise descritiva sozinho e depois a gente faz uma correção rapidamente. Uma coisa que pretendo pedir lá e a gente ainda não viu, por isso que eu não queria falar, vou dar uma partezinha aqui, justamente isso. A

gente não falou de médios, mas esse método permite que você faça as agregações de acordo com determinada função e usando outra variável dentro desse cruzamento, por exemplo, posso vir aqui e falar que quero usar uma função `aggfunc`, que é função agregadora, e quero usar a média, vou usar uma média porque é uma função que todo mundo já ouviu falar.

[6:06] Mas desse caso eu quero que ele calcule a média de que variável do nosso dataset? Vou colocar a renda. `Dados.renda`, vou tirar esse vezes 100, vamos rodar com percentual, não tem problema.

[6:27] Rodei, aqui, lógico, não botei aqui, desculpe, botei direto, tenho que passar o parâmetro, `values`, e digo que variável quero ali dentro, tá bom, desculpe, tá aqui.

[6:42] Ele faz aqui, cruzando calcula a renda média, passei a função média, ele calcula em cima da renda, renda média, dentro do cruzamento, o que isso quer dizer, que a cor amarela, homens da cor amarela ganham em média, desse nosso dataset, 4 mil, 758 reais. As mulheres da cor negra, ganham em média mil, 134, e por aí vai, com isso você consegue fazer uma quantidade de análise boa, você pode usar outra variável ali dentro pra calcular média, mediana, e por aí vai, beleza?

[7:16] Era isso que eu queria mostrar, agora vamos fazer também, continuar falando de distribuição de frequência, uma coisa bem importante pra entender como o dado funciona, mas vamos mexer com variáveis quantitativas.

[7:32] Vejo você no próximo vídeo.