08

Tipos de dados

Transcrição

[0:00] Beleza pessoal, dando continuidade no nosso curso de estatística, no último video eu esqueci de mostrar pra vocês como salvar o notebook que a gente acabou de criar, então desculpe, vou fazer isso agora, a gente entra em file, download.ipynb, o download do nosso notebook, clicou aqui, vai fazer o download, aqui estou fazendo com o Chrome, ele vai jogar na pasta downloads, mostrar pasta, clico aqui, ele jogou meu notebook aqui, eu copio daqui, substituo na pasta onde estou trabalhando.

[0:40] lembrem de fazer isso, assim que a gente acabar cada aula, a gente faz esse procedimento, File, download, e guarda nosso notebook. Lembrando também que se você assistiu a aula, agora não quero mais assistir, fechou, deslogou, vai começar de novo outro dia, outro momento, você tem que fazer o upload do dados.csv assim que você for iniciar o procedimento novamente, vamos lá, tipo de dados, pra entender o tipo de dados, você vai ver que com o passar do nosso curso, você vai entender melhor isso, pra cada tipo de dado, você tem um tipo de estatística que você vai trabalhar, tipo de tratamento que você vai executar em cima dessa informação, então por isso é importante a gente conhecer, identificar esses tipos de dados. Tenho o dataset, o dataframe, com os dados do nosso dataset, vou fazer um macete pra gente ter dados.head, eu vou fazendo isso, esse método do Pandas vai mostrar os 5 primeiros registros ou observações do nosso dataframe, já é o suficiente pra entender, indo e voltando aqui fica mais fácil. Tipos de dados, em estatística classificamos em basicamente dois tipos, os qualitativos e quantitativos, o próprio nome é bem intuitivo, os qualitativos expressam qualidade, um atributo dos dados, enquanto os quantitativos são quantidades, mensurações, vamos tentar entender isso com os dados que temos aqui em cima, unidade de federação, bem simples de entender, que é um dado qualitativo, você tem o nome dos estados, não é uma contagem, idem para sexo, sexo também você tem masculino e feminino, são características, você está atribuindo sexo a uma pessoa, e um sexo a outro, idade é um dado quantitativo, uma contagem.

[2:34] quantos anos você tem, tenho 23, outro 35, e por aí vai, você vai contando as idades. Cor, também exemplo de dado qualitativo, anos de estudo, nesse caso, como a gente vai ver em cima, vamos voltar anos de estudo, vamos ver que é uma variável qualitativa, o primeiro, instrução em menos de um ano de estudo, como se fosse uma classe, o último você pode ver bem claro, o 16, atribuído dado 16, uma classe de pessoa que tem 15 anos ou mais de estudo, pode ter 16, 17, por aí vai. Isso torna esse dado qualitativo.

[3:19] Renda, quantitativo, valores monetários, e altura também uma mensuração, dado quantitativo.

[3:32] esses dois dados, a gente consegue dividir em mais duas categorias cada um deles, no caso dos qualitativos, a gente tem os qualitativos ordinais e os nominais, e no caso dos quantitativos, os discretos e os contínuos. A gente pega um exemplo deles utilizando o Python junto com o Pandas pra ajudar a entender essa coisa. Vamos começar com variáveis qualitativas ordinais. Olhando aqui em cima, qual a gente acha que é qualitativa ordinal, bem simples. Sexo pode ser ordenado, você pode ordenar homem melhor que mulher ou vice-versa? Não, então, essa variável não serve.

[4:11] UF a mesma coisa, esses números foram atribuídos, mas você poderia ter atribuído de outra forma, o 11, por exemplo, podia ser Rio de Janeiro, ou São Paulo, foi atribuído por alguém, então não tem como ordenar. Mesma coisa pra cor, não tem como falar qual cor é melhor. Já no caso de anos de estudo, conseguimos fazer uma ordenação, 1 ano, depois 2, por aí vai, a gente consegue ordenar isso.

[4:43] vamos mostrar esses dados aqui, só na forma de mostrar, ficar claro pra estudar depois, quais os dados que são classificados de que forma.

[4:57] Anos de estudo, ordinal. Anos de estudo, uma forma da gente pegar os dados, criar uma Series, mas não quero isso, quero ver só quais são os valores únicos, com o Pandas eu faço Unique. Ok.

[5:19] Ele demora um pouco mas rodou, tá aqui. Deixo ordenado porque pegou uma ordem aqui, uso uma built-in function do Python, que é o sort, e vai sortear esse array e deixar cada um aqui, os valores que tenho no meu dicionário, 1, 2, 3, 4, 5, 6, 7, e por aí vai, até o 17. Essa é uma variável classificada como qualitativa ordinal como está aqui em cima, anos de estudo no nosso dataset. Indo mais a frente, vamos entrar nas qualitativas nominais, que demos exemplo agora pouco, sexo, cor, UF, são essas variáveis. Vamos printar aqui, da mesma forma que fizemos aqui em cima. Sort, vai vir do mesmo jeito que fizemos aqui, muda pra UF.

[6:15] UF também, vai plotar todos os códigos da unidade da federação aqui em cima, se você quiser saber quais são, vai em cima no nosso dicionário e vê o que significa cada um deles. A mesma coisa aqui pro nosso amigo, sexo, codificado, no caso, 0 é masculino e 1 feminino, mas podia ser ao contrário, por isso a gente não pode ordenar.

[6:43] E mais uma de exemplo, usando nosso próprio dataset, a cor. Esse dado foi, a numeração foi atribuída pelo pesquisador do IBGE, não se sabe como nem por que, mas podia ser qualquer outro tipo de numeração, podia estar a cor negra aqui em cima, branca no final, parda no meio.

[7:04] Por aí vai, escolheu indígena primeiro, depois branco, preto, amarelo.

[7:11] Então não tem como a gente ordenar e falar que um é melhor que o outro, e por aí vai.

[7:17] Isso classifica variáveis como qualitativas nominais. Indo à frente, vamos entrar nas qualitativas discretas, e no caso do nosso exemplo, a gente pode, nesse exemplo, uma coisa interessante, falar sobre idade, porque pode ser classificada de varias formas, dependendo de como foi representada na pesquisa, no dataset. No caso aqui eu posso classificar ela como quantitativa discreta. Porque ela representa uma contagem, representada por números inteiros, aqui, quando é feito a pesquisa no IBGE, pergunta quantos anos você tem completos, não quer saber se tem mais 2 meses, 2 horas, 3 minutos, não sei quantos segundos você nasceu, ele quer saber quantos anos completos você tem, informação desse jeito, crua.

[8:07] Vamos colocar, idade, consigo fazer certas contas, por exemplo, idade mínima do dataset, desculpa, botei idade, tem que pôr dados, no caso, o Pandas permite que faça isso aqui, como é o nome de variável inteiro, consigo fazer dessa forma aqui, idade, idade mínima é 13, posso fazer também a máxima de quantos anos, no máximo 99.

[8:40] Posso inclusive fazer um print aqui, mais sofisticado, print, ok, vamos botar aqui, DE, percentual S até percentual S anos. Ok? Fechou, coloco mais um percentual aqui, e coloco esses caras aqui dentro, dados Min, dados Max. Isso é uma forma de representar, esse vai ser o Min, e esse vai ser o Max, vou dar aqui, Min, esse aqui de fora sai, ok, tenho uma representação aqui, e uma string onde tenho de 13 a 99 anos.

[9:29] Isso é uma forma de tratar melhor os dados que estamos trabalhando. Esse cara eu consigo contar, 13, 14, 15, 16, um número inteiro, uma contagem. A variável idade, tem essa coisa interessante que acabei de falar sobre idade mais precisa, que pode ser discreta nesse caso aqui, anos completos. Ela pode ser contínua, como a gente pode ver na próxima variável, que eu vou falar sobre altura, porque você pode estar representando idade exata, mês, hora, minuto, segundo, e cada vez mais você vai ter um número fracionado, real. Você por exemplo tem 25,32 anos, porque você tem essas coisas fracionadas, a partir do momento que você faz 25 anos, começa a contar pra você se aproximar dos 26 e vai tendo fração da sua idade. Mais uma coisa, pode ser ordinal quando ela é colocada em faixas, você tem os dados agrupados em faixa de idade, 5 a 10 anos, 10 a 15, de 15 a 20 e por aí vai, você tem as frequências, essa variável idade nessa situação vira uma variável qualitativa ordinal.

[10:50] aqui a observação com os exemplos pra você, vamos falar das variáveis continuas, que no caso aqui a idade poderia ser, a renda também, mas como a renda veio do IBGE sem os centavos, achei melhor não falar.

[11:13] Eu criei a variável justamente pra dar o exemplo da variável que representa uma mensuração, legal, então, ela num intervalo pode assumir qualquer valor que você possa imaginar, altura, 1.60, 38, 08, pode ser sei lá quantas casas decimais.

[11:33] Altura entra como exemplo da variável quantitativa contínua, vou fazer o mesmo exemplinho que fiz aqui mas com altura, de, até metros, e vou mudar idade pra altura.

[11:52] Então temos aqui, de 1.33 metros de altura, aqueles quebrados todos que falei que não dá pra precisar, até 2.02 metros de altura, as alturas que a gente tem no nosso dataset. Lembrando só mais uma coisa, você pode colocar, pra não ficar jogado, você pode fazer, com colchete, isso aqui, a mesma coisa que isso aqui. Lógico, pra variável anos de estudo, que é uma variável que tem espaços entre as palavras no nome, isso não funciona. Então você vai ter que fazer sempre assim. Shift-enter, você tem no caso da altura, mesmo resultado. Aqui deixei uma imagem aqui embaixo, que criei pra lembrar de forma simples as variáveis que a gente tem. Variáveis todas, a gente pode ter as qualitativas ordinais, os exemplos que a gente deu ali em cima, e as nominais. Sexo, Cor, UF, aqui as ordinais, temos os anos de estudo, podemos ordenar, uma ordem nos anos, as quantitativas temos as discretas, que a gente deu o exemplo da idade cheia, anos completos, e as contínuas que a gente deu o exemplo da altura. Tendo isso na cabeça, vamos passar pras distribuições de frequência, a gente vai ver como calcula a distribuição de frequência pra determinados tipos de variáveis, uma forma um pouco diferente pra fazer isso pra cada um deles, e a gente vai utilizar o Pandas pra fazer isso também, legal?

[13:46] próximo vídeo a gente vê isso, abraço.