02

Distribuição de frequências qualitativas - Método 1

Transcrição

[0:00] Pessoal, legal, voltando aqui nosso curso, vou voltar a falar uma coisa que falei na introdução do nosso curso, que é o projeto que estamos desenvolvendo nesse curso de estatística parte 1, que é uma analise descritiva de um conjunto de dados. A gente tem um dataset, é como se a gente fosse um estatístico ou cientista de dados, ou especialista em data Science, que tem um chefe virtual que chega pra você, coloca na sua mesa esse conjunto de dados, fala, "faça pra mim inicialmente uma analise descritiva desses dados e depois a gente conversa sobre os próximos passos de análise com esse conjunto", legal?

[0:36] É sobre isso que nosso curso parte 1 fala, análise descritiva dos dados. Vamos fazer isso, nessa sessão eu vou falar sobre distribuição de frequência, por isso que estou falando que a gente vai começar realmente o nosso projeto, porque antes a gente só abriu o dataset, viu que tipos de dados a gente tem, agora a gente vai começar realmente a analisar esses dados.

[0:55] A distribuição de frequência é uma técnica de sumarização onde a gente vai começar a entender como as variáveis se distribuem. A gente vai ver se são assimétricas, se elas se distribuem como variável normal, se elas têm outlier, se consegue visualizar isso, se precisa de algum tipo de transformação, um conjunto de técnica que a gente pode abordar em cima de um conjunto de variáveis que primeiramente tem que passar por essa parte visual onde a gente faz essa analise de distribuição de frequência dos dados, tá bom?

[1:25] Nessa mesma sessão, a gente vai ver distribuição e também a forma gráfica de representar essas distribuições, que é o famoso histograma, que todo mundo já deve ter ouvido falar.

[1:37] Então, vamos lá, vamos começar e vou fazer primeiro pra variáveis qualitativas, que são naturalmente categorizadas, então a gente não se preocupa com a categorização delas, elas já vêm assim, que como vimos lá em cima, são UF, sexo, raça, entre outros.

[1:54] Vou fazer pra uma delas, começar com sexo, chamar dados, sexo, vou mostrar um método que faz a contagem pra gente, veja como fiz aqui, e execute as outras como exercício pras outras variáveis.

[2:10] Então vamos lá, um método pra fazer essa contagem automaticamente disponibilizada pelo Pandas, é o velho Count, Só isso ele vai criar pra gente uma contagem das categorias, categoria 0 tem 53 mil e uns quebrados, e a categoria 1, que é o feminino, tem 23 mil. Você pode estar pensando, caramba, tem muito mais homens do que mulheres, o que eu acho que tá errado, e tá, realmente, essa população não é assim que funciona, mas lembre, voltando aqui em cima, na descrição do que fizemos com os dados, temos o item 3 onde diz que foram considerados somente os registros das pessoas de referência de cada domicílio, ou seja, quem respondeu o questionário, o responsável pelo domicílio, o chefe, nesse caso, aqui nessa pesquisa parece que tem mais homens do que mulheres nessa situação, por isso que é importante ter tudo documentado pra gente entender o que tá acontecendo. Outra informação importante que sempre gosto de ter e nas tabelas de frequência vem, é esse cara representado de forma percentual. Como faço isso? Com esse mesmo carinha, o Value Counts, mas eu vou passar um parâmetro, o normalize, normalize igual a true, ele vai normalizar meus dados, colocar tudo na base de 1, é como se ele tivesse somando esses dois e dividindo cada um deles pela soma, tem aqui 70%, quase 70% homens, e 30% mulheres, o que posso fazer pra ficar em percentual, multiplicar por 100, depois a gente põe uma representação desse percentual.

[4:01] O que eu posso fazer agora é justamente isso, melhorar a apresentação, está esquisito isso, não quero mostrar isso pro meu chefe, mostrar uma coisa mais bem estruturada, vou copiar esse cara de cima, vou fazer aqui em baixo, chamar de frequência e colar aqui.

[4:25] rodei, coloquei dentro de uma variável, agora é uma series do Pandas. A mesma coisa pra esse cara, vou chamar de percentual. Pode ser?

[4:40] Beleza, rodei de novo, o que quero, pra ficar mais organizado, colocar tudo dentro de um novo dataframe, pra representar a tabelinha de frequência que vou mostrar pro meu chefe depois. Vou dar um nome, chamar de Dist, Distfreq, distribuição de frequência qualitativa. Abreviei pra não ficar muito grande.

[5:10] Como faço pra criar um dataframe? Chamo Pandas, Pd.dataframe, essa função cria dataframe pra gente, posso passar varias coisas, entre elas um dicionário Python, com essas duas series que acabei de criar, dando um nome, o primeiro vai ser frequência, dois pontos, isso é um arquivo JSON, como se fosse um arquivo JSON e eu passo esse cara aqui, frequência. Coloco aqui, facinho.

[5:47] uma vírgula e vou colocar o outro nome, vai ser a outra coluna do meu dataframe, porcentagem, e como multipliquei por 100, faço isso, fica mais simples, não preciso formatar. Passo esse percentual, pra esse carinha aqui. Já vou mostrar pra vocês como isso vai ficar, muito mais bem organizadinho. Tenho uma tabelinha onde tenho as frequências, percentagem.

[6:27] A única coisa que falta é organizar isso, 01, como vimos lá, não faz sentido, posso colocar 0, pode ser homem ou mulher, uma pessoa leiga não sabe o que isso significa, então vamos dar labels pra esse cara. Como faço isso? Posso fazer aqui em cima mesmo, não, vamos fazer aqui embaixo, primeira coisa que eu tenho que fazer é chamar o dataframe.rename, rename e aí eu vou renomear o index dele e posso fazer isso usando um dicionário também, dizendo que o 0 vai ser masculino, e o 1 vai ser o feminino. Tudo bem?

[7:24] vou falar que isso aqui, se eu fizer pra você que não viu o curso de Pandas, pra não ficar perdido, vou fazer isso, vai mostrar bonitinho, mas aqui ele não salvou a alteração, se eu plotar de novo esse cara, vai mostrar o 01 novamente, pra fazer isso sobrescrever o arquivo e salvar a alteração, tenho que colocar esse parâmetro, In place to, no que a gente fez aqui em cima, reparem, agora não vai nem mostrar o que ele criou, criou, agora eu tenho que pré-visualizar, aqui está salvo.

[8:10] Uma coisa que posso fazer é colocar um título aqui, coisa bem simples também, pego esse cara, faço um .rename, mas aqui vou fazer _axis, estou escrevendo com um monte de letra, passo o nome que quero pra essa coluninha, vai ser sexo, vou fazer o que, vou, lembrando também, tem um place, só pra gente não esquecer, mas antes quero mostrar pra vocês que pode renomear tanto uma coluna como pode renomear uma linha, então tenho que dizer o eixo que quero.

[8:59] Quero que renomeie, no caso do Pandas, ele faz o seguinte, se você colocar 1, ele vai entender que é coluna, se você colocar 0, vai entender que é linha, só que eu nunca lembro disso, então prefiro escrever Columns, porque fica mais simples, ou então, se for linhas, tá bom, agora ele vai fazer a modificação pra gente, fez, e eu rodo, e vejo a modificação aqui, sexo, masculino, frequências, tabelinha de frequências, posso copiar, colar, levar pro meu chefe, ele vai entender perfeitamente o que significa, legal?

[9:41] Vamos dar uma passa nesse vídeo pra não ficar muito grande e vou mostrar outra forma usando outro método do Pandas que é bem interessante de fazer a mesma coisa, legal, próximo vídeo a gente continua.