7.2 - Preparação dos dados (parte 2)

00:00:00:00 - 00:00:01:03
Olá, pessoal. Bom, na aula passada, a gente trabalhou a parte de pré-processamento ali, a gente já mexeu em algumas colunas e tudo mais. Nessa aula a gente vai continuar dando uma olhada em algumas colunas que vão ser importantes para o nosso modelo de regressão, ok? Bom, a gente fez uma coluna de tempo de experiência, a gente deu uma olhada na coluna de números de funcionários. A próxima coluna interessante é de principais motivos da insatisfação da pessoa com a empresa atual. A gente pode fazer um value counts para poder ver como é que está essa coluna. Vamos lá. Dados columns primeiro porque eu também não lembro o nome completo da coluna. Vamos ver. Qual o principal motivo da sua insatisfação com a empresa atual? Dados. Ops. Aqui. Ok. Ponto value counts.

00:00:01:03 - 00:00:02:00
Bom, a gente tem algumas respostas que apontam insatisfação com o salário, né? Por exemplo, salário atual não corresponde ao mercado. Bom, tem algumas que é só salário não corresponde ao mercado. Aí depois tem outras que é: o salário e alguma outra coisa, salário e outras três coisas... Aqui, por exemplo, o salário tem uma outra coisa antes e depois tem falando salário. A gente pode criar uma coluna chamada insatisfação porque somos muito criatives e colocar todos os valores como zero, e depois a gente pode olhar cada linha dessa coluna de qual é a insatisfação e todas as linhas que tiverem a palavra salário, a gente coloca o valor de 1 para essa coluna de insatisfação, tá? Porque aí a gente filtra insatisfação com o salário do que não é insatisfação com o salário, tá? Então vamos começar criando essa coluna de insatisfação.

00:00:02:00 - 00:00:03:00
Dados ponto insatisfação. E aí a gente vai colocar zero. Por enquanto, zero para tudo, é como se ninguém estivesse satisfeito com o salário. E aí agora a gente vai fazer um filtro usando o loc para todas as linhas não nulas da coluna de qual o principal motivo de sua insatisfação com a empresa atual. E aí a gente pode usar essa mesma coluna, assim, ó: vamos colocar dados ponto loc. Aí dentro dos colchetes do loc, a gente coloca a coluna que é dados... A gente pega o nome aqui: qual o principal motivo... Beleza. E aí a gente quer todos os dados não nulos. Então a gente coloca aqui, depois dos colchetes, o not null, que é uma função que filtra todos os dados não nulos.

00:00:03:00 - 00:00:03:58
Ok? Bom, então a gente está pegando dados, a gente está usando o loc, aí dentro do loc a gente está pegando a coluna e fazendo o filtro dos dados não nulos, e aí dentro dos colchetes do loc, a gente faz primeiro o nosso filtro e depois a gente coloca qual que é a coluna que a gente vai usar mesmo, qual que é a coluna que a gente vai estar tendo alguma ação sobre ela. Então, a gente está fazendo o filtro de não nulos da coluna de qual o principal motivo e a gente vai estar usando ela mesmo, porque é nessa coluna que a gente vai estar olhando se tem a palavra salário ou se não tem a palavra salário. Então, a gente repete essa coluna. Então eu vou copiar aqui para ficar mais fácil e coloco aqui. Beleza, vamos voltar aqui: eu estou pegando a tabela dados ponto loc, então eu estou localizando dentro da tabela dados o meu primeiro item aqui, que é todos os dados não nulos da coluna de qual principal motivo...

00:00:03:58 - 00:00:04:57
E aí depois da minha vírgula, eu estou colocando essa coluna, porque é essa coluna que eu vou estar fazendo alguma coisa, que no caso vai ser o apply agora. Depois aqui do meu último cochete, eu coloco ponto apply e aí eu coloco lambda X. Bom gente, lambda X, mas eu acho que a minha cabeça está na frente da fórmula, eu vou subir aqui um pouquinho. Ok, então voltando, a gente tem o ponto apply, dentro do apply a gente tem o lambda, lambda X, um para se tiver salário, zero para se não tiver. Então vamos lá: 1 if... E aí, a palavra salário tem que ser exatamente igual à que está aqui. Então, vamos procurar onde tem salário. Salário exatamente aqui igual. Então, copiei lá, volto aqui e coloco if salário in, ou seja, se o salário estiver in, x else zero. Ok? Bom, então primeiro a gente tem dados, que é a nossa tabela, depois ponto loc e aí o loc recebe ali os dois argumentos.

00:00:04:57 - 00:00:05:51
O primeiro é o nosso filtro, a gente está filtrando todos os dados não nulos da coluna de qual o principal motivo da sua insatisfação e o segundo item é a coluna. Inclusive, não é a coluna como dados, a gente só informa qual que é o nome da coluna. Então, eu posso apagar aqui esse dados, esses colchetes e um colchetes aqui do final. Então, aqui a gente tem um not null e depois qual que é o nome da coluna. E aí, depois do colchetes a gente coloca o apply e o lambda onde um, se tiver a palavra salário na frase, e caso contrário é zero. Vamos rodar para poder ver como é que vai ser a saída aqui. Ó, beleza. A saída são vários zeros e um. Um quando tem a palavra salário lá na frase. E aí agora a gente precisa pegar esse resultado e colocar lá na coluna de insatisfação. A gente não pode pegar direto aqui só a coluna e jogar aqui porque a gente fez esse filtro de not null.

00:00:05:51 - 00:00:06:51
Então a gente precisa fazer esse filtro de novo. Então vamos usar o loc também na parte de atribuição do resultado. Eu, como sou uma pessoa preguiçosa, vou copiar daqui para cá. E aí, só no lugar dessa coluna aqui, dessa segunda coluna que eu tinha colocado, que é qual o principal e tal, ao invés de ser essa, eu coloco insatisfação e aí vai dar tudo certo, ó. Então, a gente deixa aqui do jeito que tá, clica aqui no início, eu colo aquele negócio que eu já tinha colocado, que foi dados ponto loc... Estou filtrando os não nulos, depois da vírgula eu vou colocar qual que é a coluna que eu vou atribuir esse resultado, que no caso é a coluna de insatisfação, fecho os meu cochetes e aí sim coloco o meu igual. Beleza? Então, o que está sendo atribuído? Está sendo atribuído aqui tudo dessa coluna que não é nulo, que nessa coluna tenha a palavra salário.

00:00:06:51 - 00:00:08:20
Se tiver salário é um, se não tiver é zero e vai ser atribuído tudo que nessa coluna também seja não nulo, só que agora na coluna de insatisfação. Vamos rodar, torcer para não dar erro. Não deu erro. Vamos dar uma olhada aí nessa coluna de insatisfação no value counts dela. Vamos ver como que ficou. Agora sim, olha: a gente tem 279 com o valor 1, provavelmente tem a palavra salário, e 2453 com 0. A última coluna que vamos trabalhar é a de nível de ensino. Fazendo um value counts nessa coluna, a gente pode dar uma olhada e ver o que a gente pode fazer. Vamos lá. Eu vou copiar tudo e só mudar o nome da coluna, mais fácil. Nível de ensino. Então tá, pessoal, olhando esse value counts, a gente vê aqui a quantidade de cada um. E aí, como eu falei, por exemplo, "não tenho graduação formal" pode ser zero. Aí depois "estudante de graduação" pode ser um.

Graduação, dois. Pós-graduação, três. Mestrado, quatro. Doutorado, cinco. E aí tem esse Prefiro Não Informar, que pode ser menos um, por exemplo. E a gente pode fazer isso com o apply, que a gente acabou de fazer aqui. Como? A gente pode pegar a coluna aqui ponto apply, aqui dentro o lambda, lambda x, e aí a gente começa, né?

00:00:08:20 - 00:00:10:06
Bom, zero se x igual a não tenho graduação formal. Perfeito. E aí, caso contrário, aí a gente continua. Caso contrário, um if x igual igual, a gente combinou que vai ser estudante de graduação. E aí vai fazendo. Eu sou uma pessoa, como eu já disse, preguiçosa, já deixei prontinho aqui embaixo. Vou copiar e vou colar aqui. Então vamos lá. O que aconteceu aqui? Eu peguei a coluna nível de ensino, coloquei o apply e o lambda e aí comecei: zero se o X for igual a não tenho graduação formal. Caso contrário: um se o X for igual a estudante de graduação. Caso contrário: dois se o X for igual a graduação ou bacharelado. Caso contrário: três se o X for igual a pós-graduação. Caso contrário: quatro se o X for igual a mestrado. Caso contrário: cinco se o X for igual a doutorado ou PhD. Caso contrário: menos um. Bom, eu passei por todos os itens. Não é nenhum desses aqui, só pode ser o prefiro não informar. Então a gente combinou que vai ser o menos um, por isso aqui no final é else menos um, beleza? Vamos rodar para ver se vai dar certo. Ó, a gente já tem os numerozinhos aqui, deu bom. Vamos, ao invés de criar uma nova coluna, vamos atribuir a mesma coluna mesmo. E, foi...