4.5 - Featuring engineering: criando novas variáveis para facilitar a análise (parte 2)

00;00;00;19 - 00;00;24;27
Bom, a gente tem uma segunda tabela que faz parte dos nossos dados. Só que a gente dividiu elas em duas tabelas: uma que a gente está usando aqui e uma segunda tabela com outras colunas. Vamos aprender agora como juntar essas duas tabelas, beleza? A gente vai aprender a usar o "merge". Primeiro eu preciso ler essa tabela, então vamos lá:

00;00;25;22 - 00;00;48;26
dados 2 (porque eu sou muito criativa) pd ponto read excel (porque essa tabela está em Excel). E aí a gente vai procurar ela. Essa tabela vai estar disponível para vocês tranquilo para vocês fazerem o download e subir lá na pastinha de vocês, mas é bom relembrar que a gente tem o drive, a gente vai lá no nosso drive, eu vou lá na pastinha que eu criei só para o curso da PrograMaria.

00;00;49;27 - 00;01;21;28
E aí eu acho aqui a planilha parte dois e aí eu copio o caminho dessa planilha. Beleza! Executando. Vou fechar aqui. Uma nova linha de código... A gente pode até dar um: dados 2 ponto read pra gente poder ver o início dessa tabela. Olha, já de início a gente tem um id, que é interessante porque na nossa tabela que a gente está usando até agora, também tem um id, né?

00;01;22;13 - 00;01;54;10
Olha só. E a gente tem umas outras colunas aqui. Essas colunas são basicamente de perguntas que as pessoas responderam quando estavam fazendo o formulário. Ok. E a gente precisa juntar essa planilha, essa tabela, com a tabela que a gente já está usando. E aí a gente pode usar o merge que mergeia, que junta essas duas tabelas. Só que para juntar essas duas tabelas, elas precisam ter um dado em comum que aí, através desse dado em comum, você vai conseguir juntar uma coisa com a outra. O que a gente tem em comum nas duas tabelas?

00;01;54;10 - 00;02;17;29
A gente tem a coluna de id. E olha que legal: o id é único. Então a gente realmente consegue juntar essas duas colunas com esses valores únicos e ter essas duas informações de uma coluna e da outra. Então, a gente vai usar o merge assim. Merge é uma função do Pandas. Então, primeiro a gente coloca qual é a nossa tabela, nossa primeira tabela: dados ponto merge.

00;02;19;08 - 00;02;39;09
E aí a gente coloca qual é a segunda tabela que a gente está querendo mergear. Ok. E aí a gente tem um parâmetro que chama "on" (O-N). E esse on é pra dizer qual que é o dado em comum entre essas duas tabelas. No caso, o dado em comum é a coluna que chama id. Então a gente coloca: igual a id. Ok?

00;02;40;06 - 00;03;07;18
E aí a gente tem também um outro parâmetro, que é o “how”. Esse how é meio que pra perguntar como você quer mergear isso. Você quer mergear isso mantendo os dados da direita, os dados da esquerda, você quer ter só os valores em comum (porque pode ter valores que não estão em comum) ou você quer todos os valores juntos? No caso aqui a gente pode colocar aqui que queremos os dado da esquerda. No nosso caso aqui vai dar tudo certo, não vai ter problema.

00;03;09;01 - 00;03;37;27
A gente vai executar. E olha só, aqui no início a gente vai ter a tabela que a gente já está mexendo: idade, faixa de idade, e no final a gente vai ver que juntou com a segunda tabela, que tem as colunas de perguntas. Muito legal, né? Então a gente teve retorno nessa tabela. Vamos atribuir essa tabela aos nossos dados porque agora a gente tem uma tabela completona, robusta. Dados igual a esse merge.

00;03;37;27 - 00;04;05;20
Perfeito. Então a gente já tinha uma tabela, a gente tinha outra, juntamos e agora, continuando nessa engenharia de dados, a gente pode trabalhar com uma coluna com frases, porque nessa segunda tabela que a gente juntou, a gente tinha muita coluna com frases. A gente pode até dar uma olhada aqui: dados pontos columns e ver quais as outras colunas que a gente tem.

00;04;06;24 - 00;04;42;00
A gente tem “quanto tempo de experiência na área de dados você tem?” “Você está satisfeito com a empresa atual?” Enfim, que tal se a gente usar aqui a coluna de “você pretende mudar de emprego nos próximos seis meses?” A gente pode fazer um value counts pra ver quantas opções aí que a gente tem dentro dessa coluna. Mais uma célula de código, dados, a gente copia aqui a coluna “você pretende mudar de emprego nos próximos seis meses?”

00;04;42;11 - 00;05;11;01
E a gente coloca aqui dentro. Tem que ser exatamente do jeito que está ali a coluna, hein gente? Se a gente colocar o “meses” em maiúsculo, já vai dar ruim. A gente coloca o value counts e pronto. A gente tem aqui algumas opções. A gente pode criar duas novas colunas a partir dessa, uma coluna para dizer se a pessoa está em busca ou não de emprego e outra para dizer se a pessoa está aberta a oportunidades ou não.

00;05;11;03 - 00;05;38;23
Que tal? Isso já ajuda pra gente. Em vez de gente ter texto, a gente vai ter duas colunas com um true ou false, por exemplo. Mas como? Vejam as opções de pessoas que estão em busca de algum emprego: todas que estão em busca de algum emprego tem as palavras “em busca”. Então toda vez que estiver “em busca” na frase, podemos colocar um true em uma coluna que a gente pode criar chamada “em busca”.

00;05;39;11 - 00;06;23;02
Então, toda vez que a pessoa está em busca “true”, caso contrário, “false”, e aí a pessoa não está procurando emprego. A gente viu numa aula anterior que podemos usar o contains para procurar uma palavra ou um trecho específico. Então a gente pode fazer assim: a gente pode criar nossa coluna, dados, vamos chamar de “em busca” (de novo sendo muito criativa). E aí a gente vai fazer um filtro, mas a gente vai usar o contains. Então a gente vai colocar: dados, a nossa coluna, e aí a gente quer que toda vez que tiver o “em busca”, a gente quer que coloca true e, caso contrário, false.

00;06;23;08 - 00;06;48;17
Então a gente coloca: str ponto contains. Dentro dos parênteses a gente vai colocar o parâmetro, o parâmetro que a gente está buscando agora é “em busca”. Então “em busca” e vírgula por quê? Por que toda vez que ele encontrar um “em busca”, ele coloca true, mas no caso contrário a gente quer que coloca false.

00;06;48;17 - 00;07;16;02
Então a gente coloca um “case”, que é caso contrário, e false. Ok? E aí a gente vai criar essa coluna. Vamos executar aqui. Em uma nova célula de código, a gente pode até copiar aqui e dar um value counts só pra gente poder saber quantas pessoas que tem e quantas pessoas que não tem. Vamos lá.

00;07;21;06 - 00;07;48;08
Olha, a gente tem 1364 pessoas que deu true, ou seja, pessoas que estão em busca de uma nova oportunidade e 2332 que não estão, false. Ok? Agora a gente pode criar uma coluna chamada “aberta a oportunidades”. Observando essas frases aqui, toda vez que tem “aberto” na fase, significa que a pessoa está aberta a oportunidades. Então a gente pode só copiar o código anterior e mudar a palavra, que vai facilitar para a gente.

00;07;49;18 - 00;08;34;11
A gente copia isso aqui. Embaixo a gente cria mais uma célula de código, cola aqui e agora a gente vai mudar o nome. A gente pode colocar “aberto oportunidades”. Aberto oportunidades. Nome gigante, né? E o contains. Agora a gente não está procurando “em busca”, a gente está procurando “aberto” e ok. É na mesma coluna que a gente está procurando mesmo, certinho... E vamos lá... Executando. Vamos fazer a mesma coisa, vamos fazer um value counts aqui, copiando e colando aqui um value counts dessa coluna que a gente criou pra gente poder ter uma ideia.

00;08;35;18 - 00;09;10;07
E olha, 1354 estão abertos a oportunidades, enquanto 2342 não estão abertos a oportunidades, beleza? Criando essas duas colunas, podemos até fazer filtros para saber quem está aberto a oportunidades, mas não está buscando e coisas assim. Então, pessoal, nessa aula a gente vários exemplos de Featuring Engineering e deixo aí um desafio pra vocês darem uma olhada nas outras colunas, analisar o que pode ser melhorado, quais colunas podem ser criadas para facilitar na hora de fazer um gráfico por exemplo. Por exemplo, vou deixar um desafio aqui pra vocês: criar uma nova coluna de etnia, simplificando a categoria para branca, não branca e outros.

00;09;10;07 - 00;09;36;24
Que tal? Vocês vão criar uma coluna nova e usando a coluna de cor, raça ou etnia vocês vão criar essas novas variáveis. Bom, durante essa aula, nós exploramos o conceito de engenharia de recursos e sua importância na análise de dados. O Featuring Engineering envolve a criação, transformação e seleção de variáveis para melhor representar os padrões e relações usados.

00;09;37;09 - 00;10;01;29
Vimos algumas técnicas, como a criação de variáveis dummies, get dummy, para representar variáveis categóricas e a criação de novas variáveis com base em combinações de variáveis existentes. É crucial escolher as técnicas do Featuring Engineering que sejam adequadas ao tipo de dados que estamos usando, para garantir que os recursos criados capturem efetivamente os padrões nos dados. E é isso, pessoal, nos vemos na próxima aula.