7.2 - Preparação dos dados (parte 1)

00:00:00:00 - 00:00:00:59
Olá, pessoal. Bem-vindes a mais uma aula. Na última aula, terminamos falando sobre o fluxo de modelagem do nosso modelo. Deixa eu voltar aqui no fluxo. Aqui. Bom, a gente falou que essas duas primeiras partes fazem parte do pré-processamento dos dados. A análise exploratória a gente já fez, a gente ficou umas boas aulas fazendo a análise exploratória. A gente também teve uma pequena aula sobre o Futuring Engineering, só que a gente vai voltar nessa parte aqui para a gente melhorar os nossos dados, fazer alguns detalhezinhos. Enfim, nós conversamos bastante sobre o modelo de regressão também na aula passada, e eu dei um spoiler de que vamos fazer um modelo para predizer o salário das pessoas na área de dados. Mas não todas as pessoas. Para facilitar, vamos criar um modelo para predizer apenas as pessoas com emprego do tipo CLT. Isso porque existem muitas características diferentes entre pessoas do tipo freelancer ou PJ em relação à CLT que poderiam atrapalhar o resultado do nosso modelo. E queremos fazer um modelo inicial mais simples, né?

00:00:00:59 - 00:00:02:04
Nós disponibilizamos nos materiais dessa aula um arquivo que é muito parecido com o que a gente já salvou em outras aulas e tudo mais, mas eu peço para que vocês utilizem ele porque fizemos algumas alterações na coluna de salário para fazer mais sentido na nossa análise. Lembrando que essa coluna de salário não é original da pesquisa do Data Hackers, nós que criamos ela baseada na faixa salarial, tá? Não foi uma coisa completamente aleatória, não. Então, continuando, vamos lá. Ao invés de continuar naquele arquivo análise de dados e tal, eu vou agora aqui clicar com o botão direito na minha pasta da PrograMaria toda organizadinha, vou em + e vou em Google Collaboratory. Eu vou criar um arquivozinho aqui novo, zerado. Já vou colocar o nome aqui de regressão linear. Perfeito. E aí, a gente vai fazer aquela coisinha de conectar lá com o nosso drive para a gente poder conseguir acessar o nosso arquivo.

00:00:02:04 - 00:00:03:33
Então vamos lá. Aqui, conectar o Google Drive. Bom, agora já carregou aqui o nosso drive. Vou começar a expandir aqui as pastas. My Drive, agora eu vou lá na minha pastinha da PrograMaria e aí eu procuro o meu arquivo, que vai estar disponibilizado para vocês, é esse análise dados mod7, que é do módulo 7, vou nos três pontinhos, copiar caminho, ok. Só que eu tenho que fazer o import do pandas primeiro, import pandas spd, executo. Na linha de baixo, coloco dados igual a pd ponto read excel, porque agora a gente está lendo um Excel, e aqui dentro eu coloco o arquivo. Vou fechar aqui já para poder ter mais espaço para a gente poder ver o código, e vou executar. Na linha de baixo aqui, vamos dar uma olhada nesse código, nesse dado: dados ponto head. Ok, certinho, bonitinho. A primeira coisa que vamos fazer é filtrar apenas os dados das pessoas que são empregadas do tipo CLT. Os valores de pessoas em outras categorias de trabalho variam muito, então a gente vai fazer um modelo só para as pessoas do tipo CLT. Então vamos dar uma olhada nas colunas de uma forma geral para a gente poder pegar qual que é o nome da coluna que a gente quer filtrar.

00:00:03:33 - 00:00:05:04
Qual a situação atual de trabalho? É essa daqui. Então vamos colocar: dados, colchetes e o nome dessa tabela e vamos dar um value counts só para a gente poder ver quais são os tipos que a gente tem. Então, a gente tem o CLT, que é o que a gente vai filtrar aqui, a gente tem estagiário, servidor público, somente estudante, mas a gente quer só o empregado CLT. Então, a gente vai filtrar, vamos colocar dados igual a dados. Eu já estou atribuindo valor, mas o que vai ser atribuído a esse valor vai ser o nosso filtro, tá? Então, agora eu vou fazer o filtro, que é a planilha dados da coluna de qual situação atual de trabalho, aqui, que seja igual a empregado CLT. Opa. Então, o que eu estou fazendo aqui? Estou filtrando tudo que seja empregado CLT. Já vou executar. E se a gente dar um value counts de novo nessa coluna aqui embaixo, tem que aparecer somente empregado CLT. Perfeito. Agora vamos dar uma olhada na coluna de cor, raça e etnia. Então, vamos colocar dados, cochetes. Deixa eu olhar qual que é o nome da coluna. Cor, raça e etnia.

00:00:05:04 - 00:00:06:41
Vamos colar aqui embaixo. Vamos dar um ponto Value Counts. Aqui. Bom, o que a gente vai fazer nessa coluna é tirar essas categorias que têm pouquíssimos dados, como, por exemplo, indígena, outra e prefiro não informar. Porque se a gente for lá no nosso relatório, naquele gráfico que a gente plotou de média salarial por etnia, a gente viu que indígena e essa categoria de outra ficaram com uma média salarial super alta. Só que, querendo ou não, são pouquíssimos dados que não representam de uma forma geral todas as outras pessoas. A gente não pode pegar que todas as pessoas indígenas que trabalham na área de dados ganham exatamente como essas quatro pessoas, sabe? A amostra é muito pequena. Então vamos retirar esses com 16, 10 e 4 porque são números muito pequenos e podem acabar atrapalhando no resultado do nosso modelo, ok? Então vamos lá. Bom, para poder retirar esses valores, a gente pode criar uma lista com o nome deles. Então, vamos colocar, tipo: lista retirar, porque eu sou muito criativa. Para criar uma lista, a gente coloca colchetes e aí a gente coloca os itens dentro desses colchetes. Então, o primeiro item é o prefiro não informar vírgula para separar os itens de uma lista. O segundo item é outra. Aqui. E o terceiro item é indígena.

00:00:06:41 - 00:00:07:42
Ok. Então a gente executa aqui porque a gente já tem a lista do que a gente quer retirar. E aí a gente vai fazer naquele mesmo esquema, como se fosse um filtro, a gente vai colocar dados, que é a nossa tabela. Dentro de dados, a gente vai colocar dados e qual que é a coluna que a gente está querendo filtrar. Só que aí, um negócio legal é que a gente não vai usar o igual ou o diferente, a gente vai usar uma função, que é a função is in, que basicamente é se está em, ou seja, se está na nossa lista que a gente está querendo retirar. Então, ponto is in e a gente coloca lista, retirar. Só que esse is in do jeito que está aqui vai manter o que está na lista. Então a gente quer retirar, a gente quer o contrário do is in. Como a gente faz isso? Usando o til. Esse til antes do nosso filtro aqui, ele faz o contrário do que está a partir dele.

00:00:07:42 - 00:00:09:12
Então, por exemplo, se eu quisesse colocar que é igual, se eu quisesse tirar tudo que fosse igual a branco, eu colocaria cor, raça e etnia igual igual a branco, que filtraria tudo que é branco, mas eu colocaria o til antes. Então, ele tiraria o branco. Beleza? Ele faz o contrário do que está ali, depois do til. Então, a gente executa isso, né? Vamos já atribuir ao dados. Opa. Ok e vamos executar. Ok, agora vamos criar uma coluna chamada cor não branca, onde se a coluna de cor, raça, etnia for branca, essa coluna recebe zero. E caso contrário, recebe um. A gente pode fazer isso usando o apply que a gente já usou em outras aulas. Bom, então vamos lá: dados não branca, que é a coluna que a gente quer criar, não branca igual a dados, a coluna de cor, raça etnia, aqui dentro, bonitinho, ponto apply. Opa, não foi ponto, foi vírgula, mas é ponto: ponto apply, dentro dos parênteses, primeiro lambda, que é a função, lambda x igual. Então, o que a gente quer? É não branca, né? Então se for branca é zero, se for diferente branca é um. Então, um se, que é o if, x diferente de branca else zero, ok?

00:00:09:12 - 00:00:10:40
Então, vamos lá. Pensando aqui nessa coluna. Cada item, que é o nosso x, se esse x for diferente de branco, ou seja, se for amarelo, preto, le vai receber um. Se for igual a branco, ele recebe zero. Ok? Bom, vamos executar aqui. Um outro atributo que vamos utilizar é o tempo de experiência na área de dados. Fazendo um value counts na coluna de tempo de experiência, a gente pode dar uma olhada aqui no que a gente pode fazer, tá? Eu vou dar uma olhada nas colunas de novo porque eu não lembro qual que é o nome da coluna. Quanto tempo de experiência na área de dados você tem? Aqui mesmo, por cima, para não ficar aquele tanto de linha de código com dados ponto columns, eu já vou por cima esmo. Aqui, dados ponto value counts para a gente poder ver como está a distribuição disso. A gente tem aí de 1 a 2 anos, de 4 a 6 anos. Vamos usar um código para pegar o primeiro valor numérico de cada categoria. Por exemplo, na opção de 1 a 2 anos, a gente pega o dígito 1. A gente pode fazer isso usando uma função chamada extract, que é um método que usa expressões regulares, que são as regex, para extrair padrões específicos de strings. A gente pode deixar aí um material complementar a respeito de regex e tudo mais, mas confia que vai dar certo.

00:00:10:40 - 00:00:12:16
Bom, a gente vai pegar aqui essa coluna, a gente coloca um ponto str ponto extract. Nós colocamos o ponto str para acessar apenas a string e depois a função. Dentro da função do extract, nós precisamos passar o que estamos querendo extrair. Bom, vou colocar aqui o que vai dentro do extract e depois eu falo para vocês, mas é uma expressão, uma regex. Bom, primeiro um r, uma aspa simples, a gente abre parênteses, uma barra invertida, um d, um mais, e aí fecha o parênteses e fecha a aspa simples. O R, antes das aspas, indica uma string crua, uma string na raiz, onde caracteres especiais são tratados literalmente. Isso é útil para evitar confusão com caracteres de escape em expressões regulares. O barra invertido D, este é um metacaractere, que representa qualquer dígito de 0 a 9. Esse maizinho depois desse D é um quantificador que indica um ou mais do que o precede. Se é dez anos, ele não vai pegar só o um, ele vai pegar o um e o zero. Então, no final, essa barra invertida e o D mais significa um ou mais dígitos.

Os parênteses são usados para capturar o que está dentro deles, o que significa que a gente está capturando esses um ou mais dígitos mesmo. E aí a gente vai colocar, vai atribuir isso a uma nova coluna que a gente pode chamar de tempo de experiência.

00:00:12:16 - 00:00:13:44
Então a gente vem aqui, antes disso tudo, e coloca dados, tempo experiência. Opa, experiência. Fecha colchetes. Bom, a gente escreveu tudo, vamos dar uma revisada aqui no que a gente fez. Ok. A gente está criando uma coluna chamada tempo experiência. O que está indo nessa coluna? Está indo o quê? Da coluna de quanto tempo de experiência, quanto tempo de experiência na área de dados tem, a gente está pegando linha por linha com esse extract pegando a string, o texto, e a gente está usando a função extract, que vai extrair. O que essa função vai extrair? Aí dentro do parênteses a gente está colocando o R do radical dessa string, E aí a gente está usando uma regex, que é o quê? Essa regex tem vários tipos, mas aqui no caso, o que a gente está querendo mesmo? A gente está querendo o número, né? A gente está querendo pegar esse número aqui, o primeiro daqui, o primeiro daqui. Nessa daqui, mais de dez anos, a gente não quer só um, a gente quer um e o zero, a gente quer o dez. Então, o que a gente está colocando aqui? A gente está usando esse barra D mais para dizer que a gente está querendo pegar o primeiro dígito. A gente está querendo pegar o primeiro dígito, mas o mais quer dizer que se depois desse dígito tiver outro, a gente vai pegar esse segundo também. E se depois desse segundo tiver outro dígito, a gente vai pegar também. Se depois tiver uma letra, a gente parou, a gente pegou três dígitos.

00:00:13:44 - 00:00:15:12
Então aqui, leu nessa primeira, chegou no um, beleza, pegou. Tem outro dígito aqui logo em seguida? Não tem, ok. Aqui no 3, no 4, no 10. Chegou no primeiro, pegou o primeiro. Depois do 1, tem o 0. É dígito? É. Então a gente pega. Depois não tem mais. Ok. A gente pega o 10. Então é isso que a gente está fazendo com essa regex. Ou seja, a gente está criando esse tempo de experiência pegando o primeiro dígito de cada frase. Ok? Bom, vamos rodar aqui pra ver se vai dar certo. Putz, deu certo, que alívio. A pessoa programadora fica muito feliz quando não dá erro, tá gente? Bom, vamos dar uma olhada como que ficou aqui então, ó. Aqui numa linha de código abaixo, vamos executar. Ó, agora tem só números. A gente pode também fazer um value counts aqui. E aí a gente tem alguns números. Agora a gente não tem as frases. Perfeito. Vamos dar uma olhada na coluna de números de funcionários. Vamos dar aqui, ó. Eu não lembro mais o nome da coluna. Dados ponto columns números de funcionários. Aqui, ó. Números de funcionários. Exatamente assim. Vou excluir para não ficar essa poluição aqui. Colocar ponto values. Bom, ok. Aqui a gente tem o mesmo caso.

00:00:15:12 - 00:00:16:43
É o mesmo caso, só que é diferente. O que acontece? A gente tem aqui algumas opções que são assim, acima de 3.000, acima de 1.001. O que a gente tinha explicado lá naquela questão da regex? Ela vai olhar se tem dígito, aí encontrou o dígito, depois tem outro dígito, continua. Só que aqui no caso, por exemplo, de 3.000, tem um dígito, depois do dígito tem um ponto. Então ele pararia, ele não ia procurar os outros zeros. Então, na hora que a gente extraísse, ia encontrar o primeiro dígito e ia parar ali. Então, essa opção aqui, ao invés de extrair 3.000, ele apenas extrairia o 3.

E olha, assim, de 3.000 para 3 funcionários tem uma diferença muito grande. Então, a gente vai ter que fazer alguma coisa aí. A gente pode fazer isso de substituir esse ponto usando a função replace. Praticamente vai substituir esse ponto por nada, que no caso a gente quer substituir por nada. Então, a gente pega aqui essa coluna, coloca aqui. A gente vai atribuir o resultado nela mesmo, tá? Para não ter problema. Então, vai ser: essa coluna ponto str porque a gente tá pegando ali a string, a gente vai colocar o replace, e aí a gente vai colocar que a gente quer substituir esse ponto por nada, tipo, aspas simples, sem nada no meio, ok? E vamos rodar. Aí, eu vou fazer esse value counts aqui em cima mesmo só pra gente poder verificar se o ponto vai sumir, tá?

00:00:16:43 - 00:00:17:46
Vamos ver. Putz, sumiu. Felicidade, hein? Agora sim a gente pode fazer o extract, porque vai olhar o primeiro dígito, em seguida tem outro ele vai continuar, tem outro ele vai continuar, e assim vai. Como eu sou uma pessoa um pouco preguiçosa, eu vou copiar exatamente esse extract aqui de cima. Vou copiar, vou colar lá embaixo. e vou copiar aqui, só trocar o nome da coluna. Perfeito. Vou trocar aqui também. Vou salvar por cima do número de funcionários mesmo, não vou criar uma outra coluna para isso. Vamos rodar aqui. Vamos fazer um value counts, agora diferenciado mesmo aqui, numa outra coluna, porque vai mudar bastante. Agora sim, a gente tem os números de funcionários. Ok. Bom, então a gente fez o value counts aqui, agora a gente viu que a gente conseguiu extrair o número. Só que eu estou olhando o value counts ali, mas a gente tem que ver se vai ter algum nulo nisso, tanto nessa coluna quanto na que a gente criou.

00:00:17:46 - 00:00:19:01
Então, eu vou colocar aqui um dropna false para não pular os nulos. Bom, aqui não apareceu nenhum nulo, vou copiar essa mesma linha de código, vou colocar embaixo e vou copiar o tempo de experiência. A gente tem que ter certeza que não vai ter nulo nela também. Vamos ver. Bom, aqui a gente tem 127 nulos. O que a gente faz com esses nulos? Se a gente olhar, esses nulos são frutos dessa aqui, ó: não tenho experiência na área de dados. Se não tem experiência na área de dados, a gente pode colocar que o tempo de experiência é zero. Então, a gente pode usar uma função que é a fillna. Então, a gente pode pegar essa coluna, que a gente vai atribuir a ela mesma, né? Vai ser igual a essa coluna, ponto fillna, ou seja, preencher os nulos com zero, ok? Perfeito. Então, se a gente rodar esse value counts aqui de cima de novo, a gente vai ter os 127 que eram nulos, agora eles vão ter zero. Bom, é isso nessa aula. Na próxima aula a gente volta e continua a partir daqui mesmo nessas questões de Futuring Engineering aí para a gente acertar e treinar o nosso modelo. Beleza? Até mais.