6.3 - Visualização de dados em python (parte 1)

00:00:00:00 - 00:00:01:34
Olá, pessoal. Na aula de hoje nós vamos dar continuidade no assunto de visualização de dados. Mas, dessa vez, vamos adicionar o Python. Vamos ver algumas bibliotecas principais para a criação de gráficos em Python. Bora lá. Vamos abrir o nosso notebook de análise de dados. Estou aqui na minha pastinha. E aí, meu notebook está aqui. Eu vou abrir ele. Espera carregar. Bom, com o notebook aberto a gente vai lá para o final. Bom que a gente já vê aqui o tanto de coisa que a gente já fez. E vamos criar aqui mais uma célula de texto, criar um novo índice para a gente manter tudo organizado, né? O índice de visualização de dados. Porque aí tudo que a gente fizer a partir daqui fica para essa aula de visualização de dados. Ok. Bom, só para relembrar, em módulos passados nós exploramos os dados da pesquisa realizada pelo Data Hackers sobre profissionais da área de dados. Vamos fazer a leitura da nossa tabela de novo e relembrar o que temos de informação. Então tá. Bom, primeiro vamos aqui no nosso cantinho, clicar nessa pastinha aqui. Aqui, agora apareceu. Aqui para poder montar o drive, conectar o Google Drive, Vamos esperar para poder carregar aqui. Automaticamente já criou aqui essa célula de código que é informando que fez essa conexão com o nosso Google Drive.

00:00:01:34 - 00:00:03:02
Vamos atualizar aqui para ver se aparece? Beleza. Eu cliquei aqui nessa setinha aqui, dando uma voltinha para atualizar, tá gente? Bom, aí apareceu aqui o drive, vou clicar nessa setinha para expandir, vou clicar no My Drive, que é o meu drive, vai aparecer todas as minhas pastas, aí vocês procuram aí qual é a pasta de vocês, eu já criei uma pastinha aqui com PrograMaria, para poder salvar todos os meus arquivos lá. A gente salvou esse análise dados ponto csv, aí eu clico nesses três pontinhos, vou em "copiar caminho". Perfeito, eu já tenho aí, então, o caminho do meu arquivo. Vou criar uma célula de código aqui abaixo, e aí para a gente poder fazer a leitura desse arquivo que eu copiei o caminho dele, primeiro preciso fazer o import da biblioteca que vai fazer isso, né? Que é o quê? É o pandas. Então, a gente vai colocar lá: import pandas spd, que é o apelidinho do Pandas, vamos executar, e aí na linha de baixo, de código, a gente pode fazer a leitura. Vamos chamar de dados mesmo, né, dados igual a pd, que é o Pandas, ponto read, underline csv. Aquela formulazinha que a gente sempre faz para poder fazer a leitura. Agora é read CSV por quê? Porque o arquivo é um CSV, não mais um Excel. Aí a gente coloca aqui aspas simples ou duplas, e dentro delas a gente coloca o que a gente copiou lá do caminho, ok?

00:00:03:02 - 00:00:04:38
Então beleza, a gente executa essa célula. A gente já pode fechar esse cantinho aqui dessa aba esquerda. só para não ocupar muito espaço. Aqui, fez a leitura, vamos colocar mais uma célula de código aqui e vamos colocar dados ponto read. Beleza. Nossa tabela de dados inclui várias colunas como idade, gênero, nível de experiência, escolaridade, faixa salarial, entre outras. A visualização de dados é uma ferramenta poderosa para responder perguntas e revelar padrões. Então, nós já trabalhamos nossos dados, respondemos algumas perguntas, mas pensem que agora a gente precisa responder as perguntas de forma visual, utilizando gráficos. Por exemplo, aqui na nossa tabela no dados ponto read.

Nessa coluna de gênero, será que nessa pesquisa tem a maior quantidade de pessoas que responderam que são do gênero masculino ou do gênero feminino? Um gráfico poderia facilmente responder pra gente, né? Vamos aprender a criar um gráfico de barra pra gente visualizar essa diferença aí dos gêneros. Para saber a quantidade de pessoas por gênero, a gente pode utilizar aquele comando do value counts. Então, aqui em uma célula de código, a gente dá: dados, a gente tem que informar qual que é a coluna de dados, então a gente coloca o colchete, coloca gênero. O bom que já está aqui informando para a gente quais são as colunas que começam com G. Se não tivesse essa opçãozinha aqui, se não aparecesse isso, o que a gente pode fazer?

00:00:04:38 - 00:00:06:14
Ter uma célula de código, colocar dados ponto columns. E aí vai aparecer todas as colunas, a gente pega qual que a gente precisa, porque o nome da coluna tem que ser exatamente igual, tá, pessoal? Então eu copiei ali, coloco aqui. Então aqui eu estou falando qual que é a coluna. Coloco: ponto value counts e executo. E aí a gente tem as quantidades por gênero. Inicialmente, a gente pode usar a biblioteca Matplotlib para visualização de dados. Ela é uma das bibliotecas mais populares para visualização de dados em Python. Essa biblioteca oferece uma grande variedade de gráficos, inclusive vários dos que vimos na aula passada, como gráficos de linha, de barras, de dispersão, entre outros. É uma biblioteca utilizada em diversas áreas, desde análise de dados até ciência e engenharia. A gente faz a importação dessa biblioteca assim: import matplotlib ponto pyplot e a gente dá um apelidinho para não precisar escrever matplotlib e tal, a gente coloca o "as plt". Aí toda vez que a gente for usar ela, a gente só escreve plt. Bom, para criar um gráfico em matplotlib, primeiro nós vamos criar uma figura, ou uma janela para o gráfico. Então, a gente coloca assim: plt ponto figure e os parênteses, ok? Para criar um gráfico de barras, a gente usa o seguinte comando, que é o plt ponto bar, ok?

00:00:06:14 - 00:00:07:48
Então, primeiro a gente criou uma janela, e aqui embaixo, no plt ponto bar, a gente está falando: olha, a gente vai criar um gráfico de barras. Dentro da função bar, nós temos alguns parâmetros. O primeiro deles é a contagem dos valores. No nosso caso, vai ser essa variável ali do value counts, ou seja, esses valores aqui de value counts. Então, vamos salvar isso aqui de value counts dentro de uma variável. Vamos chamar de gênero counts: gênero underline counts igual. Então, a gente está atribuindo o resultado do value counts para o gênero counts, ok? Vamos executar aqui. Então, perfeito. Esse primeiro parâmetro da função de barras vai ser esse gênero counts, que é essa contagem. Depois, nós vamos colocar os rótulos para cada fatia do gráfico. Isso vai ajudar a identificar o que é cada quantidade. Aqui nós vamos usar o parâmetro labels. Enviamos para labels os nomes de cada item da coluna gênero. A gente faz isso colocando a variável de gênero ponto index. Esse ponto index é pegando cada nome de cada categoria. Então eu coloco aqui dentro vírgula, ou seja, para separar os parâmetros, né? Primeiro parâmetro é o value counts mesmo, completo. Depois, o segundo parâmetro é o label. Label igual a gênero counts - vou ter que copiar tudo, digitar tudo - ponto index.

00:00:07:48 - 00:00:09:17
Bom, então pronto, a gente tem esses parâmetros. Para ficar bonitinho, a gente pode colocar um título no gráfico. A gente aprendeu ali nas outras aulas algumas dicas para os gráficos ficarem bacanas. E para isso a gente faz como? plt ponto title (título em inglês). E aí dentro do parênteses a gente coloca o nosso título entre aspas simples ou duplas. Vou colocar aqui simples, e vamos chamar esse gráfico de distribuição de gênero, já tá aqui. Distribuição de gênero. Ok. De gêneros mesmo, no plural. Bom, e vamos colocar os nomes dos eixos também para a gente se orientar o que é o eixo X e o que é o eixo Y. A gente faz isso, plt, opa, dá um enter, plt ponto xlabel, xlabel é para o label do eixo X, então a gente pode colocar no xlabel o gênero entre aspas simples ou duplas, gênero, e plt ponto ylabel é o label do eixo y, que a gente pode colocar aqui a contagem, a quantidade. Perfeito. Agora a gente precisa exibir esse gráfico. E para exibir ele na tela, a gente coloca um plt ponto show, ou seja, um plt ponto mostra.

00:00:09:17 - 00:00:10:48
Plt ponto show e os parênteses vazio aí, beleza? E vamos executar isso para a gente ver o nosso gráfico. Bom, pessoal, deu erro. E aí ele está falando que tem um erro aqui de argumento que é o height, de altura. E se a gente pensar em gráfico de barras, a barra vai ter uma certa altura. Então realmente a gente tem que colocar lá o parâmetro. Na verdade, a gente já colocou esse parâmetro. A gente só precisa apontar: olha, o parâmetro é esse. Então é esse primeiro aqui. E é bom que eu olhei porque eu percebi que eu mandei o resultado do value counts direto. Deixa eu colocar uma linha de código aqui. Eu mandei isso aqui, isso aqui é uma tabela, não faz sentido, eu quero os valores apenas, eu quero os números. Então, eu vou colocar ponto values para poder falar: olha, é os números que a gente está falando, a gente está falando que as alturas são os números. Então, uma barra vai ter a altura de 3.140, a outra de 1.000 e tanto e tal, beleza? E aproveitando também que eu olhei porque eu falei label, label, mas o nome do parâmetro, apesar de ser a label mesmo, o nomezinho das categorias, o nome do parâmetro na função de barra é x. Então beleza agora, a gente está falando que o primeiro parâmetro é a altura das barras, que é o valor do value counts e o segundo parâmetro é o nosso x, que é o nosso label, ok? Vamos executar de novo. Agora sim. Ficou bonitinho. Ficou grandão, mas ficou bonitinho.

00:00:10:48 - 00:00:12:22
Então, através do gráfico, já podemos responder nossa pergunta, né? Temos mais pessoas que responderam que são do gênero masculino do que do gênero feminino. Aqui a gente bate o olho e já vê que tem mais pessoas do gênero masculino. Mas voltando lá para as nossas boas práticas, a gente não consegue saber mais masculinos do que o quê, assim. Então, o que falta aqui para a nossa boa visualização? Um título que fique melhor, que clarifique a nossa pergunta. A gente vai voltar aqui e vamos colocar um título melhor. Que tal, assim: quantidade de pessoas por gênero? Eu até escrevi errado, né? Ou não, tá certo. Então, vamos colocar agora quantidade de pessoas por gênero. Vai ficar melhor. Quantidade de pessoas por gênero na área de dados. Vai ficar melhor: na área de dados. Show. Utilizando as escalas dos eixos... Deixa eu executar para poder atualizar, ficar bonitinho. Aí agora sim. Utilizando as escalas dos eixos, podemos até dizer que tem umas três vezes mais pessoas do gênero masculino do que feminino. E temos ali uma pequeníssima quantidade de pessoas que assinalaram a opção de prefiro não informar.

00:00:12:22 - 00:00:13:52
Uma outra forma de fazermos esse mesmo gráfico é utilizando a biblioteca Seaborn. O Seaborn é baseado na biblioteca Matplotlib, mas oferece uma interface mais simples e gráficos mais esteticamente agradáveis. Além de que simplifica o código, já que a gente não precisa fazer o value counts, salvar em uma variável e depois colocar no plot. A própria biblioteca já faz isso para a gente. Outra vantagem é fornecer alguns parâmetros de estilo, como por exemplo colocar linhas de grade no fundo do gráfico para facilitar a visualização. Primeiro, nós vamos fazer a importação dessa biblioteca. A gente faz o import dela assim: ó: import seaborn, e aí vamos dar um apelidinho também: sns. Ok, vamos executar. Aqui também a gente começa criando uma figura, uma janela, com o matplotlib. Então, plt ponto figure, ok. E para criar o gráfico de barras no Seaborn, usamos a função countplot, que é usada para mostrar a contagem de observações de cada categoria, que é tipo o value counts. No nosso caso, a contagem de cada categoria da coluna de gênero. Então a gente coloca sns ponto countplot. Ok. Dentro dos parênteses da função, nós vamos enviar alguns parâmetros. O primeiro é o data, né? São os dados que a gente quer visualizar. No nosso caso, data igual a dados. É a nossa tabela como um todo.

00:00:13:52 - 00:00:15:28
O segundo parâmetro é o x, que especifica qual coluna da tabela que vamos usar para fazer a contagem. Beleza? Então, vamos lá: x. Qual que é a coluna? Coluna de gênero. Ok. E por último, vamos enviar um parâmetro opcional, que se chama palete. Paleta, né? Uma paleta de cores. Eu vou usar uma paleta chamada pastel, que traz cores de tons pastéis. Mas lá no site oficial da Biblioteca do Seaborn tem todas as paletas possíveis. Inclusive, eu adoro ficar olhando lá e mudando as cores, não sei o quê e tal. Então vamos lá. Vírgula palete, com dois T's. E eu vou usar pastel. Ok? E para ficar bonito, lá vamos nós colocar os rótulos e títulos do gráfico. E aí a gente usa a mesma coisa com o PLT, com o matplotlib. Eu vou dar uma roubadinha e vou copiar aqui de cima. O título, o xlabel e o ylabel, ok? Vou colar aqui embaixo. Para visualizar isso tudo, a gente faz o quê? A gente coloca o plt ponto show e os nossos parênteses vazio, beleza? Então, vamos visualizar. Ó, já ficou bonito, gostaram dos tons pastéis? Uma coisa que a gente pode fazer é colocar as linhas de grade no fundo. Como? A gente coloca aqui, antes do show, né, porque eu quero mostrar essas linhas de grade, coloco plt ponto grid true, ou seja, para mostrar, né?

00:00:15:28 - 00:00:16:18
Então, a gente atualiza de novo e olha, apareceu as linhas. Particularmente, eu achei que nesse gráfico não precisava, né? Porque eu gosto de ver meus tons pastéis limpos, assim, sem linhas. Mas tem gráficos que às vezes a gente tem mais barras, ou está tão próximo um do outro, que essas linhas vão ajudar a gente a ver aqui se um gráfico tá acima, está abaixo e tal. O legal desse gráfico é que cada barra tem uma cor diferente, né? A gente faz isso usando essas paletas. Explorem, gente, tem umas paletas muito legais. Lá no Matplotlib também tem como inserir um parâmetro que deixa cada coluna de cores diferentes. Fica a dica também de vocês procurarem no site oficial do Matplotlib as diferentes formas e opções de criar gráficos de barras. Não só com o Matplotlib, explorem aí o Seaborn também. Beleza?