6.3 - Visualização de dados em python (parte 2)

00:00:00:00 - 00:00:01:30
Bom, uma outra questão que podemos nos perguntar é se o salário varia de acordo com a idade. Podemos fazer a média do salário por idade e plotar um gráfico para visualizarmos se há diferenças. Vamos fazer esse gráfico inicialmente com o Matplotlib. Primeiro vamos criar uma variável que vai receber a média de salário por idade. Então vamos lá, vamos colocar o nome dessa variável de "salário por idade" porque eu acho que é um negócio bem criativo, diferenciado e tal. Salário por idade. Beleza? E aí a gente pode fazer um groupby. Então, dados ponto groupby. E aí a gente está querendo a média da coluna de idade, então a gente vai colocar: idade... Opa, é tudo maiúsculo. E depois o que a gente quer a média em si? É o salário. Então entre colchetes a gente vai colocar salário. E por fim, a função que a gente está querendo, que é a média min. Ok? Beleza? Vamos executar aqui. Colocar uma linha por baixo. A gente pode até dar uma visualizada nesse salário por idade. Ficou bacana, né? Aí tem a idade e na outra coluna qual que é a média de salário por essa idade. Fechou? Beleza. Bom, ótimo. Agora vamos iniciar nosso gráfico criando o quê? Uma janela: plt ponto figure. Ok. Depois nós vamos criar o gráfico.

00:00:01:30 - 00:00:02:59
Vou colocar aqui e aí eu vou explicando para vocês cada item, tá? plt ponto plot. Aí dentro eu vou colocar o salário por idade. Deixa eu copiar para ficar mais fácil. Salário por idade. Quem mandou, né, colocar um nome tão grande de variável: salário por idade ponto index, porque eu quero os valores, senão vai as colunas. A gente já teve erro por isso. Depois a gente coloca o values, que a gente já também viu quando a gente foi criar o de barras, copiar, colar, tirar o index, colocar o values. E depois eu vou colocar um marcador. Eu vou explicar direitinho para vocês. Um marcador. E por último, eu vou colocar um Line Style, que é o estilo da minha linha. Beleza, é assim, vamos lá explicar passo a passo. O primeiro parâmetro contém os rótulos que serão usados no eixo X, né? Que é o meu index, que são os rótulos. Ou seja, todas as possíveis categorias de idade. 18 anos, 19 anos e tal. O segundo parâmetro são os valores de fato que vão preencher o nosso eixo Y. Neste caso, são as médias dos salários para cada idade. Estamos usando o Marker, que é um marcador, que define o estilo dos marcadores nos pontos de dados. Cada ponto será marcado com um círculo. Então, sei lá, um marcador é 18 anos, então lá nos 18 anos vai ter um círculo. Existem vários tipos de marcadores, mas eu optei pelo O.

00:00:02:59 - 00:00:05:00
Então, só a minha opinião importa. Vocês podem consultar o site do Matplot e pesquisar por Marker e testar outros estilos. E por último, aqui a gente tem o Line Style, que é basicamente o estilo da linha. Estamos definindo ali uma linha contínua. Também existem outros estilos e vocês podem pesquisar por eles, ok? Aí a gente vai colocar o quê? Os rótulos, os eixos, né? Então, plt ponto xlabel, que no caso é a nossa idade. Idade. plt ponto ylabel, que no caso é a média do salário, ok? Então, vou colocar média de salário. Ok, e vamos por fim colocar um plt ponto show para visualizarmos o resultado, ok? Ah, está faltando uma coisa importante, né? É bom que já deu aqui uma dica, que é o título, né? Imagina se eu rodasse aqui sem título? Vamos colocar um título. Bora lá: plt ponto title, e média de salário por idade.

E agora sim a gente pode visualizar esse gráfico: plt ponto show. Vamos executar. Aqui. Bom, uma outra coisa é criando algumas linhas de grade no fundo do gráfico também para ajudar a entender que, como tem alguns pontos muito próximos, às vezes uma linha de grade vai ajudar a gente a diferenciar. A gente já sabe que é um plt ponto grid true. Olhem bem assim e vamos colocar grade para ver se ajuda aqui. Eu acho que nesse gráfico vai ajudar, viu? Ó, eu achei que deu uma ajudada, né?

00:00:05:00 - 00:00:06:35
Então assim, por exemplo, aqui nos 20 anos está aqui embaixo, mas quando tem essas linhas aqui, a gente consegue ver claramente que os outros pontos estão bem acima e tal. Achei bacana. Antes de analisar o resultado, vamos plotar esse mesmo gráfico, mas usando a biblioteca Plotly, que é uma biblioteca poderosa para a criação de gráficos interativos e dinâmicos. Além disso, oferece suporte para a criação de visualizações 3D e mapas geoespaciais, sendo uma escolha popular para projetos que exigem visualizações mais avançadas e interativas. A gente faz o import dessa biblioteca assim, ó.: import plotly. E aí a gente vai dar um apelido para ela, mas é plotly ponto express as px. Show? Vamos lá. Então, a gente faz o quê? A gente vai criar uma variável que vai receber o gráfico. Essa variável a gente pode chamar de fig. Essa biblioteca já é diferenciada, né, gente? É uma variável que vai receber esse gráfico. O gráfico que a gente vai criar é o px da biblioteca line, porque é um gráfico de linha, né? Aqui, ó, uma linha. Então, a gente vai colocar assim.... Chamamos a função Line porque a gente está criando esse gráfico de linha e dentro dos parênteses a gente vai colocar alguns parâmetros. O primeiro é o salário por idade.

00:00:06:35 - 00:00:08:11
Salário por idade ponto reset index, no caso, que também são os nossos indexes, só diferenciado porque é o Plotly. Aí, o segundo parâmetro é qual que é o eixo X, qual que é o nome do eixo X, no nosso caso é idade, depois a gente coloca qual que é o nome do eixo Y, que no caso é salário, qual que é a variável do eixo Y. A gente já vai colocar o título, que é a variável title. Aí eu não vou digitar tudo não, hein? Vou copiar aqui, hein? Ó, copiei em cima, colei aqui na frente de title. E a gente vai colocar os markers, os marcadores igual a true. Então, markers... Ops, aqui tá atrapalhando. Markers igual a true. Ok? Vocês viram que já foi diferenciado aqui? Porque aqui a gente já colocou o nome dos rótulos, o nome do eixo x, y, colocou o título, tudo dentro da função do line do Plotly, né? Então, vamos... Opa. Lá no final vou colocar uma linha abaixo e para visualizar vou colocar o plt ponto show. Ok? Vou executar. Bom, pessoal, rodei aqui, deu erro, vamos dar uma olhada aqui. O valor x não é o nome da coluna lá na nossa tabela. Eita, o que rolou?

00:00:08:11 - 00:00:09:38
Aqui no x é o nome do eixo X e do Y, só que precisa ser exatamente igual o nome da coluna na nossa tabela. E eu coloquei idade aqui, ó, com algumas letras minúsculas. Na nossa tabela, idade é tudo maiúsculo. Então, eu vou colocar exatamente igual ao da nossa tabela. Salário também, exatamente igual ao da nossa tabela, ok? Vou rodar de novo. Não apareceu nada. Por que não apareceu nada? Aqui eu coloquei plt ponto show, mas eu não quero que mostre o plt, a gente não está usando matplotlib mais. A gente está usando o plotly. E aqui a gente colocou essa variável fig para receber o gráfico. Então agora eu quero exibir o fig.

Então no lugar do plt a gente vai colocar fig. Beleza? Agora a gente vai executar. E olha o nosso gráfico aí. Gente, eu acho o Plotly muito diferenciado, tá? Por quê? Ele é muito interativo, ele é muito bonito. Ó, comigo conquistou com pouca coisa, viu? Esse gráfico ele é bem legal porque podemos perceber que o gráfico cresce com a idade até certo ponto, né? Podemos ver que até os 40 anos o salário ele realmente vai aumentando aqui, ó. Se a gente olhar, vai aumentando. Teve uma queda depois dos 40 e depois dos 45 e depois dos 50 anos. Mas claro que não podemos dizer que o salário varia apenas com a idade. Outros fatores também influenciam. Porque à medida que a idade aumenta, a tendência é que aumente a experiência também, né?

00:00:09:38 - 00:00:11:13
O legal do gráfico com o Plotly é que, por ser interativo, a gente pode passar o mouse nos pontos do gráfico e ver o valor exato da média salarial. Eu acho isso fantástico. A gente passa aqui, com 40 anos de idade, a média salarial é 15 mil. Vamos fazer um gráfico do tipo scatter plot, que é um gráfico do tipo de dispersão utilizando esses mesmos dados, até para a gente conseguir comparar qual o tipo de gráfico deixa a interpretação melhor. Primeiro com o matplotlib. Então vamos lá. Bom, e já podemos ser mais rápidos, porém para fazermos um gráfico do tipo scatter plot, a função agora é plt ponto scatter. Então, vamos lá: plt ponto figure, que é para a nossa janela, e agora não é plt ponto bar, é plt ponto scatter plot. É plt ponto scatter, não plot. E o resto para criar os rótulos é como a gente já viu mesmo. Na função scatter plot, nós enviamos um parâmetro chamado alfa, além dos parâmetros que a gente já viu, que define a transparência dos pontos no gráfico de dispersão. Bom, vamos colocar aqui os nossos parâmetros. Então, o primeiro é a nossa coluna de idade. Dados idade. O segundo é a nossa coluna de salário: salário. Por que isso? Primeira idade que é o nosso eixo x, depois o salário que é o nosso eixo y. E como eu estava falando, a gente tem um parâmetro alfa. Esse parâmetro alfa vai definir a transparência dos pontinhos no gráfico de dispersão.

00:00:11:13 - 00:00:13:21
Quanto mais próximo de zero, mais transparente, mais invisível os pontos vão ficar. Quanto mais próximo de 1, mais opacos, mais completamente preenchidos pela cor os pontinhos vão ficar. Vamos usar o 0.5, vai estar ali no meio, mas façam esse teste aí, depois coloquem 1, coloquem 0.1 para ver como é que vai ficar. Então, vamos colocar aqui 0.5. Beleza? E aí, aquela coisinha de sempre, que é o label para o... Label X, label Y, vamos colocando. plt ponto... Opa. plt ponto xlabel. E aí, no eixo X a gente vai ter a idade, plt ponto ylabel. Vai ser o salário. A gente vai colocar um título, né, pra ficar bacaninha. Como é que a gente vai chamar isso aqui. Que tal Dispersão de idade por salário. Achei bacana, achei chique. Eu acho a palavra dispersão muito chique. Dispersão de salário por idade. Ok? Vamos colocar o grid, para poder ver como é que vai ficar. E aí vamos visualizar o gráfico, que é o plt ponto show. Vamos executar. Uma coisa legal é que se a gente quiser aumentar o tamanho do gráfico, o meu aqui já está estourado a tela, mas se a gente quiser aumentar, a gente aumenta o tamanho da janela que a gente cria aqui no início. E a gente pode fazer isso colocando um parâmetro dentro desses parênteses que se chama figsize. Vou colocar aqui: figsize. E aí a gente coloca, né, figsize igual parênteses e aqui dentro desse parênteses a gente coloca dois valores. Esses valores são em polegadas.

00:00:13:21 - 00:00:14:51
O primeiro é referente ao eixo X, a largura, e o segundo é a altura, beleza? Eu vou colocar... Sei lá, está vindo na minha mente que é muito forte 5.5 porque meu gráfico já está muito grande. Eu vou executar, mas se aparecer um negócio pequenininho... Ah, achei que ficou melhor, hein? Ó, ficou horroroso, hein, gente? Ficou horroroso. Vou aumentar. Vamos lá. Ficou da hora, ficou da hora agora. Ficou mais larguinho, separou mais as idades e tal, acho que ficou melhor. Bom, uma coisa também interessante é que eu estou colocando o plt ponto show aqui porque é com Python que a gente está trabalhando e tal, mas como a gente está aqui no Jupyter, um ambiente do collab e tal, não necessariamente precisa do ponto show. Se eu colocar uma hashtag aqui que é para comentar as linhas, comentei, ou seja, quando executar essa linha vai ser ignorada, mas o gráfico vai aparecer mesmo assim, porque já está executando essa linha, tudo que está nessa linha. Então, vai mostrar o gráfico. O plt ponto show é importante quando a gente estiver escrevendo, sei lá, um script, como a gente aprendeu lá no iniciozinho de algum módulo, bem no início do curso. Aí, a gente precisaria do plt ponto show. Beleza? Bom, vamos fazer esse mesmo gráfico usando o plot. Vamos ver se vai ficar chique ou não. Bora lá. Ao invés de lines, a gente vai usar o scatter.

00:00:14:51 - 00:00:16:18
Então, vamos colocar fig, ou seja, uma figura que vai receber o gráfico, px, que a gente está chamando o plot de px, scatter, a gente tinha feito aquele gráfico de linha lá, agora a gente está fazendo o scatter, e a gente vai colocar os parâmetros. Primeiro a tabela dados, aí a gente vai colocar o X e o Y, que é o nome das nossas colunas, idade, depois o Y, que é o salário, ok, Beleza? E aí a gente vai colocar um título bonitinho que eu não vou digitar, que eu vou vir aqui em cima e copiar, porque sou uma pessoa preguiçosa. Aqui. Fechou, né? A gente informou qual que é a nossa tabela, qual que é a nossa coluna para o eixo X, qual que é a nossa coluna para o eixo Y e um título bacana. O que falta para a gente poder visualizar esse gráfico é fazer o show, mas não é plt ponto show, a gente vai exibir o fig, fig ponto show. E vamos executar para ver como que vai ficar. Bom, vamos entender esse gráfico de dispersão. A primeira coisa é que quanto mais dispersos os dados estiverem, maior a variabilidade e quanto menos, mais homogêneos nossos dados são. Também precisamos lembrar que no eixo Y temos os salários e no eixo X temos a idade. Esse tipo de gráfico mostra a dispersão dos dados.

00:00:16:18 - 00:00:17:45
É bom para entendermos possíveis outliers, né? Então, muito dispersos, muito longe da concentração. Por exemplo, esses dados aqui mesmo que são soltos, longe da concentração da base, são possíveis outliers, pois eles se afastam do padrão. O padrão é onde tem a maior quantidade de bolinhas, ou seja, maior concentração de dados. Então, vamos lá. Temos uma concentração muito grande por volta aqui de pessoas que recebem até 20 mil. Está bem cheinho aqui. Uma coisa que a gente espera é que a medida que aumenta a idade, aumente o salário. Porém, nesse gráfico podemos dizer que não é isso que acontece. Já que muitas pessoas acima de 40 recebem até 20 mil. Então, a gente não esperava essa concentração aqui de idade muito alta, de idade acima de 45 com um salário de até 20K, a gente esperava que já estivesse aqui pra cima. Bom, a gente pode observar que a gente não tem tantos dados de pessoas acima de 45 anos, ou seja, o público que respondeu esse questionário é um público mais jovem.

No gráfico com o Plotly podemos até passar o mouse por cima de alguns pontos e obter informações como a idade e o salário. Mas gente, realmente o gráfico de linha ficou melhor do que o de dispersão para a gente entender o que está acontecendo. Fica aí a dica do que a gente já viu em aulas passadas, que é sempre melhor escolher um gráfico mais claro e simples.

00:00:17:45 - 00:00:18:44
Nesse exemplo, o de dispersão não ficou legal, o de linha ficou melhor, mas pode ter algum outro momento que o gráfico de dispersão vai conseguir transmitir melhor o que a gente está querendo dizer. Tudo depende dos dados, da história e do objetivo. Uma outra coisa que eu aconselho vocês é a irem no site oficial do Plotly e ver outras aplicações ou parâmetros para os gráficos que a gente usou aqui. Além disso, vocês podem plotar outros gráficos também, seja com outras colunas da tabela, seja refazendo os mesmos gráficos aqui, mas com outros tipos, de pizza e tal. E é isso, né? A visualização de dados é uma ferramenta essencial para entender melhor os dados e responder perguntas importantes. Ao começar com perguntas claras, podemos utilizar gráficos para obter insights valiosos. Hoje vimos como criar gráficos simples e interativos utilizando diferentes bibliotecas de Python. Cada biblioteca tem suas próprias vantagens.

00:00:18:44 - 00:00:19:49
O Matplotlib, por exemplo, é excelente para criar gráficos estáticos e altamente personalizáveis. O Seaborn facilita a criação de gráficos estatísticos e possui uma integração forte com o Pandas. O Plotly é ideal para gráficos interativos, que podem ser explorados dinamicamente. Entender o contexto dos dados e formular as perguntas certas são passos cruciais para uma análise eficaz. Esperamos que esta aula tenha ajudado a ilustrar como diferentes visualizações podem ser usadas para responder questões específicas sobre um conjunto de dados. Aqui nós vimos algumas das principais bibliotecas para geração de gráficos em Python. Aconselho que naveguem pelos sites oficiais de cada biblioteca para ver quais outros tipos de gráficos podem ser feitos. Tentem plotar outros gráficos a partir de outras cores. Enfim, explorem as ferramentas mostradas para vocês hoje. Além disso, explorem os recursos que cada site oferece, aprendendo sobre outros parâmetros. Sempre tem alguns exemplos em cada página e tal. Vai aí se divertindo, os gráficos são muito legais, muitas cores. E é isso, pessoal. Nos vemos na próxima!