**Olá, estudante!**

Imagine que um produtor de cinema esteja analisando o desempenho dos seus filmes. Ele compara os valores de orçamento dos filmes com as bilheterias e nota que existe uma tendência geral: quanto maior o orçamento maior a bilheteria. Claro, existem exceções. Alguns filmes tiveram grande investimento e foram fracassos, enquanto outros poucos tiveram pouco investimento e alcançaram grandes bilheterias.

Na análise de dados, é importante descobrir como variáveis quantitativas se relacionam entre si, encontrando tendências e exceções, aumentando a compreensão dos dados. Essa tarefa é simplificada através de técnicas de visualização de dados.

Nesta aula, você deve aprender como analisar a relação entre variáveis quantitativas e como criar visualizações, para evidenciar essas relações através de gráficos de dispersão e gráficos de bolhas, utilizando as bibliotecas pandas e seaborn de Python.

**Objetivos**

* Compreender o conceito de [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) entre variáveis quantitativas;
* Aprender a criar e a modificar gráficos de dispersão com seaborn;
* Aprender a criar e a modificar gráficos de bolhas com seaborn.

## Tópico 1 – Correlação e gráfico de dispersão

**OBJETIVOS**

* Compreender o conceito de [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) entre duas variáveis;
* Aprender a criar gráficos de dispersão de duas variáveis, utilizando seaborn;
* Aprender a criar gráficos de dispersão com reta de regressão, utilizando seaborn.

Ao analisar uma variável quantitativa de um conjunto de dados, é útil conhecer a sua distribuição. Porém, muitos conjuntos de dados possuem mais de uma variável quantitativa. Nesses casos, muitas vezes será interessante descobrir se essas variáveis possuem alguma relação entre si. Por exemplo, em dados de altura e peso de pessoas, existe uma tendência estatística de que quanto maior a altura maior o peso de um indivíduo. Porém, isso não é uma regra, pois podem haver indivíduos altos e com baixo peso, e vice-versa. Portanto, chamamos isso de [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao). Nessa perspectiva, uma maneira bastante útil de descobrir se duas variáveis estão correlacionadas é através de gráficos de dispersão.

Neste tópico, você deverá conhecer mais sobre o conceito de [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) e como criar gráficos de dispersão e retas de regressão, para visualizar a relação entre duas variáveis quantitativas. Para isso, vamos seguir.

Para compreender os conceitos estudados neste conteúdo, deve-se explorar um conjunto de dados reais de gorjetas em um restaurante.

Os exemplos serão executados no já conhecido ambiente Jupyter Notebooks, necessitando que as bibliotecas pandas, matplotlib e seaborn estejam devidamente instaladas. Se você utiliza a distribuição Anaconda, que é fortemente recomendada, elas já vêm instaladas por padrão.

Então, em primeiro lugar, você irá criar um notebook, importar e configurar as bibliotecas que serão utilizadas. Confira no código a seguir:

1 import pandas as pd

2 import matplotlib.pyplot as plt

3 import seaborn as sns

4 %matplotlib inline

5 sns.set(style="darkgrid")

Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915159613561/aula/code/codet1p2s4.txt)

Das linhas 1 a 3, são importadas as bibliotecas, atribuindo a elas os respectivos pseudônimos que são utilizados por padrão. O comando, na linha 4, chamado palavra mágica, é específico do Jupyter Notebook, e apenas diz que imagens geradas pelo matplotlib devem ser exibidas na página do Notebook. A linha 5 é opcional, pois configura o estilo dos gráficos gerados com a biblioteca seaborn.

Agora você já pode carregar os dados, que iremos utilizar, e começar a explorá-los, é importante saber que o conjunto de dados chama-se tips (gorjetas, em inglês), e ele pode ser carregado diretamente através do seaborn, sem a necessidade de baixar os dados para a sua máquina. Confira, no código a seguir, como carregar os dados em um DataFrame de pandas, através do seaborn, e exibir as 5 primeiras linhas da tabela:

1 tips = sns.load\_dataset("tips")

2 tips.head()

|  | **total\_bill** | **tip** | **sex** | **smoker** | **day** | **time** | **size** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |

Note que o DataFrame criado com os dados foi passado para a variável tips e exibido com o método head().

Com isso, as colunas representam, da esquerda para a direita, o valor total da conta; o valor da gorjeta; o sexo do cliente, se ele é fumante ou não; o dia da semana em que ocorreu; a refeição; e o número de pessoas na mesa. Assim, existem muitas perguntas que poderiam ser feitas a esses dados. Por isso, digamos que o dono do restaurante deseja saber se quanto maior o valor da conta maior será o valor da gorjeta. Um método muito prático e poderoso de visualizar essa relação é pelo gráfico de dispersão. Confira a seguir.

O gráfico de dispersão, em inglês, scatterplot, representa os dados como pontos em um plano cartesiano. O valor de uma variável é mapeado na posição do eixo horizontal, enquanto outra variável é mapeada na posição vertical. É um tipo de gráfico bastante utilizado pela sua simplicidade de compreensão e por tornar evidente padrões presentes nos dados. Dessa forma, criar gráficos de dispersão, utilizando seaborn é muito simples, principalmente, através da função scatterplot(). Então, os parâmetros básicos da função são iguais aos de outras funções do seaborn, que você já deve ter utilizado, sendo necessário passar a coluna que será mapeada no eixo x, a que será mapeada no eixo y, além do conjunto de dados. No exemplo a seguir, iremos criar um gráfico de dispersão, mostrando a relação entre o valor da conta (total\_bill) e a gorjeta (tip) dos dados analisados. Confira o exemplo de código e seu resultado:

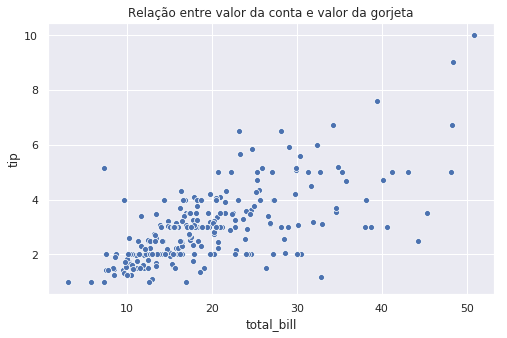
1 plt.figure(figsize=(8,5))

2 plt.title('Relação entre valor da conta e valor da gorjeta')

3 sns.scatterplot(x="total\_bill", y="tip", data=tips)

4 plt.show()

Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915159613561/aula/code/codet1p4s6.txt)

Figura 1 – Relação entre valor da conta e valor da gorjeta

Perceba que a linha 1 configura o tamanho do gráfico em polegadas, e a linha 2 configura o título. Na linha 3, é chamada a função que desenha o gráfico, passando a coluna ‘total\_bill’ para o eixo x; a coluna ‘tip’ para o eixo y; e o DataFrame tips para data, indicando que as colunas passadas são deste DataFrame.

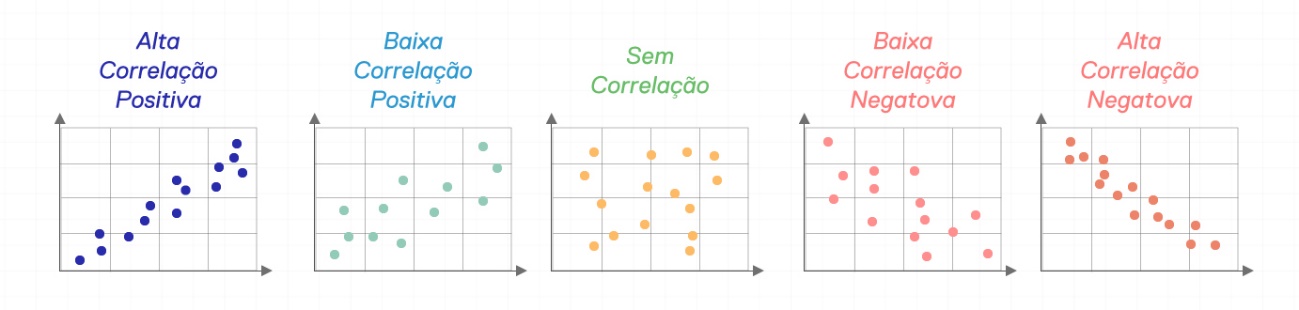
Em seguida, analise o resultado. Realmente existe uma tendência do valor da gorjeta crescer, quando o valor da conta cresce. Entretanto, note que existem pontos que foram exceções, pois existem clientes que deram gorjetas muito generosas, mesmo com o valor da conta baixo, enquanto outros foram mão-de-vaca, mesmo com contas altas. Você consegue identificar esses pontos? Ótimo!

Agora, verifique, de forma geral, que o valor da conta influencia o valor da gorjeta, mas não unicamente, porque existem fatores que não estão representados nos dados e que muitas vezes dependem do acaso, como a simpatia do atendente, o humor do cliente naquele dia, a qualidade da comida, e inúmeros outros motivos. Por isso, não é uma regra rígida que uma alta conta terá uma alta gorjeta, mas há uma tendência estatística, entre as duas variáveis, que chamamos de [**correlação**](https://leadfortaleza.com.br/ead/glossary/Correlacao).

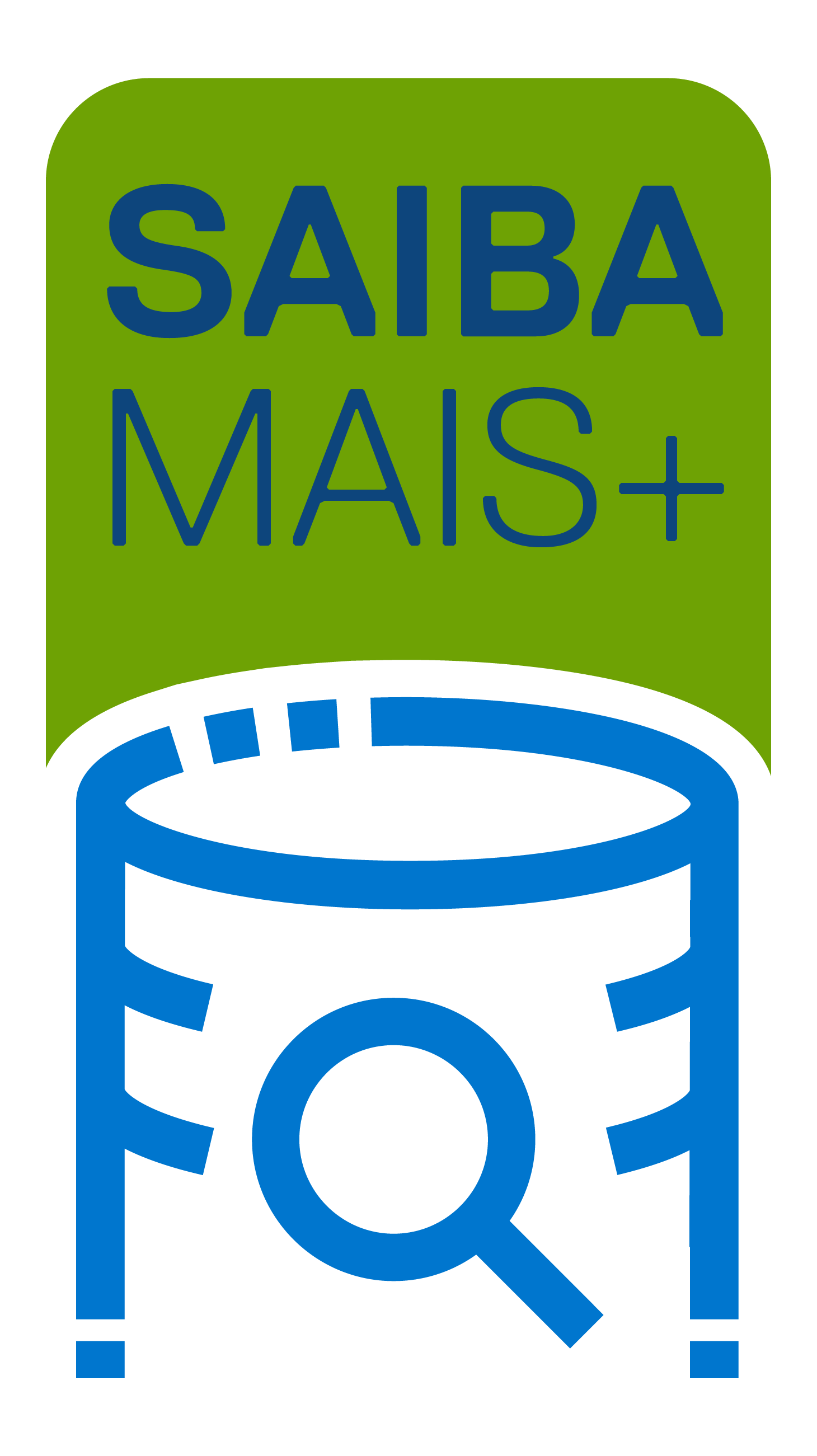
A [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) pode ser positiva ou negativa. No exemplo anterior, a [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) entre a conta e a gorjeta é positiva, pois, conforme uma cresce, a outra tende a crescer também. Do mesmo jeito que, quando uma decresce, a outra tende a decrescer. Ou quando a [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) é negativa, o inverso ocorre: quando uma variável cresce a outra tende a diminuir, e vice-versa.

Alguns exemplos de correlações negativas são: quando os pneus têm menos ar, o carro tende a gastar mais combustível por km rodado; quando uma pessoa se exercita mais, ela tende a ter menos risco de doenças cardíacas; quando a temperatura diminui, a venda de aquecedores tende a aumentar etc. Você consegue pensar em outros exemplos? Provavelmente sim, não é mesmo?

Sabendo disso, a correlação pode ainda ser alta ou baixa: quanto mais alta a correlação, mais a relação entre as variáveis é **linear**. Ou seja, mais os dados se encaixam em uma reta. Por isso, quanto mais baixa, menos os dados se aproximam de uma reta. Por fim, pode não haver correlação entre as variáveis, de modo que não há relação alguma ou tendência entre as variáveis. A figura abaixo ilustra as possibilidades de correlação entre os dados, mostrando exemplos de correlação positiva alta, baixa; e correlação negativa alta e baixa e dados sem correlação. Confira:

Figura 2 – Tipos de [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao)

Nesse momento, você deve estar se perguntando, “mas, como saber se uma [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) é alta ou baixa?” Ou, “ela é percebida apenas visualmente?” Calma, vamos descobrir mais informações, conferindo o box Saiba mais na próxima página, ok?



[Correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) não é apenas uma noção que é percebida visualmente, mas um valor que pode ser calculado a partir dos dados. Portanto, o índice de correlação varia de -1 (correlação negativa perfeita) até 1 (correlação positiva perfeita). Assim, o valor 0 indica inexistência de correlação. Se você ficou curioso para saber como calcular o índice de correlação, consulte a bibliografia recomendada no fim da aula.

Apesar de ser um indicativo muito útil na análise de dados, é muito comum que a [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) leve a conclusões equivocadas. Para conferir uma dica bem importante sobre isso, confira o box a seguir:

#### Ícone Fique Atento!



Tenha muita cautela ao interpretar o significado de uma correlação entre duas variáveis! O fato de duas variáveis estarem correlacionadas não quer dizer que, necessariamente, elas têm uma relação de causa e efeito entre si. Esse tipo de interpretação equivocada é tão comum que se criou uma máxima em análise de dados: [**correlação**](https://leadfortaleza.com.br/ead/glossary/Correlacao)**não implica em causalidade**! Por exemplo, a quantidade de vendas de óculos de sol pode estar fortemente correlacionada com a venda de sorvetes em uma praia, mas não é verdade que um cause o outro. Muito provavelmente ambos têm uma causa em comum, como a alta temperatura no verão. Esse exemplo mostra como, às vezes, é necessário considerar o contexto dos dados, além dos valores brutos.

Agora, o que você acha de continuar aprendendo? Vamos prosseguir!

Você já notou que quando variáveis estão fortemente correlacionadas, elas mais se aproximam de uma reta? Para melhor visualizar essa relação, pode-se calcular uma reta que melhor representa os dados. Essa reta é chamada reta de regressão. O seaborn possui uma função chamada regplot(), que calcula e representa essa reta em um gráfico de dispersão. Dessa maneira, ela pode ser utilizada de forma idêntica a já utilizada função scatterplot(). Confira no exemplo a seguir:

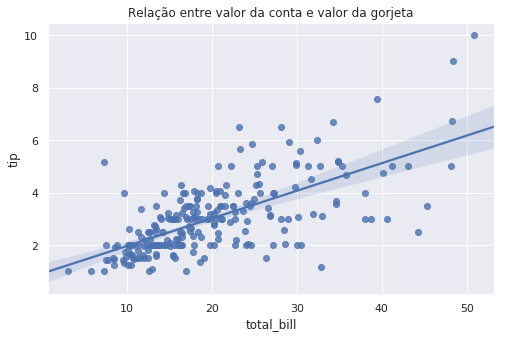
1 plt.figure(figsize=(8,5))

2 plt.title('Relação entre valor da conta e valor da gorjeta')

3 sns.regplot(x="total\_bill", y="tip", data=tips)

4 plt.show()

Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915159613561/aula/code/codet1p7s9.txt)

Figura 3 – Criando uma reta de regressão

Note que o código é quase idêntico ao do exemplo anterior, mudando apenas a função utilizada na linha 3, que substitui scatterplot() por regplot(). Contudo, até os parâmetros são os mesmos.

A reta de regressão permite uma melhor percepção da [correlação](https://leadfortaleza.com.br/ead/glossary/Correlacao) positiva que existe entre as duas variáveis.

Até o momento, você trabalhou comparando duas variáveis quantitativas. No próximo tópico, você aprenderá como visualizar mais variáveis, quantitativas e categóricas, no mesmo gráfico de dispersão.