**Olá, estudante!**

Vamos para a nossa aula sobre visualização de distribuições conjuntas.

Bancos de dados de empresas contêm dados sobre as vendas, as compras, os lucros e prejuízos mensais. Conjuntos de dados imobiliários possuem informações sobre o valor, a área, a quantidade de cômodos, de banheiros e de vagas de garagem. Conjuntos de dados sobre carros fabricados por uma indústria podem conter informações sobre consumo médio de combustível, cavalos de força, aceleração e cilindradas dos veículos.

Você já deve ter notado que muitos conjuntos de dados possuem diversas variáveis, sendo conhecidos como dados multivariados. Compreender as relações existentes entre as múltiplas variáveis e distribuições é um desafio que nem sempre é fácil, mas permite uma compreensão ampla dos padrões existentes nos dados.

Para essa tarefa, visualizações de distribuições conjuntas são uma ferramenta poderosa. Assim, nesta aula, você aprenderá a representar múltiplas distribuições e relações entre variáveis de conjuntos de dados através da biblioteca seaborn.

**Objetivos**

* Representar distribuições conjuntas bivariadas;
* Representar distribuições conjuntas par-a-par em conjuntos de dados multivariados.

## Tópico 1 – Distribuições bivariadas

**OBJETIVOS**

* Representar a distribuição conjunta de duas variáveis utilizando gráficos de dispersão e [histograma](https://leadfortaleza.com.br/ead/glossary/Histograma);
* Representar a distribuição conjunta de duas variáveis utilizando gráficos de hexbins;
* Representar a distribuição conjunta de duas variáveis utilizando gráficos de densidade.

Como você já deve saber, para visualizar a distribuição de uma variável quantitativa, costuma-se utilizar [histograma](https://leadfortaleza.com.br/ead/glossary/Histograma) ou gráfico de densidade. Para representar a relação entre duas variáveis quantitativas, pode-se utilizar gráficos de dispersão. Porém, o gráfico de dispersão por si não dá ênfase nas distribuições individuais de cada variável. A distribuição conjunta de duas variáveis chama-se distribuição bivariada e pode ser visualizada através de uma combinação entre [histograma](https://leadfortaleza.com.br/ead/glossary/Histograma)s e gráfico de dispersão.

Neste tópico, você aprenderá a criar gráficos de distribuição conjunta para representar distribuições bivariadas.

Para que você entenda melhor os assuntos estudados nesta aula, você irá explorar um conjunto de dados reais botânicos, contendo medidas de três espécies de flores íris.

Os exemplos serão executados no já conhecido ambiente Jupyter Notebooks e necessitam das bibliotecas pandas, matplotlib e seaborn devidamente instaladas. Primeiramente, você irá criar um notebook e 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_102915187578976/aula/code/codigo-aula-4-2-topico1-pagina2.txt)

Nas 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 Notebooks, e apenas diz que imagens geradas pelo matplotlib devem ser exibidas na página do Notebook. A linha 5 é opcional, e configura o estilo dos gráficos gerados com a biblioteca seaborn.

Pronto! Agora você já pode carregar os dados que iremos utilizar e começar a explorá-los.

Esse conjunto de dados é muito famoso no mundo de análise de dados e 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 da função load\_dataset() do seaborn e exibir as 5 primeiras linhas da tabela de dados.

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

2 iris.head()

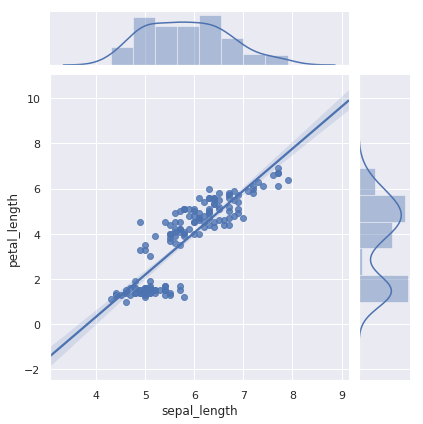
|  | **sepal\_length** | **sepal\_width** | **petal\_length** | **petal\_width** | **species** |
| --- | --- | --- | --- | --- | --- |
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |

A função load\_dataset() pode carregar diversos conjuntos de dados fornecidos pelo seaborn. No caso, passando o valor ‘iris’, o conjunto de dados de flores Iris foram carregados em um DataFrame e passados para a variável iris. A função head() exibe as 5 primeiras linhas do DataFrame. O conjunto de dados contém um conjunto de 150 linhas com cinco atributos - comprimento da sépala (parte constituinte da flor), largura da sépala, comprimento da pétala, largura da pétala, em centímetros e a espécie.

Esses dados foram coletados pelo estatístico Ronald Fisher. O conjunto de dados consiste em 50 amostras de cada uma das três espécies de Iris: Iris setosa, Iris virginica e Iris versicolor. A ideia desse conjunto de dados é, a partir das medidas das pétalas e das sépalas, distinguir a espécie de flor Iris. Agora, para começar a nossa análise, vamos comparar as distribuições de duas variáveis: o comprimento da sépala e o comprimento da pétala.

Como já dito, o histograma permite visualizar a distribuição de uma única variável, mas não favorece a comparação em uma distribuição bivariada. Já o gráfico de dispersão permite visualizar e comparar a distribuição conjunta, mas não favorece a percepção das distribuições individuais de uma distribuição conjunta, também chamadas marginais. Para sanar essa limitação, o seaborn fornece a função jointplot(), que permite visualizar uma distribuição conjunta através da combinação do histograma e do gráfico de dispersão. Confira, no exemplo a seguir, como visualizar a distribuição conjunta do comprimento da sépala e do comprimento da pétala:

1 sns.jointplot(x="sepal\_length", y="petal\_length", kind = 'reg', data=iris)

Figura 1 – Distribuição conjunta do comprimento da [sépala](https://leadfortaleza.com.br/ead/glossary/S%C3%A9pala) e da pétala com gráfico de dispersão e marginais com gráfico de densidade

Utilizar a função joinplot() é simples, bastando passar os valores que serão mapeados no eixo x e os valores que serão mapeados no eixo y, assim como se faria para criar um gráfico de dispersão com a função scatterplot(). A seguir, confira um pouco mais sobre essa função aplicada aos dados.

Um detalhe importante sobre a função jointplot() é que ela é baseada na classe do seaborn JointGrid, que controla a grade de visualizações do histograma com o gráfico de densidade e o gráfico de dispersão. Por isso, não é possível configurar o visual dessa função utilizando funções do matplotlib.pyplot, como é possível com outras funções do seaborn.

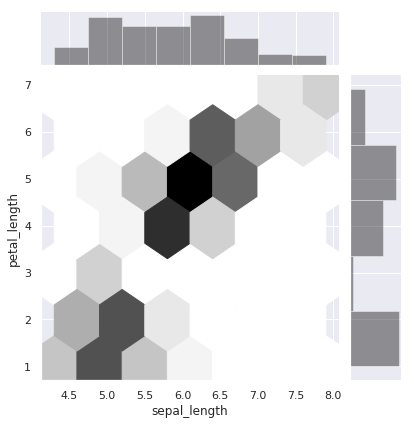
Note que, além dos eixos x e y, foi passada para o parâmetro kind o valor “reg”, indicando que a função deve desenhar a reta de regressão. Se nenhum valor for passado, a função desenha apenas o gráfico de dispersão e histograma.

Analise o resultado da função. No topo observa-se um [histograma](https://leadfortaleza.com.br/ead/glossary/Histograma) e gráfico de densidade da distribuição marginal dos comprimentos das sépalas e na lateral à direita, há a distribuição marginal do comprimento das pétalas. Em uma única visualização, pode-se perceber que existe uma forte correlação entre os comprimentos da pétala e da sépala, e que a distribuição dos comprimentos das pétalas possuem dois “picos”, no linguajar estatísticos chamados modas, sendo portanto uma distribuição bimodal. Esse tipo de gráfico é útil por poupar tempo e mostrar bastante informação de forma clara.

O seaborn oferece ainda outras variações para o jointplot(), que podem ser controlados através do parâmetro kind. Vamos continuar e conferir mais sobre isso.

Uma das variações é o gráfico de hexbin, que é análogo ao histograma para duas dimensões. Nessa variação, ao invés de representar a distribuição conjunta por pontos, o plano cartesiano é dividido em “bins” hexagonais de dimensões iguais. A quantidade de dados dentro de um determinado bin é mapeada para uma escala de cor. Ou seja, quanto mais escuro o bin, mais dados há nele, formando um tipo de gráfico de densidade. Para criar um gráfico de hexbin, basta passar o valor ‘hex’ para o parâmetro kind. Confira:

1 sns.jointplot(x="sepal\_length", y="petal\_length", kind = 'hex', color = 'black', data=iris)

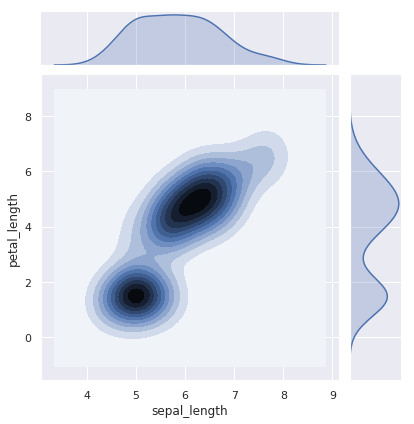
Figura 2 – Distribuição conjunta do comprimento da sépala e da pétala com gráfico de hexbin e marginais com histograma

Note que foi passado o valor ‘black’ para o parâmetro color, indicando que será utilizada uma escala de tons de cinza, onde quanto mais próximo do branco há menos dados, e quanto mais próximo do preto há mais dados.

Em relação ao gráfico de dispersão com reta de regressão, o gráfico de hexbin foca menos na correlação entre as duas variáveis e mais na distribuição espacial entre elas, chamando atenção para as regiões onde os dados são mais frequentes.

Uma outra variação é o gráfico de densidade, onde são estimadas a probabilidade de haver elementos em uma dada região do gráfico. A principal diferença deste gráfico para o de hexbin é que a distribuição conjunta é representada por contornos que possuem uma continuidade, enquanto as marginais são representadas por curvas de probabilidade. Essas curvas de probabilidade são estimadas através de um método conhecido como **Estimativa de Densidade de Kernel** (em inglês, a sigla KDE é utilizada). Para utilizar esse tipo de gráfico, basta passar o valor ‘kde’ para o parâmetro kind. Confira:

1 sns.jointplot(x="sepal\_length", y="petal\_length", kind = 'kde', data=iris)

Figura 3 - Distribuição conjunta e marginais do comprimento da [sépala](https://leadfortaleza.com.br/ead/glossary/S%C3%A9pala) e da pétala com gráficos de densidade

Assim como o gráfico de hexbin, o gráfico de densidade foca na distribuição espacial dos dados. Porém, ele representa essa distribuição na forma de probabilidades, onde as áreas mais escuras possuem maior probabilidade de conter dados e as mais claras possuem menos. Comparando, é a diferença entre um histograma e uma curva de densidade de probabilidade para uma distribuição univariada. O gráfico de densidade mostra a distribuição conjunta de forma mais suave e pode ser mais indicado quando o conjunto de dados é muito grande, ajudando a identificar padrões espaciais. Mas é importante lembrar que ele é uma estimativa criada a partir dos dados. Já o gráfico de hexbin mostra a distribuição real dos dados, mas a sua precisão, assim como no histograma, depende da escolha do tamanho dos bins.

Compare as duas visualizações. Qual você achou que transmitiu a distribuição conjunta de forma mais clara? A seguir, verifique um box “Fique atento!” e confira uma dica bacana.

Para comparar distribuições de uma variável em relação a um ou mais atributos categóricos, os gráficos de caixa e violino são muito úteis. Porém utilizar esses gráficos para comparar diretamente duas ou mais variáveis quantitativas distintas pode não ser indicado, pois é comum que as variáveis estejam em escalas diferentes e a comparação pode não fazer sentido. Nesses casos, é preferível gráficos de distribuição conjunta, como o exemplo apresentado anteriormente.

Gráficos de distribuição conjunta são muito úteis para analisar a distribuição entre duas variáveis. Porém, é comum que conjuntos de dados possuam muitas variáveis quantitativas. Nesse caso, pode ser cansativo analisar uma a uma, todas as possíveis combinações de variáveis à procura de padrões significativos. Para facilitar esse trabalho, o seaborn possui funções para comparar múltiplas distribuições conjuntas em uma única visualização. Esse método será abordado no próximo tópico.