**Tópico 2 – Distribuições conjuntas par-a-par**

**OBJETIVOS**

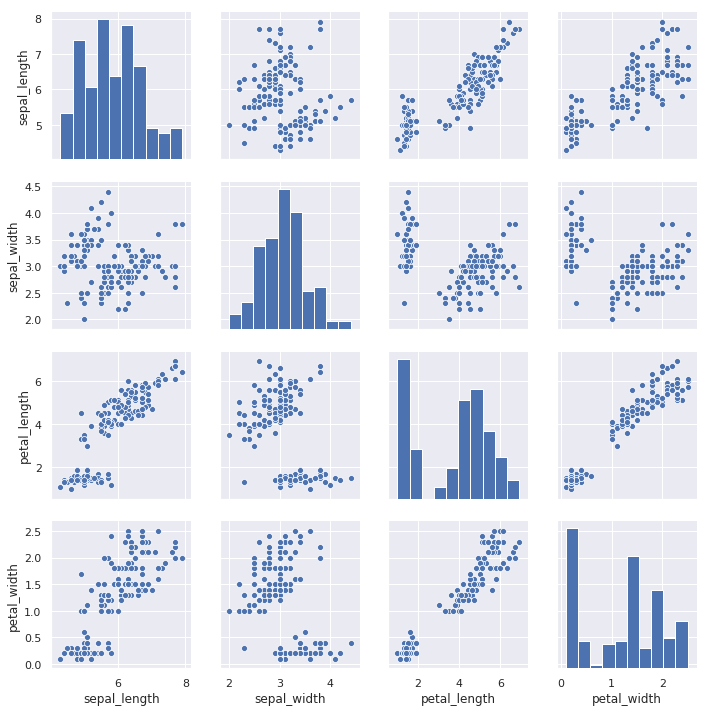
* Representar a distribuição conjunta de múltiplas variáveis utilizando visualizações em grade, par-a-par;
* Representar a distribuição conjunta de múltiplas variáveis distinguindo categorias.

Ao explorar conjuntos de dados com muitas variáveis, muitas vezes não sabemos que padrões estamos procurando e explorar as relações entre as diferentes variáveis pode ser cansativo. Nesses casos, é muito útil um método que permita uma visão geral do conjunto de dados, possivelmente, revelando padrões que podem ser investigados com outras visualizações mais específicas. A biblioteca seaborn fornece uma função muito útil para este fim, chamada pairplot().

Neste tópico, você irá aprender a utilizar a função pairplot() para visualizar múltiplas distribuições, simultaneamente, e entenderá como diferenciar as distribuições por categorias.

A função pairplot() é uma das mais práticas do seaborn. Passando como parâmetro apenas a variável que contém o DataFrame, a função identifica todas as variáveis quantitativas e exibe uma matriz de visualizações par-a-par, mostrando a relação entre cada combinação de variáveis. Por padrão, também exibe a distribuição univariada de cada variável na diagonal. Confira como utilizar a função no exemplo a seguir:

1 sns.pairplot(iris)

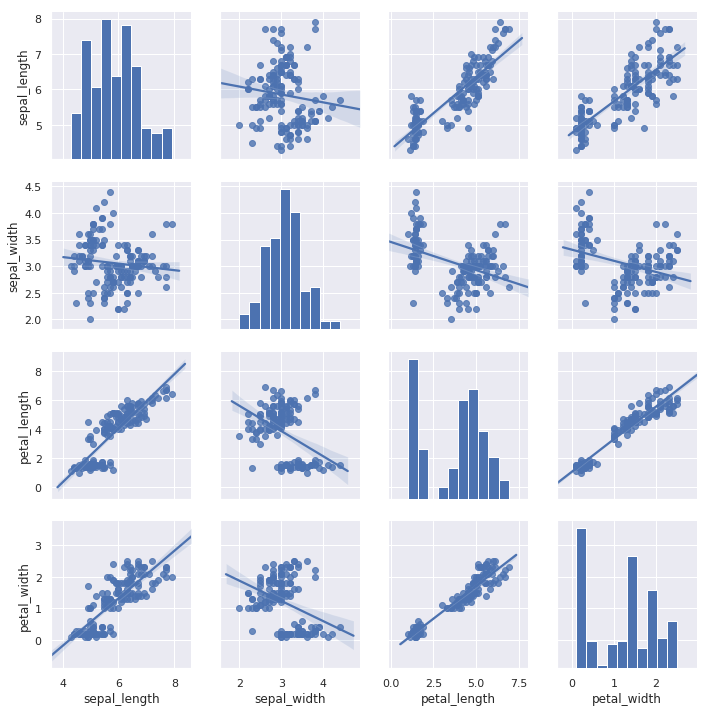
Figura 4 – Distribuições conjuntas de todas as variáveis

Essa função é baseada na classe do seaborn PairGrid, que configura o visual das múltiplas visualizações. Por esse motivo, não pode ser configurada com funções do matplotlib.pyplot.

Observe o resultado. Cada linha e cada coluna da matriz representa uma variável. Quando a variável da linha e da coluna são distintas, o elemento da matriz mostra um gráfico de dispersão das duas variáveis. Quando é a mesma, mostra o [histograma](https://leadfortaleza.com.br/ead/glossary/Histograma) da variável. Em uma mesma visualização, é possível analisar as relações entre as variáveis possíveis, além das distribuições univariadas. Como mostra muita informação em uma única imagem, ela não favorece a percepção de detalhes em particular, mas sim uma visão geral do conjunto de dados que pode ser aprofundada posteriormente. Conferindo rapidamente, pode-se notar, por exemplo, que existe uma forte correlação positiva entre o comprimento e a largura da pétala.

ambém é possível utilizar o parâmetro kind para alterar a visualização. Por exemplo, para criar uma reta de regressão, basta passar o valor ‘reg’ para o parâmetro kind, da mesma forma que foi feito com a função jointplot(). Confira:

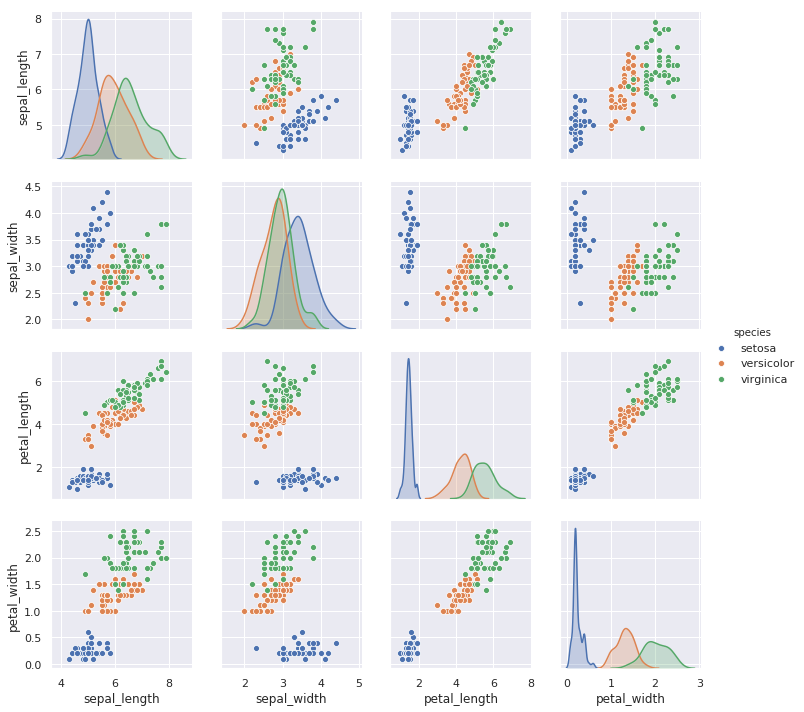
1 sns.pairplot(iris, kind = 'reg')

Figura 5 – Distribuições conjuntas de todas as variáveis com retas de regressão

Essa visualização enfatiza as correlações existentes entre as diferentes variáveis. Ela é bastante útil para conhecer as variáveis quantitativas. Mas em relação a variáveis categóricas? No caso desse conjunto de dados em particular, o maior interesse é em como distinguir as três espécies de Íris a partir das medidas das pétalas e sépalas. Confira a seguir como isso pode ser feito.

Para distinguir as três espécies, a função pairplot() consegue distinguir uma variável categórica em distribuições conjuntas, mapeando a categoria para uma cor. Para isso, basta passar a coluna com a variável categórica, nesse caso, basta passar a coluna ‘species’ para o parâmetro hue. Confira no exemplo a seguir:

1 sns.pairplot(iris, hue=’species’)

Figura 6 – Distribuições conjuntas de todas as variáveis distinguindo as espécies

Observe que nos gráficos de dispersão os pontos azuis são Iris da espécie setosa, os laranjas são da espécie versicolor e os verdes da espécie virginica. Além disso, na diagonal as distribuições univariadas são representadas por gráficos de densidade, onde é feita uma separação por espécie.

Note como a função pairplot() é prática: Através de uma única função, conseguimos uma visão geral de todas as variáveis do conjunto de dados! Para o objetivo de distinguir as espécies, pode-se notar, por exemplo, que o comprimento e largura da pétala são bastante informativos. Note que as distribuições tanto do comprimento como da largura são bastante separáveis entre as espécies, com poucas áreas de interseção. Além disso, na distribuição conjunta do comprimento e largura da pétala, nota-se que as espécies possuem uma separação razoável, onde as setosas possuem menores comprimentos e larguras, as versicolor possuem valores intermediários e as virginicas possuem valores maiores, com alguma interseção. A seguir, confira um box “Saiba mais!” com uma informação bem interessante.

Ao se explorar um conjunto de dados com múltiplas variáveis quantitativas, é uma boa prática começar a explorá-lo utilizando a função pairplot() e depois criar visualizações que aprofundem as relações mais interessantes. Além disso, essa função é muito flexível, assim como outras do seaborn, e pode ter o visual configurado de muitas formas através dos seus parâmetros. Para conhecer todas as possibilidades, consulte a documentação oficial da função neste link:  
<https://seaborn.pydata.org/generated/seaborn.pairplot.html#seaborn.pairplot>.

Nessa aula, você aprendeu a utilidade de visualizações de distribuições conjuntas na análise exploratória de dados e teve a oportunidade de criar essas visualizações com a biblioteca seaborn. Com essa ferramenta, será muito mais fácil obter um entendimento geral dos dados e depois aprofundá-los com visualizações mais específicas. Lembre-se de exercitar o que você aprendeu e de estudar as possibilidades de customização visual oferecidas pela biblioteca seaborn.

Espero que você tenha aprendido bastante. Até a próxima aula!

### Referências

* WILKE, Claus O. Fundamentals of Data Visualization. O’Reilly, 2019.
* Matplotlib. <https://matplotlib.org/index.html>. Acessado pela última vez em Maio de 2020.
* Seaborn. <https://seaborn.pydata.org/tutorial.html>. Acessado pela última vez em Maio de 2020.