**Tópico 3 – Gráficos de violino**

**OBJETIVOS**

* Interpretar gráficos de violino;
* Aprender a criar gráficos de violino, utilizando a biblioteca seaborn;
* Aprender a criar gráficos de violino para representar múltiplas distribuições categóricas.

O gráfico de violino é um tipo de gráfico de densidade. Ou seja, ele representa uma distribuição de dados por uma curva de probabilidade, indicando os valores mais prováveis para cada elemento. Essa curva é estimada através de um método conhecido como **Estimativa de Densidade de Kernel** (em inglês, a sigla KDE é utilizada). A curva é, então, espelhada, gerando uma forma simétrica que lembra um violino, de onde o gráfico tira o seu nome. Por isso, gráficos de violino combinam a visualização compacta do gráfico de caixa com a nuance da forma da distribuição do gráfico de enxame.

Neste tópico, você irá entender como funcionam e como criar gráficos de violino, utilizando a biblioteca seaborn.

A criação de gráficos de violino com seaborn é simples. Basta utilizar a função violinplot() com os mesmos parâmetros que foram usados nas funções swarmplot() e boxplot(). Confira no exemplo a seguir:

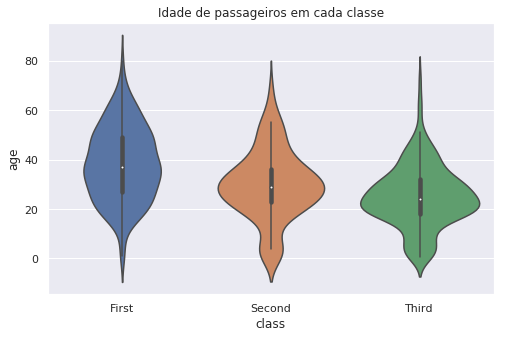
1 plt.figure(figsize=(8,5))

2 plt.title('Idade de passageiros em cada classe')

3 sns.violinplot(x="class", y="age", data=titanic)

4 plt.show()

Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915070709558/aula/code/codet3p2s19.txt)

Figura 11 – Idade de passageiros em cada classe: gráfico de violino

Note que o formato dos violinos lembra bastante os gráficos de enxame. Ou seja, com eles, pode-se visualizar as nuances da distribuição que, com os gráficos de caixa, não era possível. Porém, diferente do gráfico de dispersão, o gráfico de violino mostra probabilidades, e não os dados diretamente, como pontos. Isso significa que ele não sofre a limitação de precisar de mais espaço conforme aumentam os dados.

Além disso, o gráfico de violino também mostra uma versão reduzida do gráfico de caixa, representado dentro de cada violino. Dessa maneira, o ponto branco, no meio, representa a mediana; a linha mais espessa representa a caixa; e as linhas menos espessas representam os bigodes.

Também é possível combinar o gráfico de violino com o gráfico de enxame simplesmente, chamando as duas funções! Confira no exemplo:

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

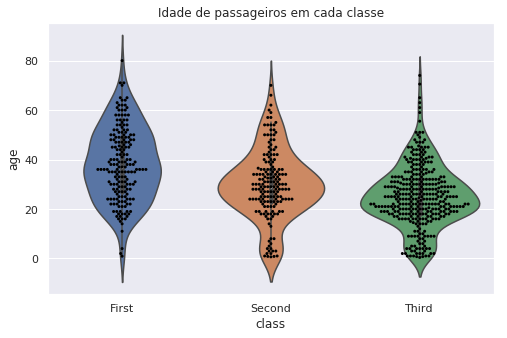
2 plt.title('Idade de passageiros em cada classe')

3 sns.violinplot(x="class", y="age", data=titanic)

4 sns.swarmplot(x="class", y="age",color = 'black', size = 3, data=titanic)

5 plt.show()

Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915070709558/aula/code/codet3p3s20.txt)

Figura 12 – Idade de passageiros em cada classe: gráfico de violino com gráfico de enxame

Observe que não houve alteração nas linhas 1, 2 e 3. Na linha 3, inclusive, foi chamada a função violinplot() de forma idêntica ao que já foi feito. Na linha 4, é chamada a função swarmplot(), alterando apenas os parâmetros color e passando o valor ‘black’; e o tamanho, através do parâmetro size, passando o valor 3, para tornar a visualização mais limpa.

Para visualizar também a distribuição dos sexos, basta passar a coluna ‘sex’ para o parâmetro hue, assim como feito nas funções swarplot() e boxplot(). Porém, existe a possibilidade de “partir” o violino e representar uma curva de probabilidade de cada sexo no mesmo violino, ao invés da forma simétrica. Para isso, basta passar o valor True para o parâmetro split. Confira no exemplo:

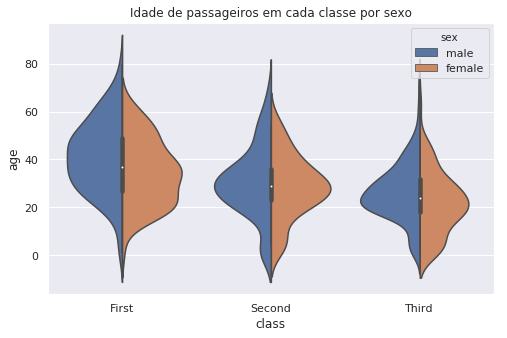
1 plt.figure(figsize=(8,5))

2 plt.title('Idade de passageiros em cada classe por sexo')

3 sns.violinplot(x="class", y="age", hue = 'sex', split = True, data=titanic)

4 plt.show()

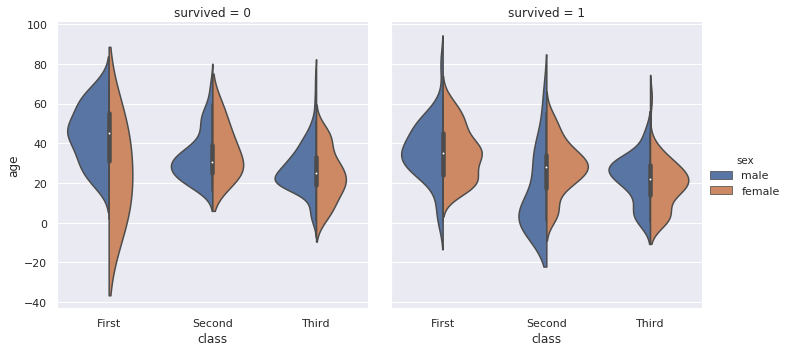
Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915070709558/aula/code/codet3p4s21.txt)

Figura 13 – Idade, sexo e classe de passageiros: gráfico de violino

Partir os violinos permite um uso mais eficiente do espaço, representando mais informação, mas sem poluir a visualização. Vamos seguir?

Por fim, assim como feito anteriormente, podemos utilizar a função catplot() para criar uma grade e visualizar os passageiros que sobreviveram e os que não sobreviveram.

1 sns.catplot(x="class", y="age", hue = 'sex', col = 'survived', kind = 'violin', split = True, data=titanic)

Figura 14 – Idade, sexo, classe de passageiros que sobreviveram ou não: gráfico de violino em grade

Esse gráfico chama a atenção para a curva de probabilidade feminina da primeira classe dos não-sobreviventes, indicando que houveram crianças do sexo feminino da primeira classe que não sobreviveram. Perceba que o gráfico de violino, assim como o gráfico de caixa, não indica quantidades. Quando existem muitos dados, isso é desejado, mas quando há um menor número, pode ser mais interessante visualizar com o gráfico de enxame.

Ao longo desta aula foram apresentados 3 tipos de gráficos para visualizar distribuições entre categorias. Esses três tipos foram utilizados para visualizar os mesmos dados e cada um revelou ou deu ênfase a diferentes padrões existentes nos dados. Perceba que cada um deles foca em questões diferentes, e o mais apropriado irá depender do objetivo da visualização. Então, isso é uma característica fascinante da área de visualização de dados, pois visualizações diferentes revelam alguns aspectos dos dados e ocultam outros. Portanto, é preciso avaliar qual é a mais vantajosa, especialmente, quando os dados possuem muitas variáveis que se relacionam entre si de diferentes formas.

Muitas vezes é útil utilizar mais de um tipo de visualização sobre os mesmos dados para descobrir ou transmitir a informação que se deseja, como foi feito nesta aula.

Caro estudante, você chegou ao fim desta aula, percorrendo um longo caminho e conhecendo técnicas muito poderosas para visualizar distribuições categóricas. Além disso, foi conhecido o gráfico de dispersão categórica, o gráfico de caixa e o gráfico de violino, estudando os seus pontos fortes e fracos.

Agora, você possui ótimas ferramentas para analisar conjuntos de dados com múltiplas variáveis. Procure sempre experimentar diferentes visualizações para o mesmo problema, avalie o que cada uma lhe permite perceber nos dados e procure sempre conhecer mais sobre as ricas possibilidades de customizar os gráficos, consultando a documentação oficial do seaborn.

Espero que você tenha aprendido e experimentado bastante! Até a próxima!

### Referências

* 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.
* WILKE, Claus O. **Fundamentals of Data Visualization. O’Reilly**, 2019.