**Tópico 2 – Gráficos de caixa**

**OBJETIVOS**

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

O gráfico de caixa é um tipo de gráfico muito popular em estatística, pois fornece uma visualização compactada de uma distribuição de dados, além de ser bastante útil para representar múltiplas distribuições simultaneamente. Porém, a interpretação deste gráfico não é tão intuitiva. Assim, é necessário compreender o seu funcionamento.

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

Lembra do método describe() em pandas? Então, esse método retorna alguns valores estatísticos que fornecem uma noção sobre a distribuição de uma variável. Dentre elas, o método diz o valor mínimo e máximo e os quartis. O primeiro quartil (Q1) é o número que deixa 25% dos dados abaixo e 75% acima. O segundo quartil (Q2), também chamado mediana, divide os dados em duas partes iguais, e o terceiro quartil (Q3) deixa 75% das observações abaixo e 25% acima. Portanto, a ideia do gráfico de caixa é uma representação visual padronizada dessas informações, através da dimensão de uma “caixa” em um eixo.

Para exemplificar, analise as informações obtidas com o método describe() aplicado sobre a idade dos passageiros do Titanic, em particular o min, 25%, 50%, 75% e max:

count 714.000000  
mean 29.699118  
std 14.526497  
min 0.420000  
25% 20.125000  
50% 28.000000  
75% 38.000000  
max 80.000000

Agora, vamos comparar essas informações com um gráfico de caixa e um gráfico de dispersão categórico.

A seguir, confira as informações obtidas e as duas imagens com o gráfico de caixa e o gráfico de dispersão, respectivamente.

min 0.420000  
25% 20.125000  
50% 28.000000  
75% 38.000000  
max 80.000000

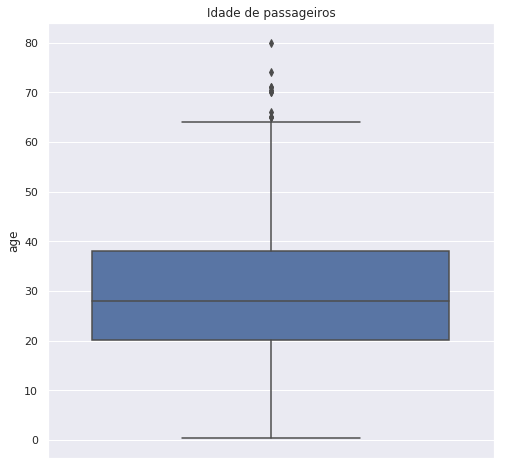
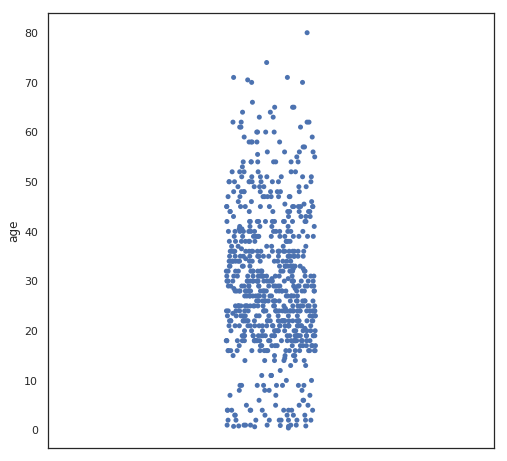
Figura 5 – Gráfico de caixa

Figura 6 – Gráfico de dispersão

Em um diagrama de caixa, é desenhada uma caixa do primeiro ao terceiro quartil. A reta que passa pela caixa representa o segundo quartil, ou mediana. Perceba que, dentro do intervalo da caixa, estão 50% dos dados, e esse intervalo é conhecido como [**amplitude interquartil**](https://leadfortaleza.com.br/ead/glossary/Amplitude%20interquartil) (AIQ). As retas verticais acima e abaixo da caixa são chamados bigodes, e indicam a cauda da distribuição, ou seja, onde existem menos dados. Os limites dos bigodes são calculados por 1,5\*AIQ e contêm cerca de 99% dos dados. Já os dados fora dos limites dos bigodes são dados raros, ou seja, exceções. Por isso, esses dados são representados por pontos e são chamados [**outliers**](https://leadfortaleza.com.br/ead/glossary/Outliers), ou pontos discrepantes. Compare o diagrama de caixa com o gráfico de dispersão categórico. Note que é uma representação mais compacta.

Agora que você já entende como funciona o gráfico de caixa, está na hora de aprender a criá-los utilizando seaborn.

Vamos representar os mesmos dados dos gráficos criados anteriormente para fins de comparação. Você perceberá que o gráfico de caixa é criado de forma muito similar, passando os mesmos parâmetros, e a função que cria gráficos de caixa no seaborn é boxplot(). Confira, no exemplo a seguir, como visualizar as distribuições de idades por classes, passando as classes para o eixo x, e as idades para o eixo y:

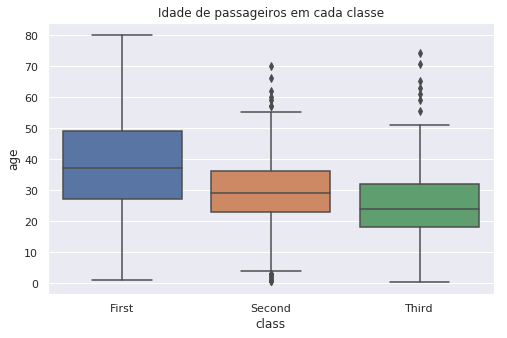
1 plt.figure(figsize=(8,5))

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

3 sns.boxplot(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/codet2p4s14.txt)

Figura 7 – Idade de passageiros em cada classe em gráfico de caixa

Note que a única diferença, para o primeiro gráfico de dispersão categórico, é que foi trocado o nome da função. Então, o que você achou do resultado em relação ao gráfico de dispersão categórico? Consegue perceber algum novo padrão? Uma das vantagens dos gráficos de caixa, além da sua simplicidade, é dar destaque na representação de [outliers](https://leadfortaleza.com.br/ead/glossary/Outliers). Porém, a grande vantagem em relação ao gráfico de dispersão categórico se mostra ao se representar mais de uma variável categórica, como o sexo.

O sexo pode ser representado no boxplot() de forma idêntica ao que foi feito com o swarmplot(), passando a coluna ‘sex’ para o parâmetro hue. Confira:

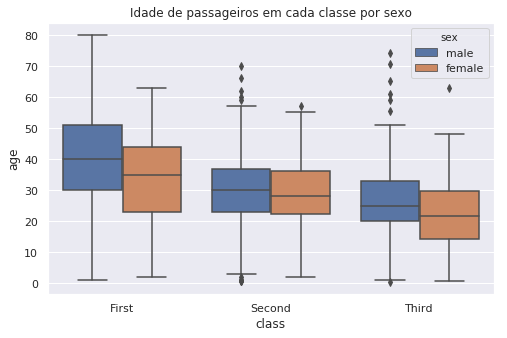
1 plt.figure(figsize=(8,5))

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

3 sns.boxplot(x="class", y="age", hue = 'sex', 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/codet2p5s15.txt)

Figura 8 – Idade e sexo de passageiros em cada classe em gráfico de caixa

Essa visualização permite ter uma noção comparativa sobre as distribuições de idade por sexo em cada classe, enquanto que a visualização gerada pelo swarmplot() permite apenas comparar a proporção entre os sexos. Por essa visualização, pode-se perceber, por exemplo, que as mulheres são mais jovens do que os homens em todas as classes.

O seaborn disponibiliza ainda uma função que é uma variação do gráfico de caixa, chamada boxenplot(). A interpretação desse gráfico é a mesma, a diferença é que ele representa os bigodes com espessuras que permitem uma melhor percepção visual das quantidades de dados nas caudas. A função pode ser utilizada com os mesmos parâmetros de boxplot(). Confira no exemplo:

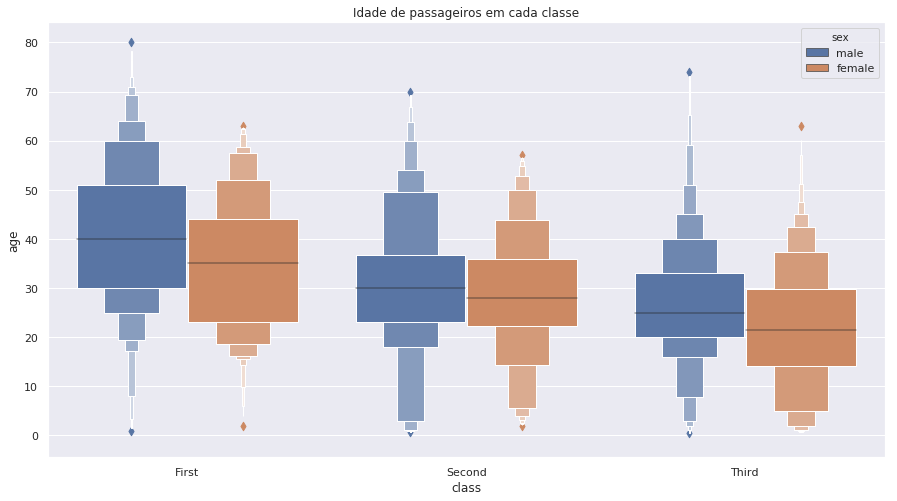
1 plt.figure(figsize=(15,8))

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

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

4 plt.show()

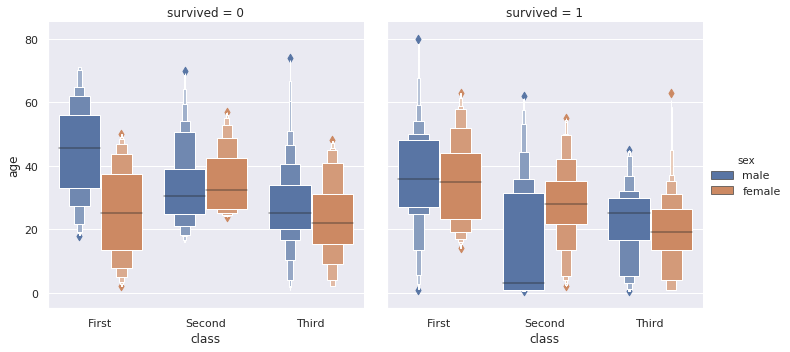
Download do código sem numeração no link a seguir: [Download código](https://leadfortaleza.com.br/ead2pcd/conteudo/tmp/myopenolat_1_102915070709558/aula/code/codet2p6s16.txt)

Figura 9 – Idade e sexo de passageiros em cada classe em gráfico de caixa com a função boxenplot()

Note que a ideia é a mesma do gráfico de caixa, apenas dando uma maior ênfase visual aos bigodes. Você acha essa visualização melhor? Conseguiu perceber um novo padrão a partir dela?

E para visualizar também os passageiros que sobreviveram e os que não sobreviveram através do gráfico de caixa? Assim como já foi feito, pode-se criar uma grade de gráficos através da função catplot(). Lembre-se de que essa função pode representar diferentes tipos de gráficos. O tipo é escolhido pelo parâmetro kind e, para utilizar gráfico de caixa, basta passar o valor ‘box’. Ou, para o resultado do estilo do boxenplot, basta passar o valor ‘boxen’. Confira no exemplo:

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

Figura 10 – Idade, sexo, classe de passageiros que sobreviveram ou não: gráfico de caixa em grade

Lembre-se de que não é possível configurar os gráficos dessa função através de funções do matplotlib. Esse resultado não dá uma ideia clara das quantidades, mas fornece uma melhor noção sobre as distribuições do que a visualização equivalente com gráfico de dispersão categórico. Por exemplo, pode-se perceber que metade dos sobreviventes da segunda classe eram crianças.

Perceba que o gráfico de caixa fornece uma melhor noção geral sobre os intervalos das distribuições, onde se encontram os dados e os dados discrepantes. Já o gráfico de enxame fornece uma noção sobre as quantidades e o formato da distribuição, que não é apresentada no gráfico de caixa. No próximo tópico, você irá conhecer o gráfico de violino, que combina algumas das características do gráfico de caixa e do gráfico de enxame em uma mesma visualização.