Visualização de dados

Unidade III

Cristiane Neri Nobre

PUC Minas Virtual

Distribuições

PUC Minas Virtual

Relacionamentos

Vamos explorar agora os gráficos para visualizar **distribuições**

Visualizando distribuições

Qual os tipos de gráficos mais comuns para visualizar **distribuição dos dados**?

1. Histograma
2. Gráficos de densidade
3. Boxplot
4. Violino

Visualizando distribuições

Histograma

Frequentemente encontramos a situação em que gostaríamos de entender como uma determinada variável é **distribuída** em um conjunto de dados

Número de passageiros com idade conhecida no Titanic.

Faixa etária	Contar
0–5	36
6 a 10	19
11-15	18
16-20	99
21-25	139
26-30	121

Faixa etária	Contar
31-35	76
36-40	74
41-45	54
46-50	50
51-55	26
56-60	22

Faixa etária	Contar
61-65	16
66-70	3
71-75	3

Fonte: (Wilke, 2019)

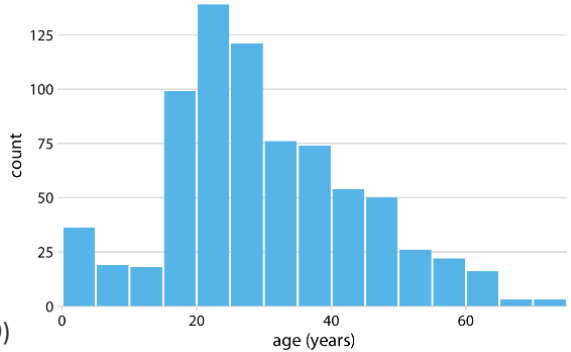
Visualizando uma única distribuição

Número de passageiros com idade conhecida no Titanic.

Faixa etária	Contar
0–5	36
6 a 10	19
11-15	18
16-20	99
21-25	139
26-30	121

Faixa etária	Contar
31-35	76
36-40	74
41-45	54
46-50	50
51-55	26
56-60	22

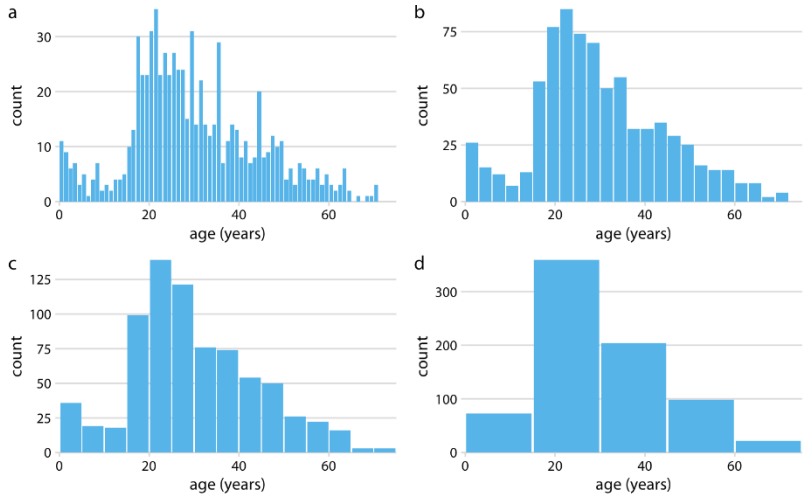
Faixa etária	Contar
61-65	16
66-70	3
71-75	3



Fonte: (Wilke, 2019)

Visualizando uma única distribuição

Neste tipo de gráfico é importante observar a **amplitude de classe**



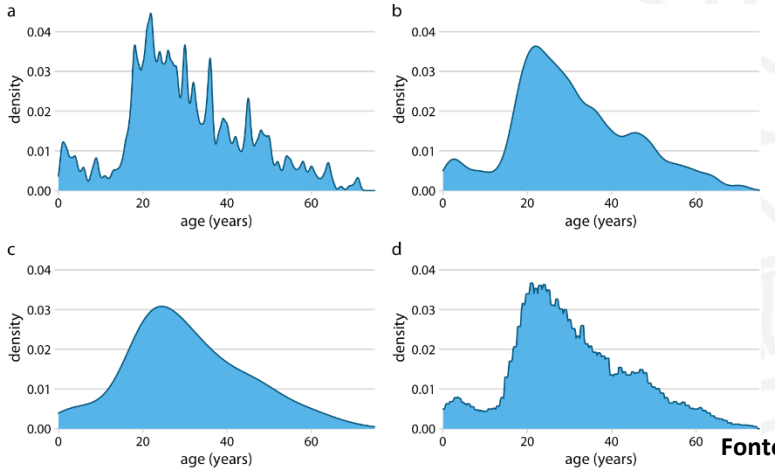
Amplitude de classe de (a) **um** ano; (b) **três** anos; (c) **cinco** anos; (d) **quinze** anos. **Fonte**: (Wilke, 2019)

Para a distribuição de idade dos passageiros do Titanic, podemos ver que amplitude de classe de 1 ano é **muito pequena** e de 15 anos é **muito grande**, enquanto de amplitude de três a cinco anos **funcionam bem**

Visualizando uma única distribuição

Podemos também fazer gráficos de densidade

- Importante também analisar a amplitude de classe, além da kernel (gaussiano, retangular, etc)



Fonte: (Wilke, 2019)

Visualizando uma única distribuição

No site: http://shiny.leg.ufpr.br/walmes/density/

Você consegue visualizar o efeito de trocarmos a função e a amplitude da classe.

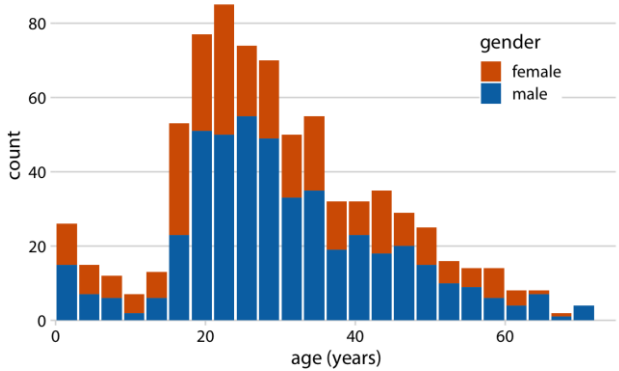
Experimente!

Visualizando várias distribuições ao mesmo tempo

Imagine a situação:

Gostaríamos de ver como as idades dos passageiros do Titanic são distribuídas entre homens e mulheres

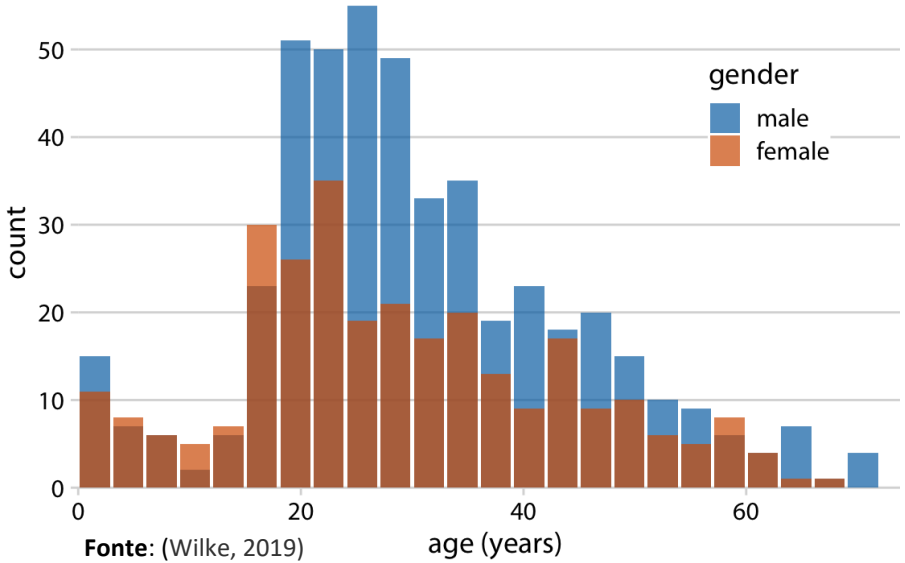
- Homens e mulheres eram da mesma idade ou havia diferença de idade entre os sexos?



Fonte: (Wilke, 2019)

Neste gráfico, onde exatamente as barras começam. Elas começam onde a cor muda ou devem começar do zero?

Visualizando várias distribuições ao mesmo tempo



Fonte: (Wilke, 2019)

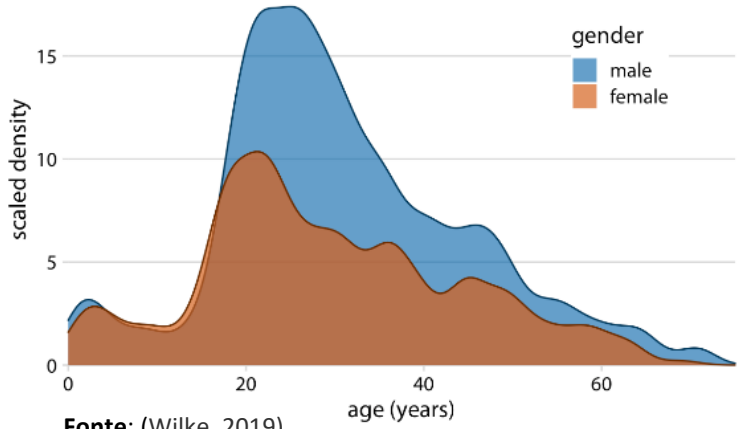
Podemos resolver esses problemas fazendo com que todas as barras comecem em zero e tornando-as parcialmente transparentes

Mas nesta solução, onde as barras azuis começam?

Há dois grupos ou 3?

Visualizando várias distribuições ao mesmo tempo

Uma solução é fazer gráfico de densidade

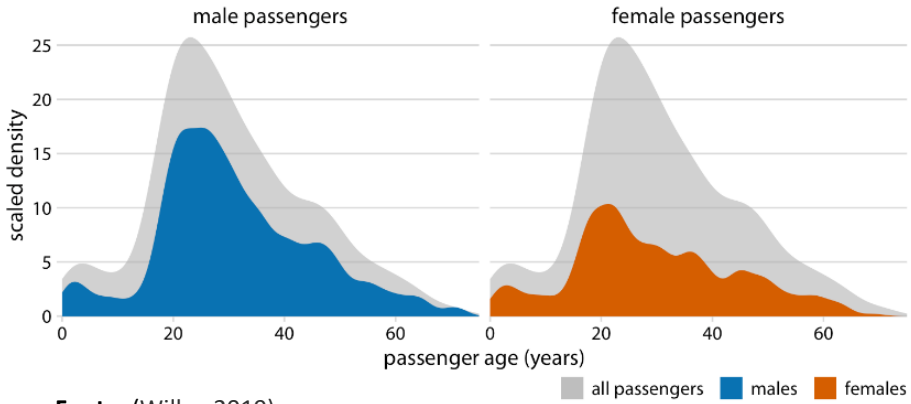


Fonte: (Wilke, 2019)

Mas ficará difícil observar que as distribuições de idade para passageiros do sexo masculino e feminino são **quase idênticas** até por volta dos **17 anos**

Visualizando várias distribuições ao mesmo tempo

Uma solução que funciona bem para este conjunto de dados é mostrar as distribuições de idade de passageiros do sexo **masculino e feminino separadamente**, cada um como uma proporção da **distribuição geral** de idade



Fonte: (Wilke, 2019)

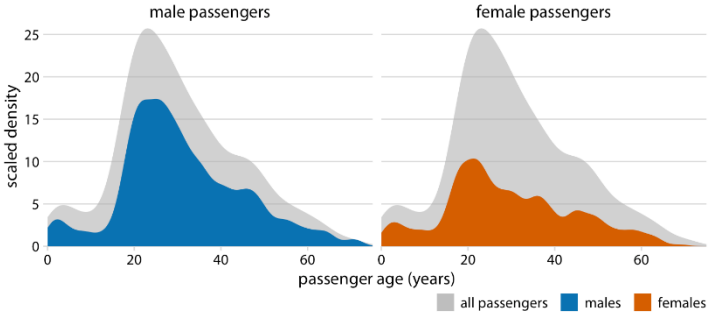
Esta visualização mostra de forma **intuitiva** e **clara** que havia muito menos mulheres do que homens na faixa etária de **20 a 50 anos** no Titanic

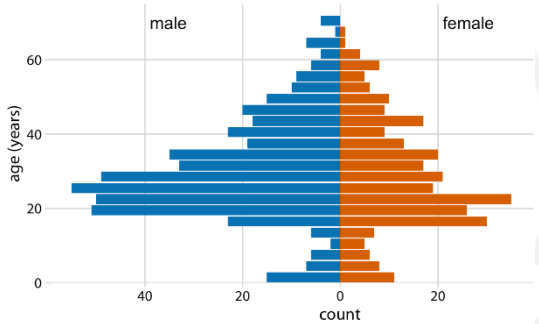
Visualizando várias distribuições ao mesmo tempo

Age pyramid

No entanto, quando queremos visualizar exatamente duas distribuições, podemos também fazer dois histogramas separados, girá-los 90 graus e fazer com que as barras de um histograma apontem na direção oposta do outro

- Este gráfico é chamado de *Age Pyramid*





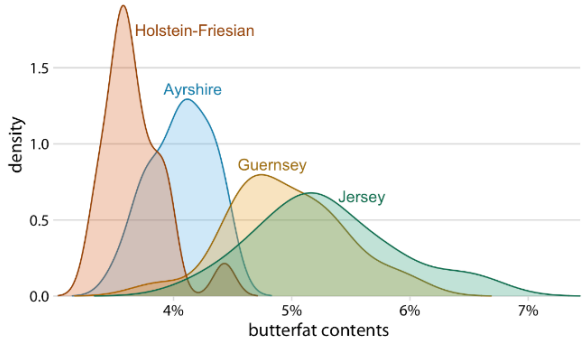
Fonte: (Wilke, 2019)

Visualizando várias distribuições ao mesmo tempo

Finalmente,

- Histograma não é adequado para distribuições múltiplas, a menos que sejam feitos separados (utilize o recurso de **pequenos múltiplos**)

- Neste caso, é melhor usar o gráfico de densidade, desde que as distribuições sejam um tanto distintas e contíguas.



Porcentagem de gordura da manteiga entre vacas de quatro raças diferentes de gado. **Fonte**: (Wilke, 2019)

Visualizando muitas distribuições de uma vez

Existem ainda muitos cenários em que desejamos visualizar várias distribuições ao mesmo tempo

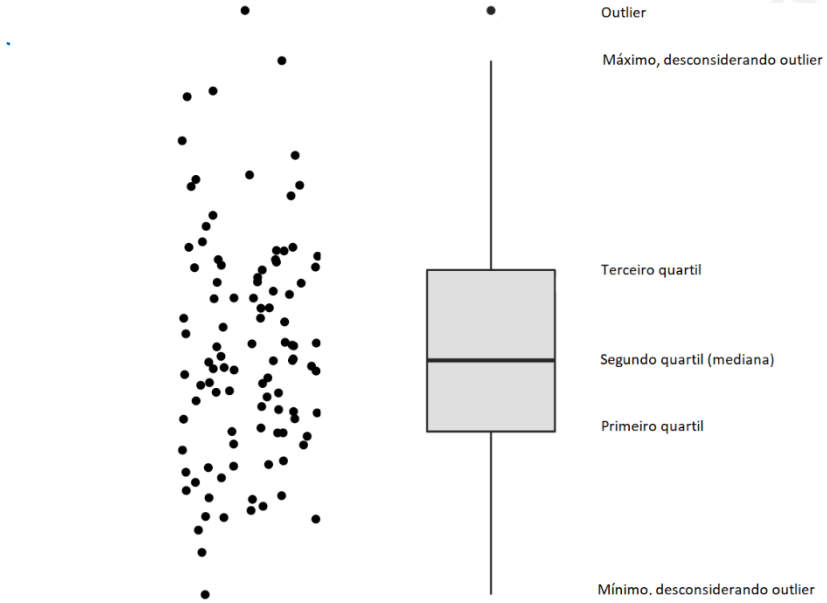
- Dados meteorológicos

Gráficos indicados:

1. Boxplots - criado por John Tukey em 1970
2. Plotagens de violino

Visualizando muitas distribuições de uma vez

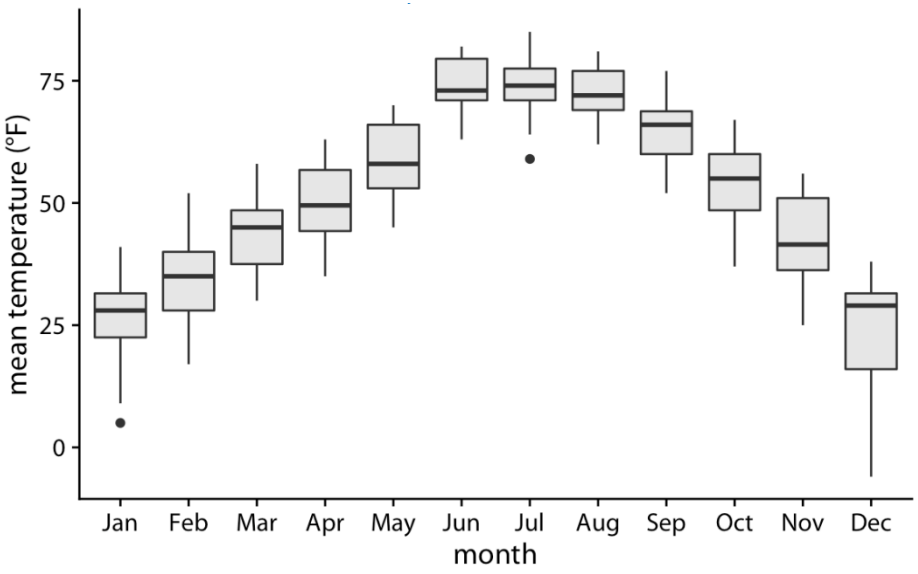
Boxplots



Fonte: Adaptada de (Wilke, 2019)

Visualizando muitas distribuições de uma vez

Boxplots



Fonte: (Wilke, 2019)

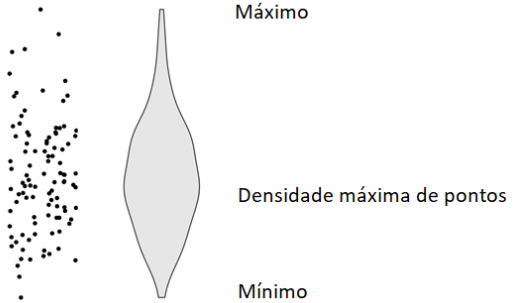
Em **dezembro**, podemos ver que a temperatura está altamente distorcida (a maioria dos dias são moderadamente frios e alguns são extremamente frios) e não muito distorcida em alguns outros meses, por exemplo, em **julho**

Visualizando muitas distribuições de uma vez

Violino

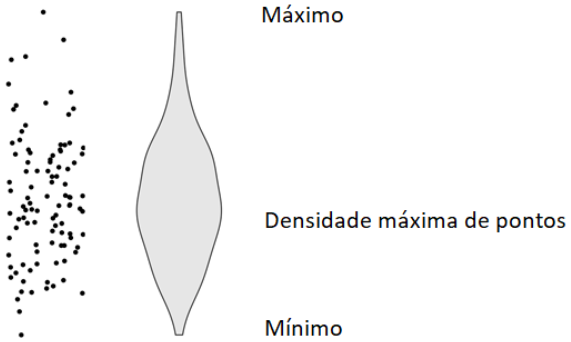
Gráficos violino são equivalentes às estimativas de densidade, giradas em 90 graus e depois espelhada

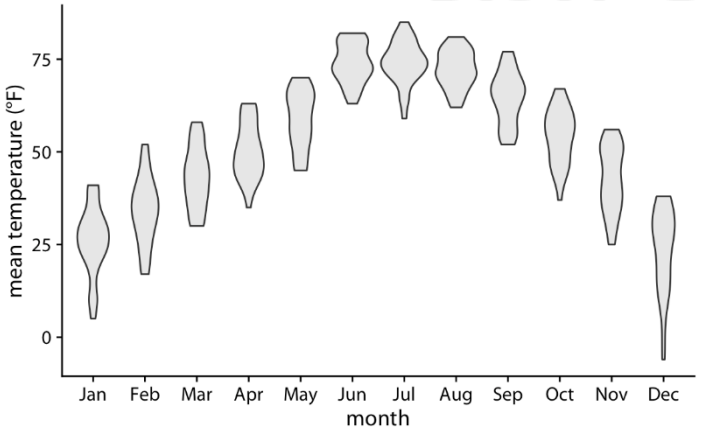
- A parte mais grossa do violino corresponde à **densidade de pontos mais alta** no conjunto de dados
- Em particular, os gráficos de violino representarão com **precisão os dados bimodais**, enquanto um boxplot não
- Muito apropriado para **grande quantidade de dados**



Visualizando muitas distribuições de uma vez

Violino

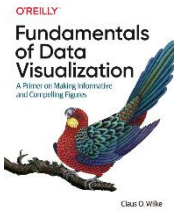




- o mês de **novembro** parece ter tido dois clusters de temperatura, um em torno de 50 graus e outro em torno de 35 graus

Fonte: (Wilke, 2019)

Leitura recomendada



Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures, 2019

Livros e Links:

1. CLAUS O. Wilke, Fundamentals of Data Visualization. https://clauswilke.com/dataviz/
2. R visualization workshop: https://stulp.gmw.rug.nl/26-04-2018/ggplotworkshop/
3. http://www.sthda.com/english/articles/32-r-graphics-essentials/129-visualizing-multivariate-categorical-data/
4. https://www.arbelatech.com/insights-resources/white-papers/zero-to-beautiful-choosing-charts-for-data-visualization
5. https://www.klipfolio.com/resources/articles/what-is-data-visualization
6. https://aniruhil.org/courses/mpa6020/handouts/module02.html
7. https://coggle.it/diagram/X2VVUk4r0hGyP1Ic/t/s%C3%A9ries-temporais-1