GENE7033 – Tópicos Especiais em Genética I:

Visualização de dados para publicações científicas

Profª Drª Chirlei Glienke

Drª Desirrê Petters-Vandresen

Visualizando múltiplas distribuições

Drª Desirrê Petters-Vandresen

17/11/2022

Finalidade

- Visualização de como uma variável se distribui e se comporta em diversos conjuntos de dados simultaneamente

- Por exemplo, variação de temperatura ao longo de um ano
  - Representar adequadamente as distribuições de temperaturas observadas dentro de cada mês
  - Comparar adequadamente as distribuições entre diferentes meses

- Inviabilidade de utilizar histogramas, gráficos de densidade, ecdfs ou gráficos Q-Q

- Alternativas: diagrama de caixa (boxplot), gráfico de violino, gráfico ridgeline

Visualizando múltiplas distribuições

- Raciocínio útil: variável de resposta vs. variáveis de grupo, cada uma em um eixo

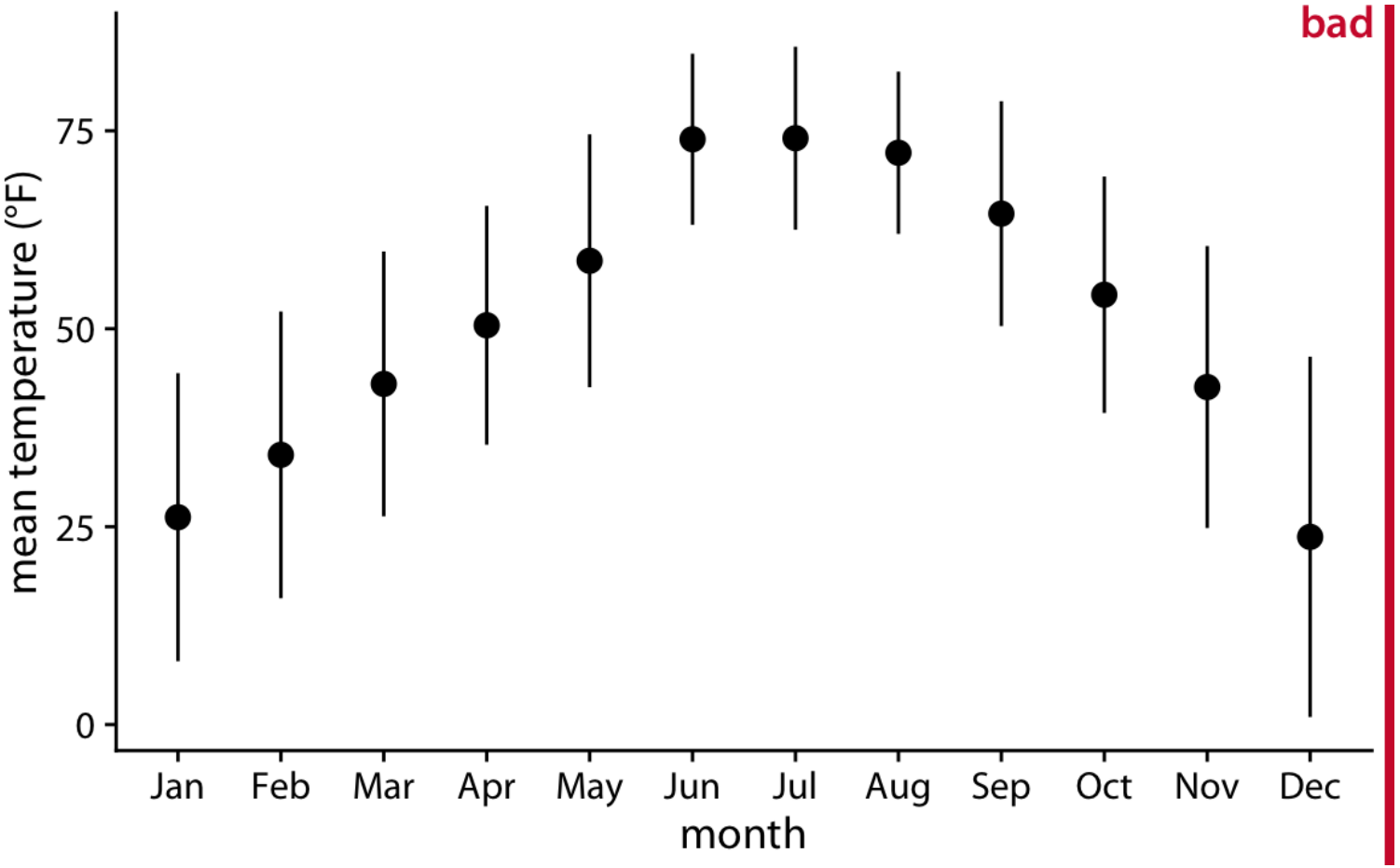
- Variável de resposta: variável relacionada às distribuições que queremos representar (temperatura)

- Variável de grupo: subconjuntos dos dados com distribuições distintas da variável de resposta (meses do ano, estações do ano)

- Em geral:
  - Variável de resposta no eixo vertical: boxplot e gráfico de violino
  - Variável de resposta no eixo horizontal: gráfico ridgeline

Eixo vertical

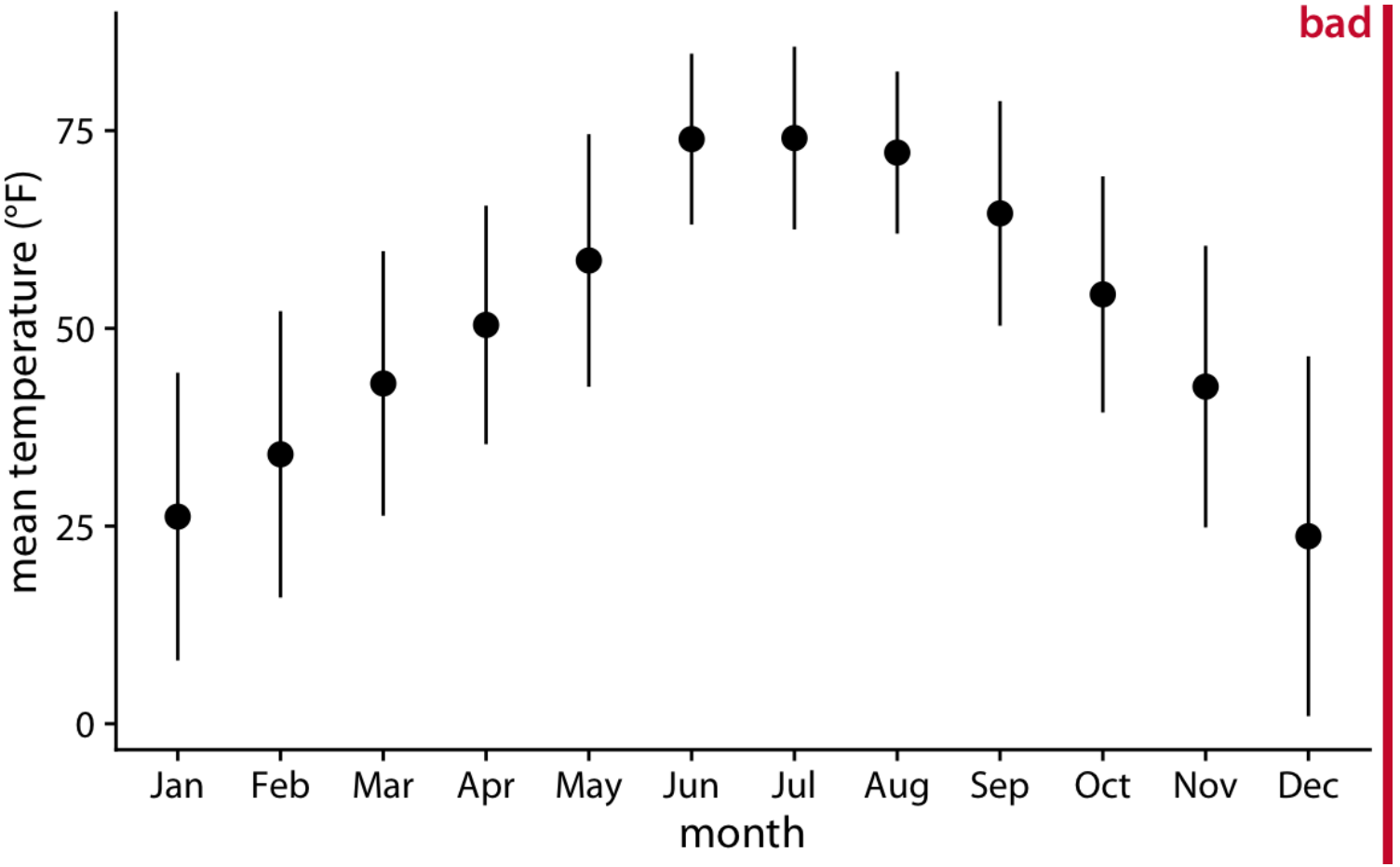
- Abordagem mais simples: apresentar a média ou mediana como informação principal, e indicar a variação em torno do valor com barras de erro



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo vertical

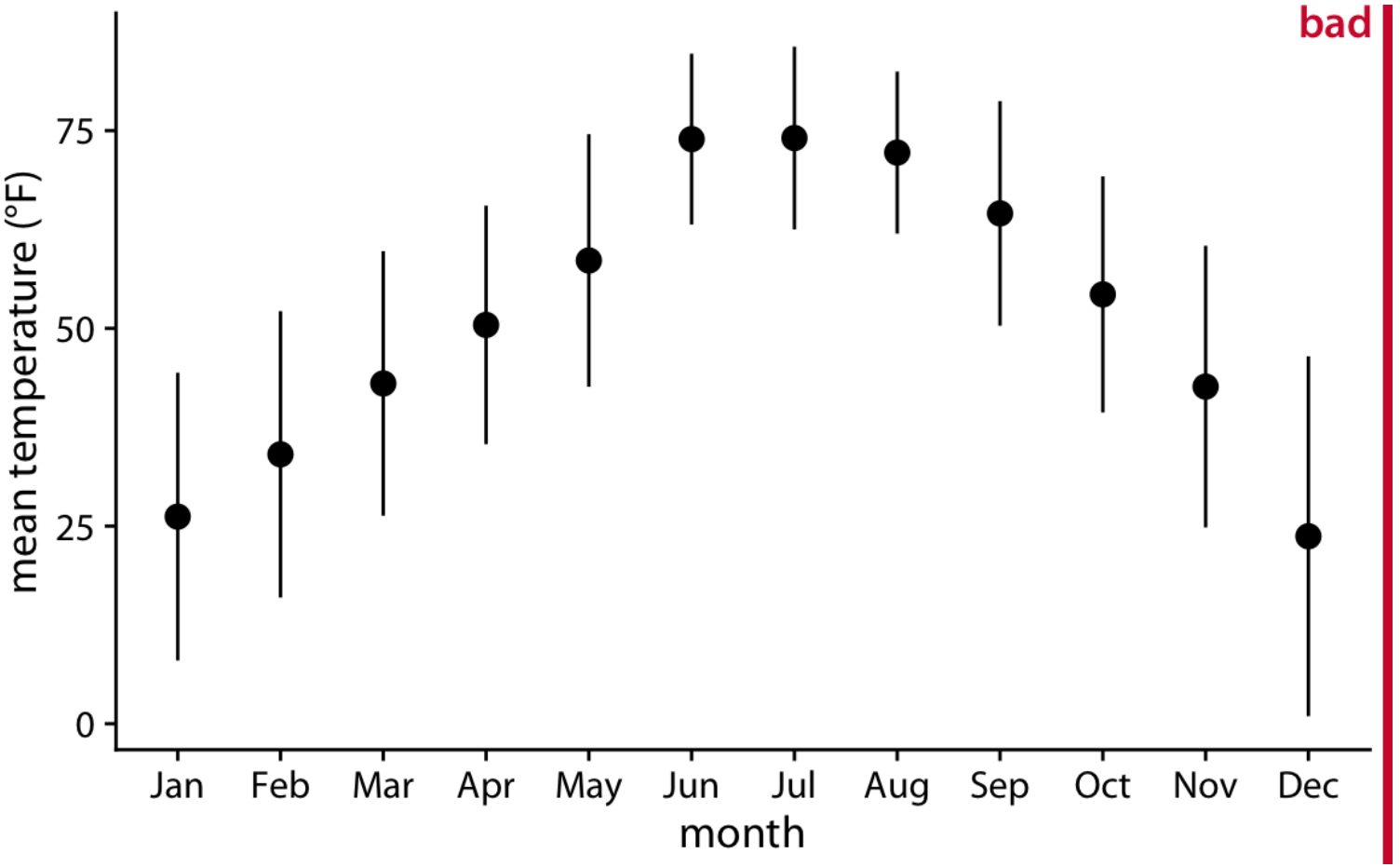
- Problemas
  - Representação muito simplificada dos dados
  - O que os pontos representam média ou mediana?
  - O que as barras de erro representam: desvio padrão ou erro padrão, intervalo de confiança?
  - Erros de interpretação se houver outliers (o que quase sempre acontece)



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo vertical

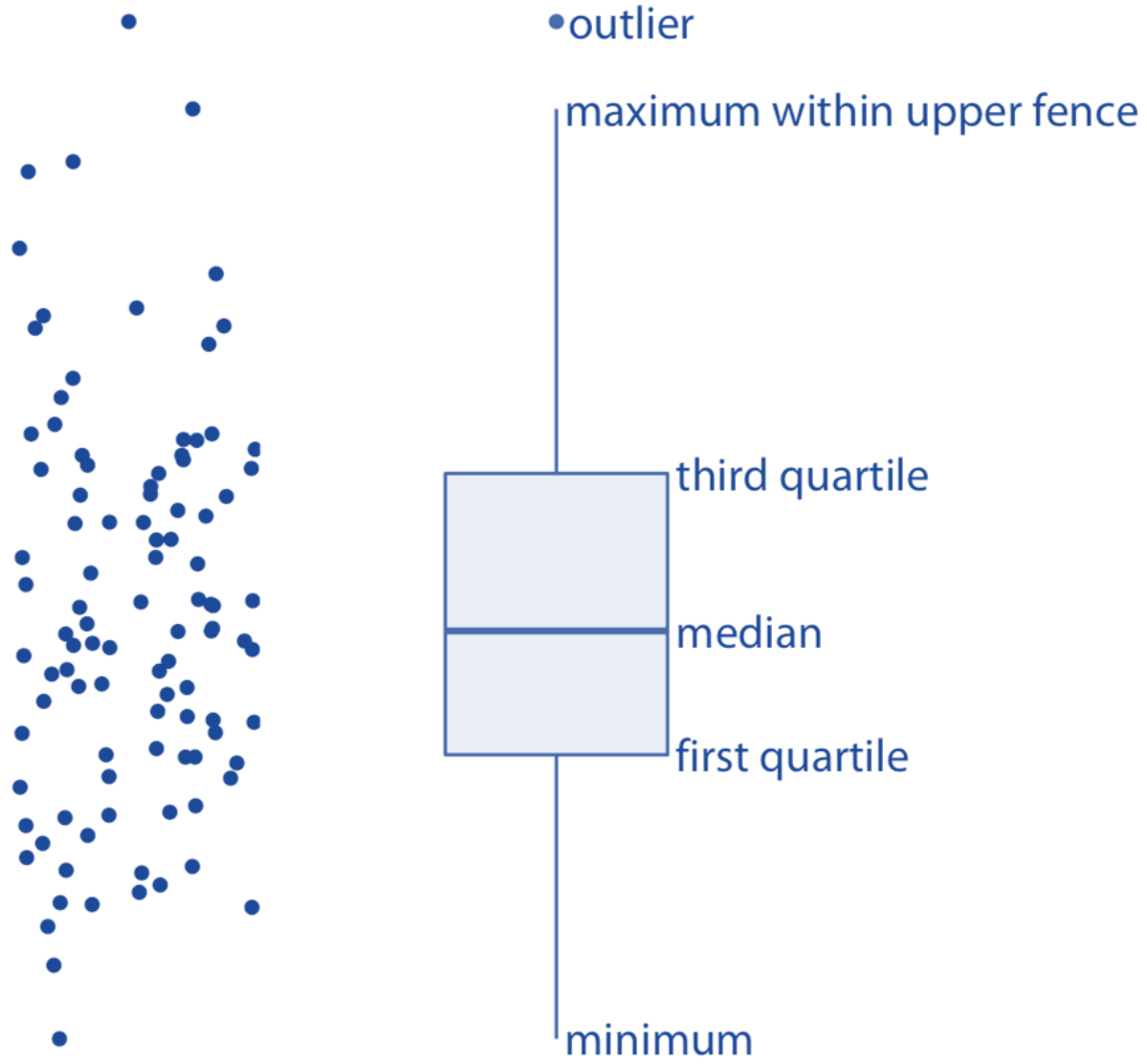
- Em geral, barras de erro indicam erro padrão, facilmente confundidas com desvio padrão
- Desvio padrão: dispersão ao redor da média
- Erro padrão: quão precisa é a estimativa da média
- Conjuntos de dados podem ter erro padrão baixo e desvio padrão alto



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

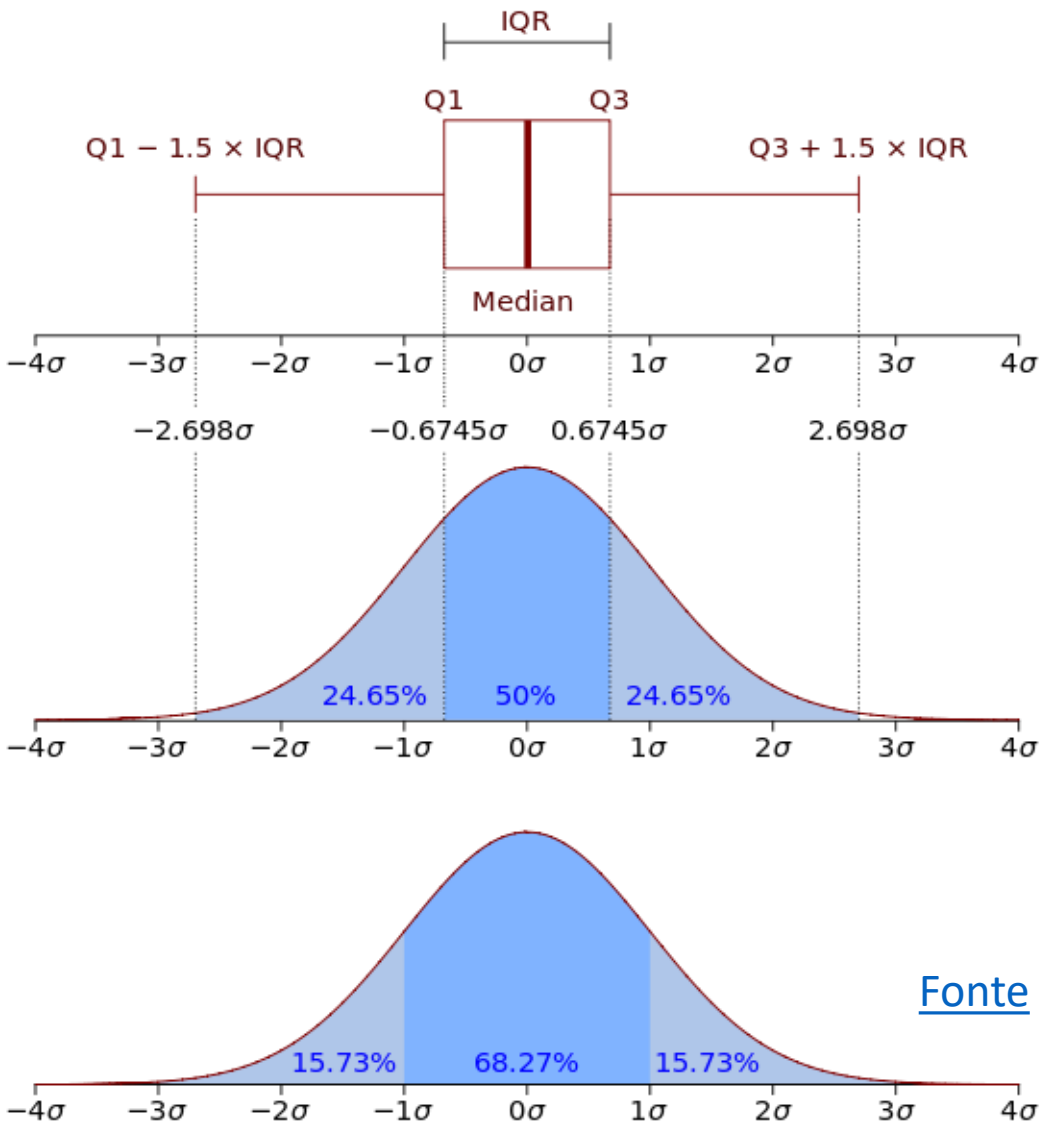
Eixo vertical: boxplot

- Divisão do conjunto de dados em quartis (quatro partes com 25% das observações em cada)
- Mediana (Q2): valor central, 50% das observações serão iguais ou maiores, 50% serão iguais ou menores
- Amplitude interquartil (Q1 – Q3): 50% das observações
- Limite inferior teórico: Q1 – 1,5 x AIQ
- Limite superior teórico: Q3 + 1,5 x AQI

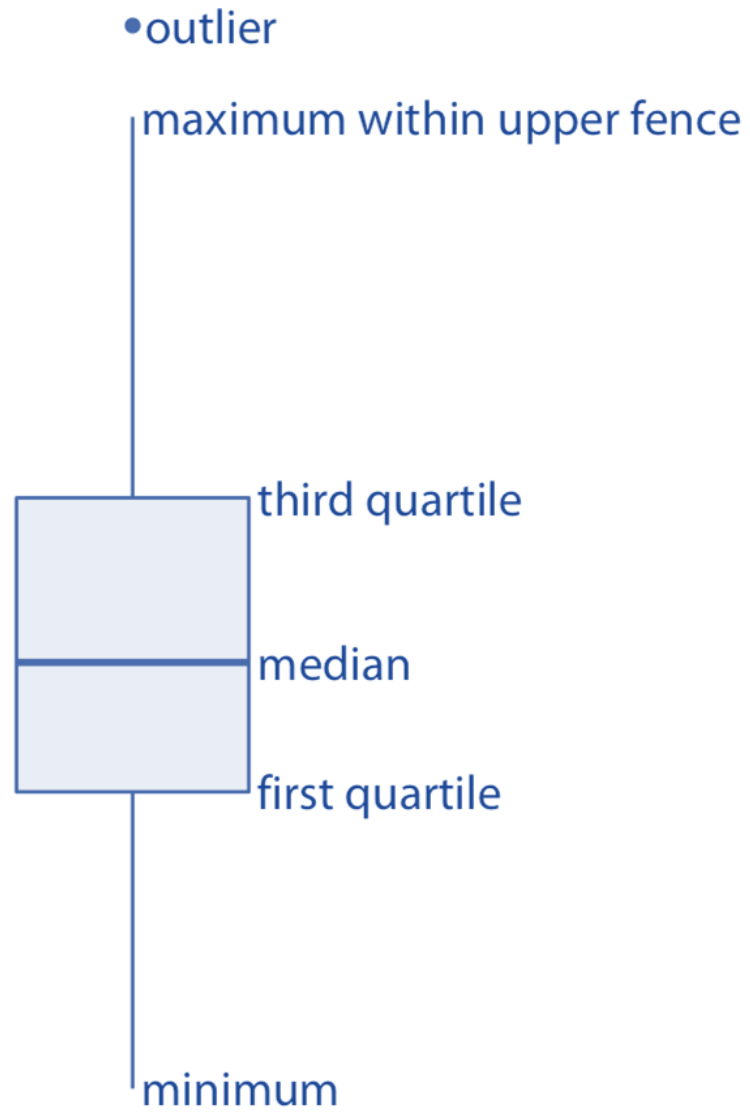


Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo vertical: boxplot

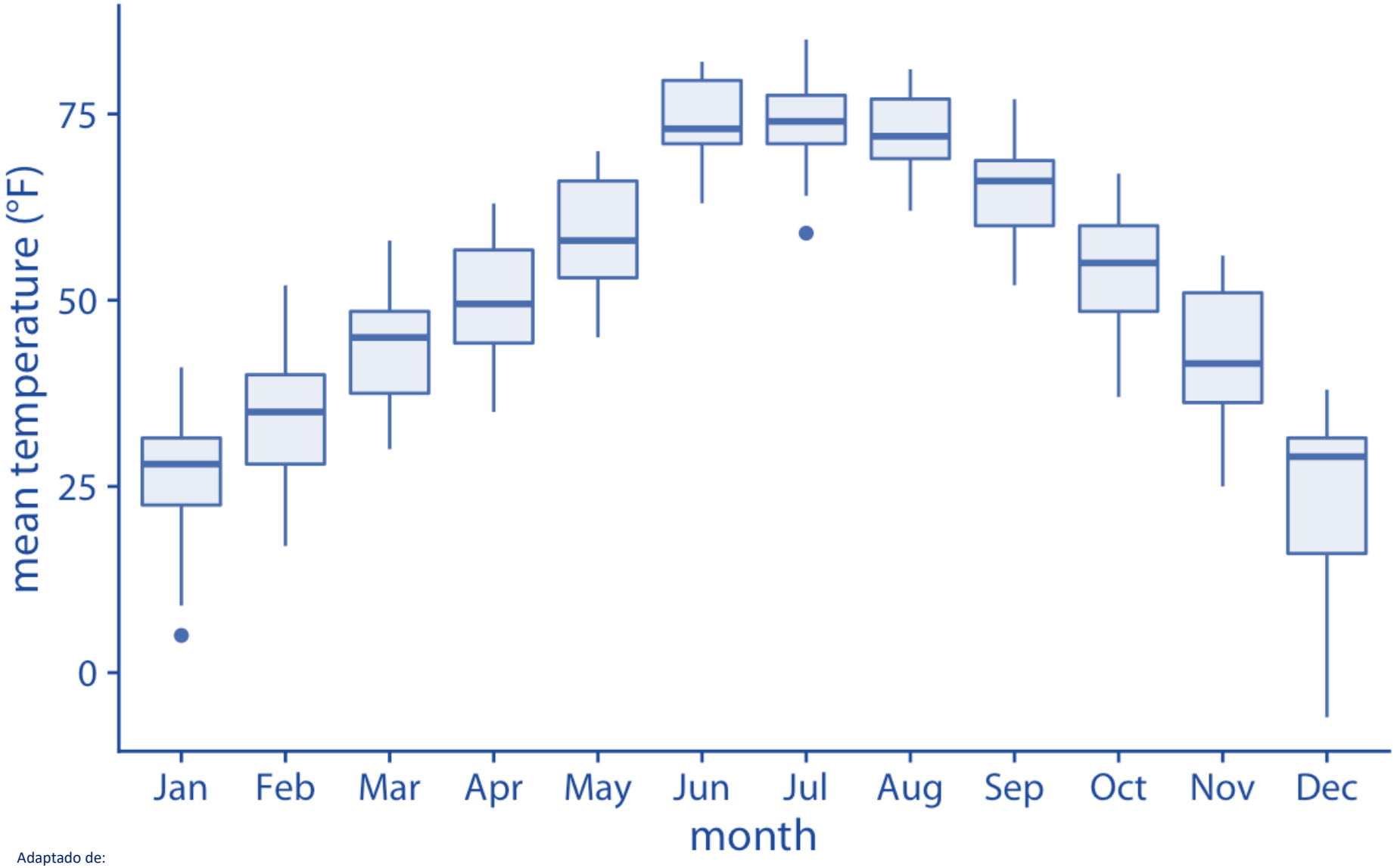


Fonte



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo vertical: boxplot



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

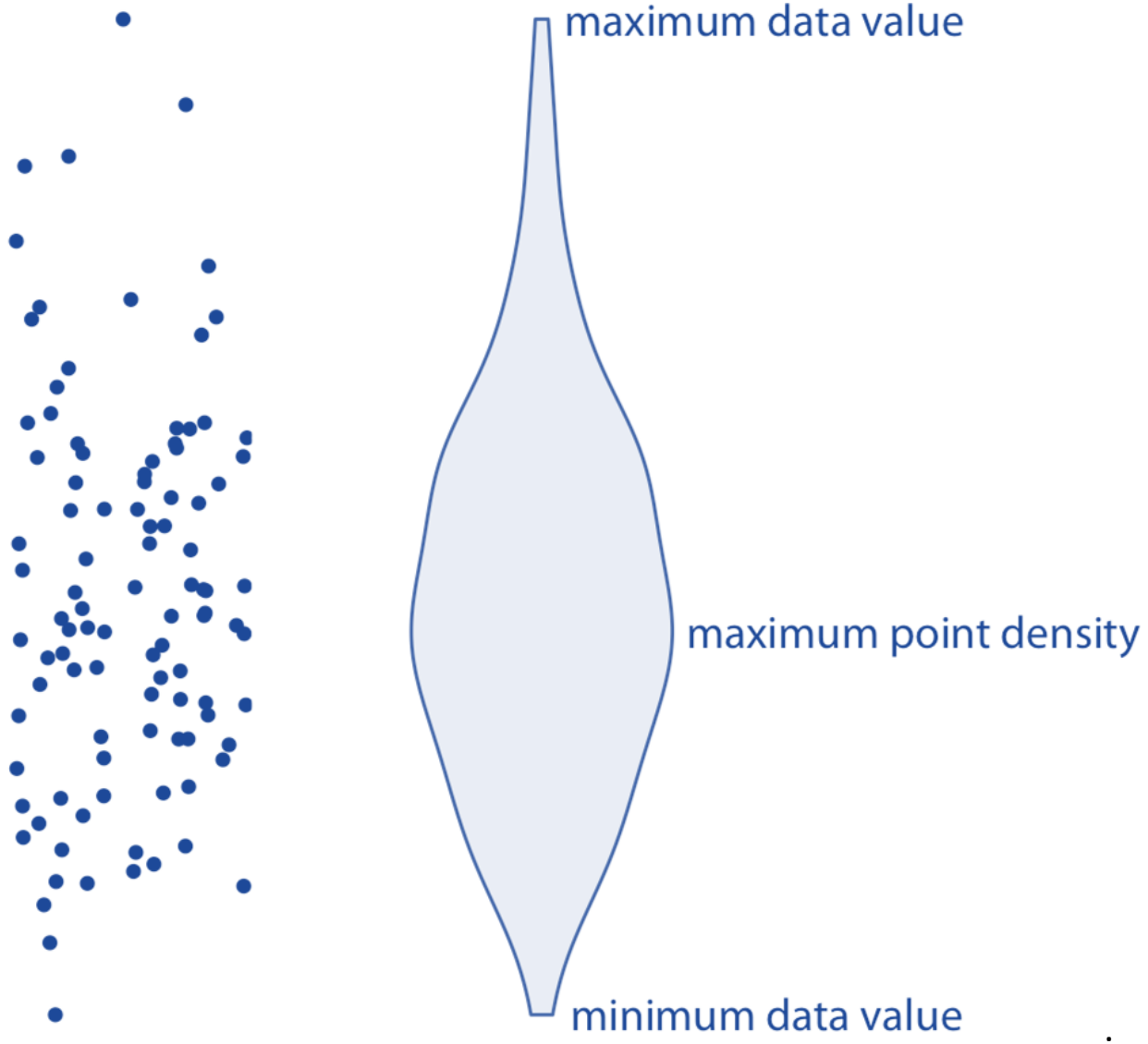
Eixo vertical: boxplot

- Grande popularidade principalmente pela facilidade de construção do gráfico:
  - Inventados pelo estatístico John Tukey na década de 1970, podiam ser facilmente desenhados à mão

- Altamente informativos:
  - Observar o quanto a distribuição varia ao redor da mediana
  - Detectar a presença de dados altamente enviesados

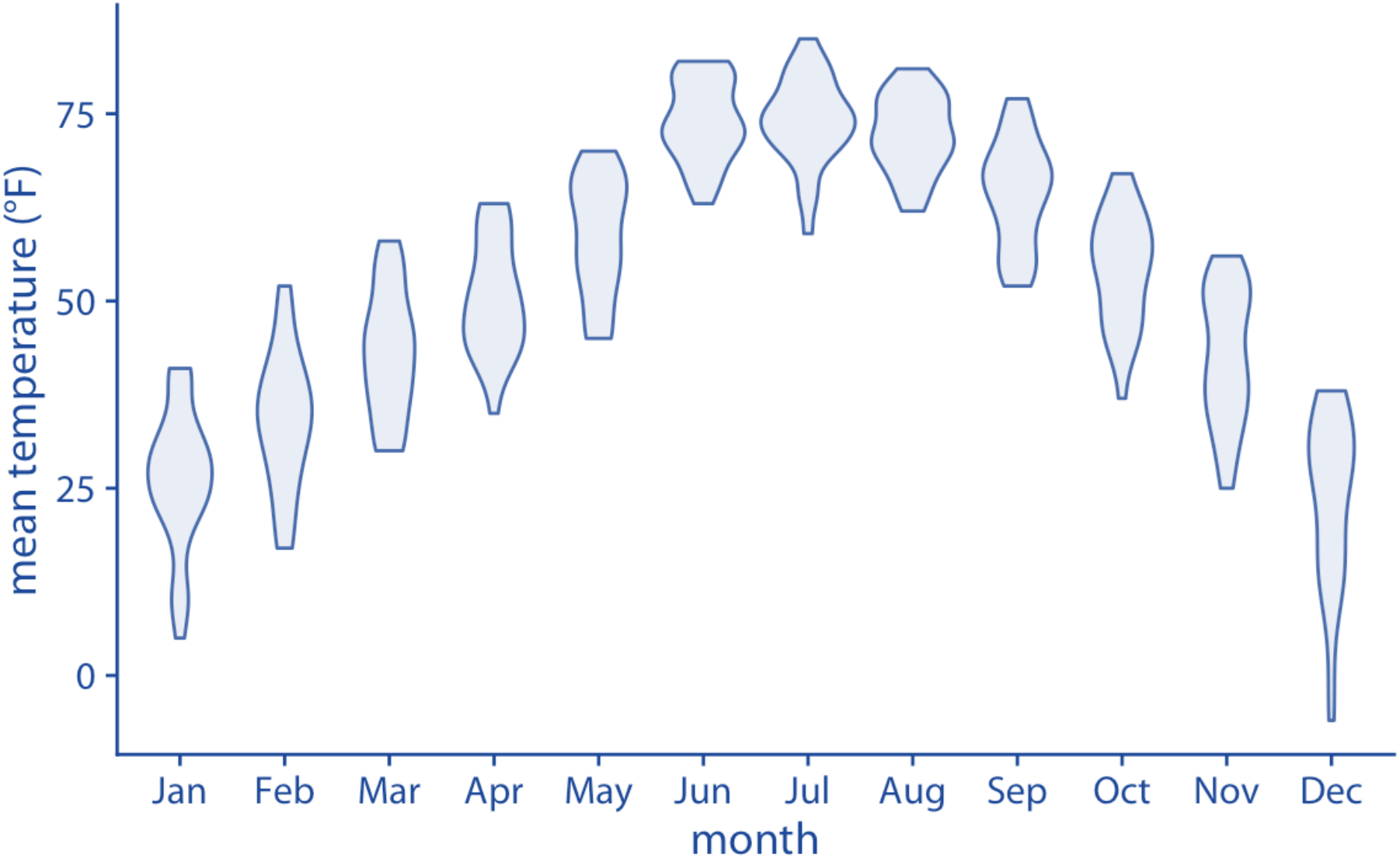
Eixo vertical: gráfico de violino

- Semelhante à gráficos de densidade dispostos de forma perpendicular e espelhados, simétricos
- Utilizados nos mesmos contextos que o boxplot, mas de forma mais suavizada
- Vantagem: representação fiel de dados bimodais, que não é feita adequadamente em um boxplot
- Desvantagens similares às dos gráficos de densidade:
  - Aparência de existência de dados quando não existem (caudas)
  - Aparência de conjunto de dados denso, quando é esparso



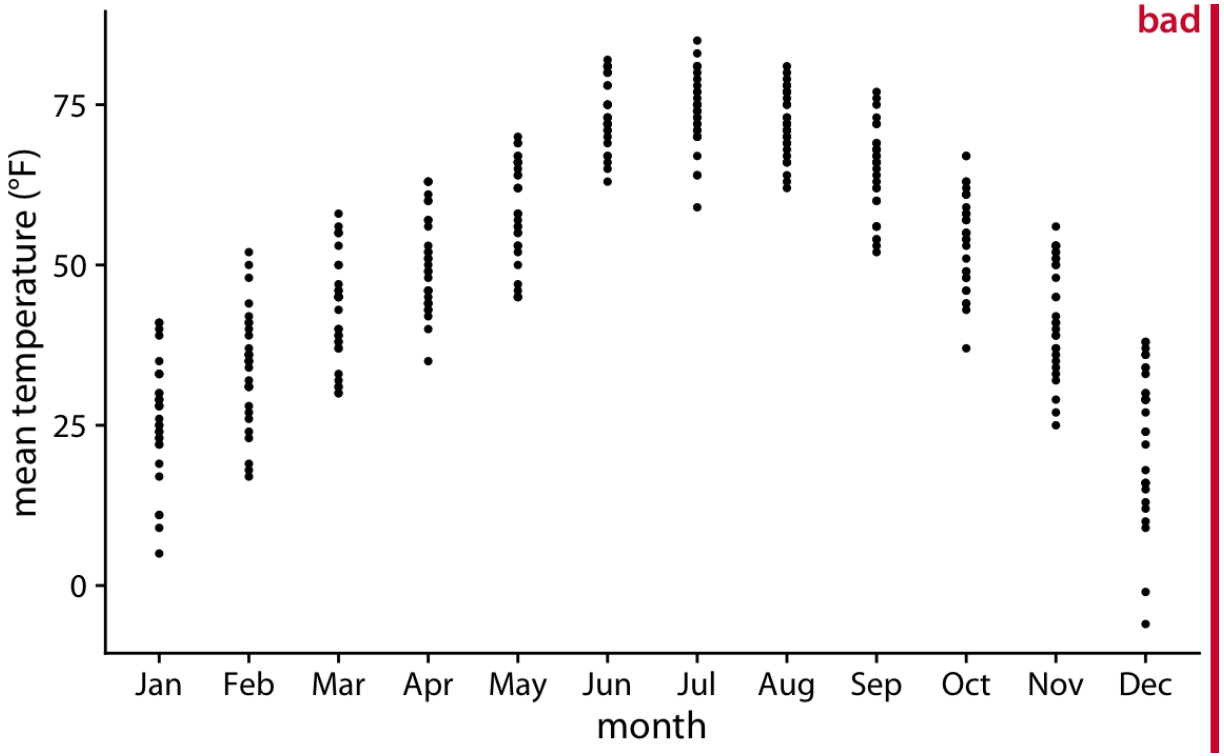
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

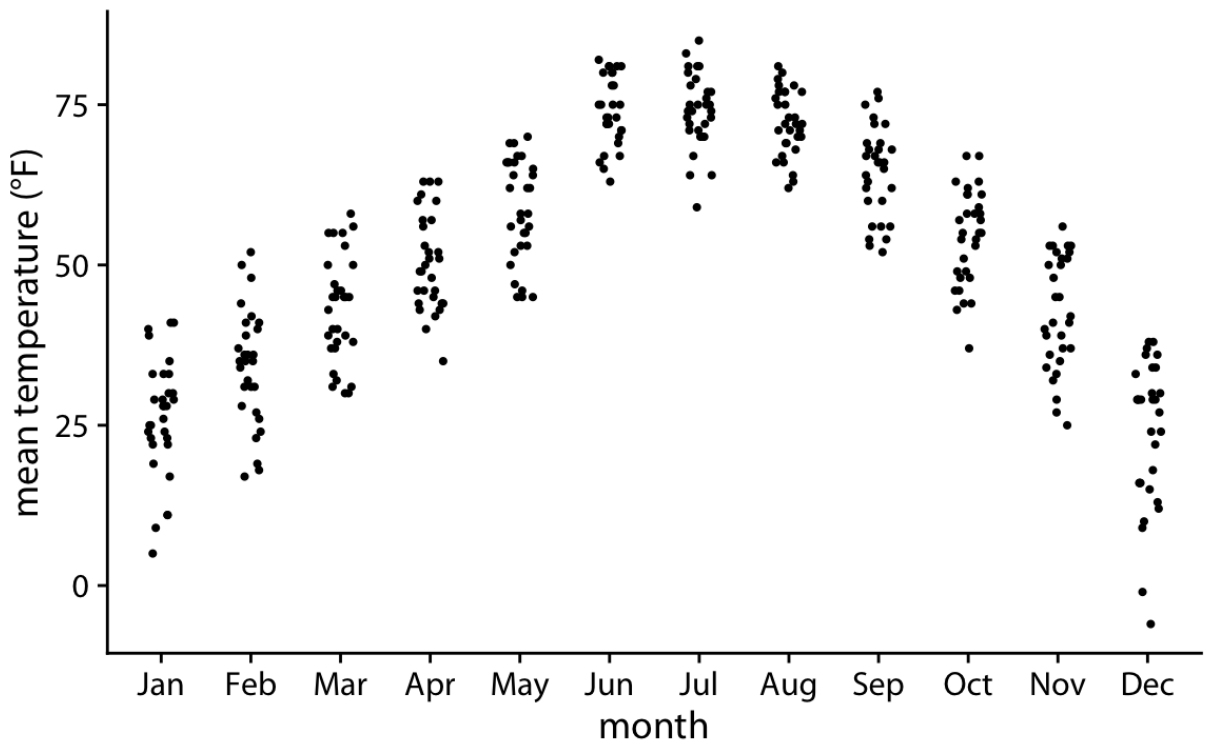
Eixo vertical: gráfico de violino



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo vertical: gráfico de fitas (*strip chart*)





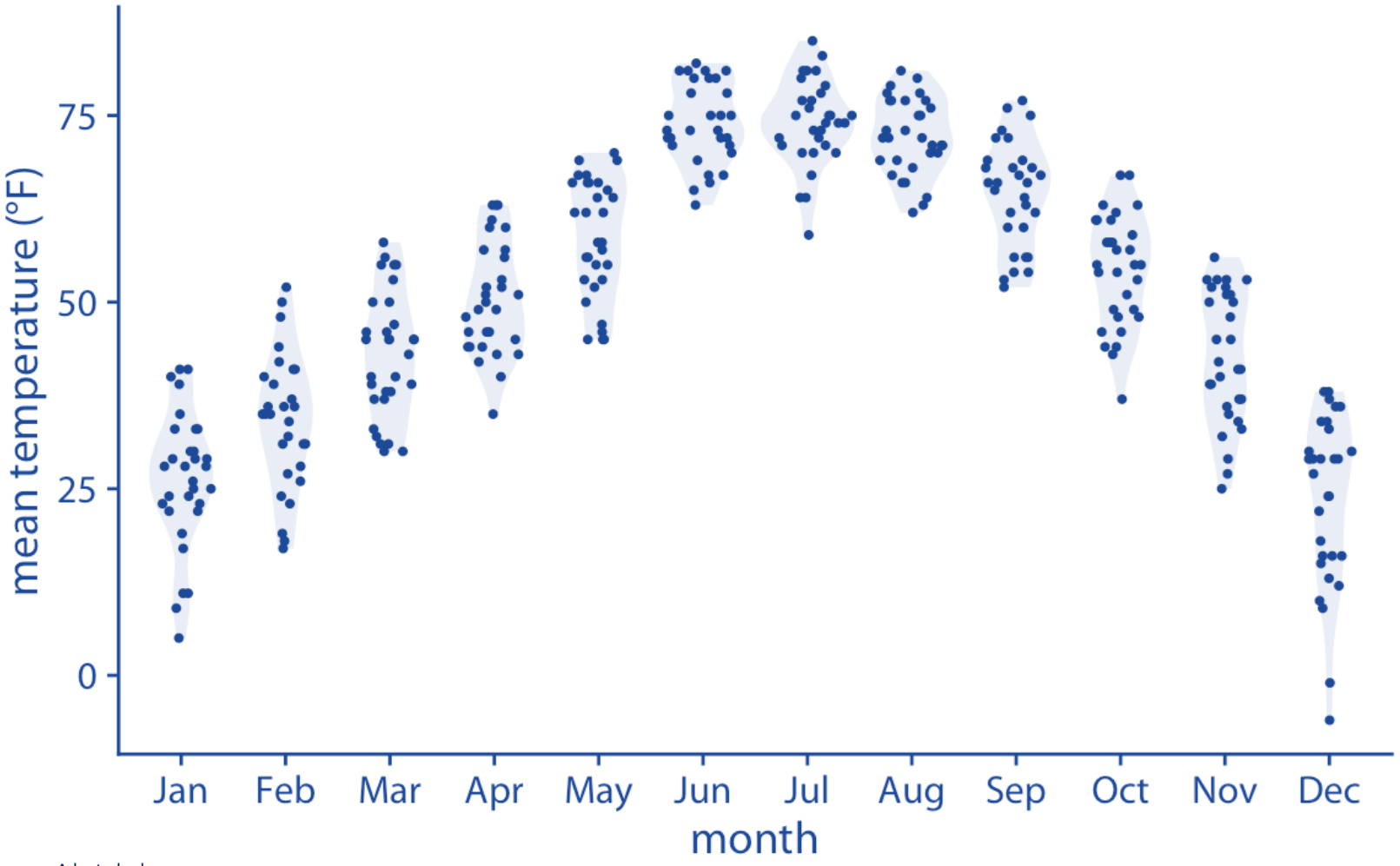
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

- Evitar sobreposição de pontos com espalhamento dos pontos aleatoriamente ao longo do eixo x (jittering)

Eixo vertical: sina plot

- Sina plot: homenagem à Sina Hadi Sohi, primeira versão do código utilizado para produzir os gráficos sina

- Híbrido entre gráficos de violino e espalhamento de pontos proporcional à densidade dos pontos naquela posição



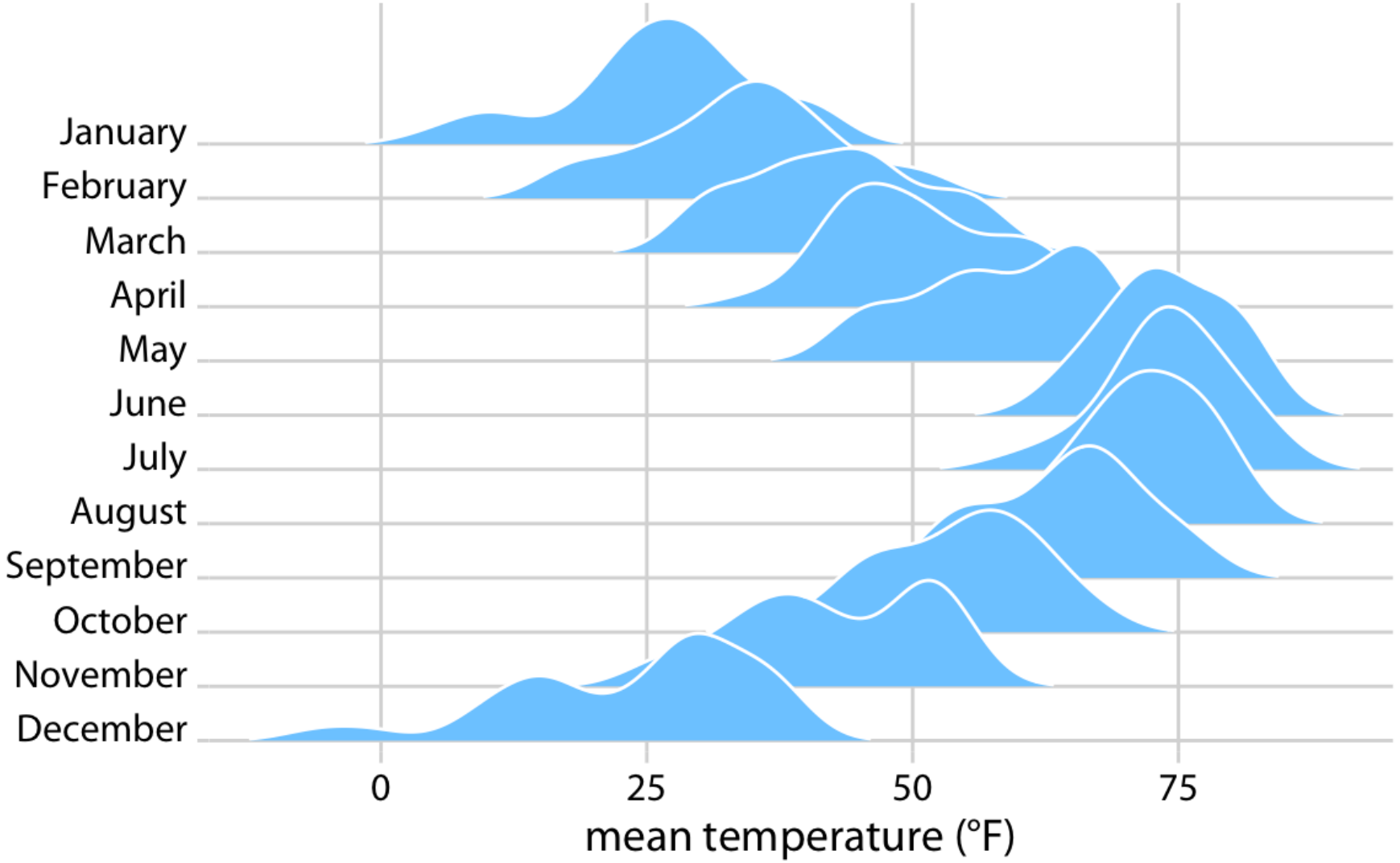
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo horizontal: ridgeline plot

- Utilização de gráficos de densidade dispostos verticalmente, com aspecto de encostas de montanhas

- Úteis para demonstrar tendências nas distribuições ao longo do tempo

Eixo horizontal: ridgeline plot



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

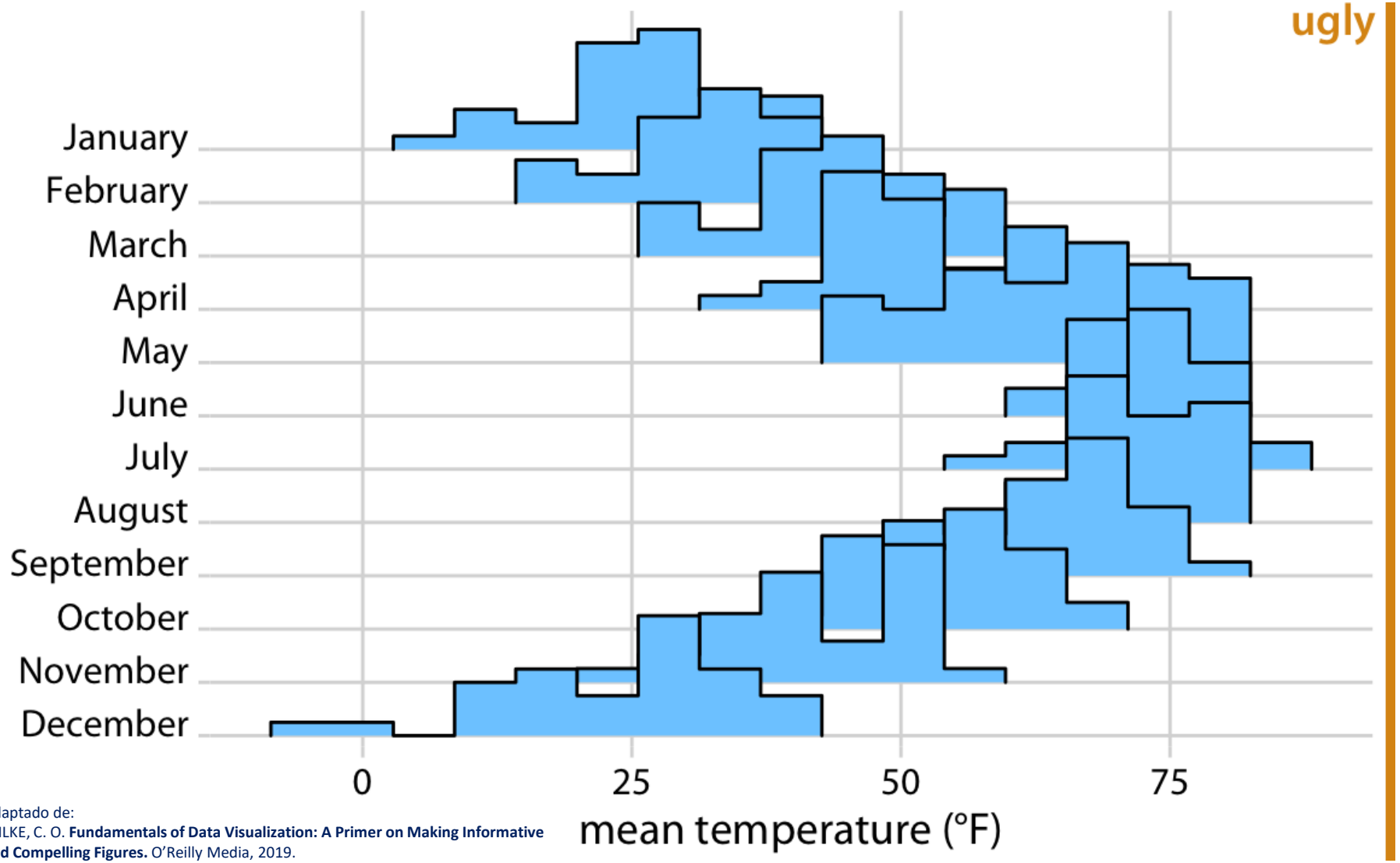
Eixo horizontal: ridgeline plot

- Eixo x: variável de resposta
- Eixo y: variável de grupo

- Sem eixo adicional/escala explícita para a magnitude das densidades observadas, são plotadas proporcionalmente dentro do espaço de cada variável de grupo

- Foco na comparação entre as formas e alturas relativas das densidades entre os grupos, e não em valores específicos de densidade

Eixo horizontal: ridgeline plot com histogramas



Onde começa e termina cada histograma quando há sobreposição?

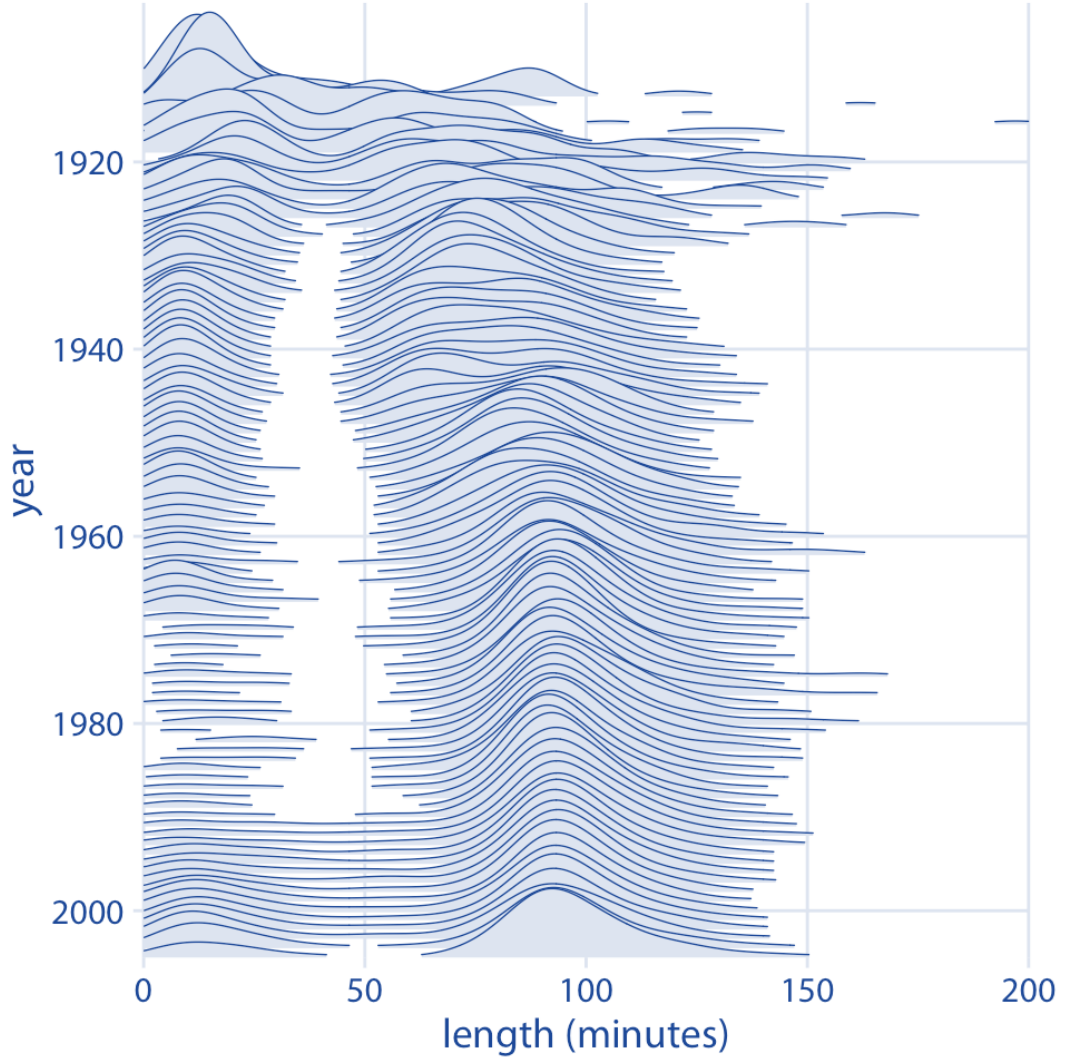
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo horizontal: ridgeline plot

- Facilidade de apresentar uma quantidade imensa de distribuições simultaneamente, sem prejuízo de interpretação

- Distribuição da duração de filmes ao longo das décadas

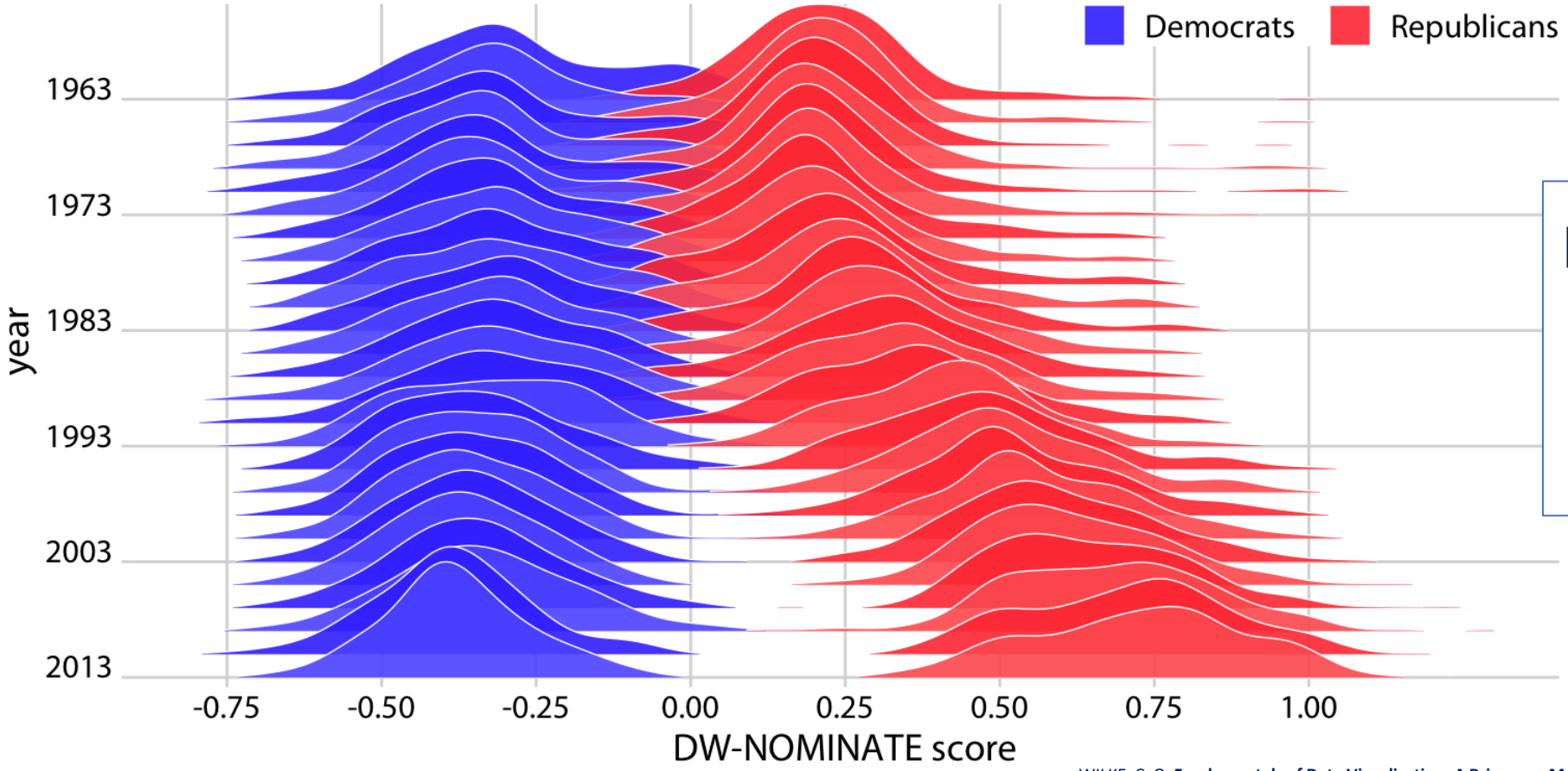
  - 1920: filmes de duração bastante variada
  - 1960 em diante: duração em torno de 90 minutos



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Eixo horizontal: ridgeline plot

- Comparação de duas tendências simultaneamente ao longo do tempo: codificação por cor



Padrão de votação dos membros do congresso dos EUA

Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.