Data Science II

- Introduction to Data Visualization -

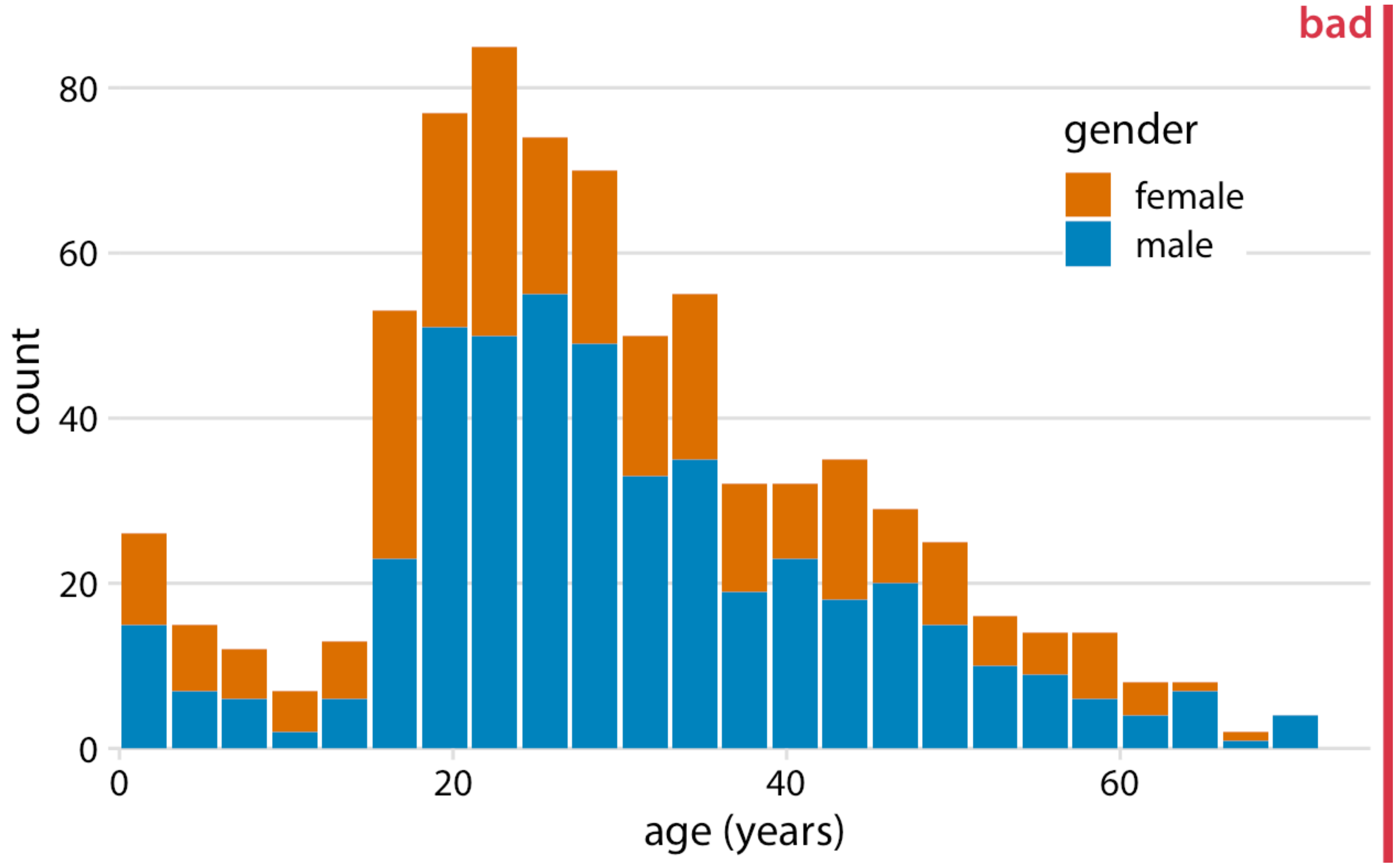
Visualizing Multiple Distributions Simultaneously

Prof. Dr. Eduard Kromer
Summer Semester 2024
University of Applied Sciences Landshut

Visualizing Multiple Distributions Simultaneously

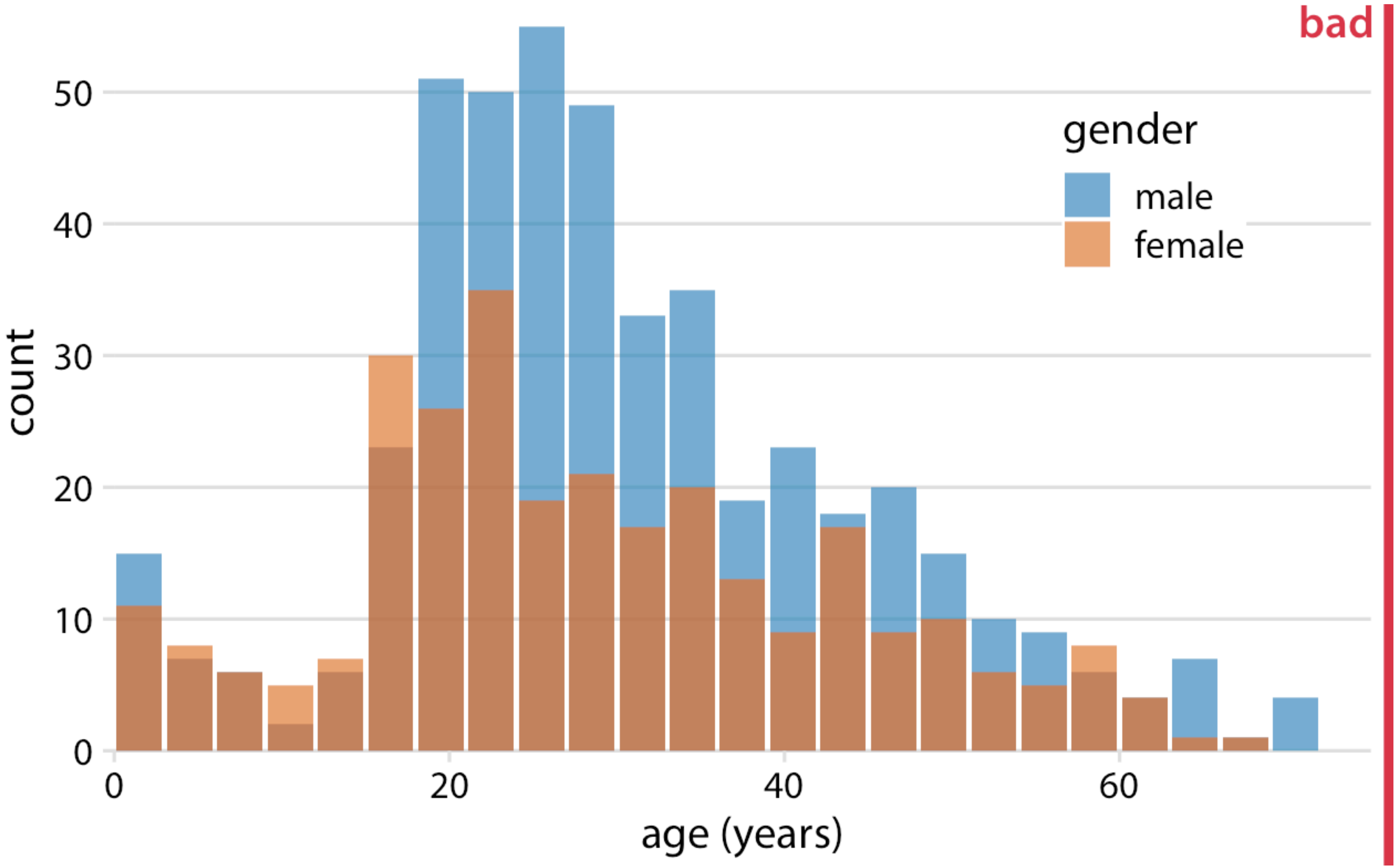
Visualizing Mutliple Distributions Simulateneously

One commonly employed strategy for the simultaneous visualization of multiple distributions are stacked histograms:

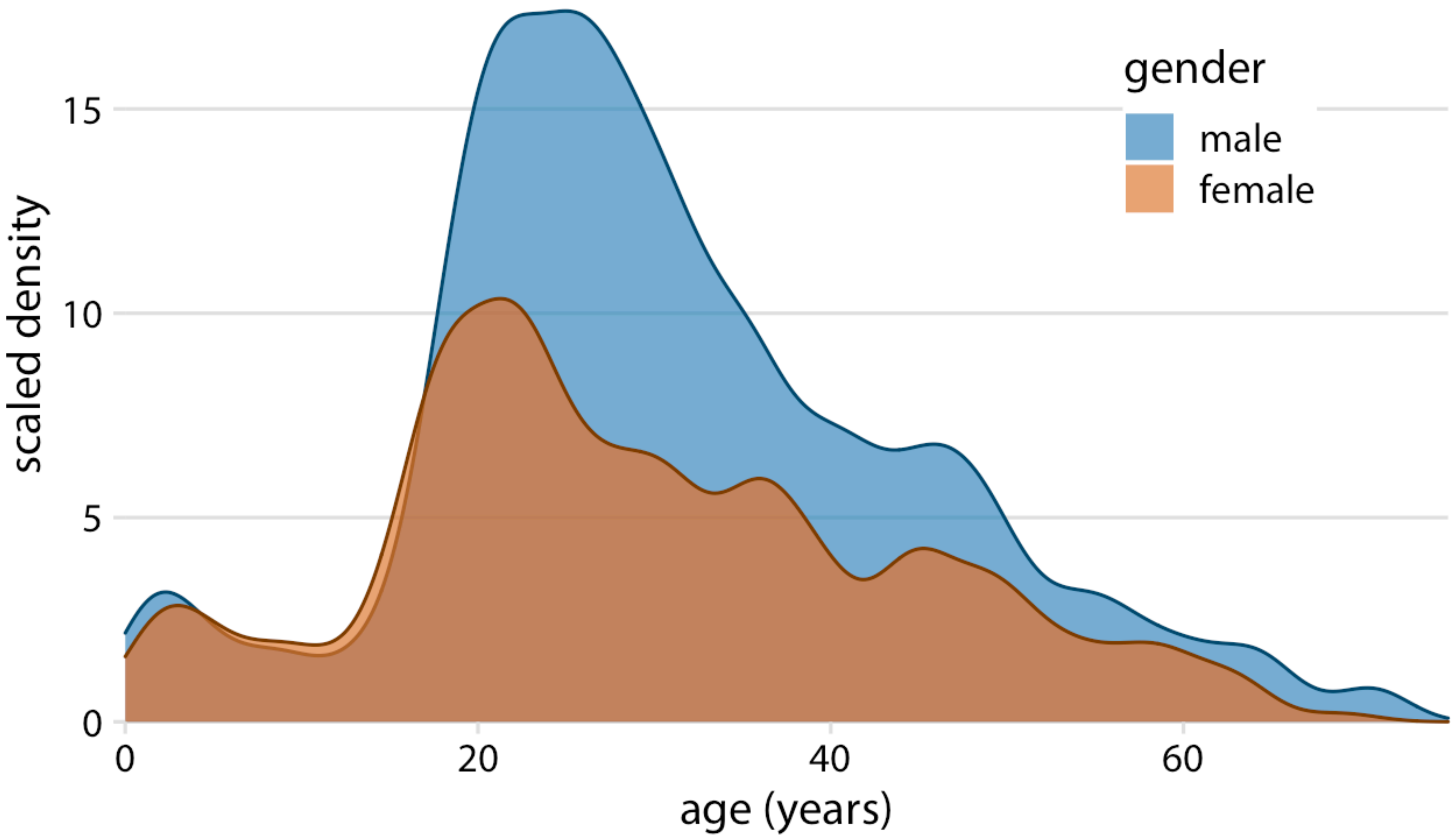


- there are several problems with this approach:
    - not clear where exactly the bars begin (where the color changes or at zero)
    - bar heights of orange bars cannot be compared to each other

Why is this a bad visualization?



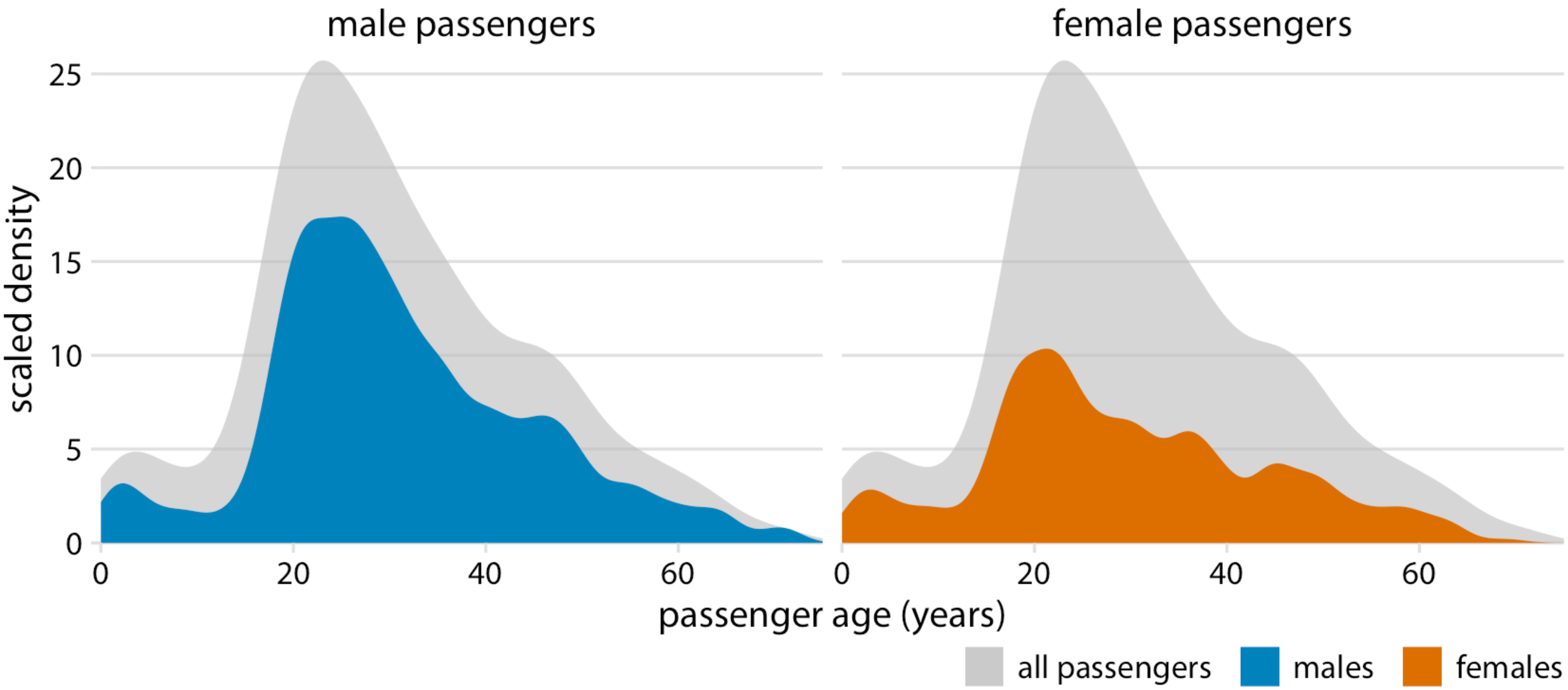
Overlapping Density Plots



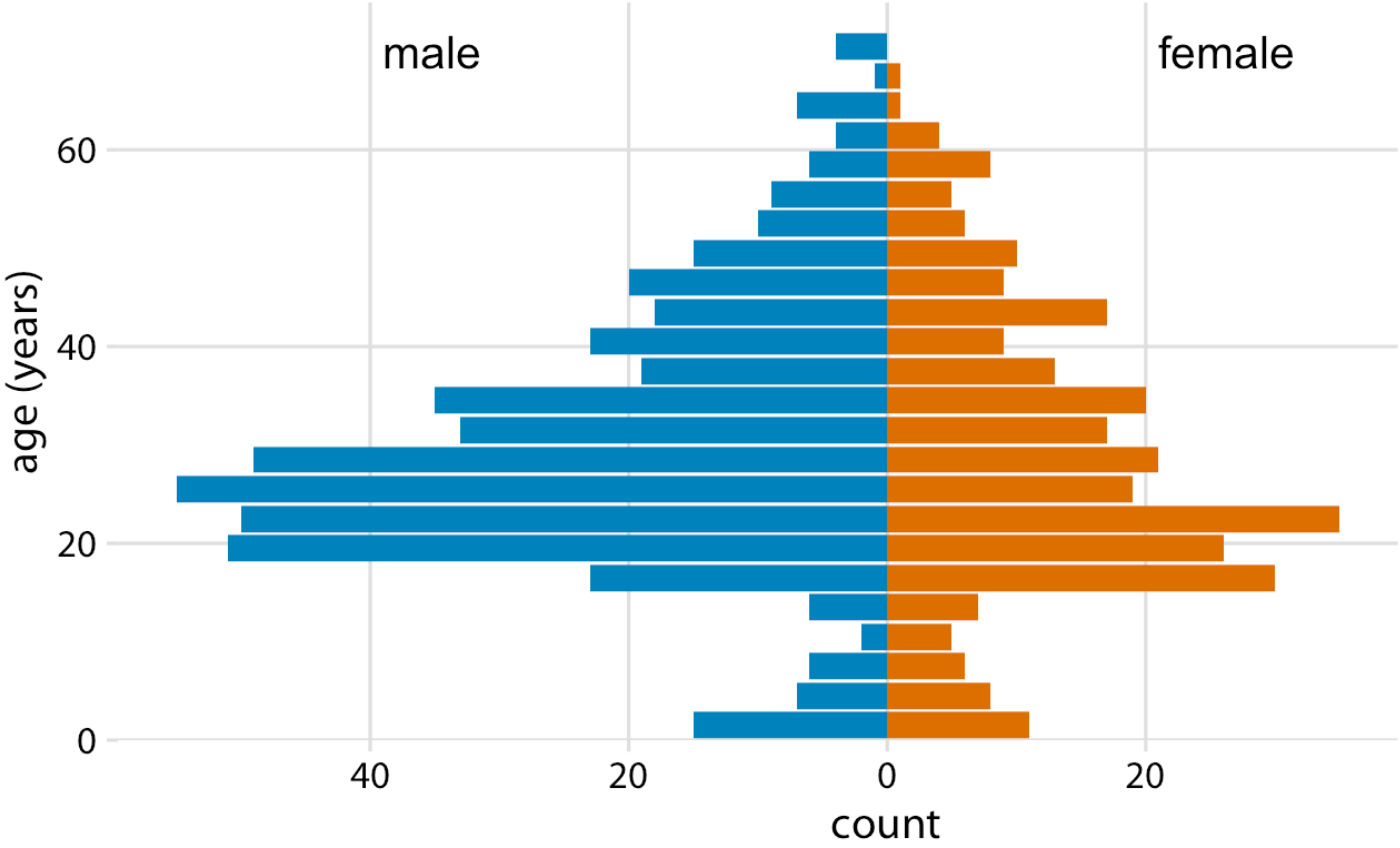
- continuous density lines help the eye keep the distributions separate

Why is this visualization still not ideal?

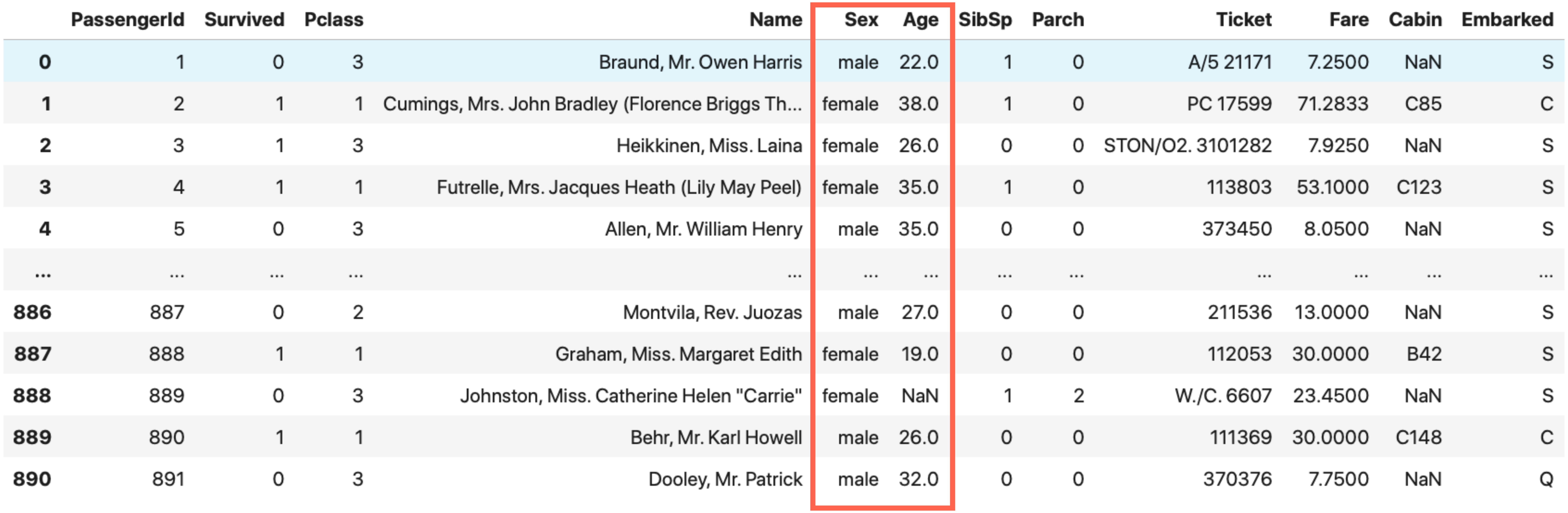
Overlapping Density Plots



The Age Pyramid



Generating the Age Pyramid with matplotlib



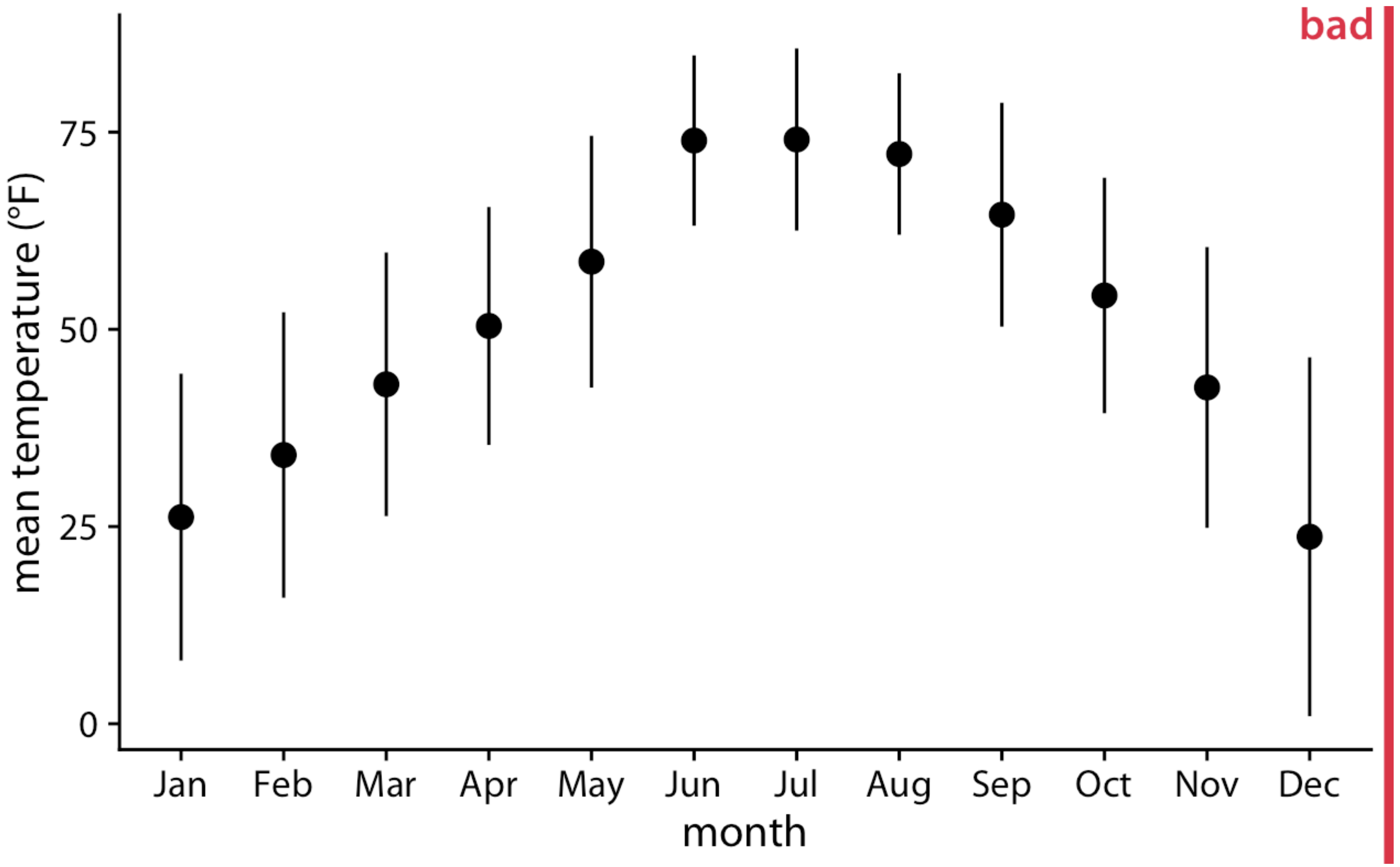
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

- use the data from columns ‘*Sex*’ and ‘*Age*’ of the Titanic data set and generate the corresponding age pyramid with `seaborn.barplot`

Visualizing Distributions Along the Vertical Axis

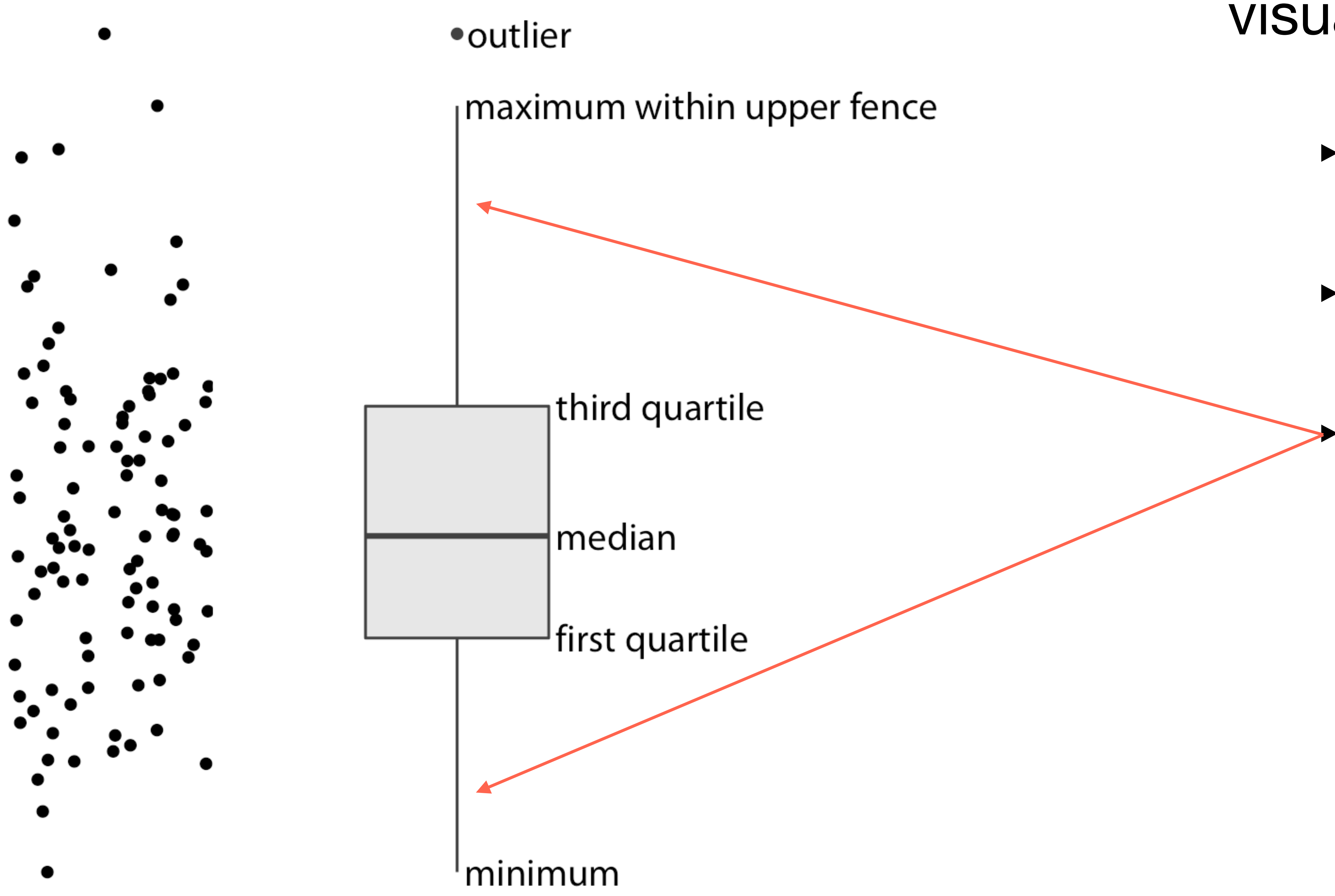
Visualizing Distributions Along the Vertical Axis

Simplest approach: show mean / median as points and variation around mean / median by error bars.



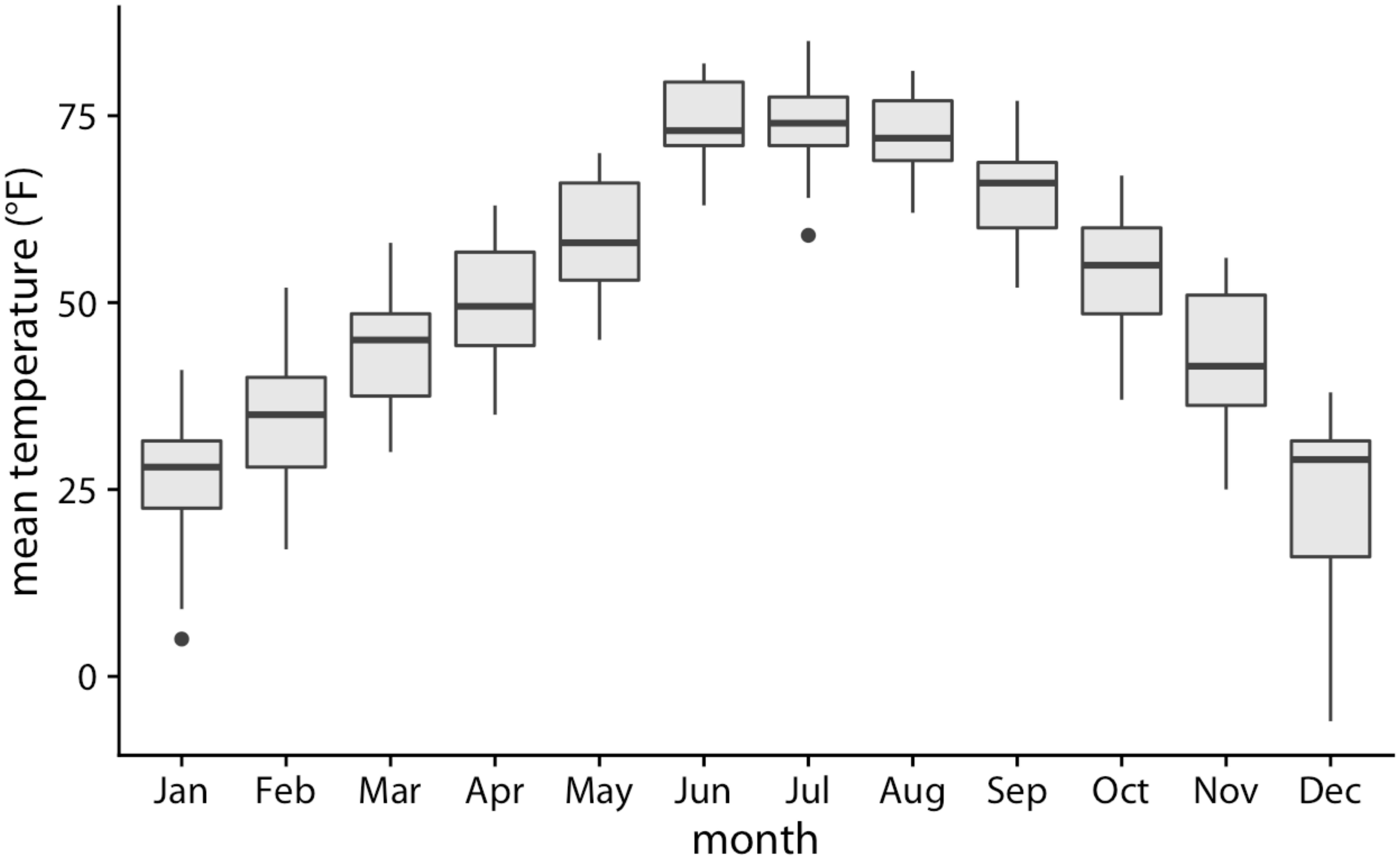
Why is this a bad visualization?

Boxplots

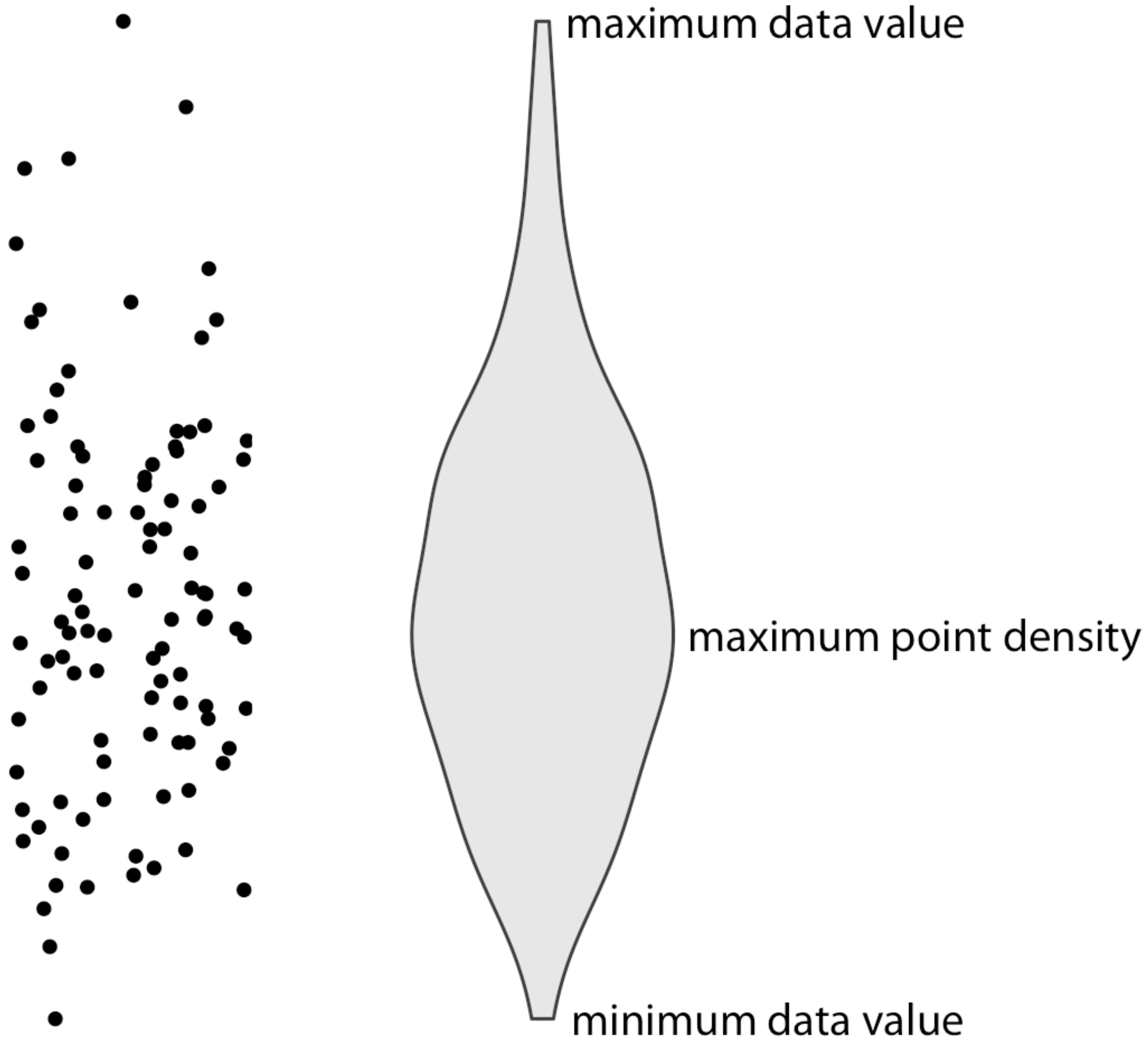


- a boxplot divides the data into quartiles and visualizes them in a standardized manner
  - median: line in the middle
  - middle 50% of data: box
  - whiskers extend either
    - to the maximum / minimum values of the data
    - to the maximum / minimum values that fall within 1.5 times the height of the box
  - outliers: data points that fall beyond the fences

Visualizing Distributions Along the Vertical Axis



Violin Plots



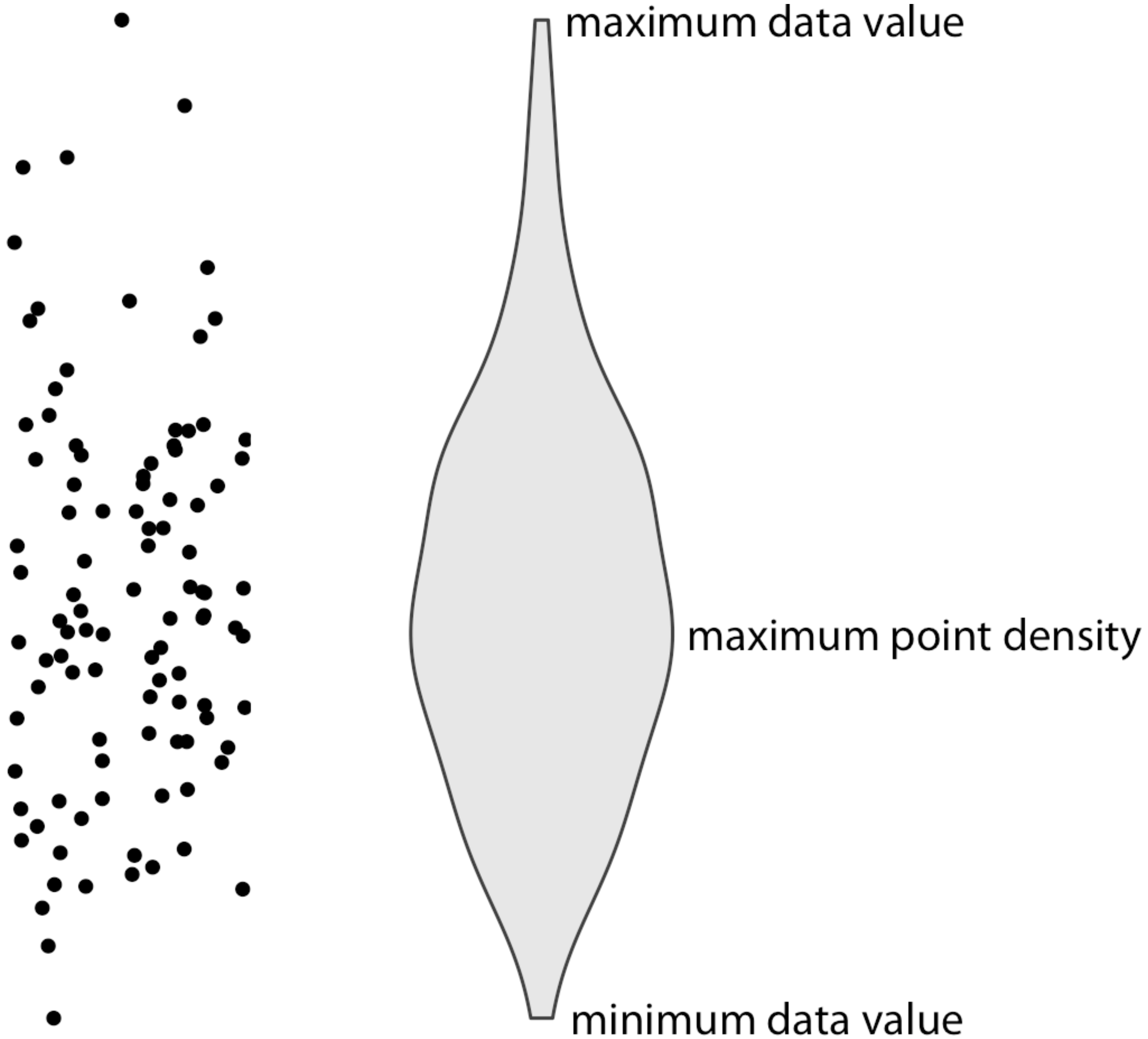
- violin plots provide a more nuanced picture of the data

- width of the violin plot represents the point density at that y value

  - violin plot is a density estimate rotated by 90 degrees and then mirrored

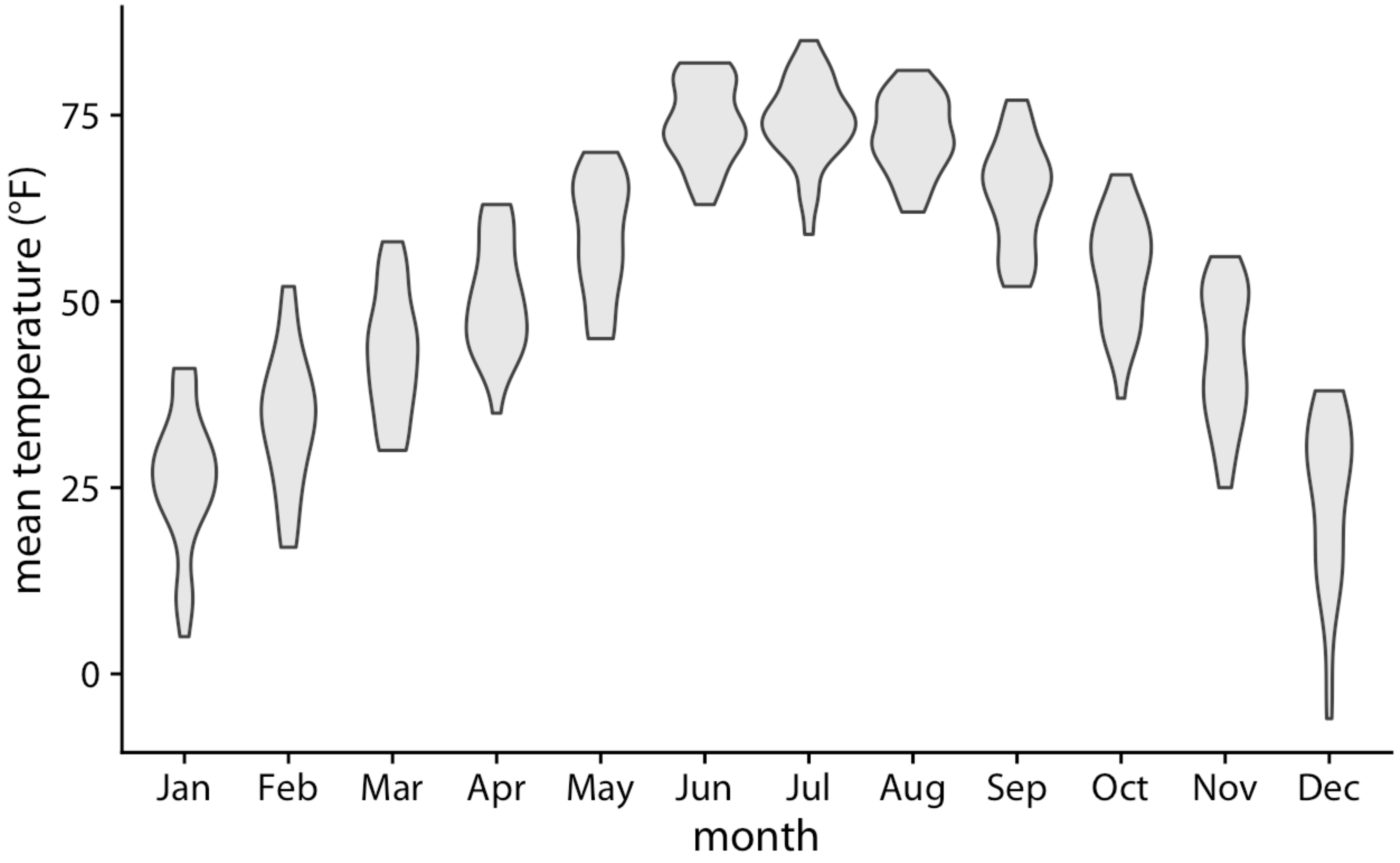
- violins begin and end at the minimum and maximum data values

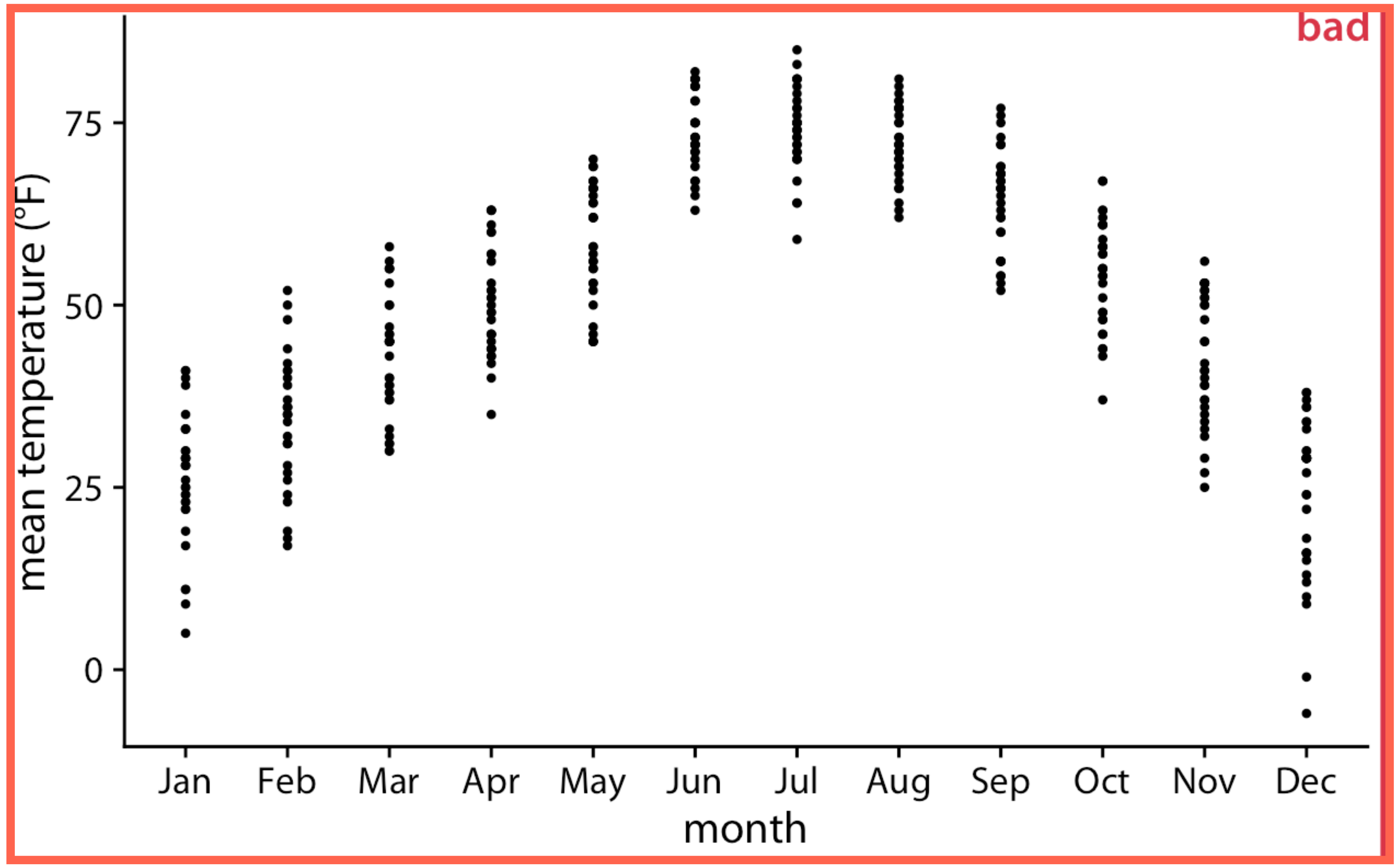
Violin Plots



- violin plots have the similar shortcomings as density estimates
- they can generate the appearance that
  - there is data where none exists
  - the dataset is very dense when actually it is quite sparse

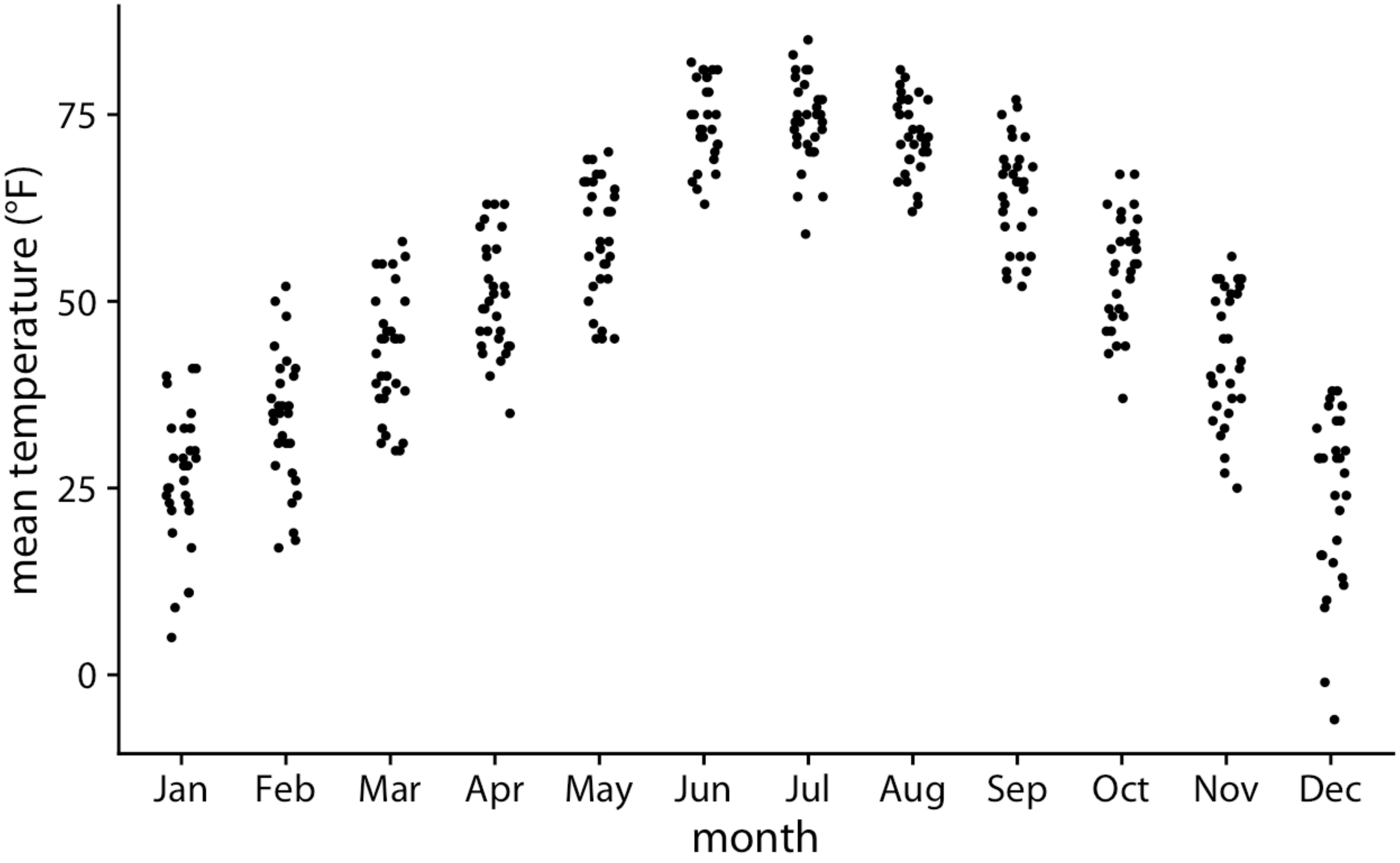
Strip Charts



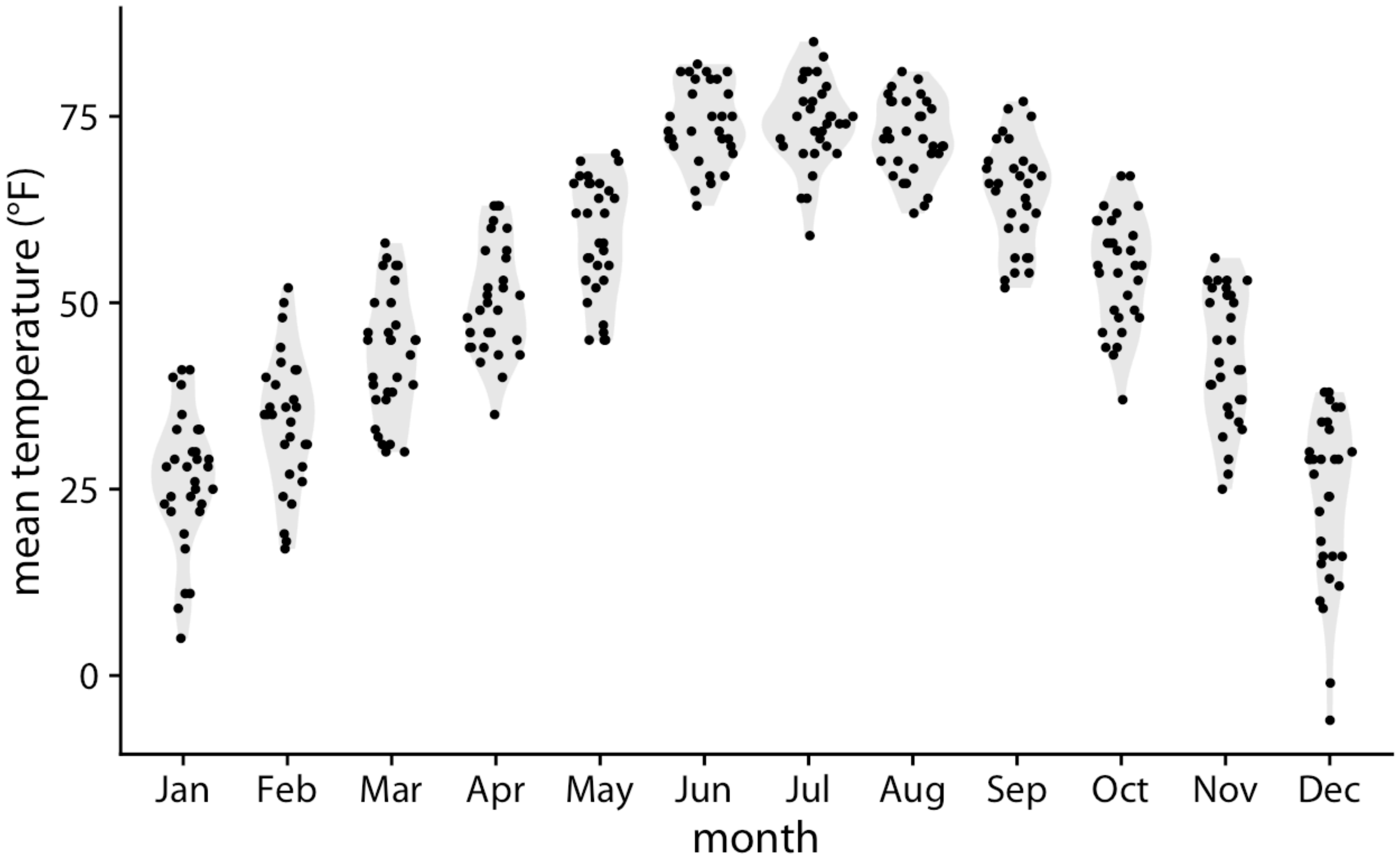


Why is this visualization bad?

Strip Charts - Jittering



Sina Plots



- spread out the dots in proportion to the point density at a given y coordinate
- here: sina plots are superimposed on violin plots

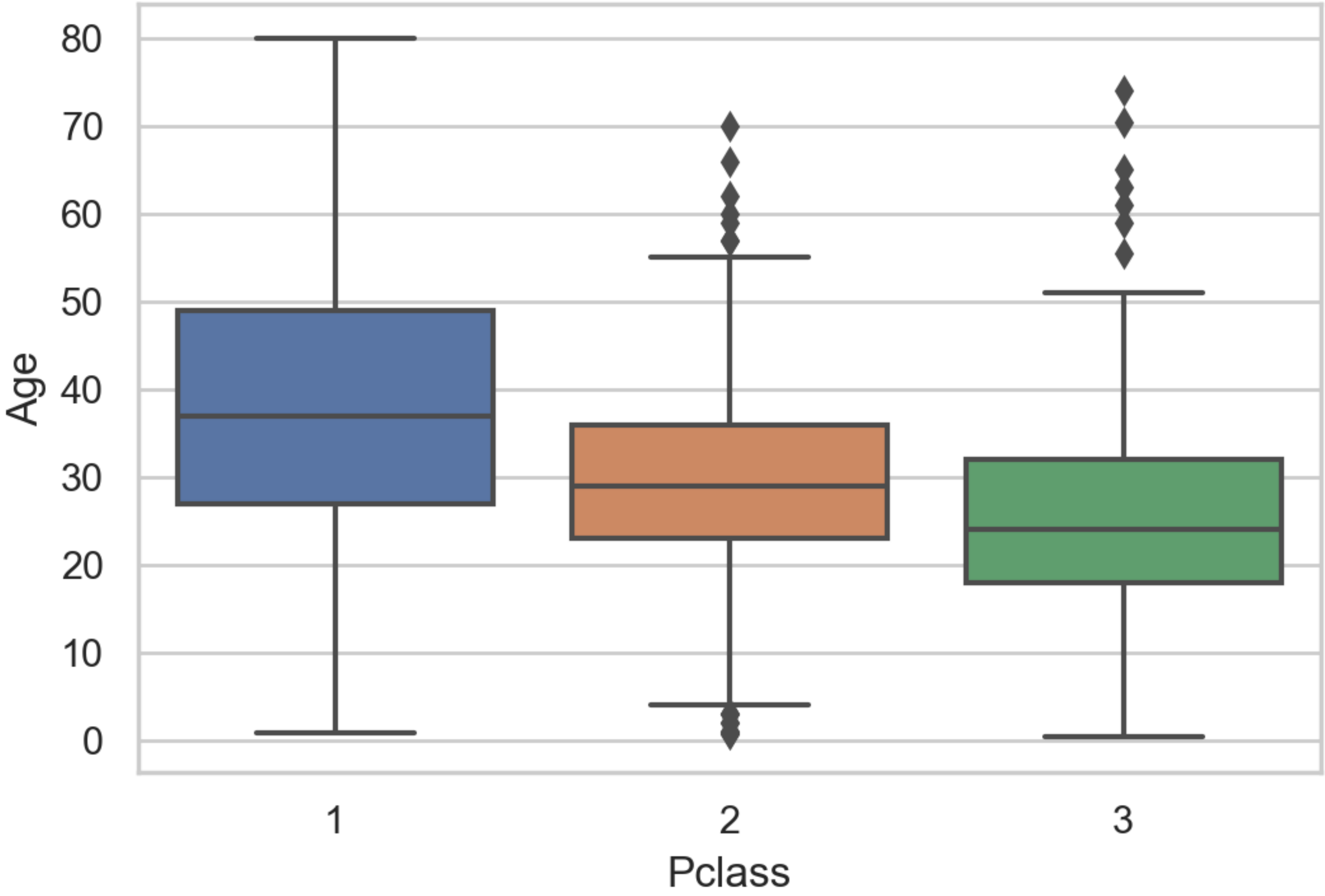
Violin Plots, Strip Charts and Sina Plots with seaborn

- read the documentation of `seaborn.boxplot`, `seaborn.violinplot` and `seaborn.stripplot` and generate the corresponding visualizations for the Titanic data set:

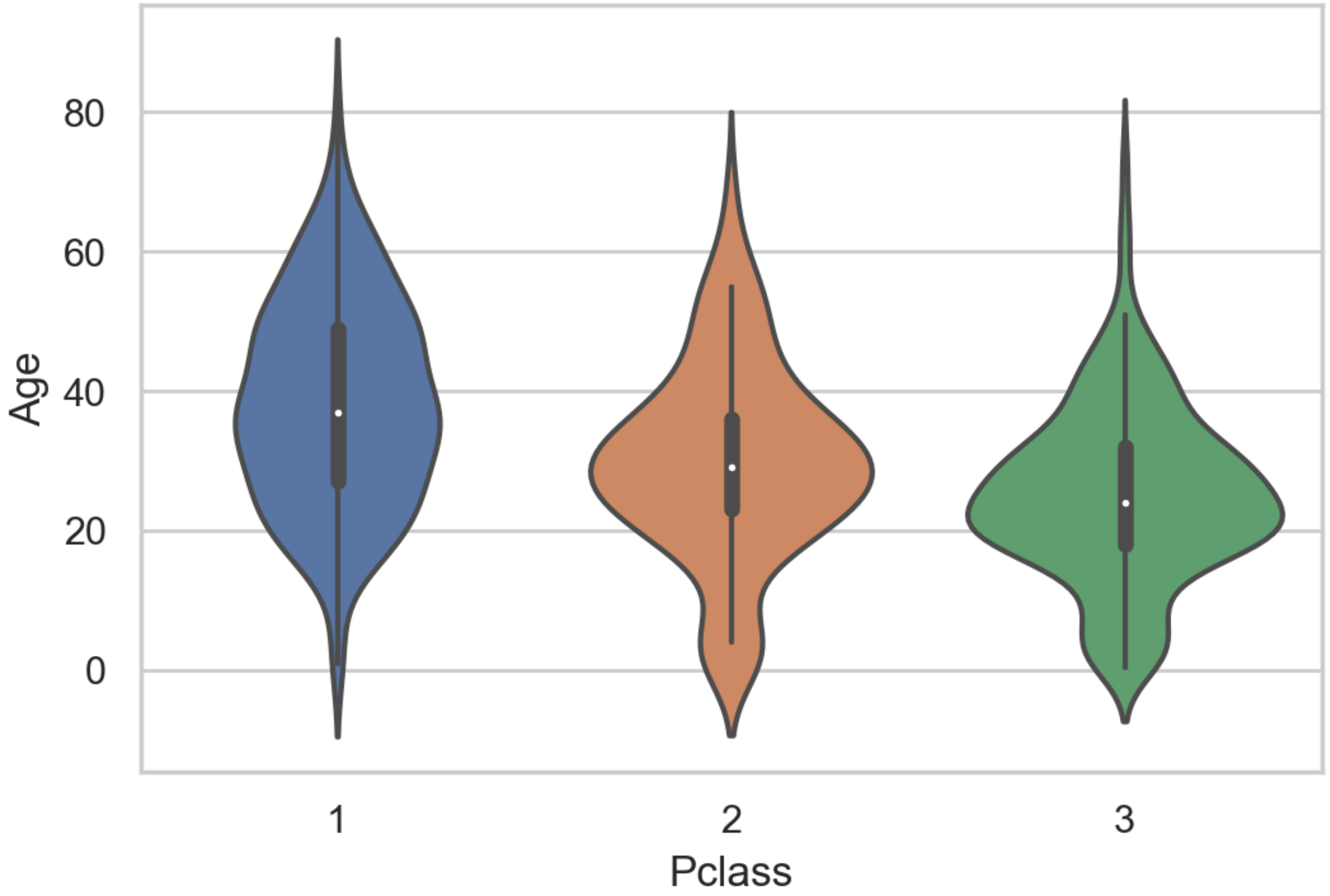
    - use “Pclass” as grouping variable (x-axis) and “Age” as response variable (y-axis)

- check out the Github repository https://github.com/mparker2/seaborn_sinaplot for a sina plot function and generate a sina plot according to the specification above

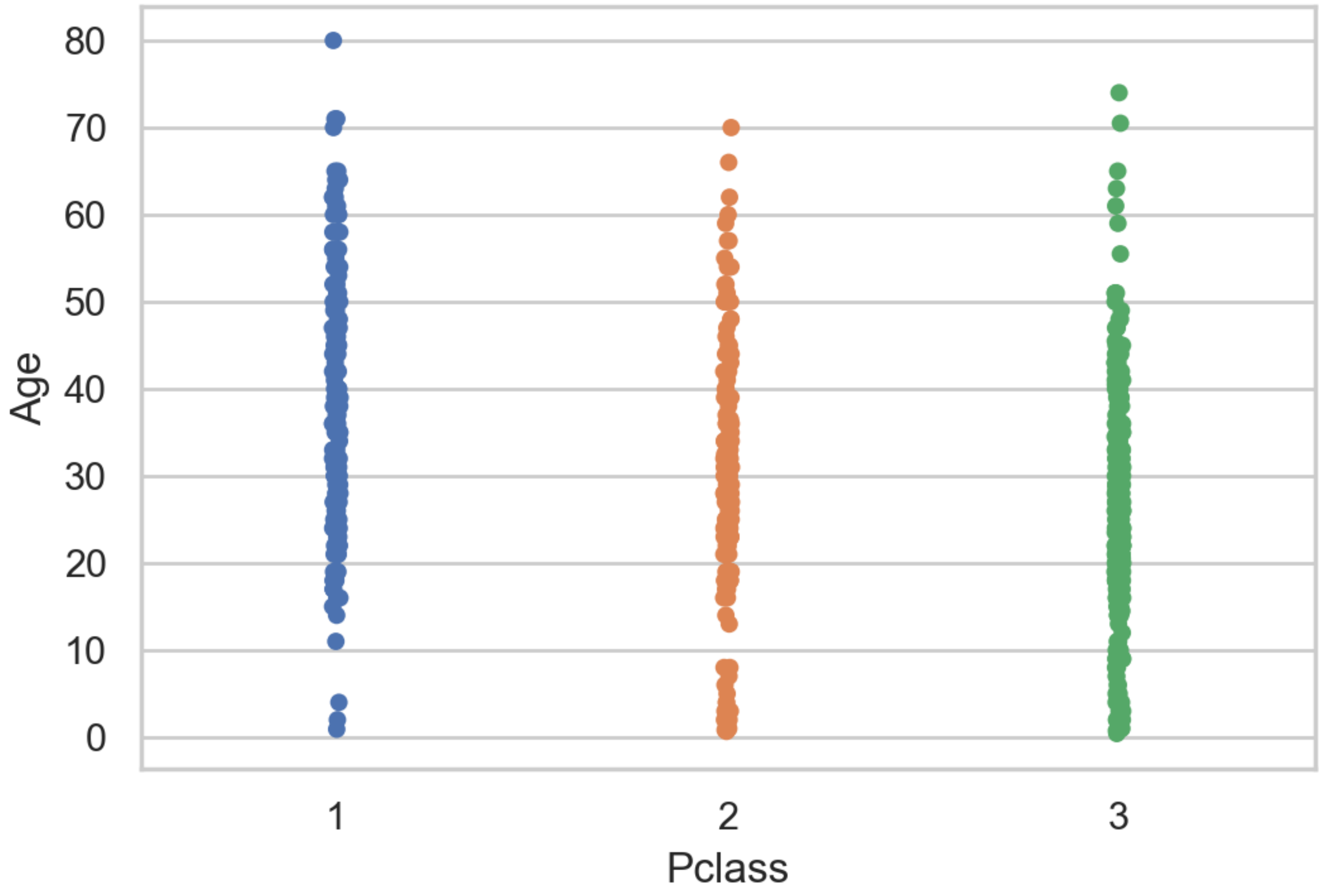
Boxplot with seaborn



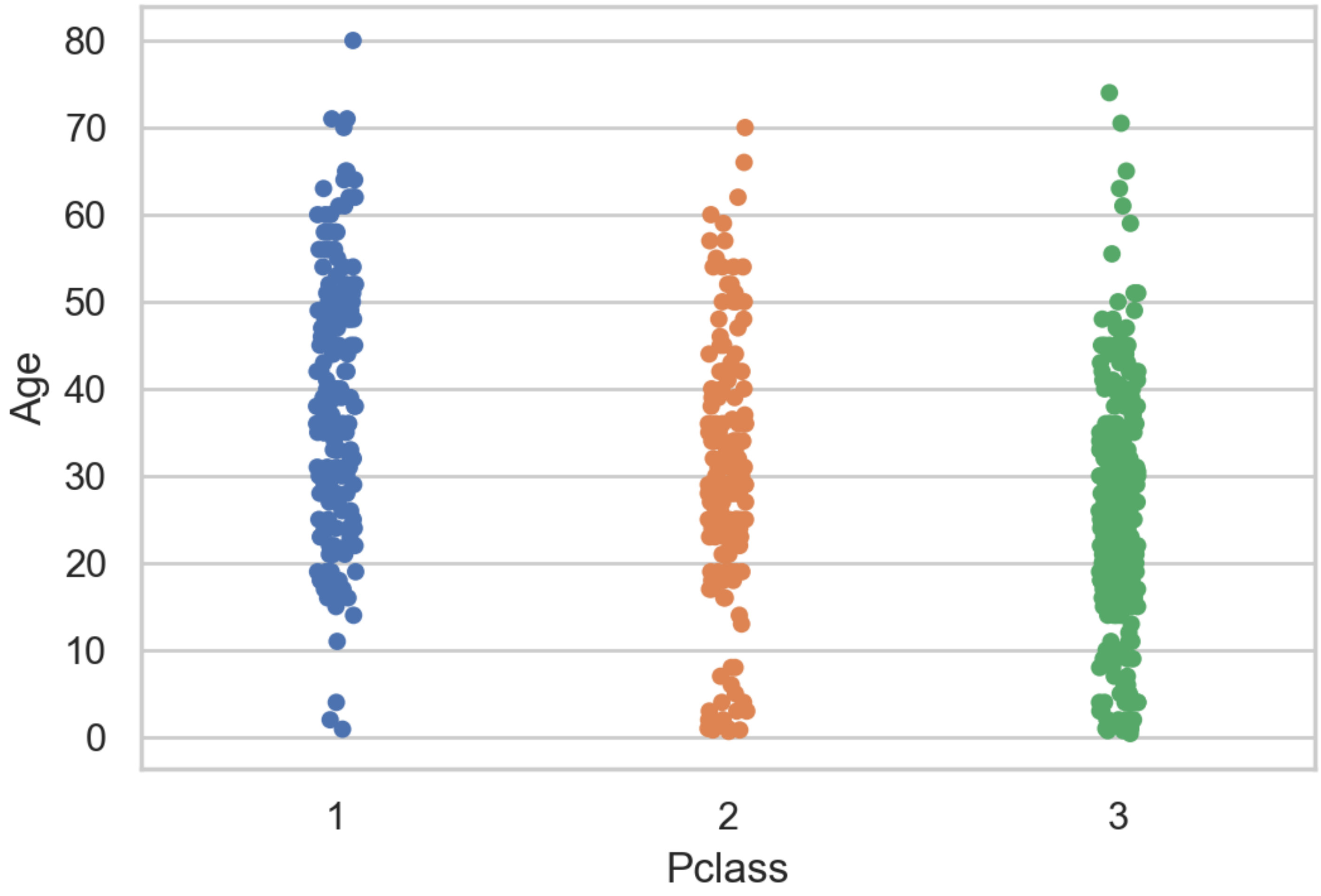
Violin Plot with seaborn



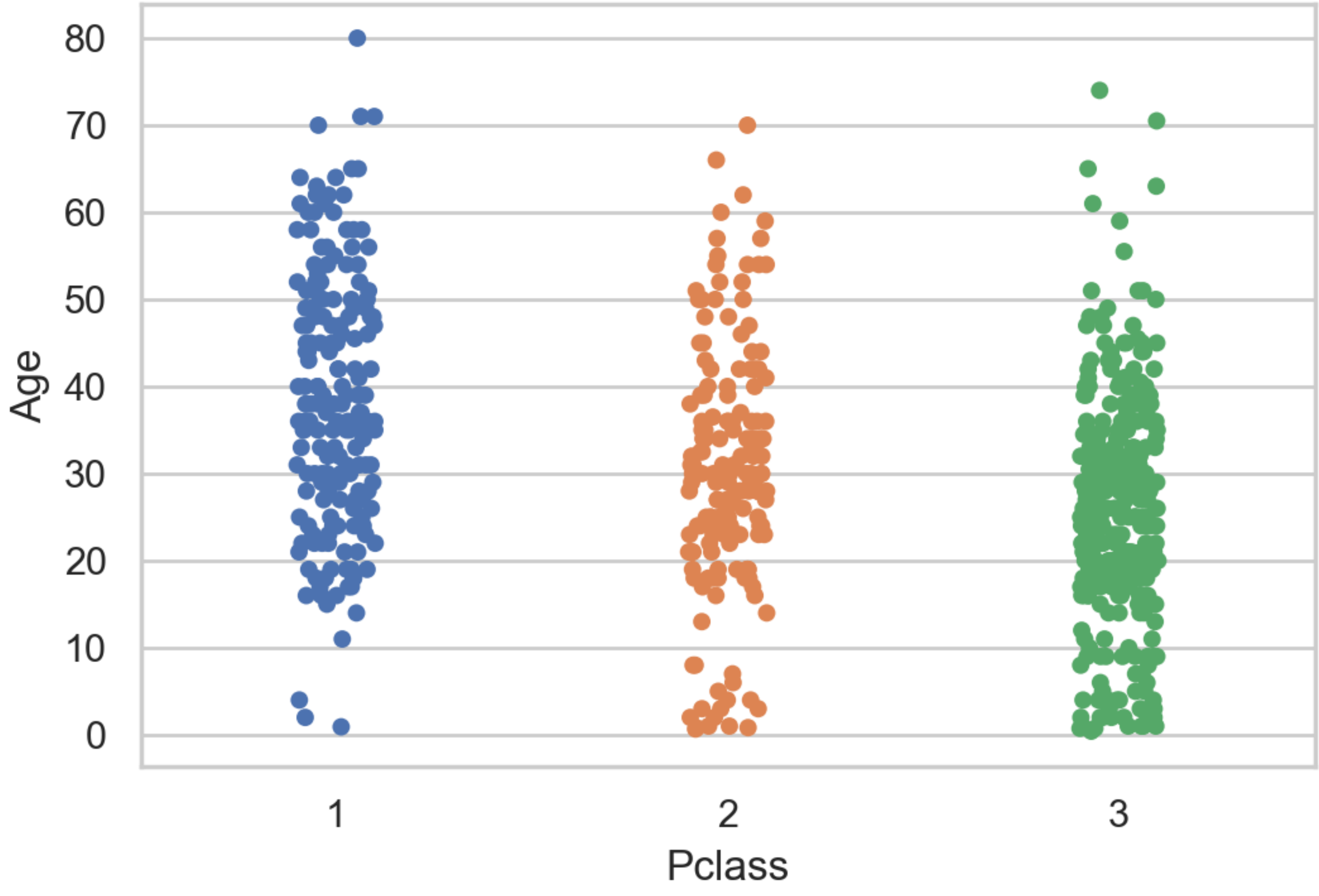
Strip Plot with seaborn



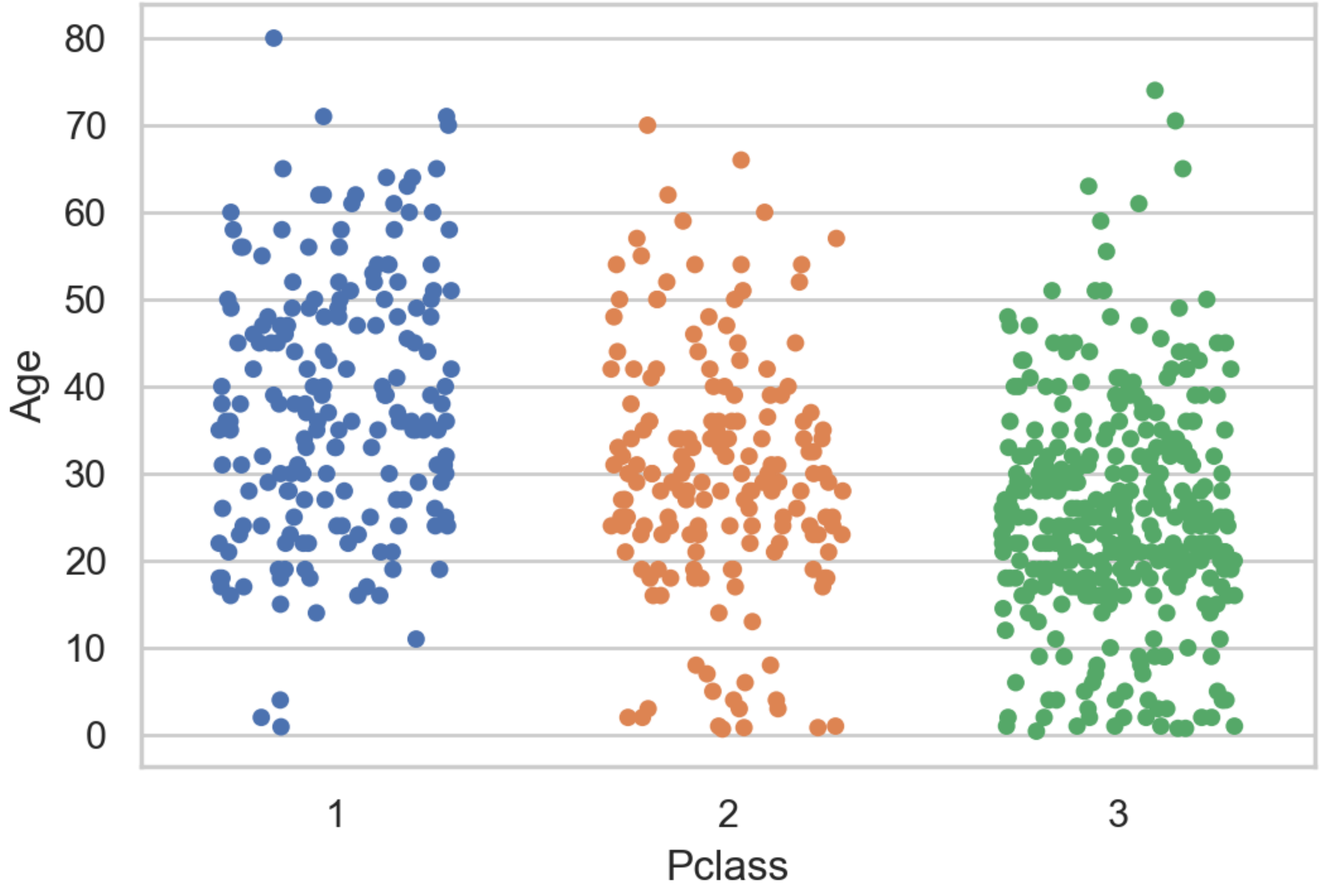
jitter = 0.01



jitter = 0.05

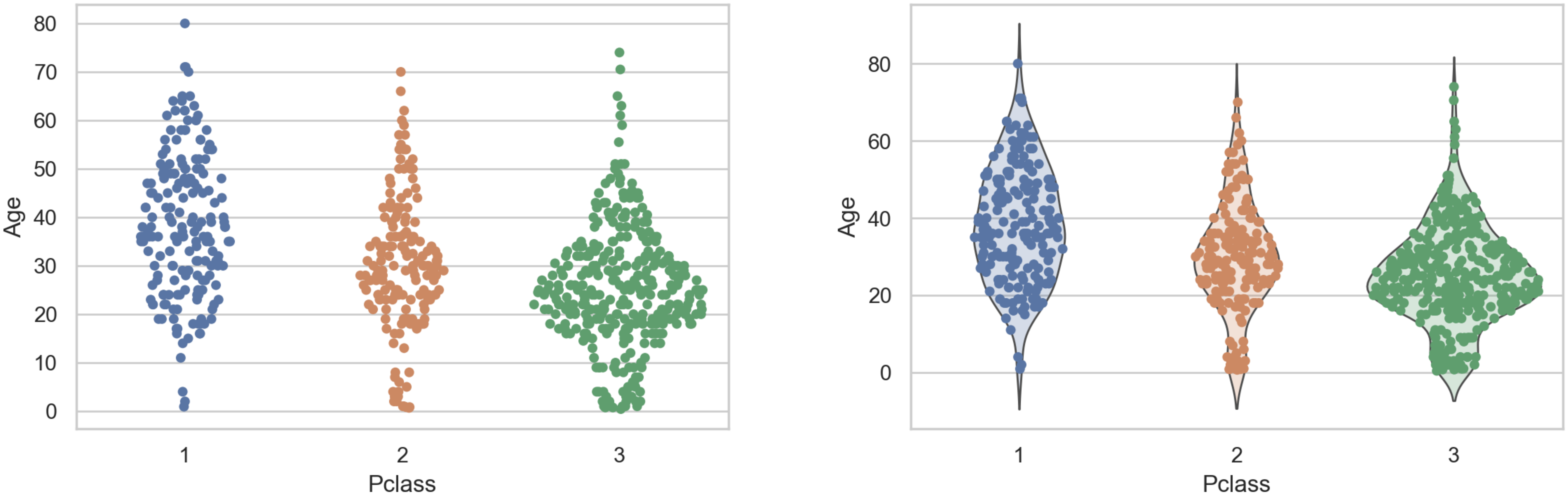


jitter = 0.1



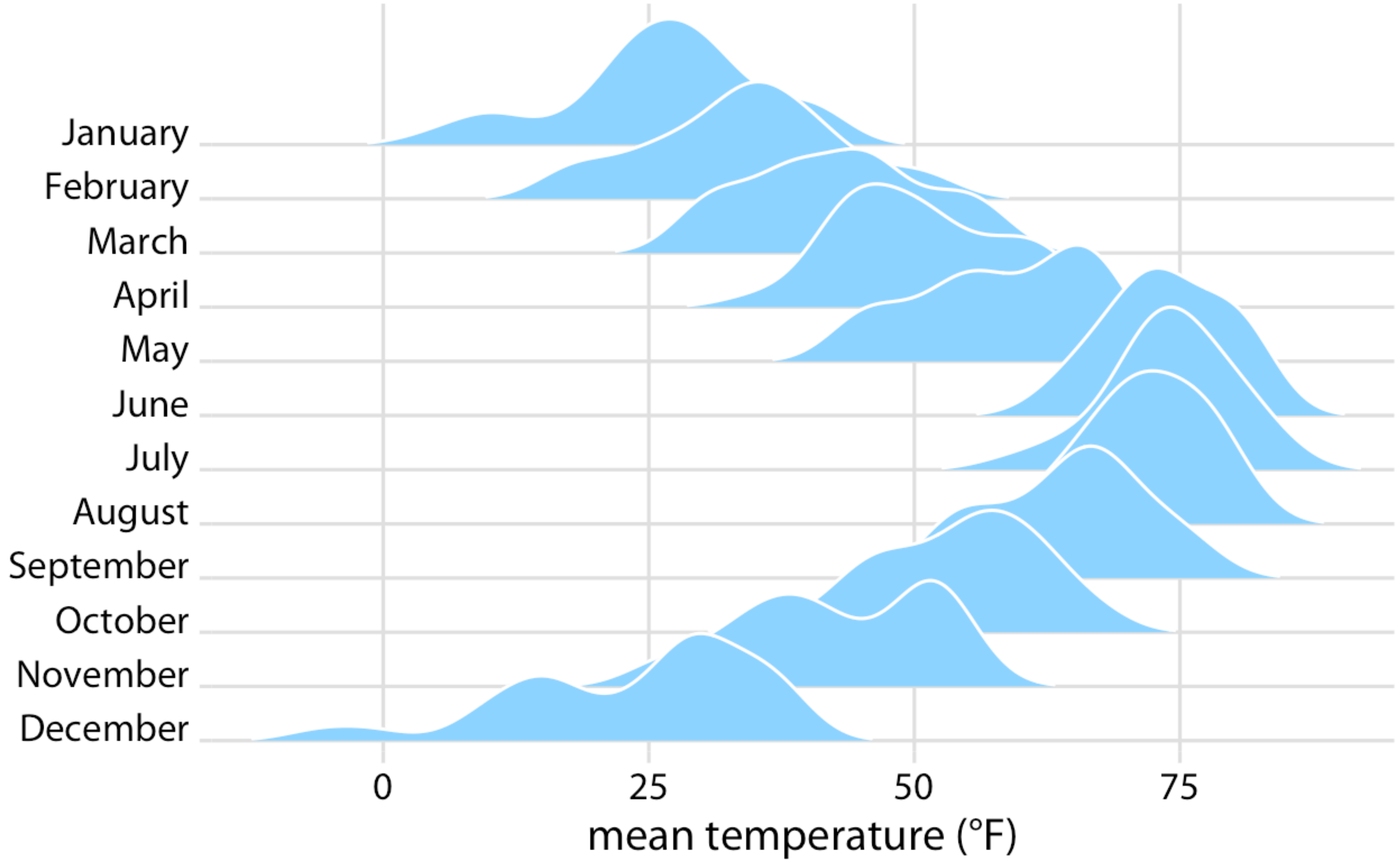
jitter = 0.3

Sina Plot with seaborn



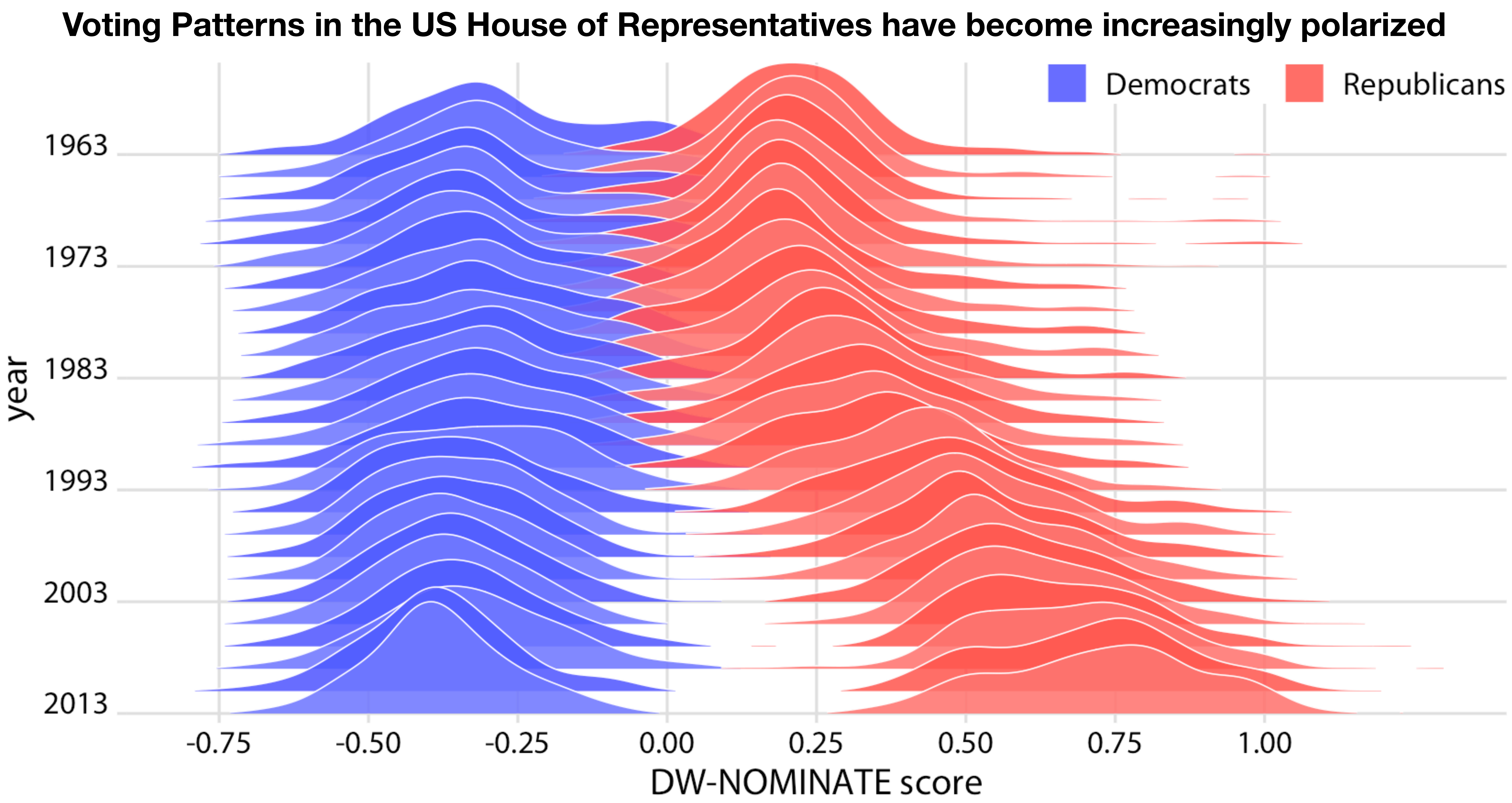
Visualizing Distributions Along the Horizontal Axis

Ridgeline Plots



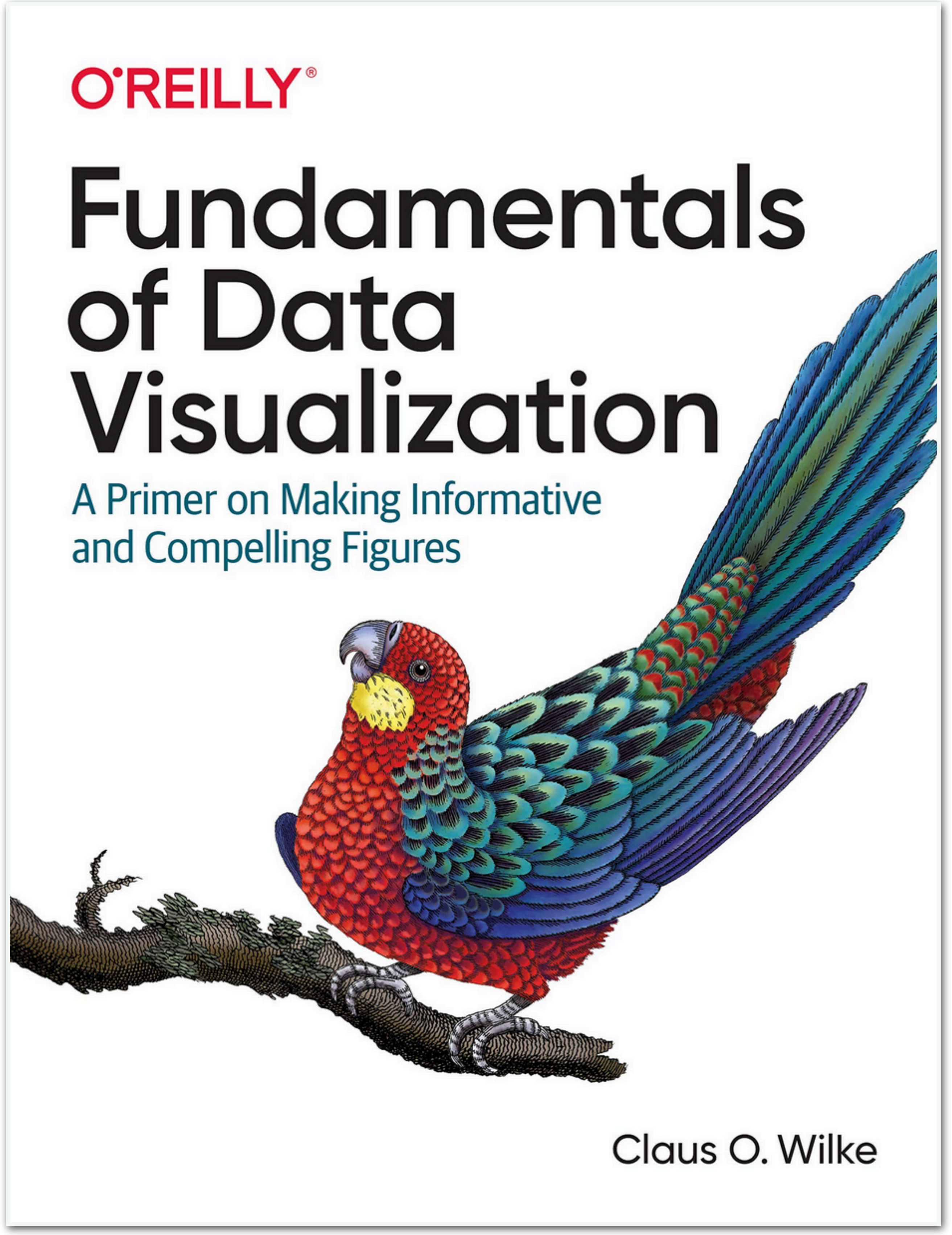
- work particularly well if you want to show trends in distributions over time

Ridgeline Plots



Voting Patterns in the US House of Representatives have become increasingly polarized

Literature



O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

Claus O. Wilke

References

- Slide 3-7, 10-17, 24, 25; Image Source: Claus O. Wilke - Fundamentals of Data Visualization, O’Reilly