Visualizing distributions 2

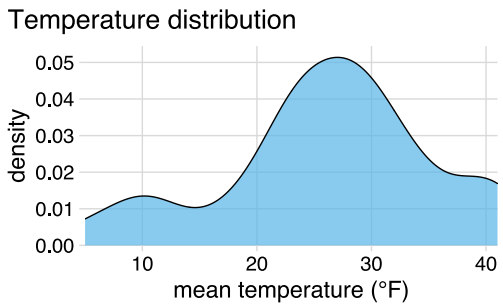
Claus O. Wilke

last updated: 2021-01-18

Reminder: Density estimates visualize distributions

Mean temperatures in Lincoln, NE, in January 2016:

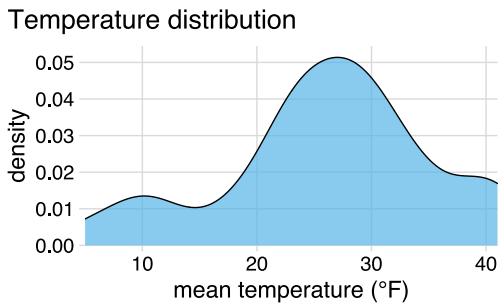
date	mean temp
2016-01-01	24
2016-01-02	23
2016-01-03	23
2016-01-04	17
2016-01-05	29
2016-01-06	33
2016-01-07	30
2016-01-08	25



Reminder: Density estimates visualize distributions

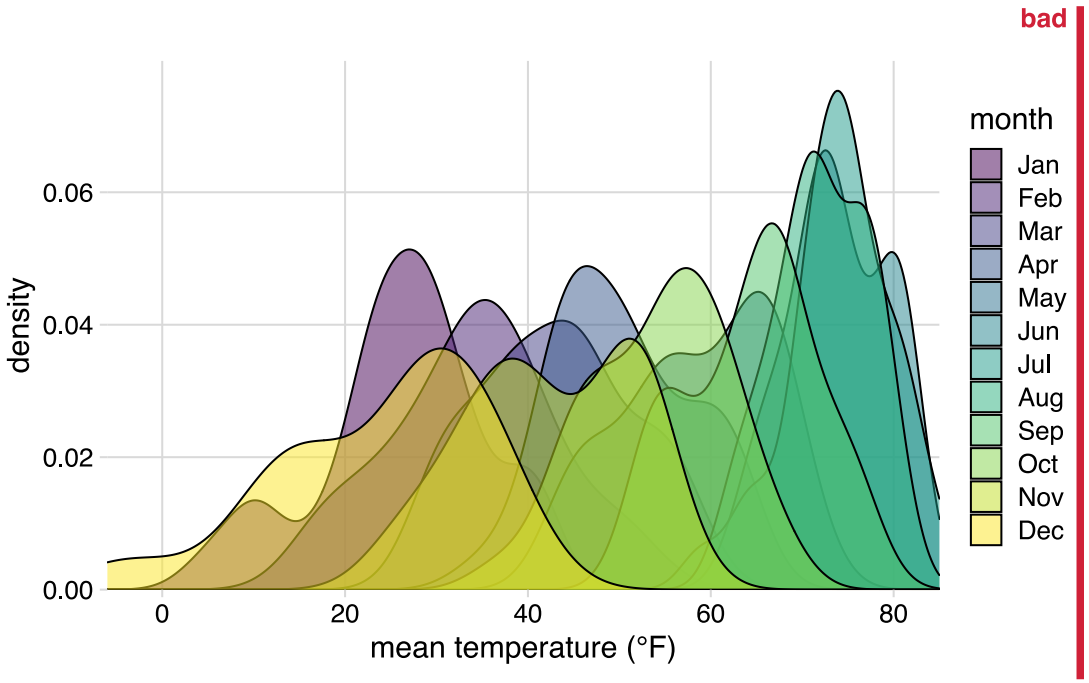
Mean temperatures in Lincoln, NE, in January 2016:

date	mean temp
2016-01-01	24
2016-01-02	23
2016-01-03	23
2016-01-04	17
2016-01-05	29
2016-01-06	33
2016-01-07	30
2016-01-08	25

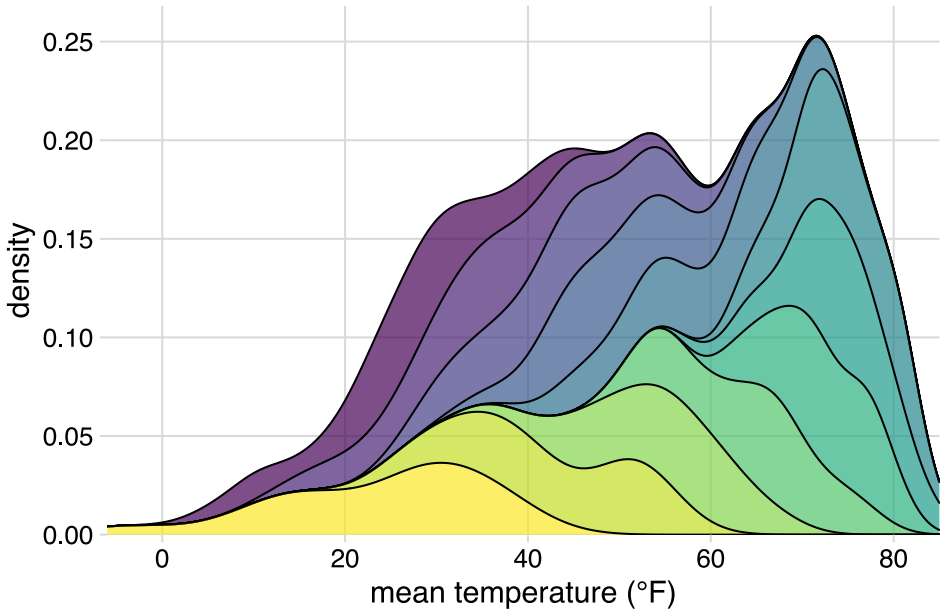


How can we compare distributions across months?

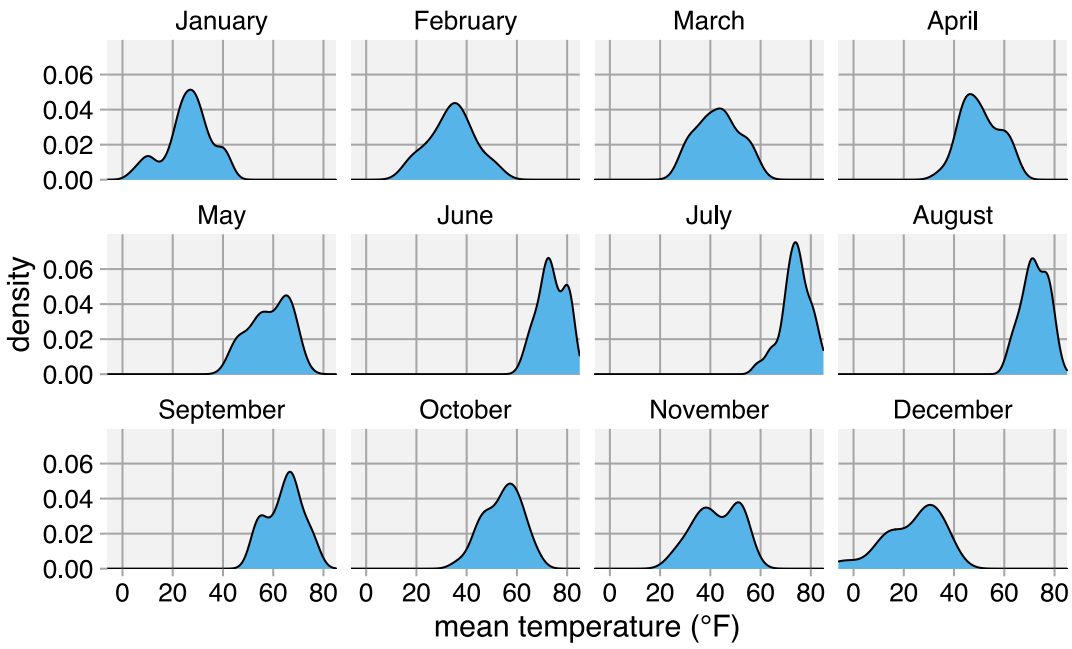
A bad idea: Many overlapping density plots



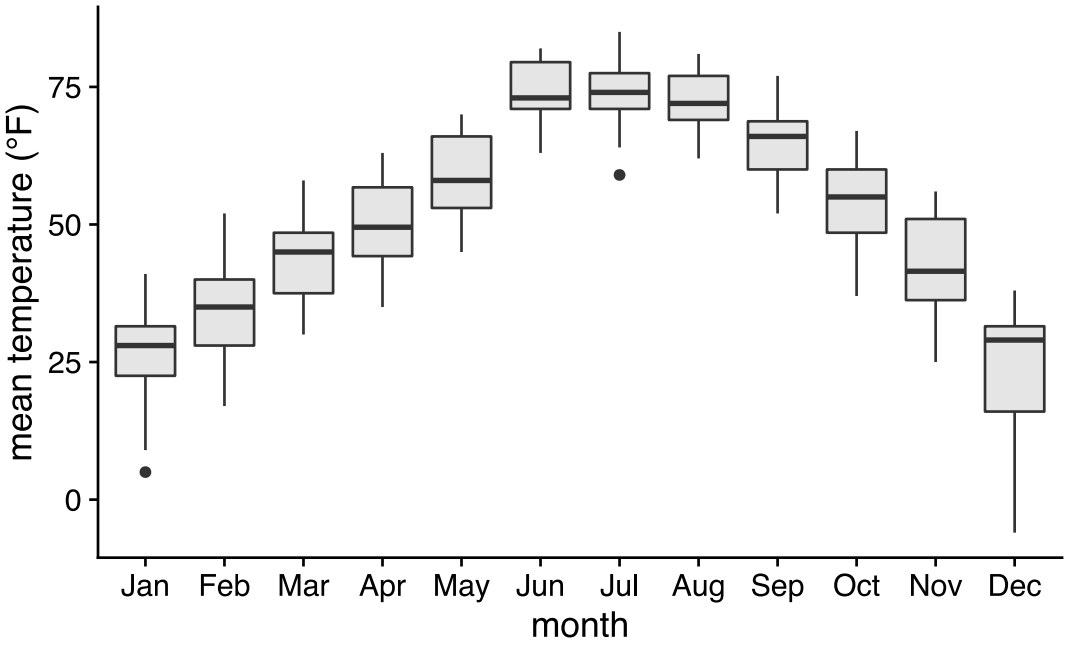
Another bad idea: Stacked density plots



Somewhat better: Small multiples

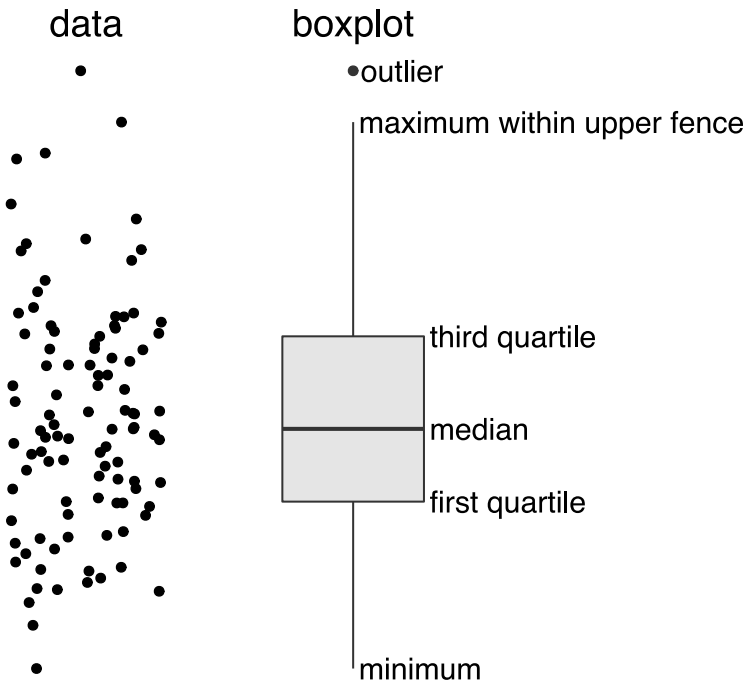


Instead: Show values along y, conditions along x

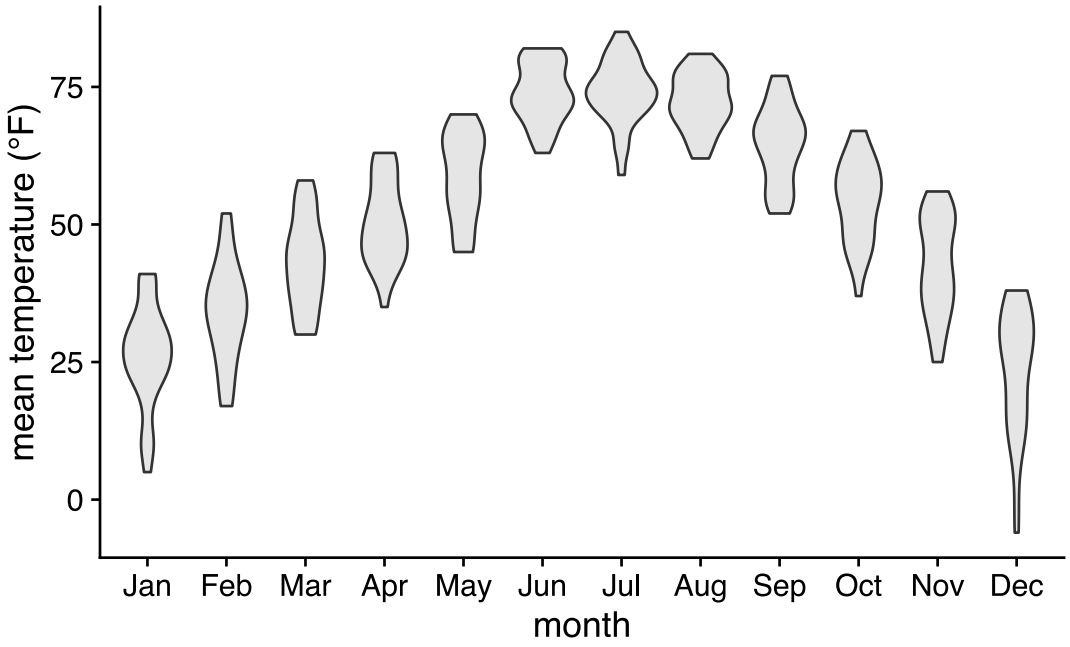


A boxplot is a crude way of visualizing a distribution.

How to read a boxplot

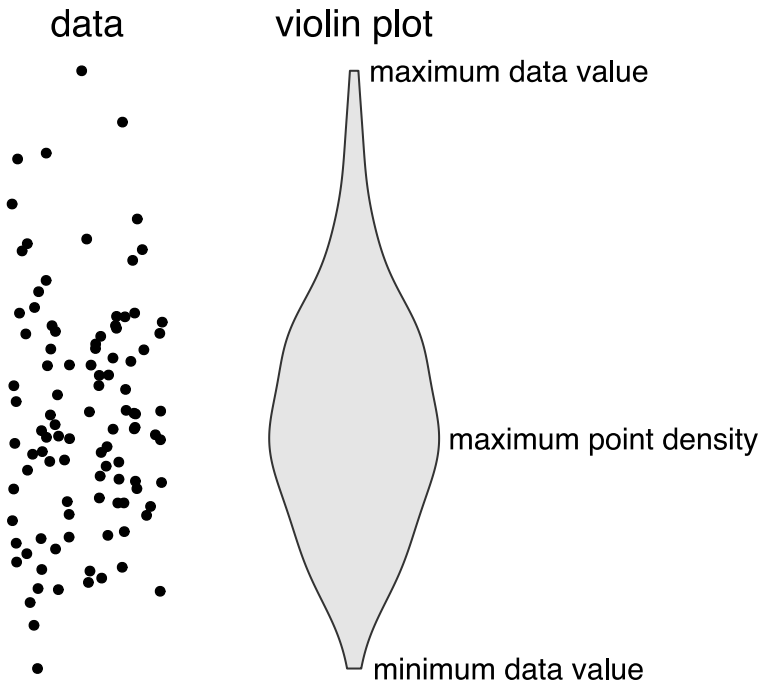


If you like density plots, consider violins



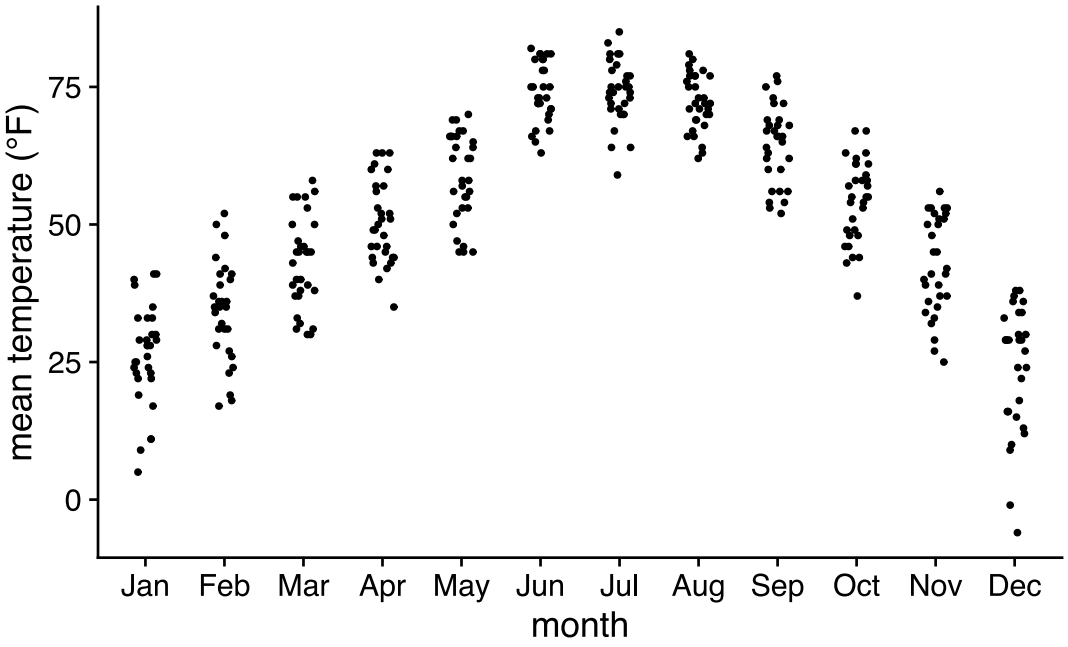
A violin plot is a density plot rotated 90 degrees and then mirrored.

How to read a violin plot



For small datasets, you can also use a strip chart

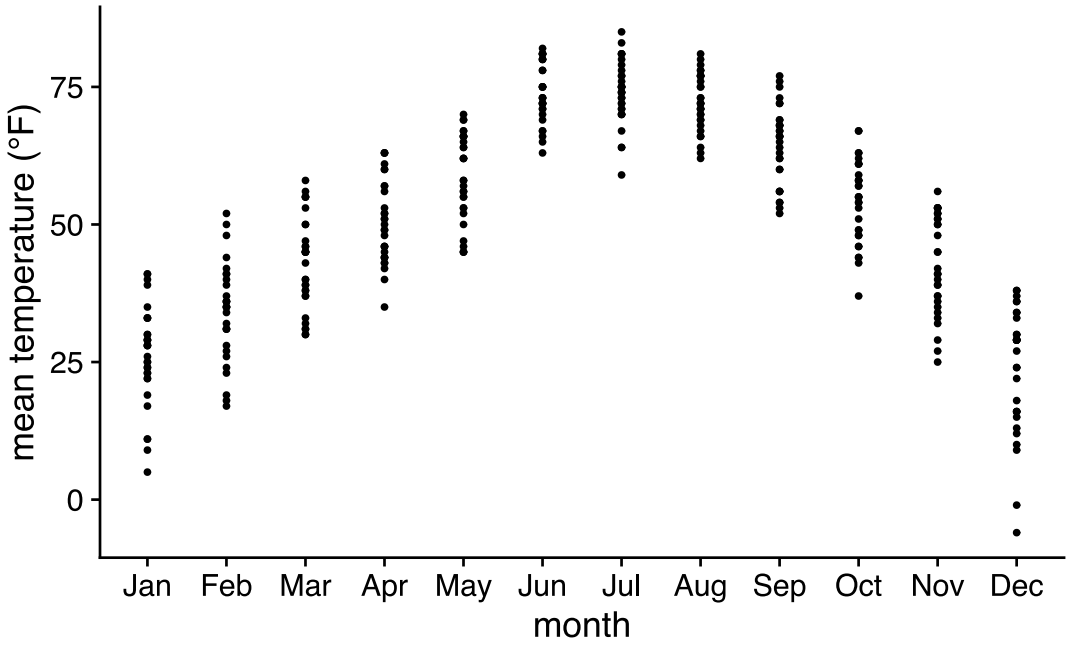
Advantage: Can see raw data points instead of abstract representation.



Horizontal jittering may be necessary to avoid overlapping points.

For small datasets, you can also use a strip chart

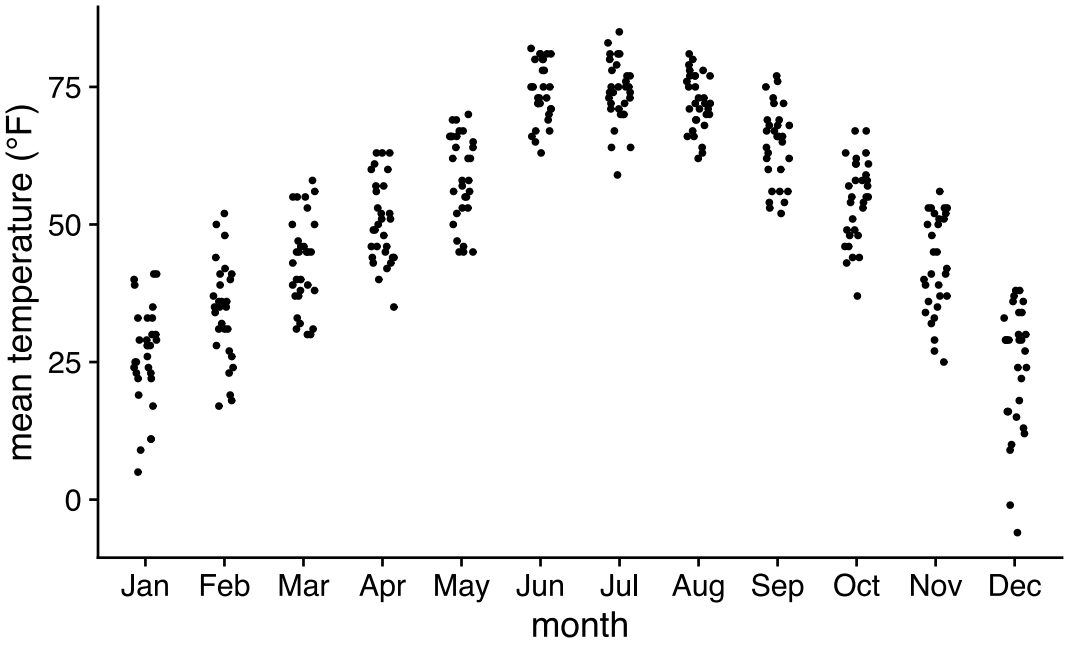
Advantage: Can see raw data points instead of abstract representation.



Horizontal jittering may be necessary to avoid overlapping points.

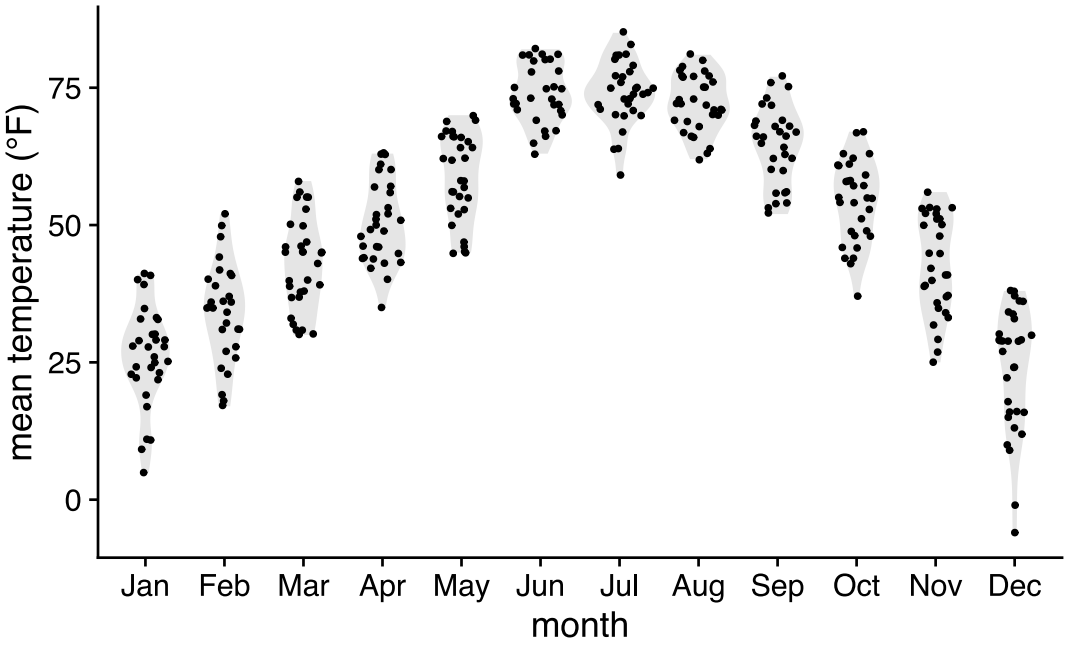
For small datasets, you can also use a strip chart

Advantage: Can see raw data points instead of abstract representation.



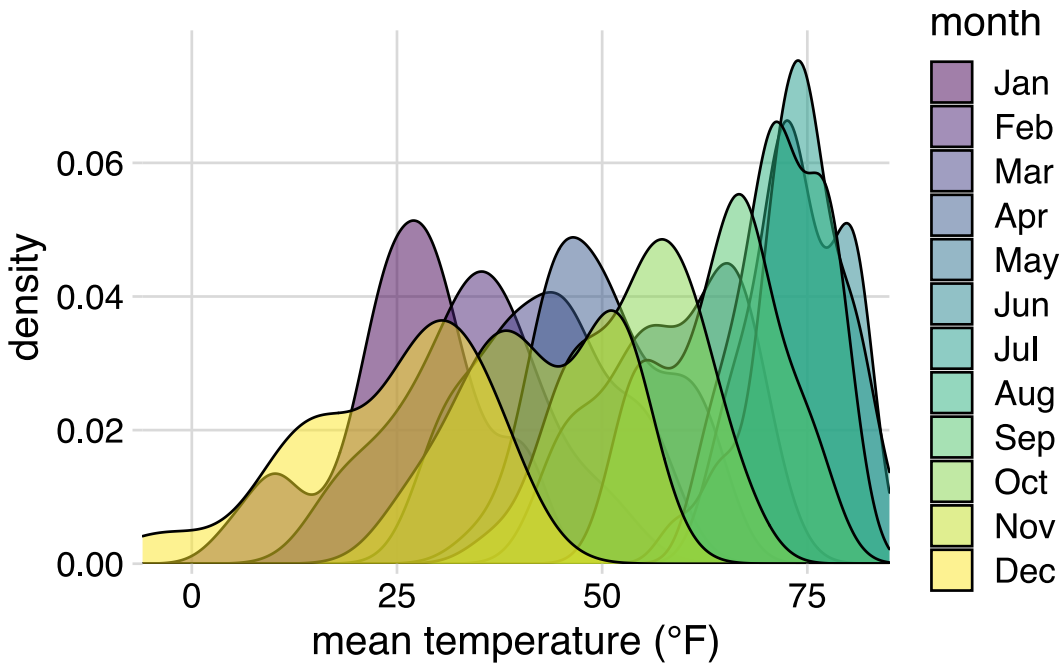
Horizontal jittering may be necessary to avoid overlapping points.

We can also jitter points into violins



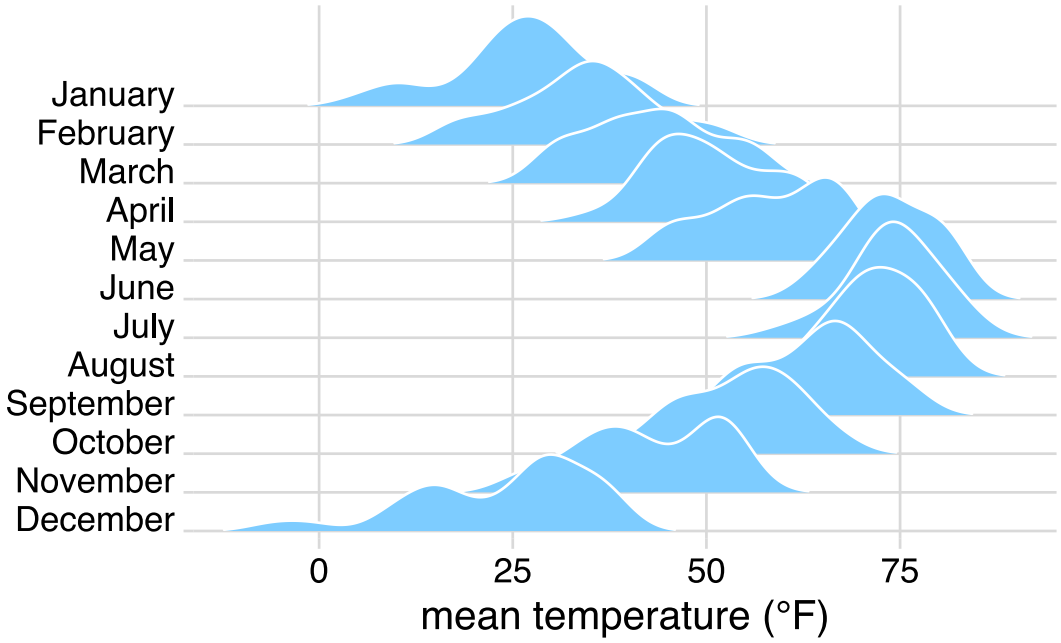
Such plots are called sina plots, to honor Sina Hadi Sohi.

But maybe there's hope for overlapping density plots?



How about we stagger the densities vertically?

Vertically staggered density plots are called ridgelines



Notice the single fill color. More colors would be distracting.

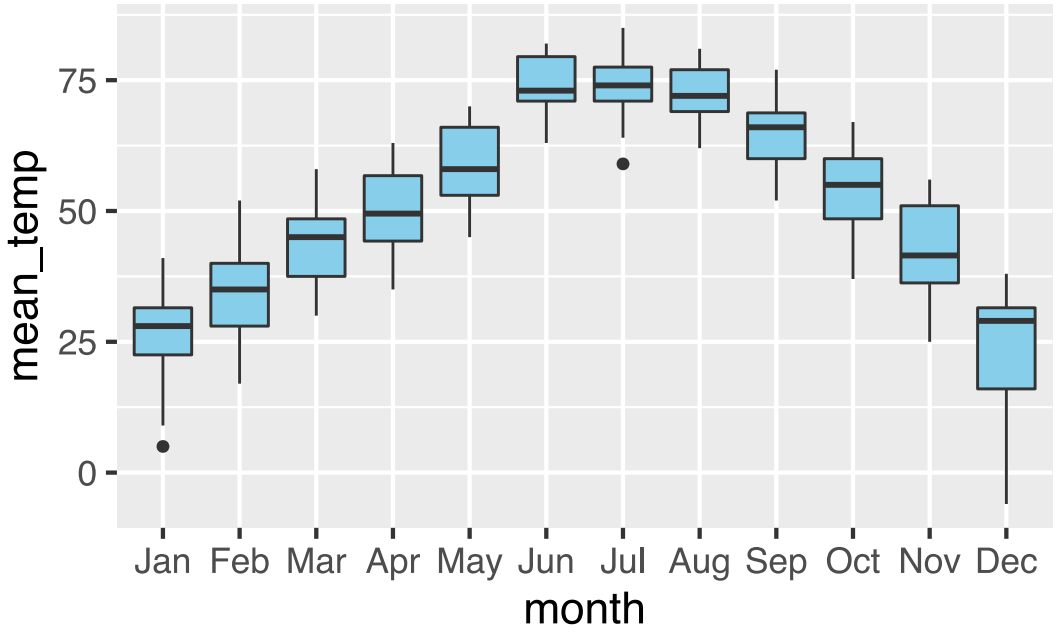
Making these plots in ggplot

Plot type	Geom	Notes
boxplot	`geom_boxplot()`	
violin plot	`geom_violin()`	
strip chart	`geom_point()`	Jittering requires `position_jitter()`
sina plot	`geom_sina()`	From package **ggforce**
ridgeline	`geom_density_ridges()`	From package **ggridges**

all others but the ridgeline: both categorical and interval variables are exchangeable for x and y axis
ridgeline : categorical on y, numerical on x

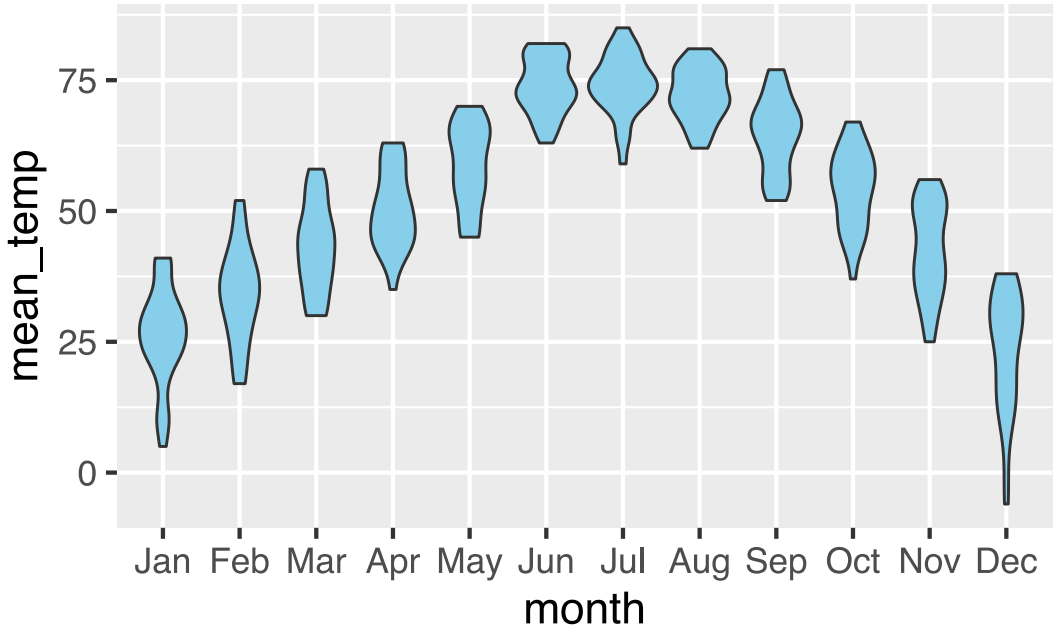
Examples: Boxplot

```
ggplot(lincoln_temps, aes(x = month, y = mean_temp
  geom_boxplot(fill = "skyblue")
```



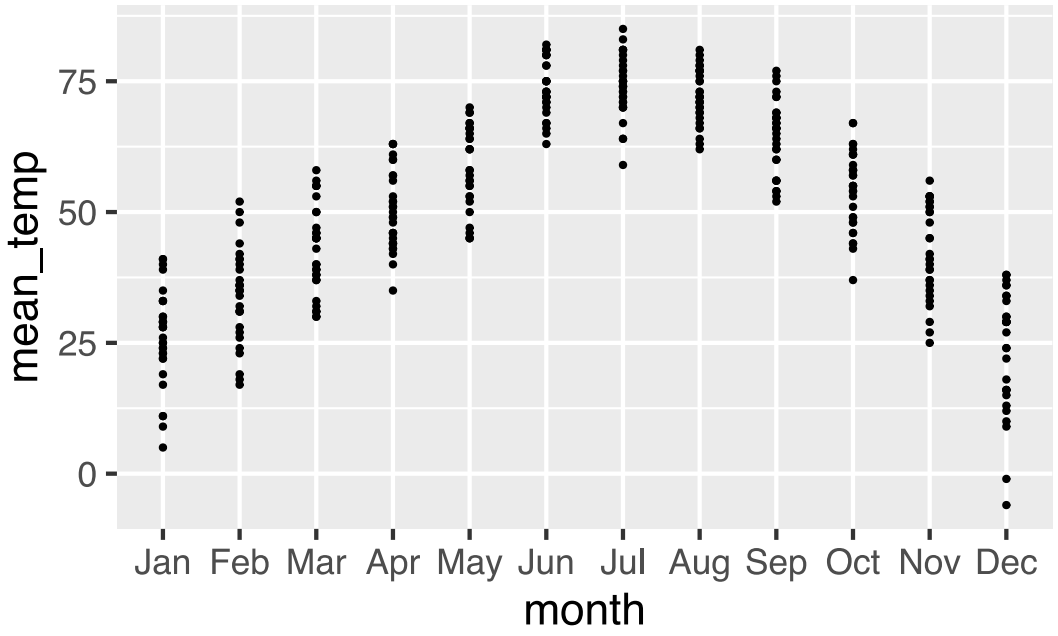
Examples: Violins

```
ggplot(lincoln_temps, aes(x = month, y = mean_temp
  geom_violin(fill = "skyblue")
```



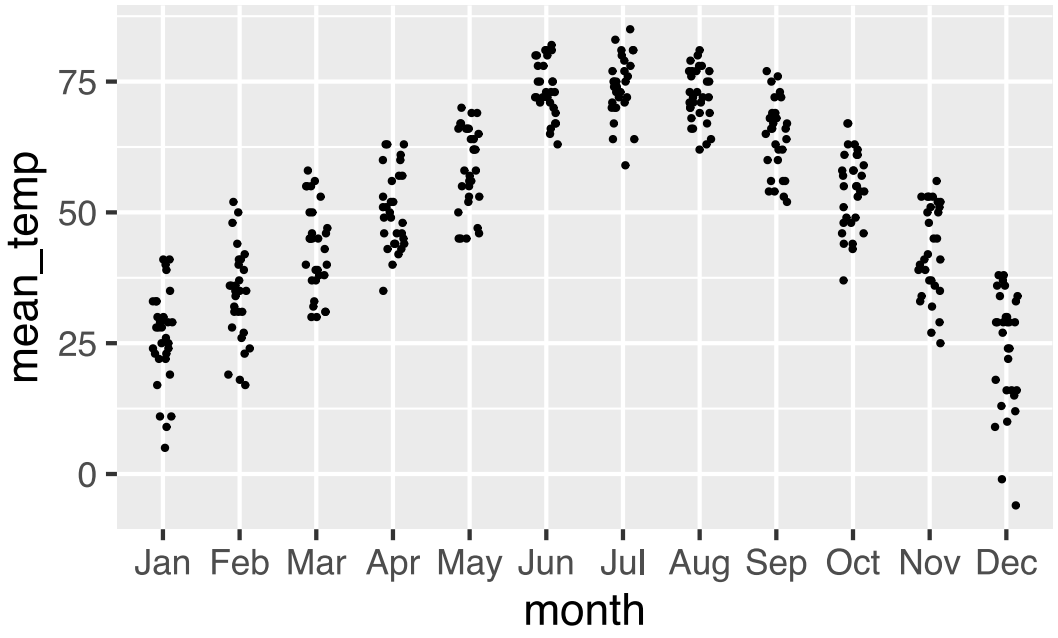
Examples: Strip chart (no jitter)

```
ggplot(lincoln_temps, aes(x = month, y = mean_temp
  geom_point(size = 0.75)  # reduce point size to
```



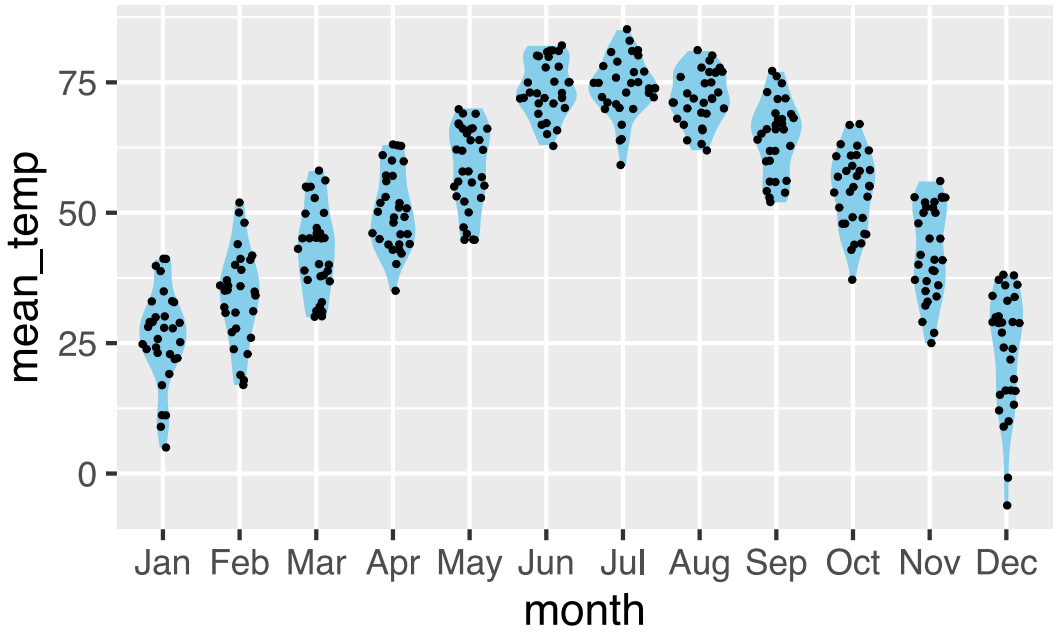
Examples: Strip chart (w/ jitter)

```
ggplot(lincoln_temps, aes(x = month, y = mean_temp
  geom_point(size = 0.75,  # reduce point size to
    position = position_jitter(
      width = 0.15,  # amount of jitter in horizo
      height = 0     # amount of jitter in vertic
    )
  )
```



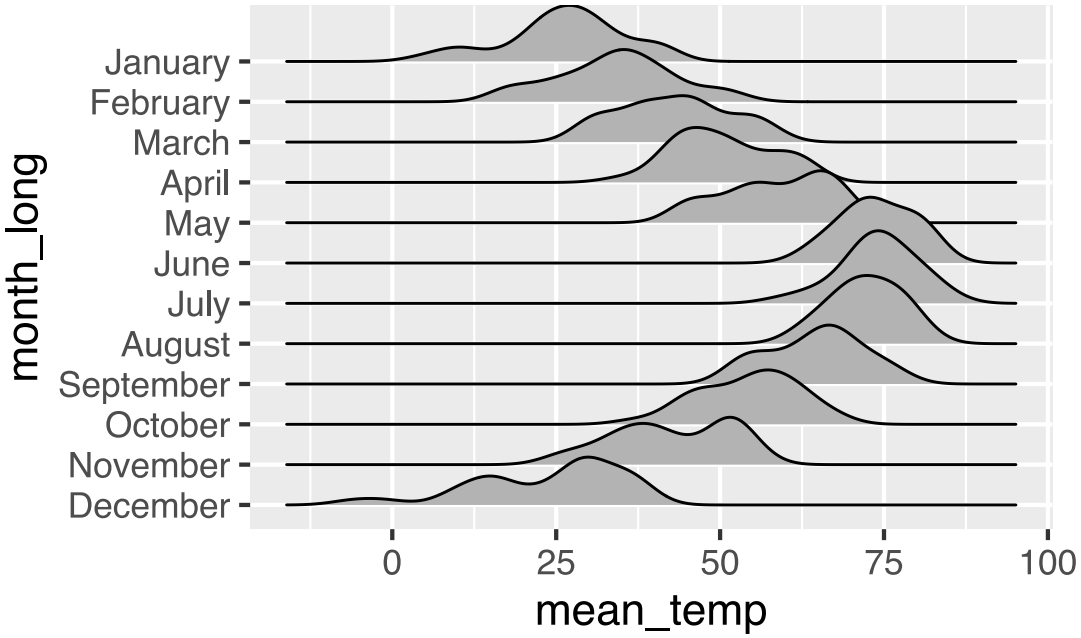
Examples: Sina plot

```
library(ggforce) # for geom_sina()

ggplot(lincoln_temps, aes(x = month, y = mean_temp
  geom_violin(fill = "skyblue", color = NA) + # v
  geom_sina(size = 0.75) # sina jittered points i
```



Examples: Ridgeline plot

```
library(ggridges) # for geom_density_ridges

ggplot(lincoln_temps, aes(x = mean_temp, y = montl
  geom_density_ridges()
```



Further reading

- Fundamentals of Data Visualization: Chapter 7: Visualizing many distributions at once
- **ggplot2** reference documentation: `geom_boxplot()`, `geom_violin()`, `position_jitter()`
- **ggforce** reference documentation: `geom_sina()`
- **ggridges** reference documentation: `geom_density_ridges()`