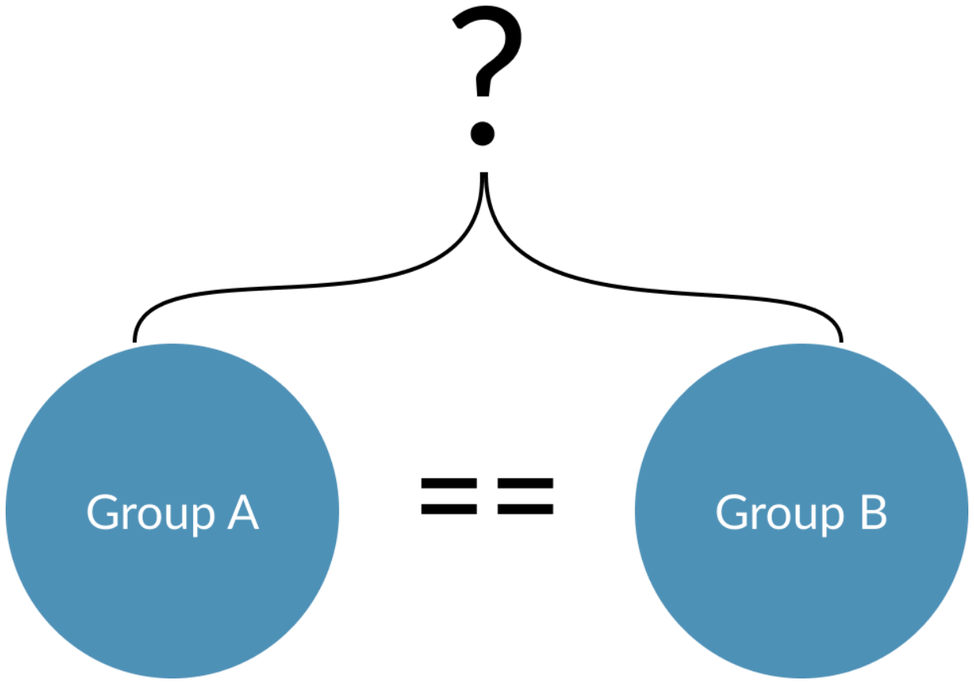
Intro to comparing distributions

VISUALIZATION BEST PRACTICES IN R

Nick Strayer
Instructor

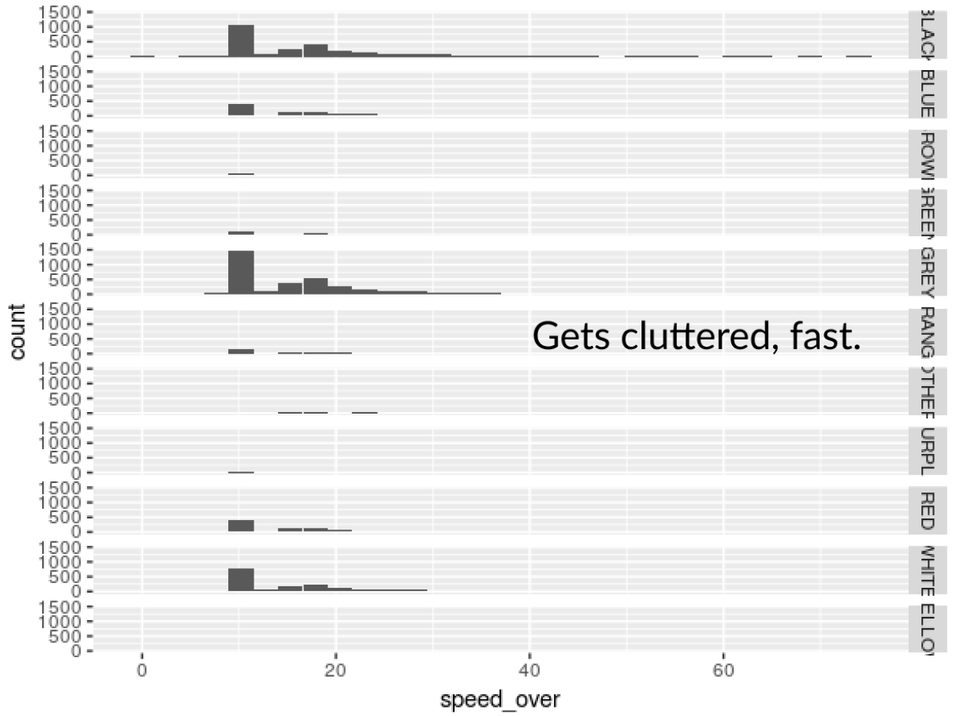
Why compare distributions?

- Verify balanced groups
- For comparison's sake

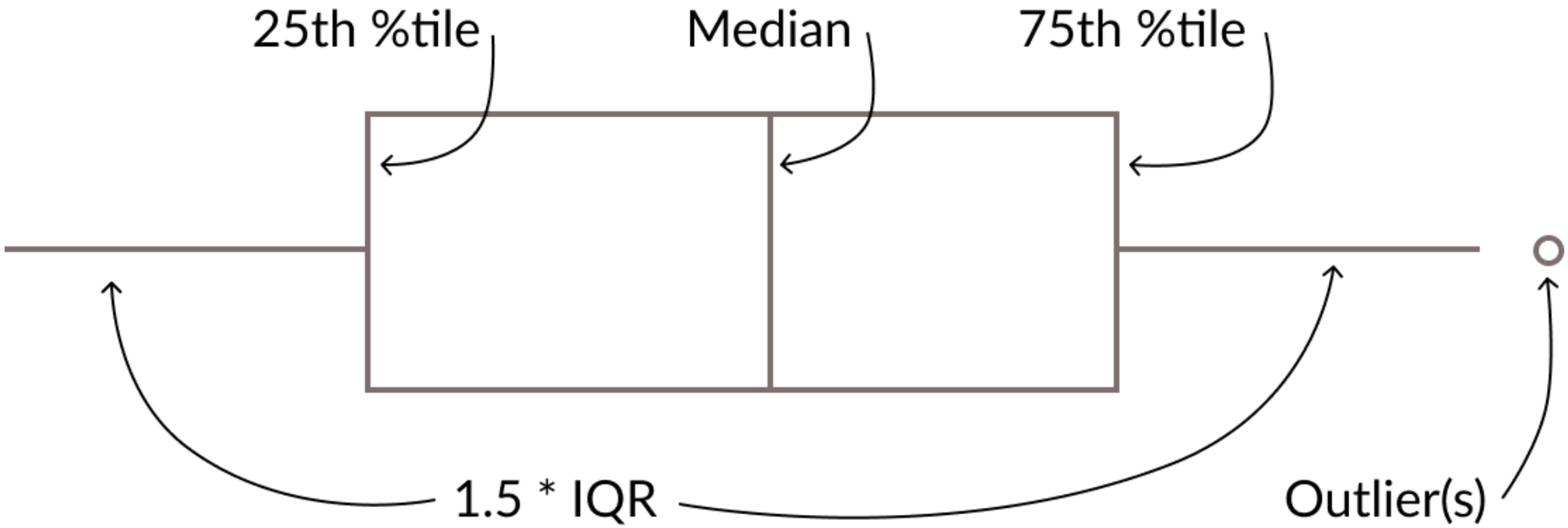


Why not facet histogams?

```
ggplot(md_speeding, aes(x = speed_over)) +
  geom_histogram() +
  facet_grid(vehicle_color ~ .)
```

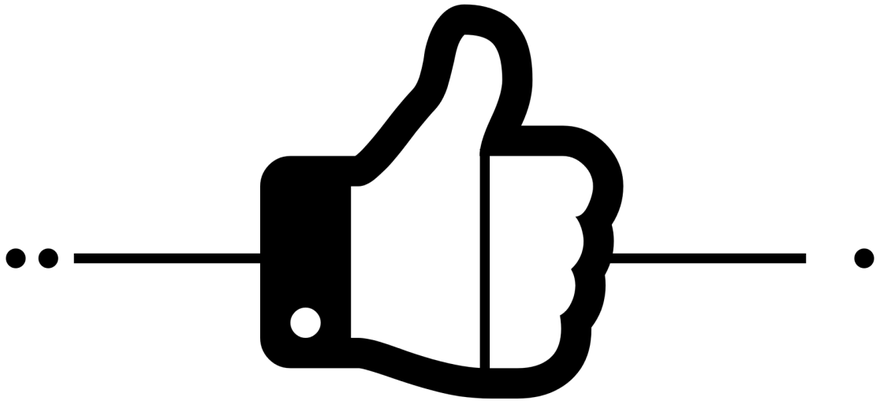


The boxplot



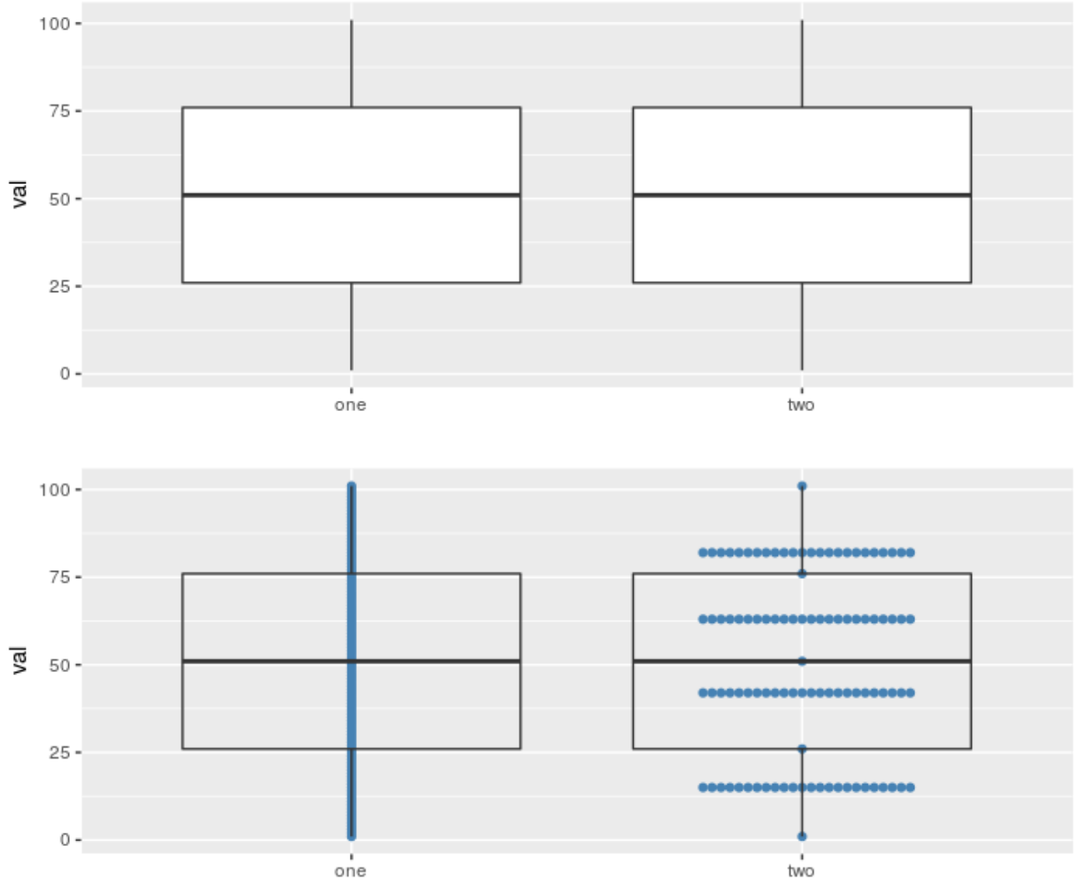
Boxplot pros

- Familiar
- Lots of good summary statistics



boxplot cons

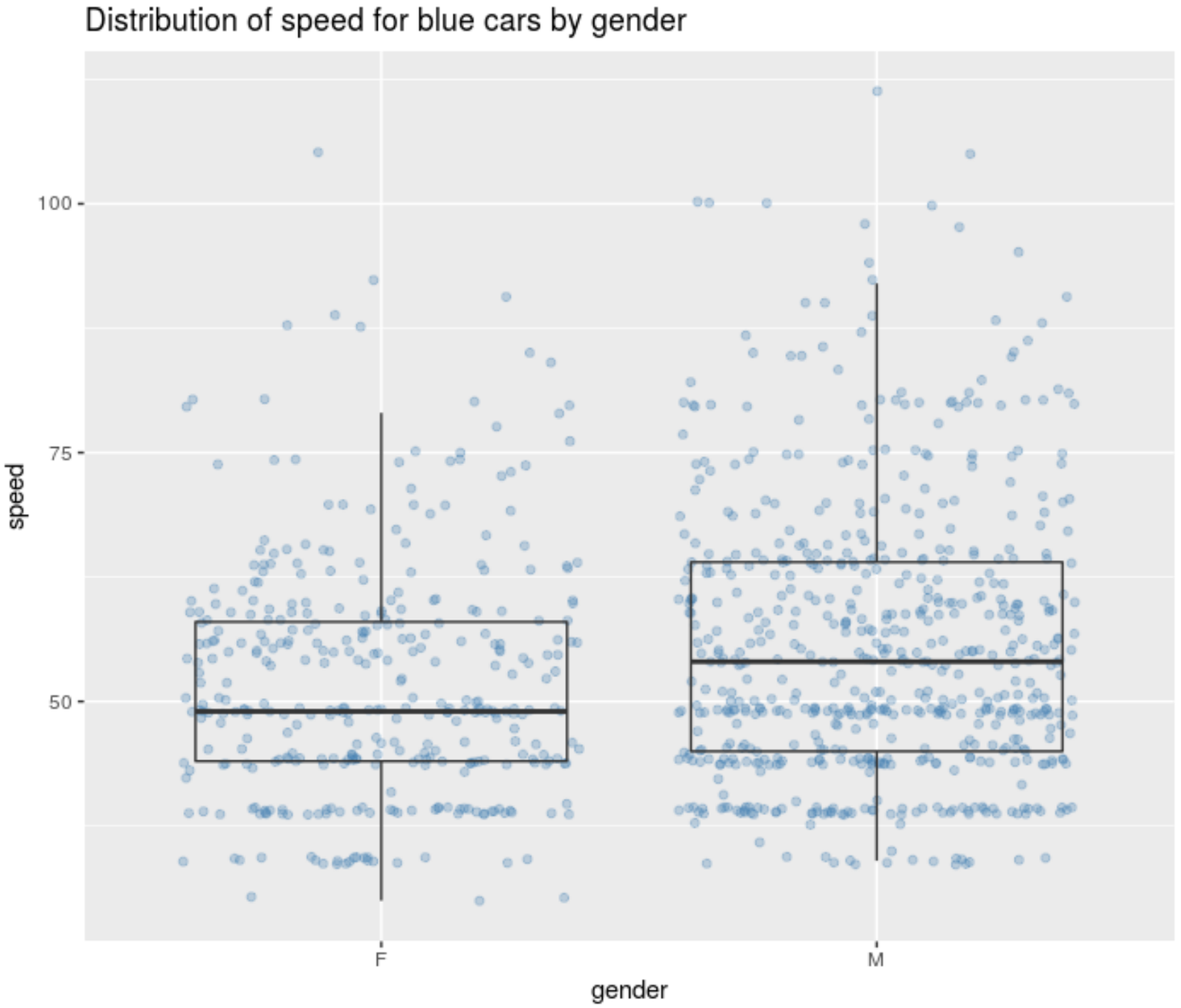
- Show me the data!



A simple addition

- `geom_jitter()` shows raw points jostled to avoid overlap.
- Layer under your `geom_boxplot()`.

```
md_speeding %>%
  filter(vehicle_color == 'BLUE') %>%
  ggplot(aes(x = gender, y = speed)) +
    # Draw points behind
    geom_jitter(alpha = 0.3, color = 'steelblue') +
    # Make transparent
    geom_boxplot(alpha = 0) +
    labs(title = 'Distribution of speed for blue cars by gender')
```


Distribution of speed for blue cars by gender
speed
100
75
50
F
M
gender

Let's compare some distributions!

VISUALIZATION BEST PRACTICES IN R

Boxplot alternatives

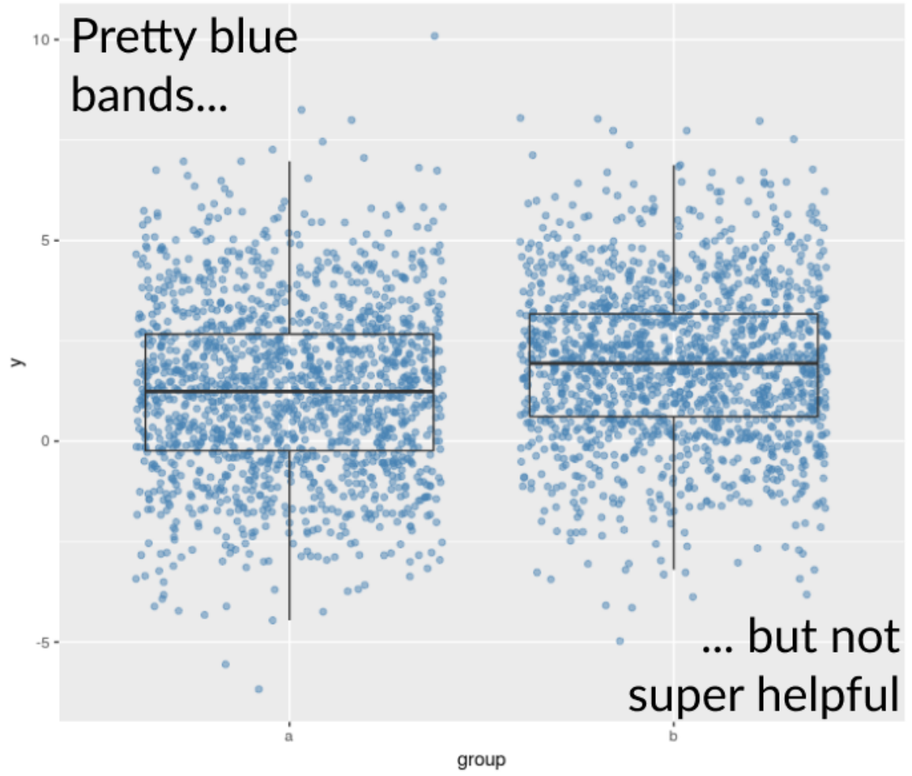
VISUALIZATION BEST PRACTICES IN R

Nick Strayer
Instructor

datacamp

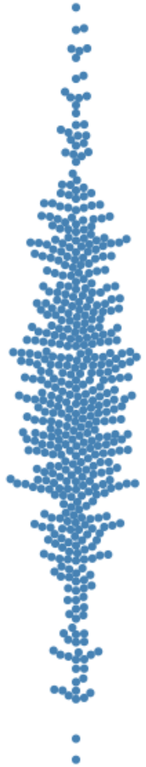
Limitations of the boxplot with jitter

- Jostling points can only deal with so much overlap
- Hard to get an idea of data density

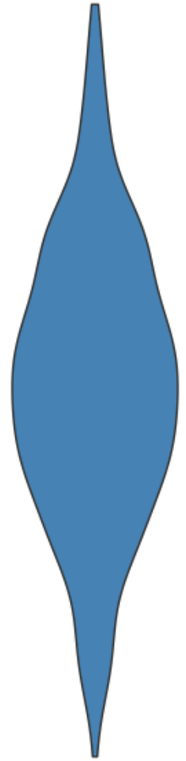


What are some other options?

Beeswarm plots



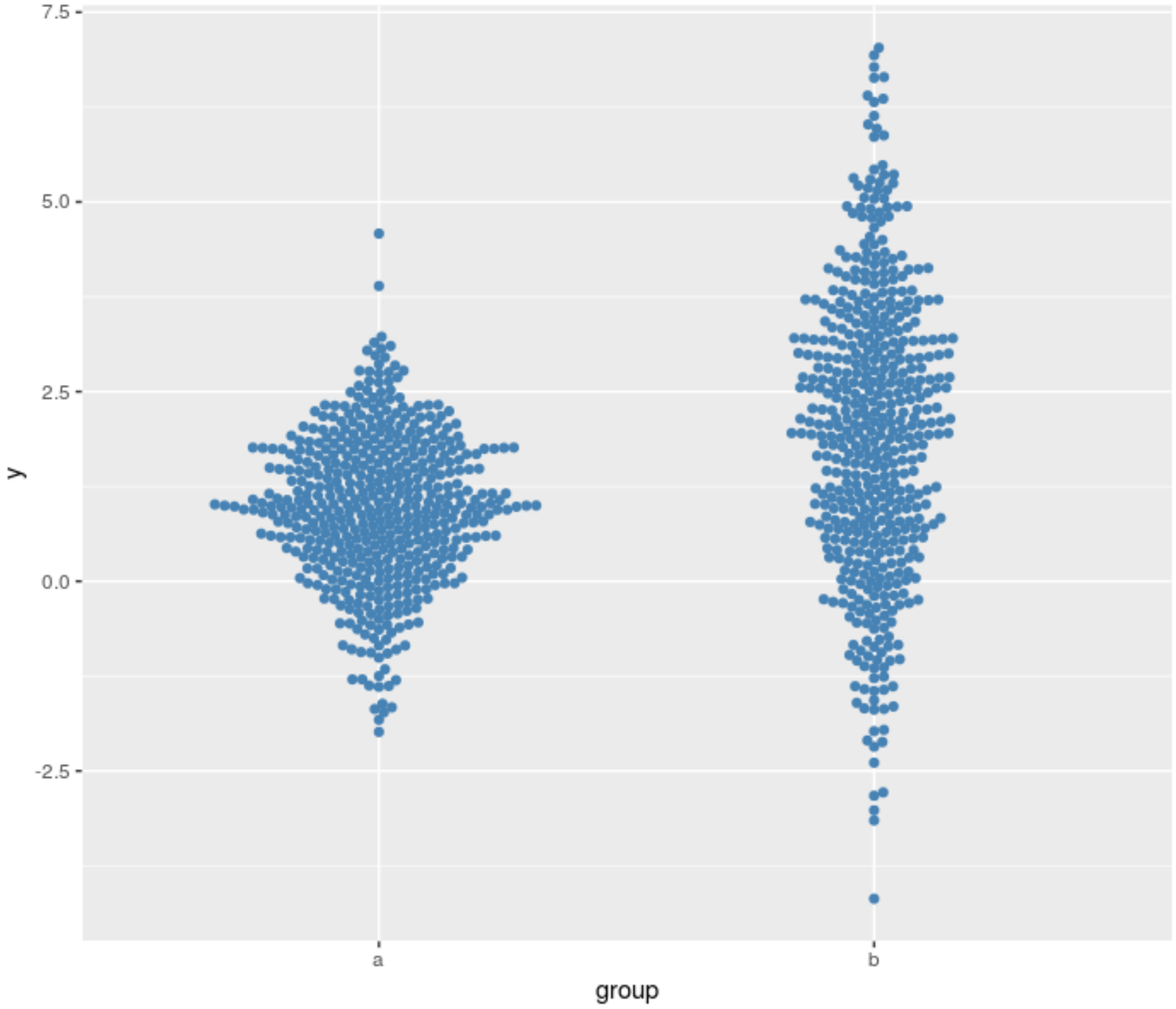
Violin plots



Beeswarm plots

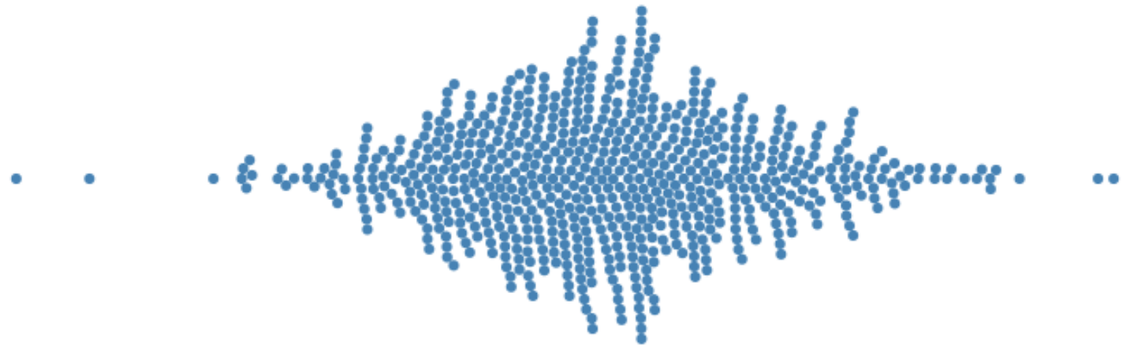
- 'Smart' jittering
- Individual points are clumped together as close to the axis as possible
- Handily included as `geom_beeswarm()` in the `ggbeeswarm` package.

```
library(ggbeeswarm)
ggplot(data, aes(y = y, x =  group)) +
  geom_beeswarm(color = 'steelblue')
```


7.5
5.0
2.5
0.0
-2.5
y
a
b
group

Beeswarm pros

- Individual data points
- Distributional shape



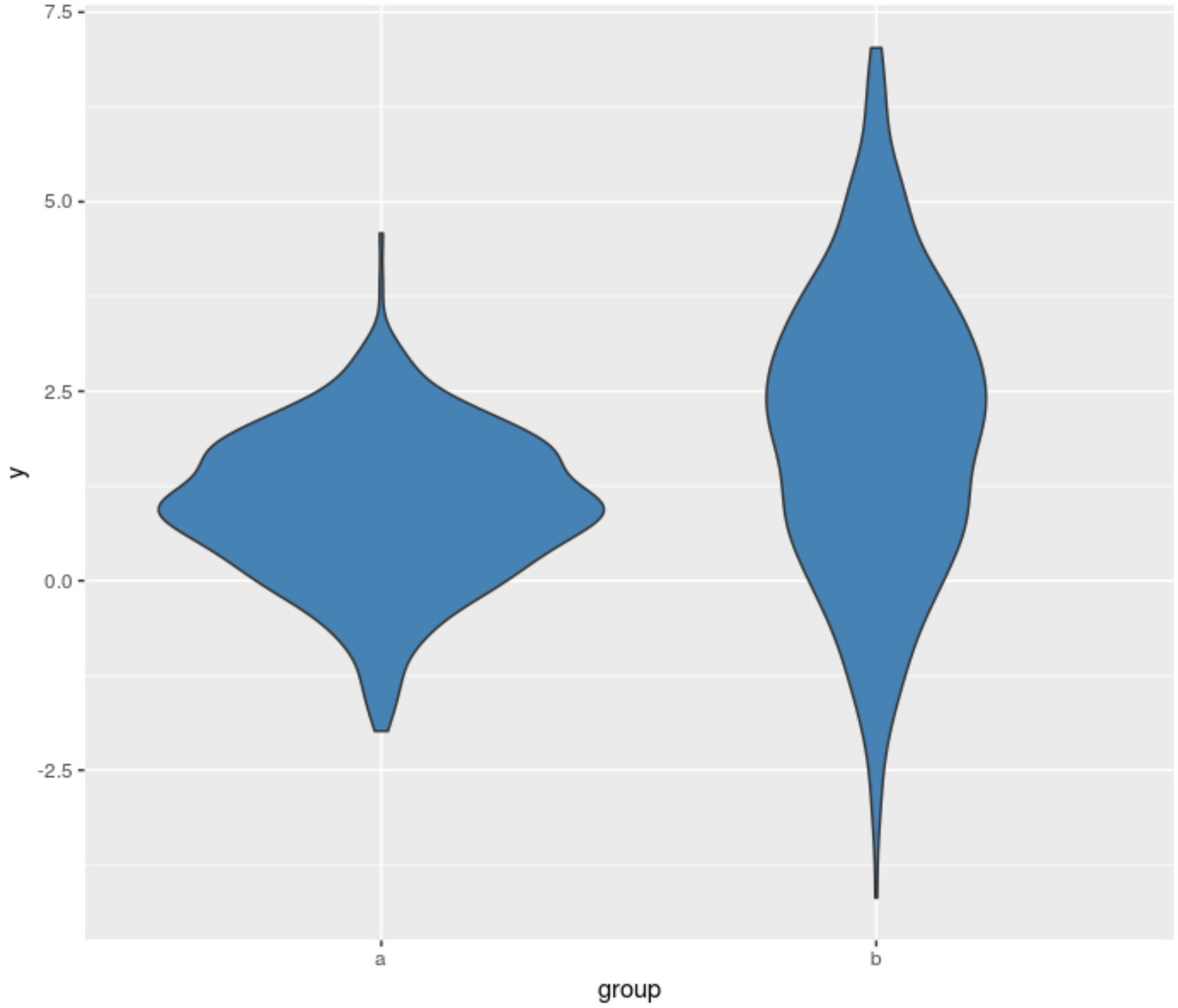
Beeswarm cons

- Get hard with lots of data
- Arbitrary stacking

Violin plots

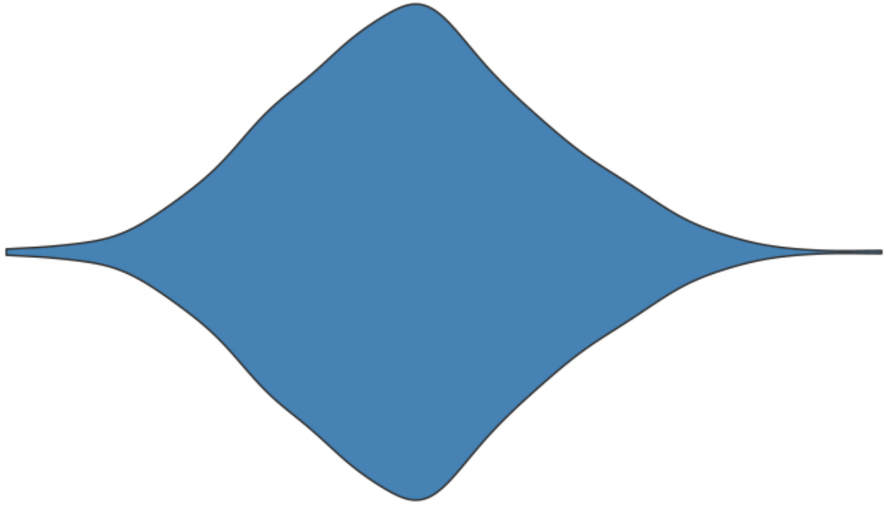
- KDE reflected to be symmetric
- Just replace `geom_boxplot()` with `geom_violin()`.

```
ggplot(data, aes(y = y, x = group)) +
  geom_violin(fill = 'steelblue')
```


7.5
5.0
2.5
0.0
-2.5
y
a
b
group

Violin pros

- Every data point is heard
- Not every data point is seen, so good for lots of data.



Violin cons

- Kernel width choice
- Not every data point is seen

Let's try some more advanced comparisons!

VISUALIZATION BEST PRACTICES IN R

Comparing spatially-related distributions

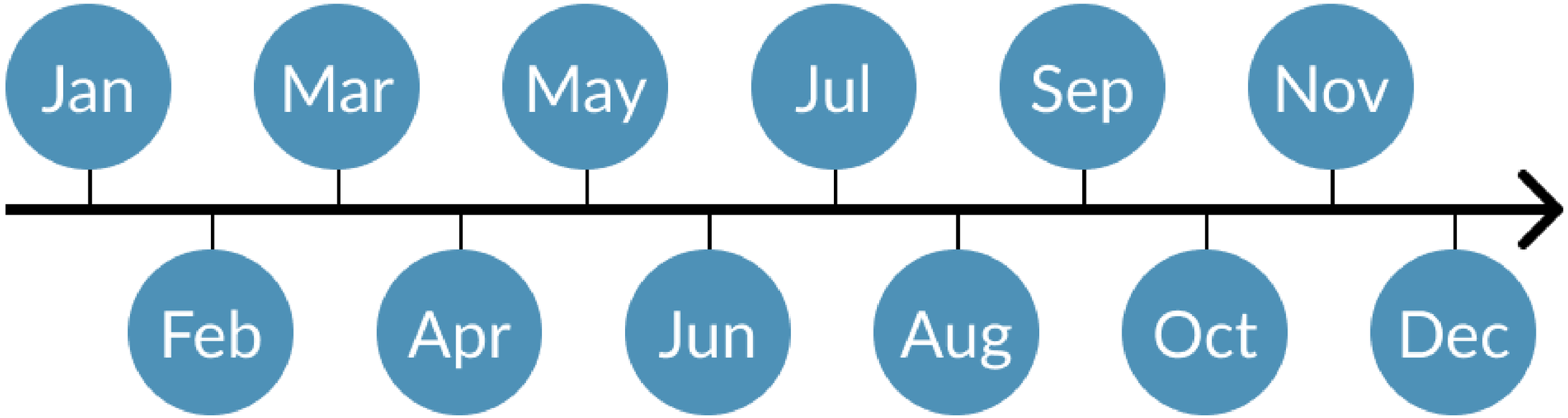
VISUALIZATION BEST PRACTICES IN R

Nick Strayer
Instructor

datacamp

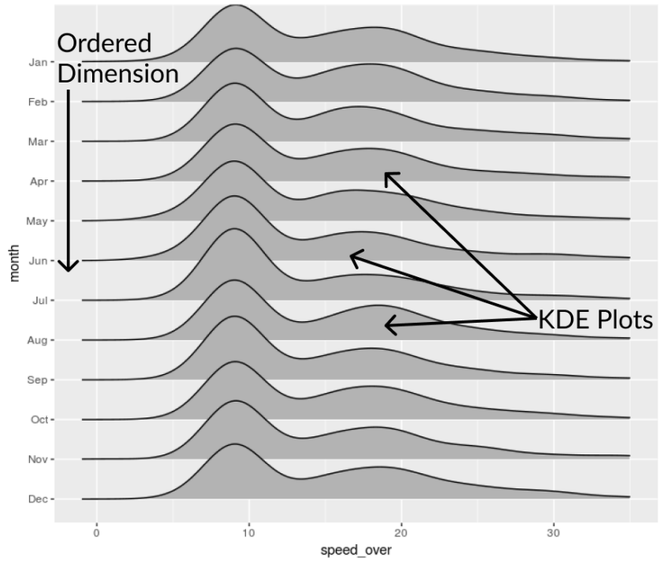
What are 'spatially connected axes'?

- There is an underlying ordering of the classes.
- E.g. months of the year: Jan < Feb < Mar < ...

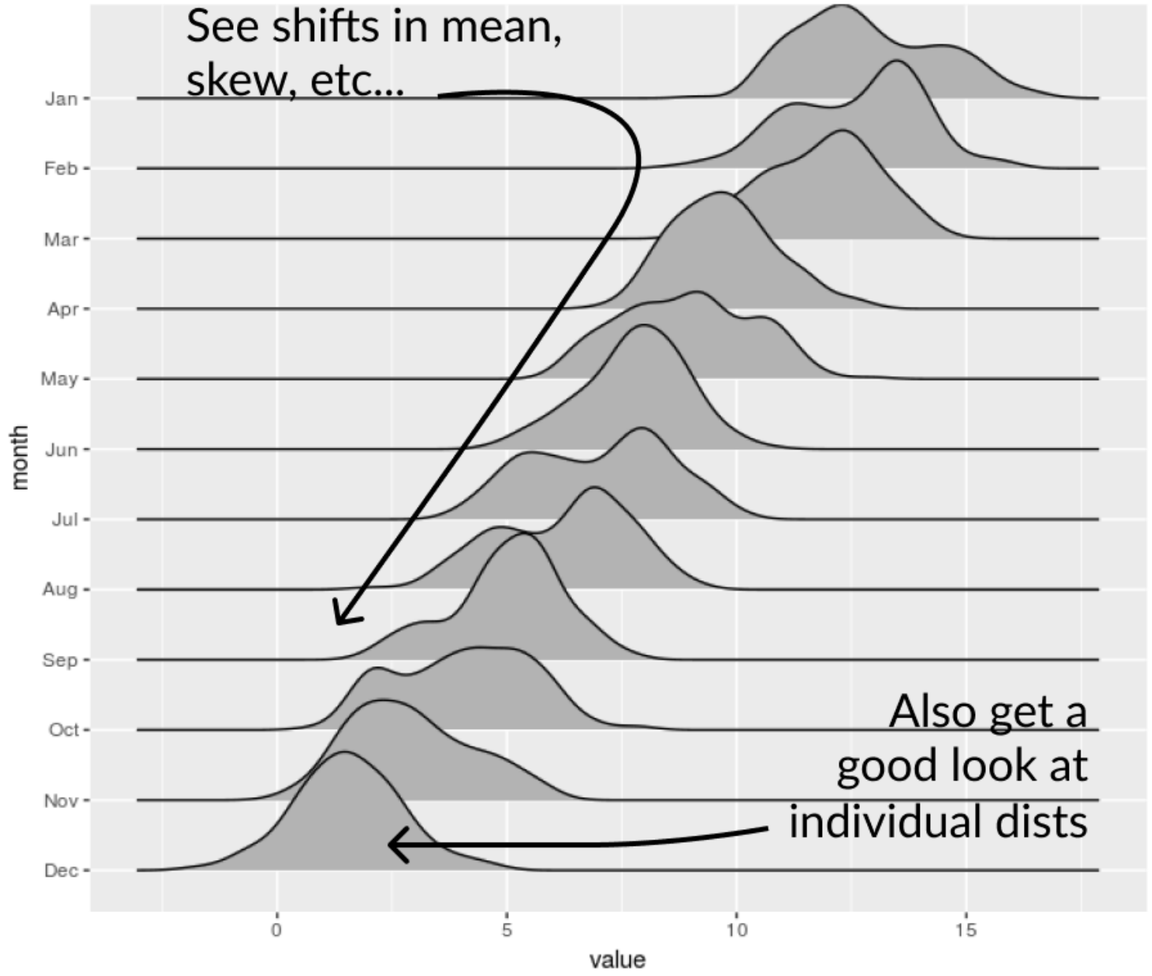


The ridgeline plot

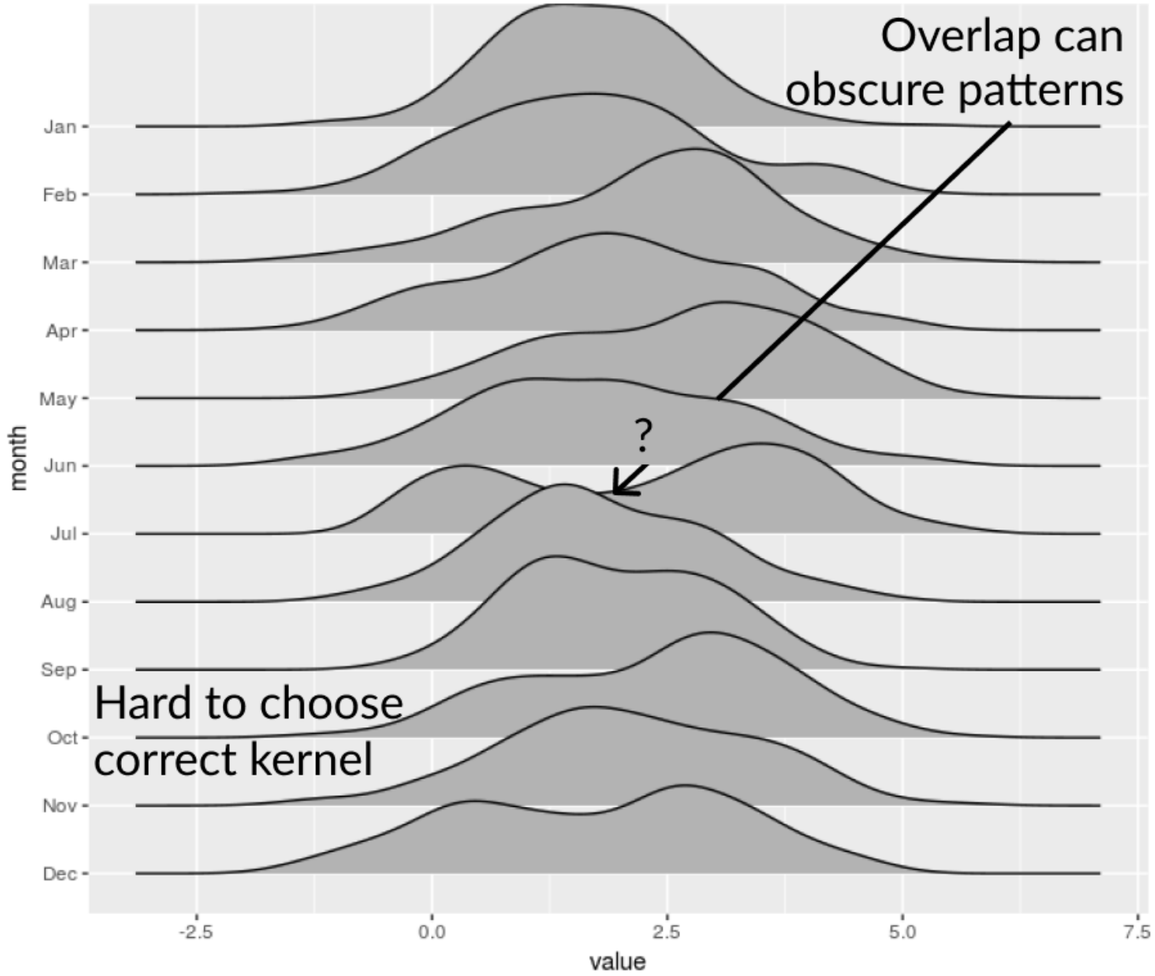
```
library(ggridges) # Gives us geom_density_ridges()
ggplot(md_speeding, aes(x = speed_over, y = month)) +
  geom_density_ridges(bandwidth = 2) +
  xlim(1, 35)
```



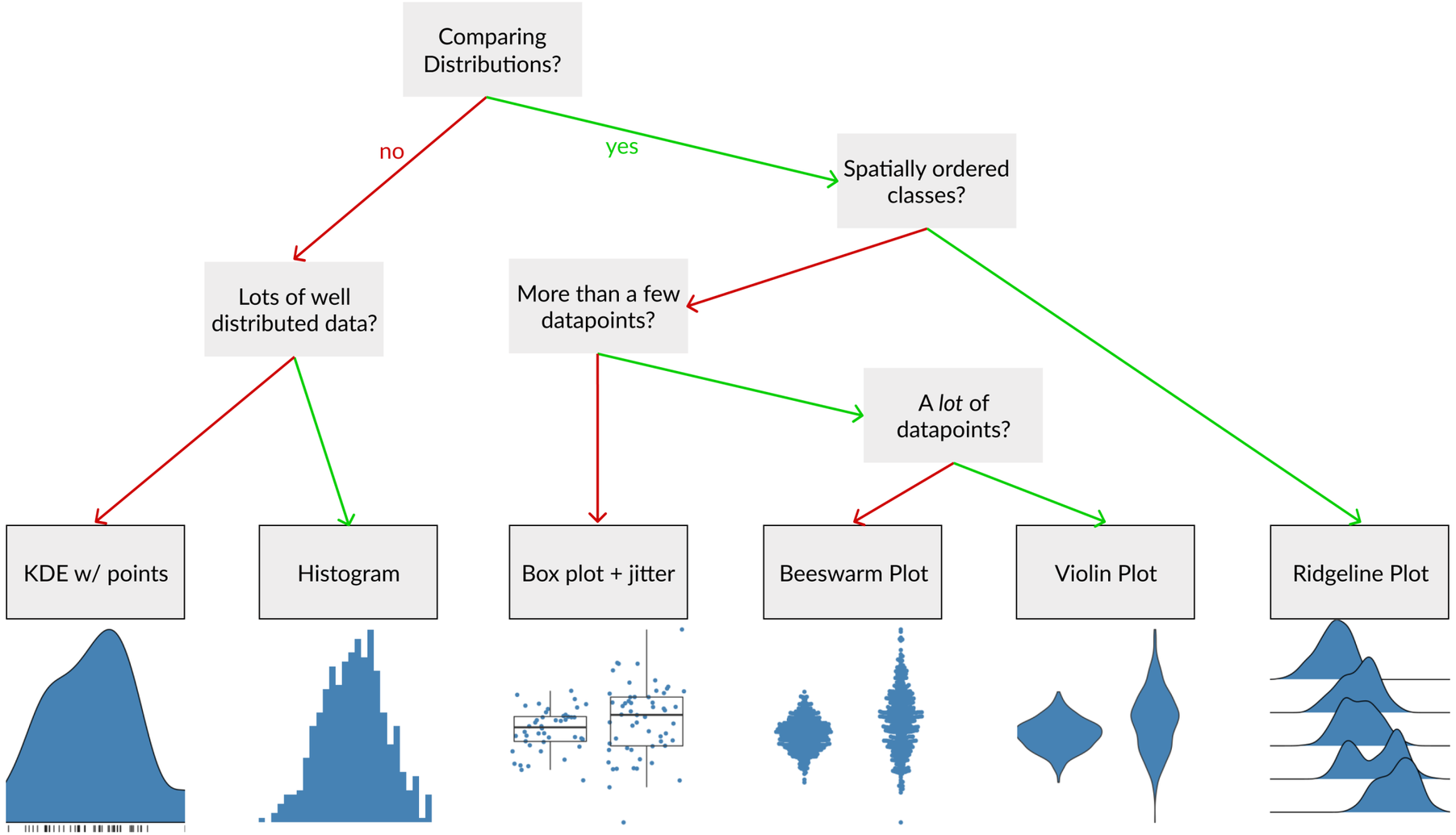
Ridgeline pros



Ridgeline cons



Overview of distribution visualizations



Let's make some ridgelines!

VISUALIZATION BEST PRACTICES IN R

Congratulations!

VISUALIZATION BEST PRACTICES IN R

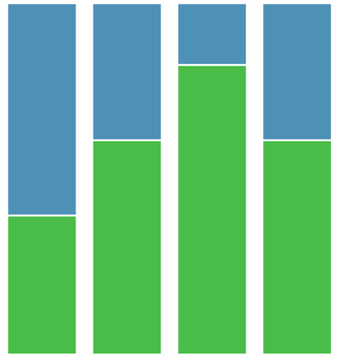
Nick Strayer
Instructor

Chapter 1: Proportions

Three or less classes and precision not important?

Need more precision and have the space?

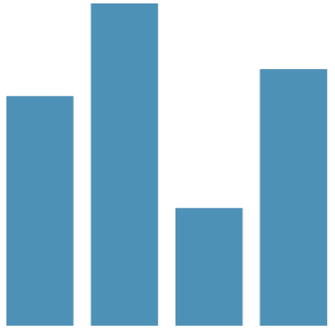
Good for comparing values *across* populations....



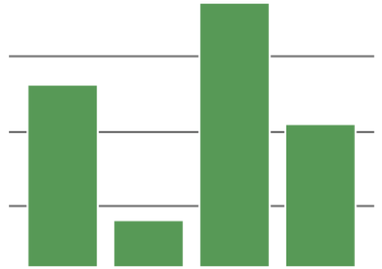
... bad for comparing values *within* populations

Chapter 2: Point data

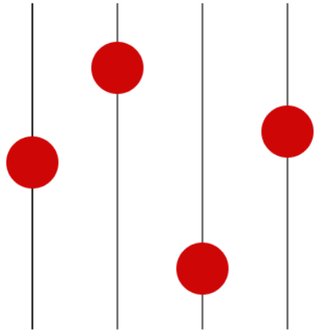
Data is a stackable quantity? E.g. dollars, counts...



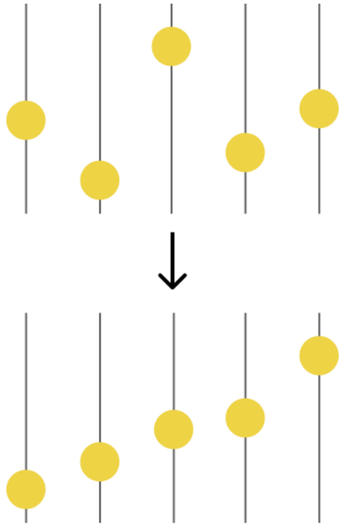
No need for vertical grid lines on bars.



Not stackable? E.g. percents, ratio...

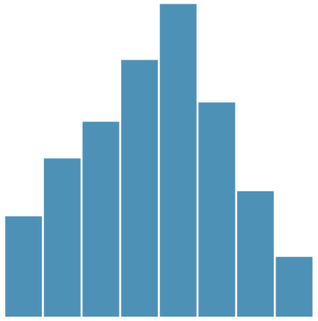


Ordering is almost always a good idea.

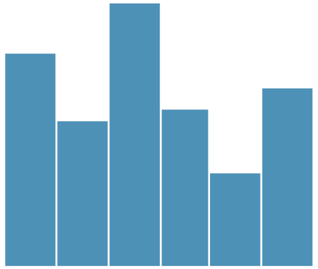


Chapter 3: Single distributions

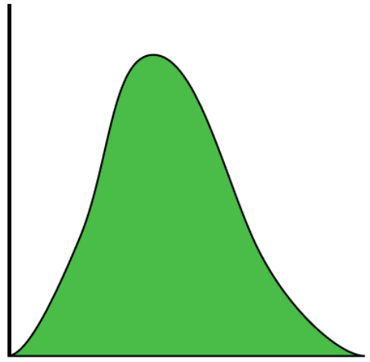
Histograms are simple and intuitive...



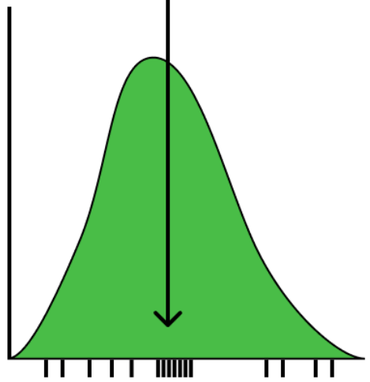
... but are sensitive to bin width and placement choices.



Kernel Density Estimators work with small data and produce nice smooth plots.

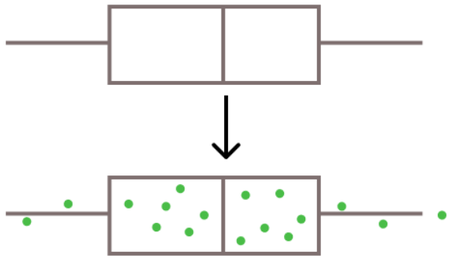


But always try and show raw data if possible to see gaps the KDE may be filling.

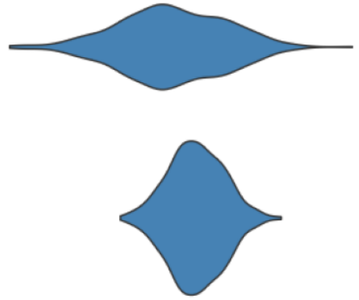


Chapter 4: Comparing distributions

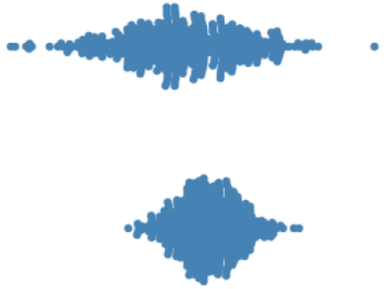
Boxplots hide a lot of data, so augment them with jittered points



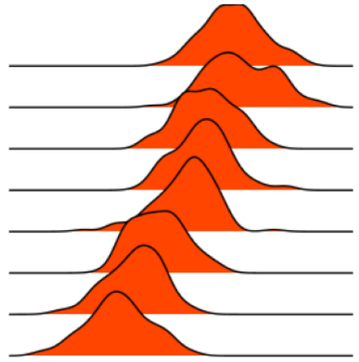
Violin plots are symmetric KDEs that work well when you have lots of points.



Beeswarm plots are an alternative 'smart' jittering that shows density.



If you have spatially ordered groups, consider the ridgeline plot.



Going further

Flowing data

Curated list of data visualizations and R-based tutorials.

Datawrapper Blog

Articles that dig deep into visualization techniques and mistakes.

Twitter (#datavis)

An ongoing stream of cool projects and inspiration.

Books!

- **Data Visualization**, Andy Kirk
- **The Functional Art and The Truthful Art** by Alberto Cairo

Thank you!

VISUALIZATION BEST PRACTICES IN R