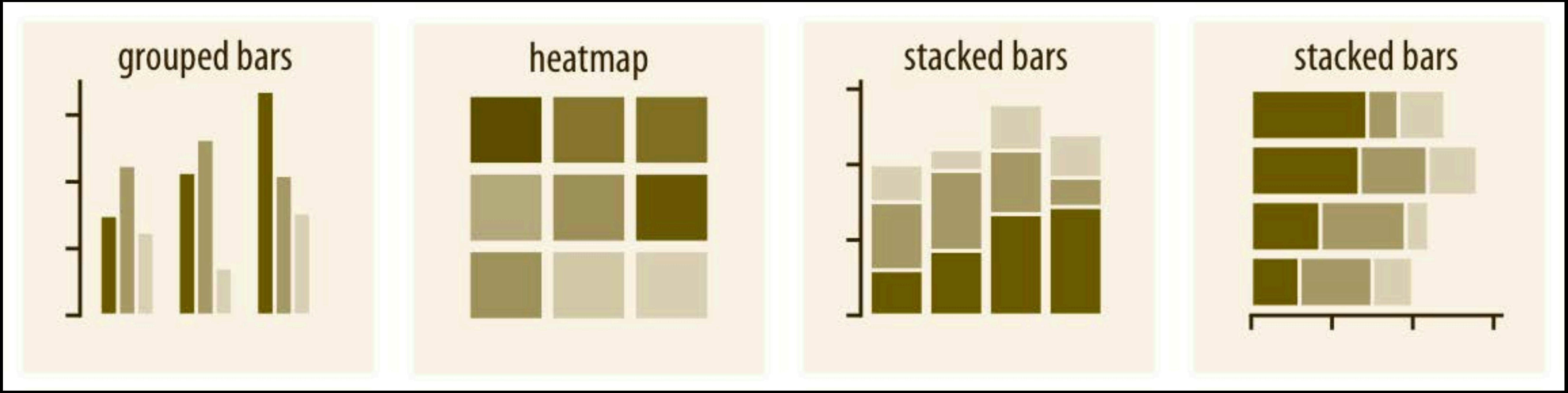
ANT 6973: DATA VISUALIZATION AND EXPLORATION

VISUALIZING DISTRIBUTIONS AND UNCERTAINTY

PREVIOUSLY…

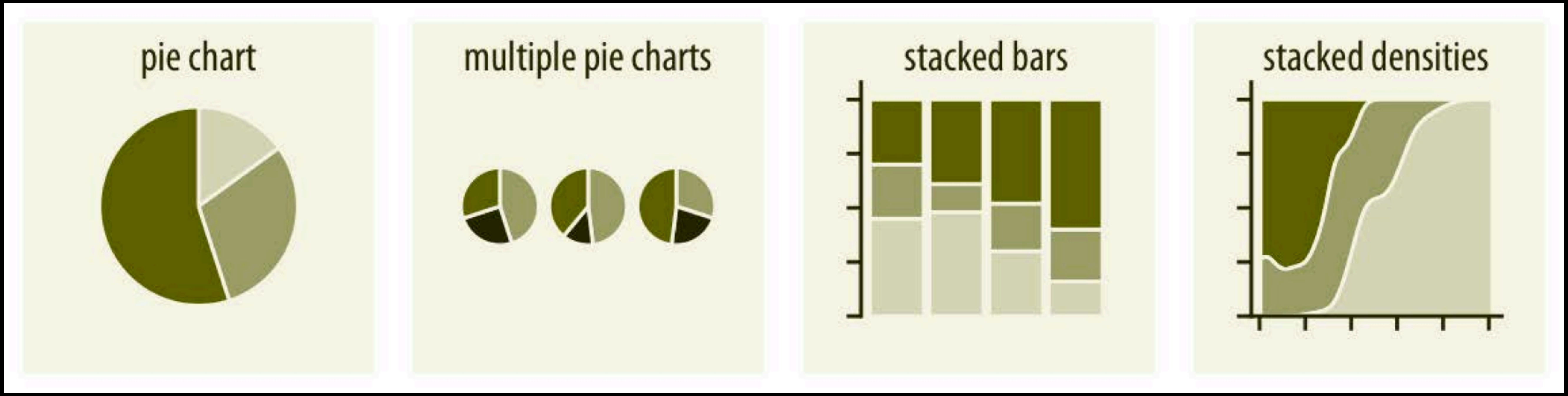
PREVIOUSLY...

- Visualizing amounts with bar charts (and variants)



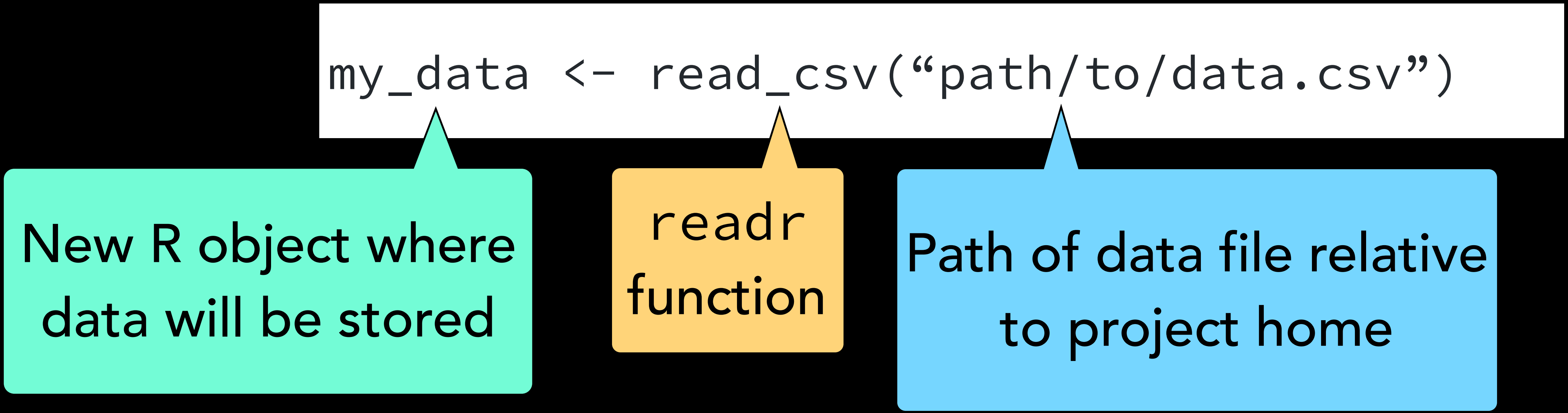
PREVIOUSLY...

- Visualizing proportions with pies, stacked bars, and stacked areas



PREVIOUSLY…

- Import from a .csv file using `read_csv()` the readr package.



```
my_data <- read_csv(“path/to/data.csv”)
```

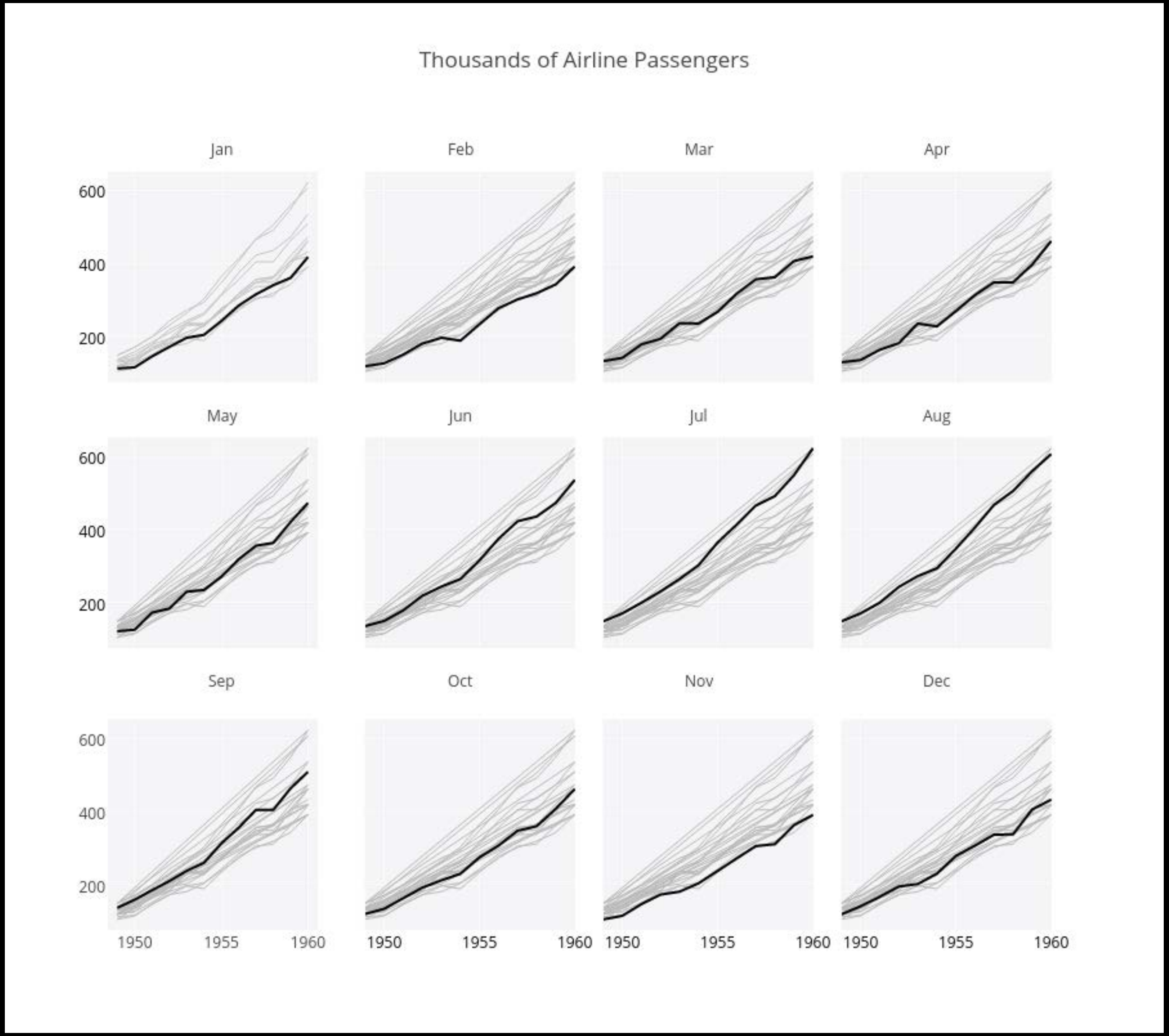
New R object where data will be stored

readr function

Path of data file relative to project home

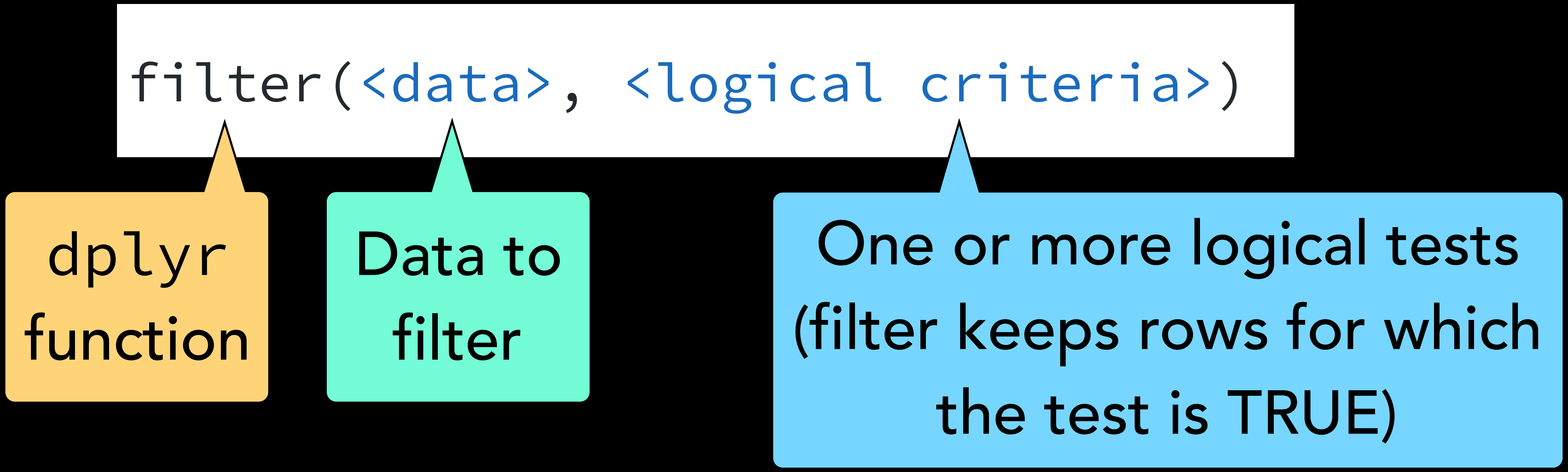
PREVIOUSLY...

- Create small multiples (plots of slices of data) using `facet_wrap` and `facet_grid()`.



PREVIOUSLY...

- Obtain subsets of data with the `filter()` function



```
filter(<data>, <logical criteria>)
```

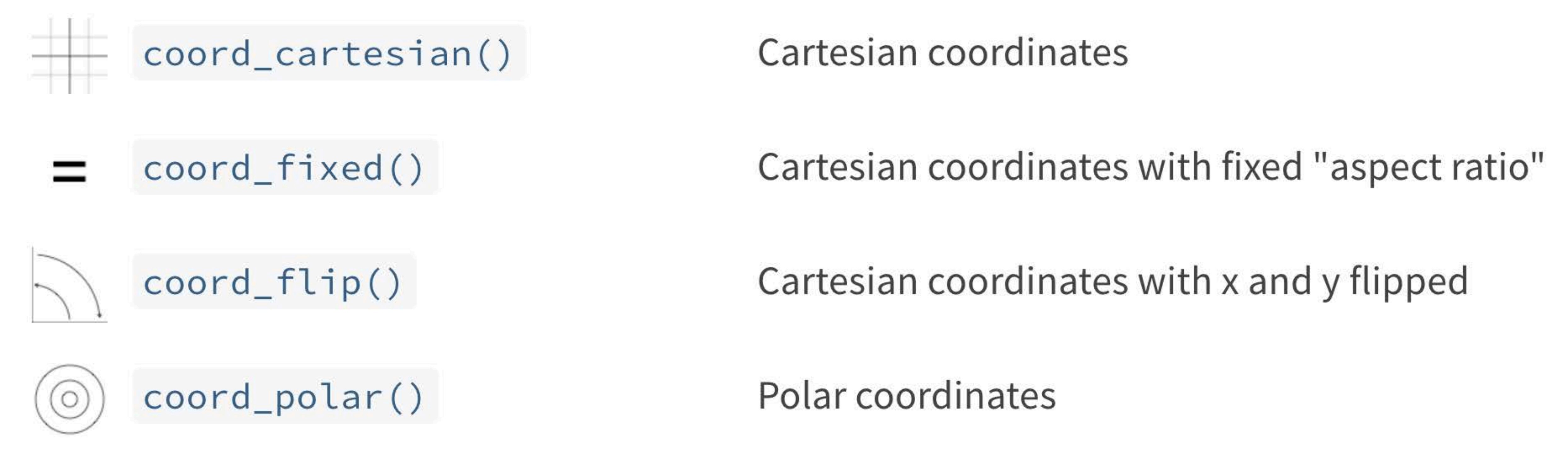
`dplyr` function

Data to filter

One or more logical tests (filter keeps rows for which the test is TRUE)

PREVIOUSLY...

- Alternate coordinate systems `coord_polar()` and `coord_flip()`.

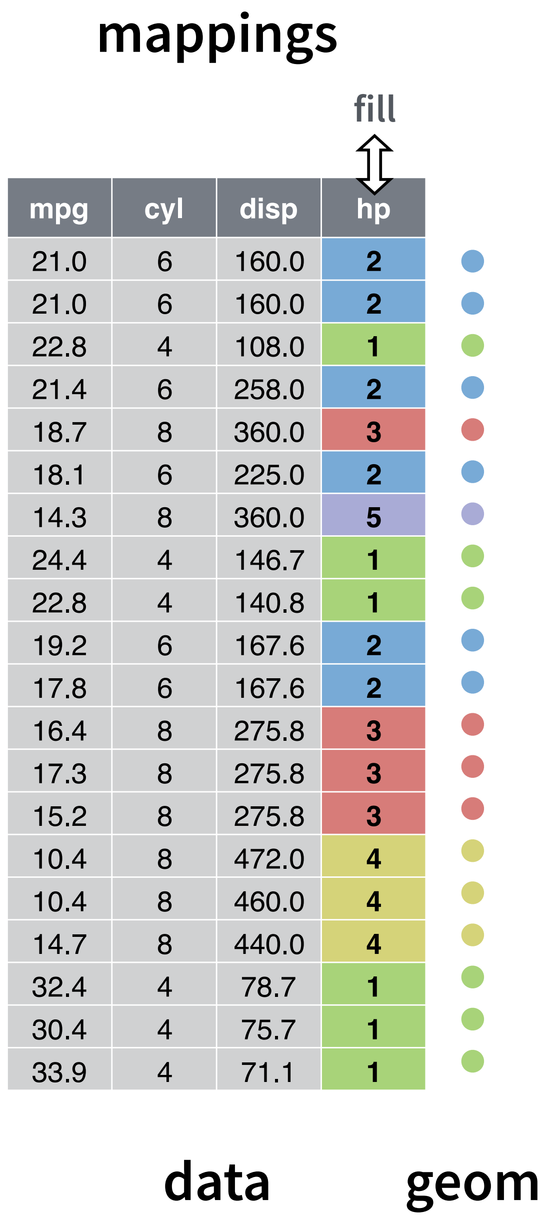


#	`coord_cartesian()`	Cartesian coordinates
=	`coord_fixed()`	Cartesian coordinates with fixed "aspect ratio"
	`coord_flip()`	Cartesian coordinates with x and y flipped
◎	`coord_polar()`	Polar coordinates

TODAY'S TOPICS

- Visualizing single distributions
- Visualizing multiple distributions
- Visualizing uncertainty
- Activities:
  - Body mass distributions of late Quaternary mammals (interactive)
  - Continue on baboon activities and any others not completed

OUR TEMPLATE



mpg	cyl	disp	hp
21.0	6	160.0	2
21.0	6	160.0	2
22.8	4	108.0	1
21.4	6	258.0	2
18.7	8	360.0	3
18.1	6	225.0	2
14.3	8	360.0	5
24.4	4	146.7	1
22.8	4	140.8	1
19.2	6	167.6	2
17.8	6	167.6	2
16.4	8	275.8	3
17.3	8	275.8	3
15.2	8	275.8	3
10.4	8	472.0	4
10.4	8	460.0	4
14.7	8	440.0	4
32.4	4	78.7	1
30.4	4	75.7	1
33.9	4	71.1	1

1. Pick a **data** set

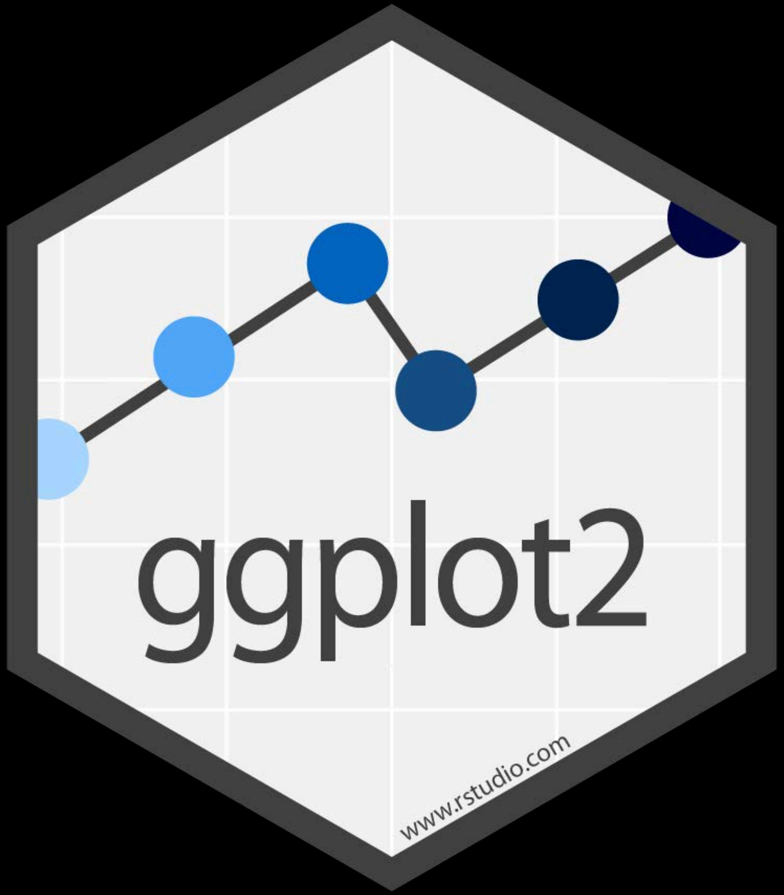
```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

2. Choose a **geom** to display cases

3. **Map** aesthetic properties to variables

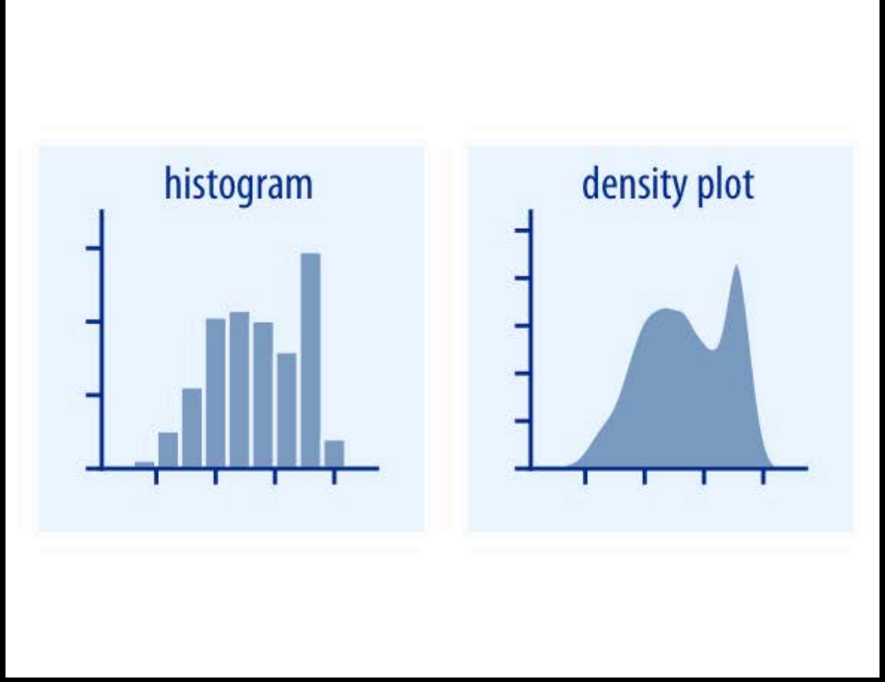
CC by RStudio

VISUALIZING DISTRIBUTIONS

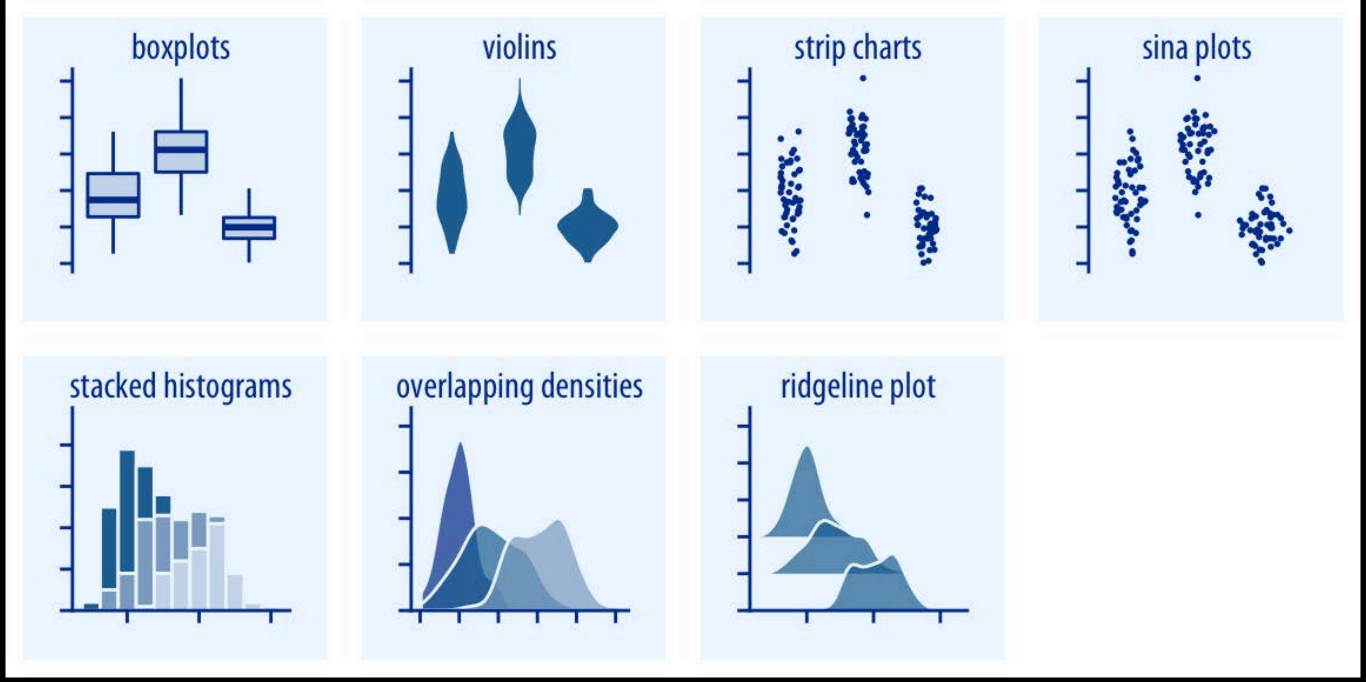


VISUALIZING DISTRIBUTIONS

Single distributions



Multiple distributions

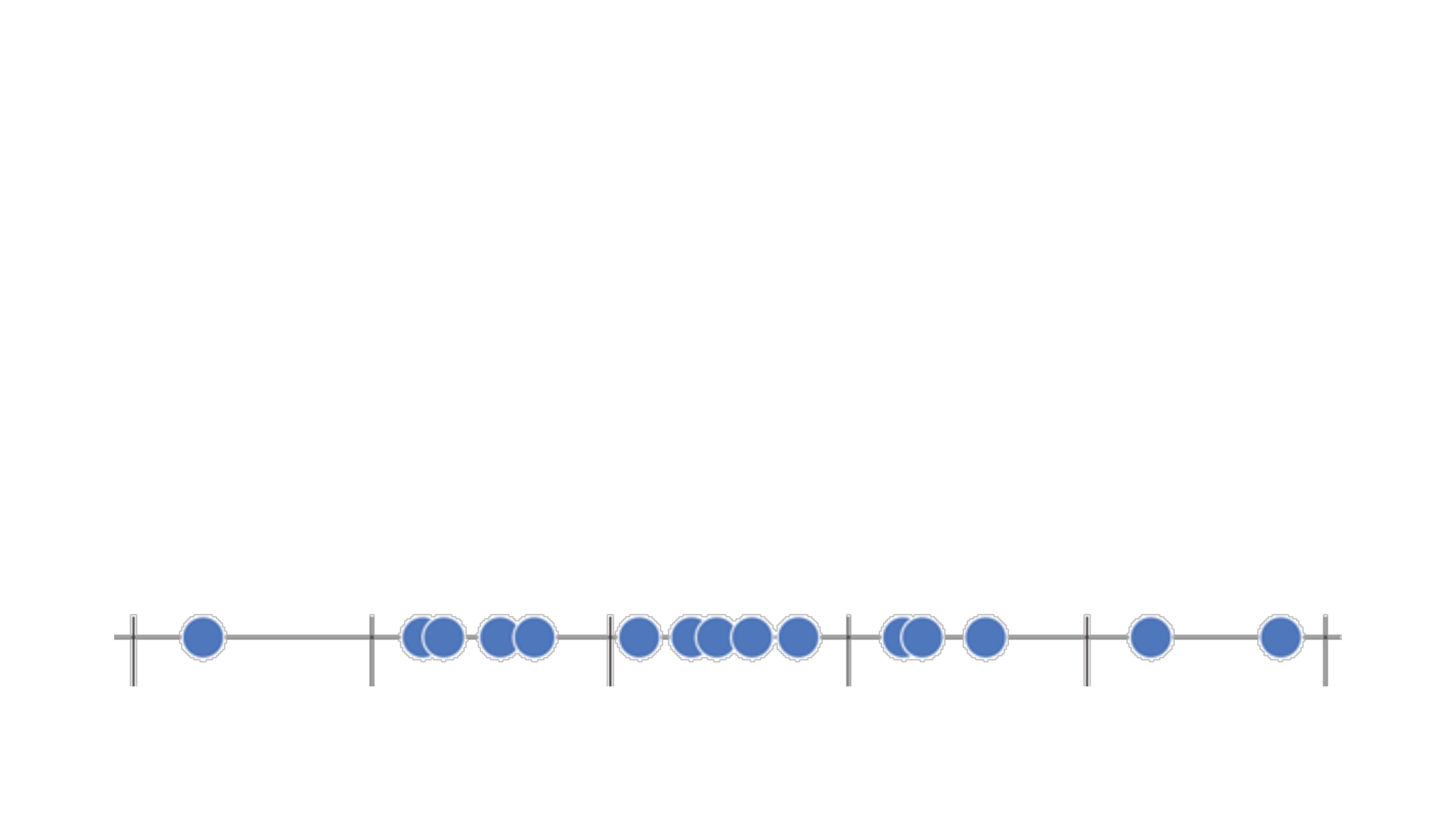


HISTOGRAMS

HISTOGRAMS

- Shows the distribution of a single variable

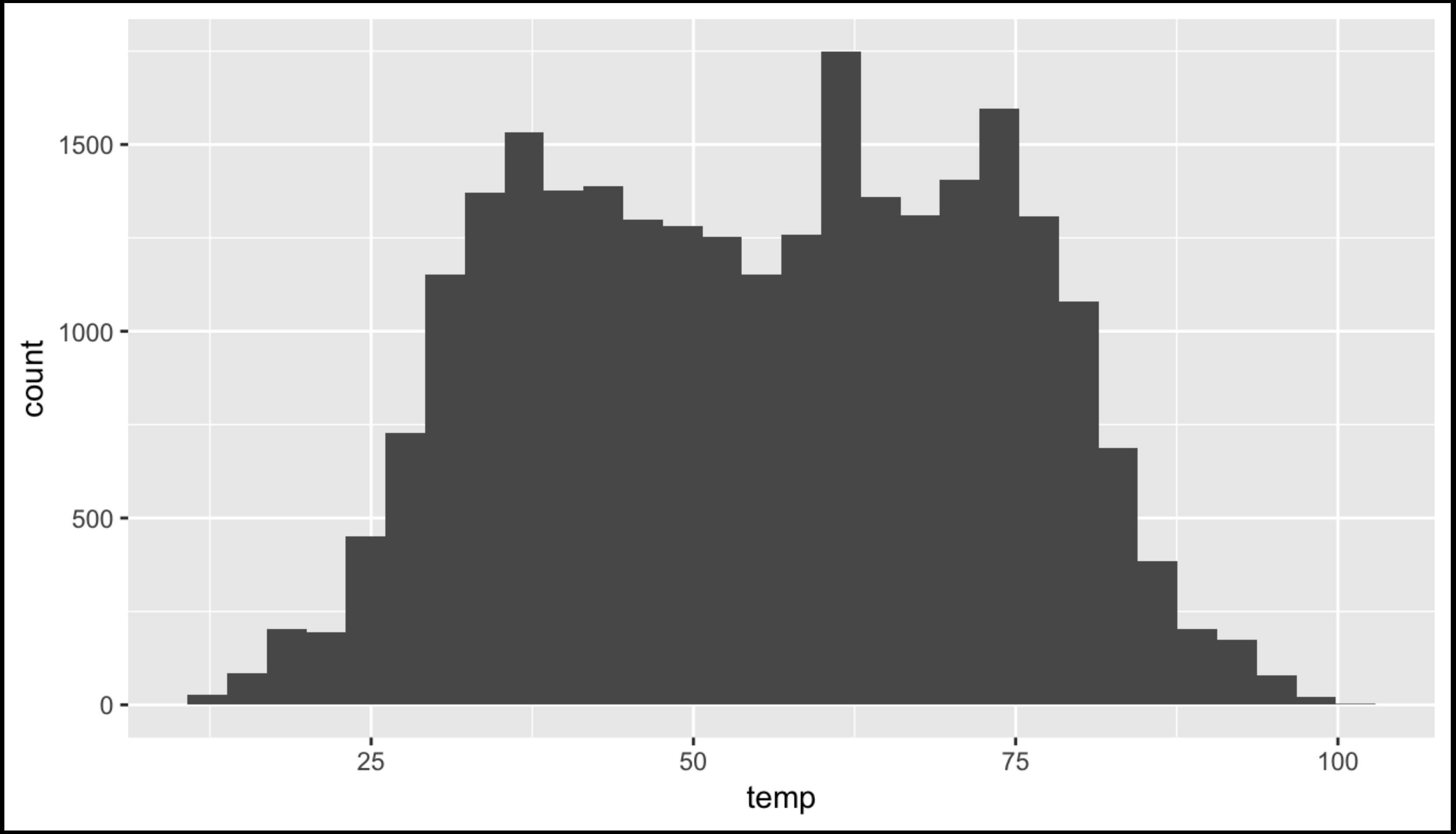
- Breaks the data apart into specific bins and shows *count* of cases in each bin



HISTOGRAMS

Major difference with bar charts

- Geometric object is `geom_histogram()`
- One numeric variable to mapped to x in `aes()`
- y-aesthetic computed automatically (don't map anything to y)



YOUR TURN

- Create a new folder for the week and a new R markdown file.

- Create a new R chunk and load the `tidyverse` package.

- Download the late Quaternary mammals data set from the course website (“late-quaternary-mammals.RDS”). This is a file format for single R objects.

- Read the file into R using:

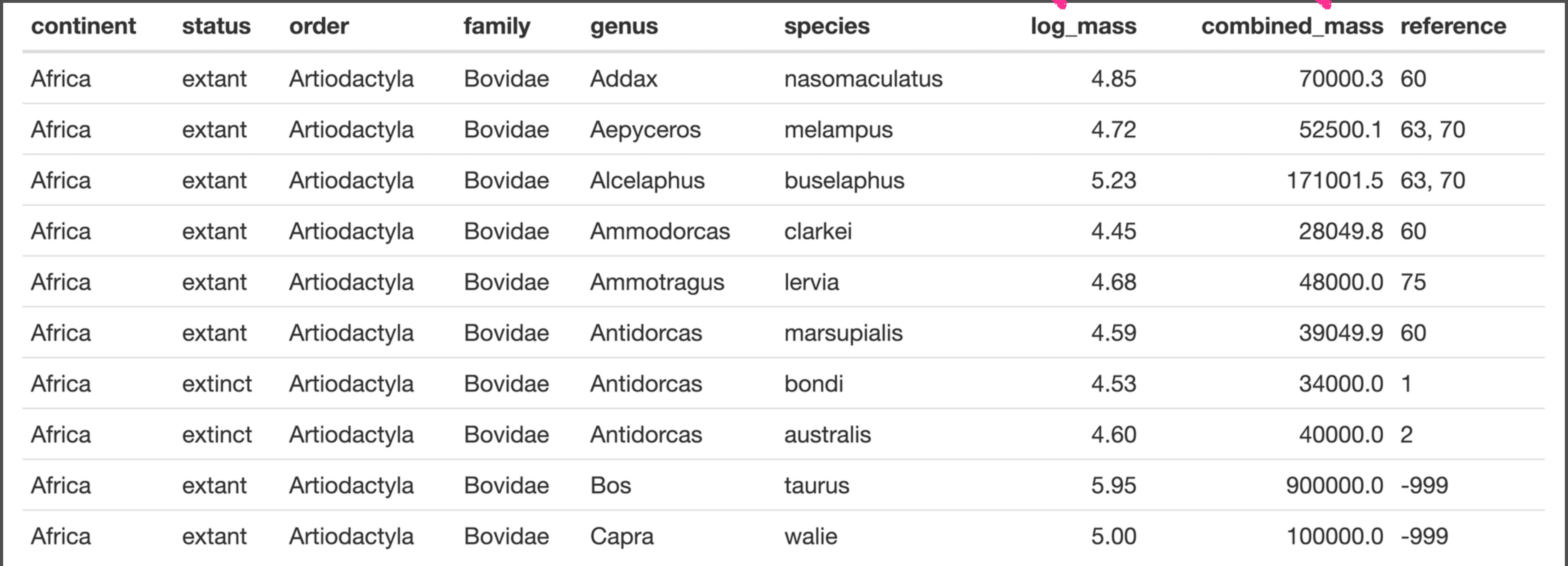
```
lqm <- readRDS(“your-path/late-quaternary-mammals.RDS”)
```

3:00

YOUR TURN

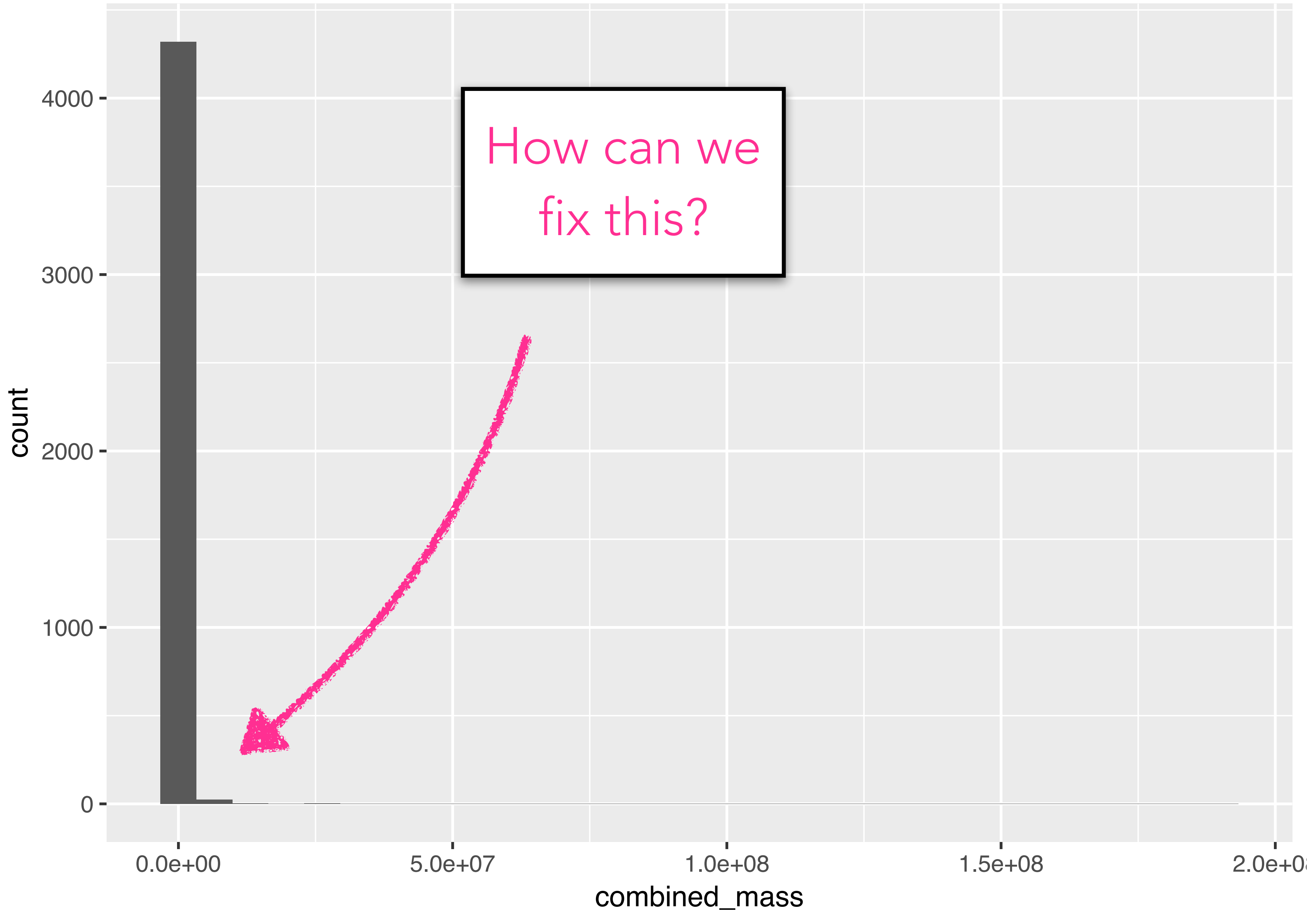
Average of M/F body masses (g)

log$_{10}$(combined_mass)



continent	status	order	family	genus	species	log_mass	combined_mass	reference
Africa	extant	Artiodactyla	Bovidae	Addax	nasomaculatus	4.85	70000.3	60
Africa	extant	Artiodactyla	Bovidae	Aepyceros	melampus	4.72	52500.1	63, 70
Africa	extant	Artiodactyla	Bovidae	Alcelaphus	buselaphus	5.23	171001.5	63, 70
Africa	extant	Artiodactyla	Bovidae	Ammodorcas	clarkei	4.45	28049.8	60
Africa	extant	Artiodactyla	Bovidae	Ammotragus	lervia	4.68	48000.0	75
Africa	extant	Artiodactyla	Bovidae	Antidorcas	marsupialis	4.59	39049.9	60
Africa	extinct	Artiodactyla	Bovidae	Antidorcas	bondi	4.53	34000.0	1
Africa	extinct	Artiodactyla	Bovidae	Antidorcas	australis	4.60	40000.0	2
Africa	extant	Artiodactyla	Bovidae	Bos	taurus	5.95	900000.0	-999
Africa	extant	Artiodactyla	Bovidae	Capra	walie	5.00	100000.0	-999

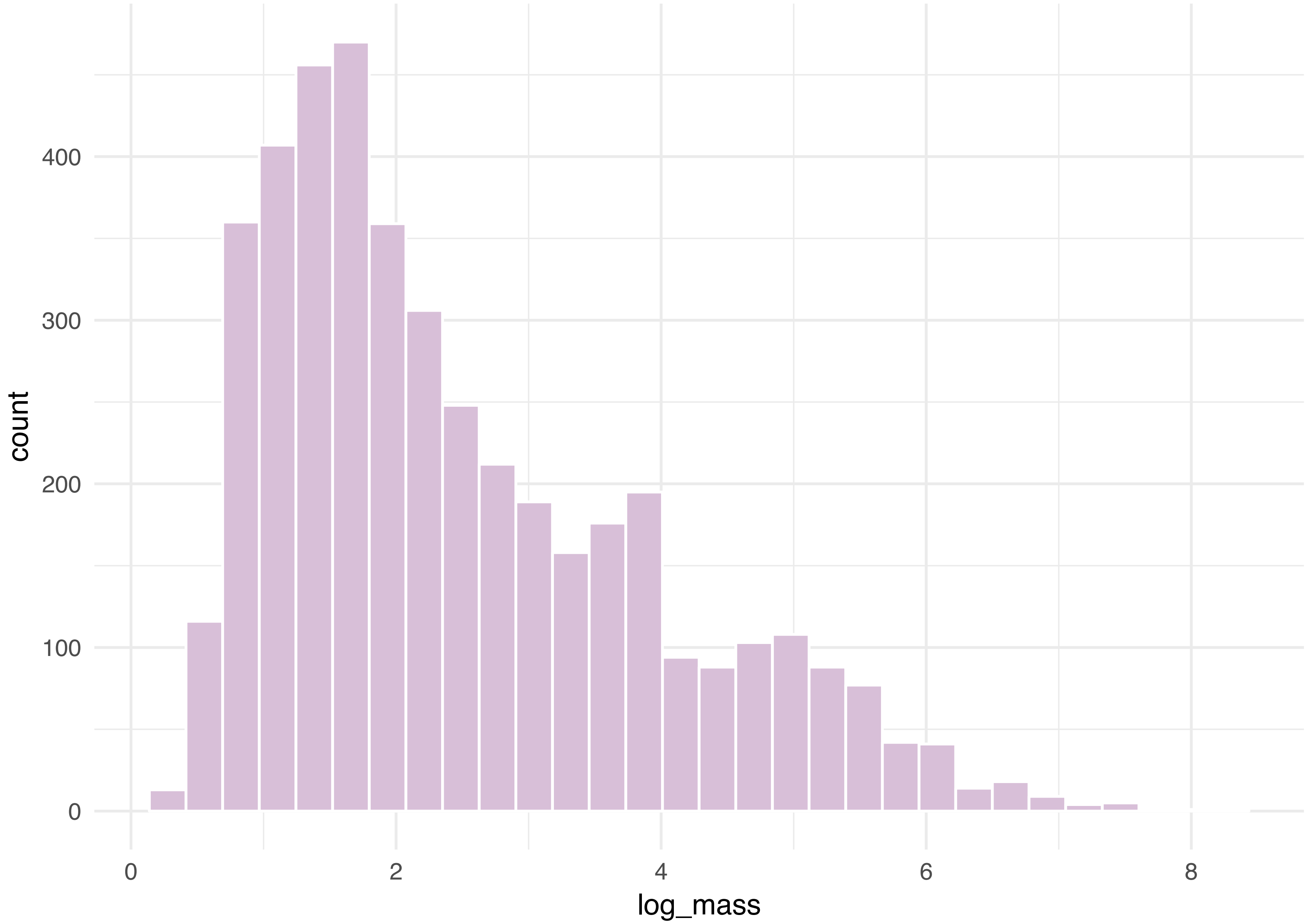
YOUR TURN



How can we fix this?

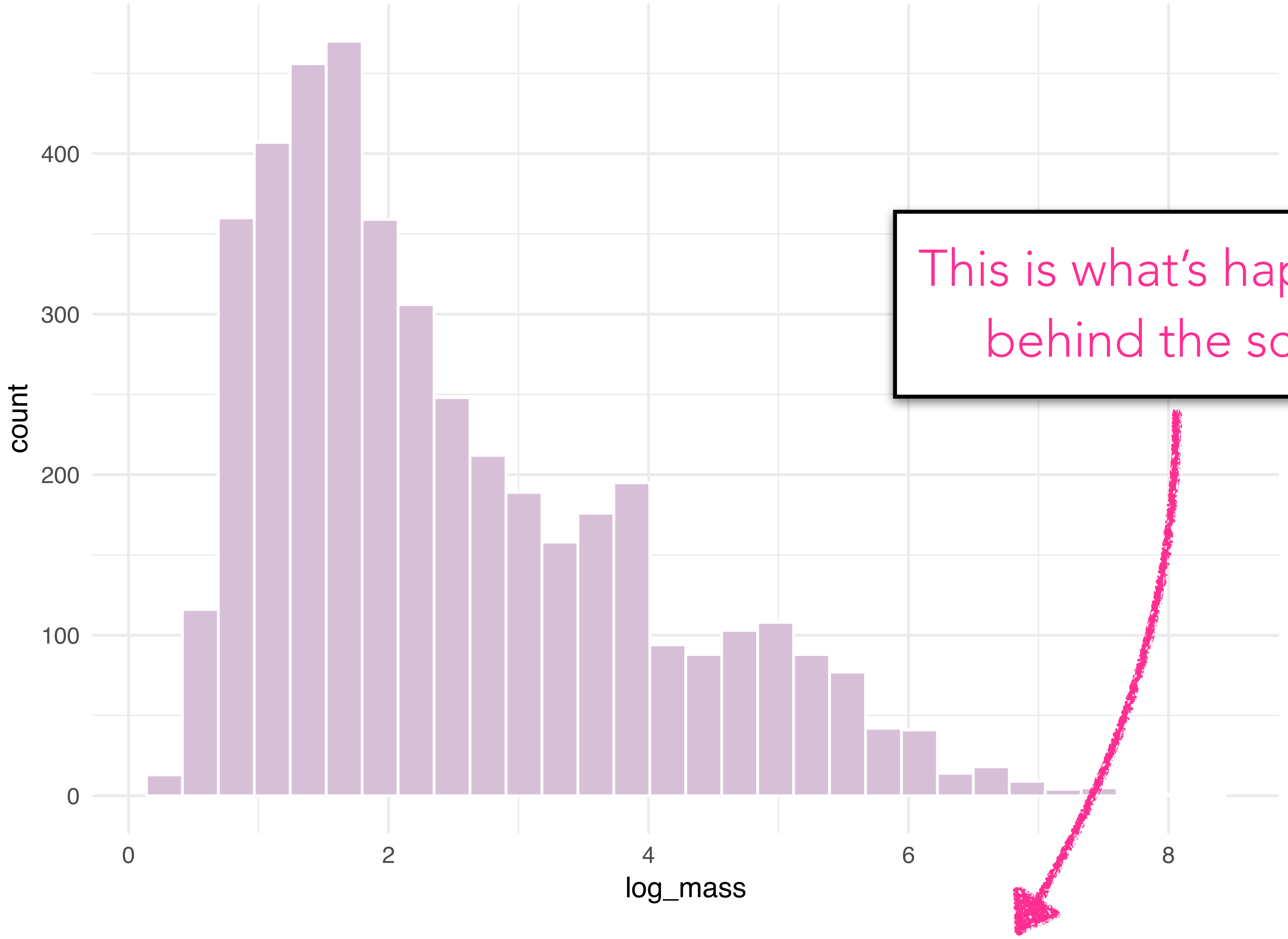
```
ggplot(lqm, aes(x = combined_mass)) +
  geom_histogram()
```

YOUR TURN



```
ggplot(lqm, aes(x = log_mass)) +
  geom_histogram(fill = "thistle", color = "white")
```

YOUR TURN



This is what's happening behind the scenes

```
ggplot(lqm, aes(x = log_mass, y = stat(count))) +
  geom_histogram(fill = "thistle", color = "white")
```

HISTOGRAMS

- *When making a histogram, always explore multiple bin widths.*
- Binning behavior can be modified using:

  # of bins

  width of one bin in units of x

  - bins: `geom_histogram(bins = 30)`
  - binwidth: `geom_histogram(binwidth = 1)`
  - breaks: `geom_histogram(breaks = 0:10)`

  sequence of bin boundaries

HISTOGRAMS

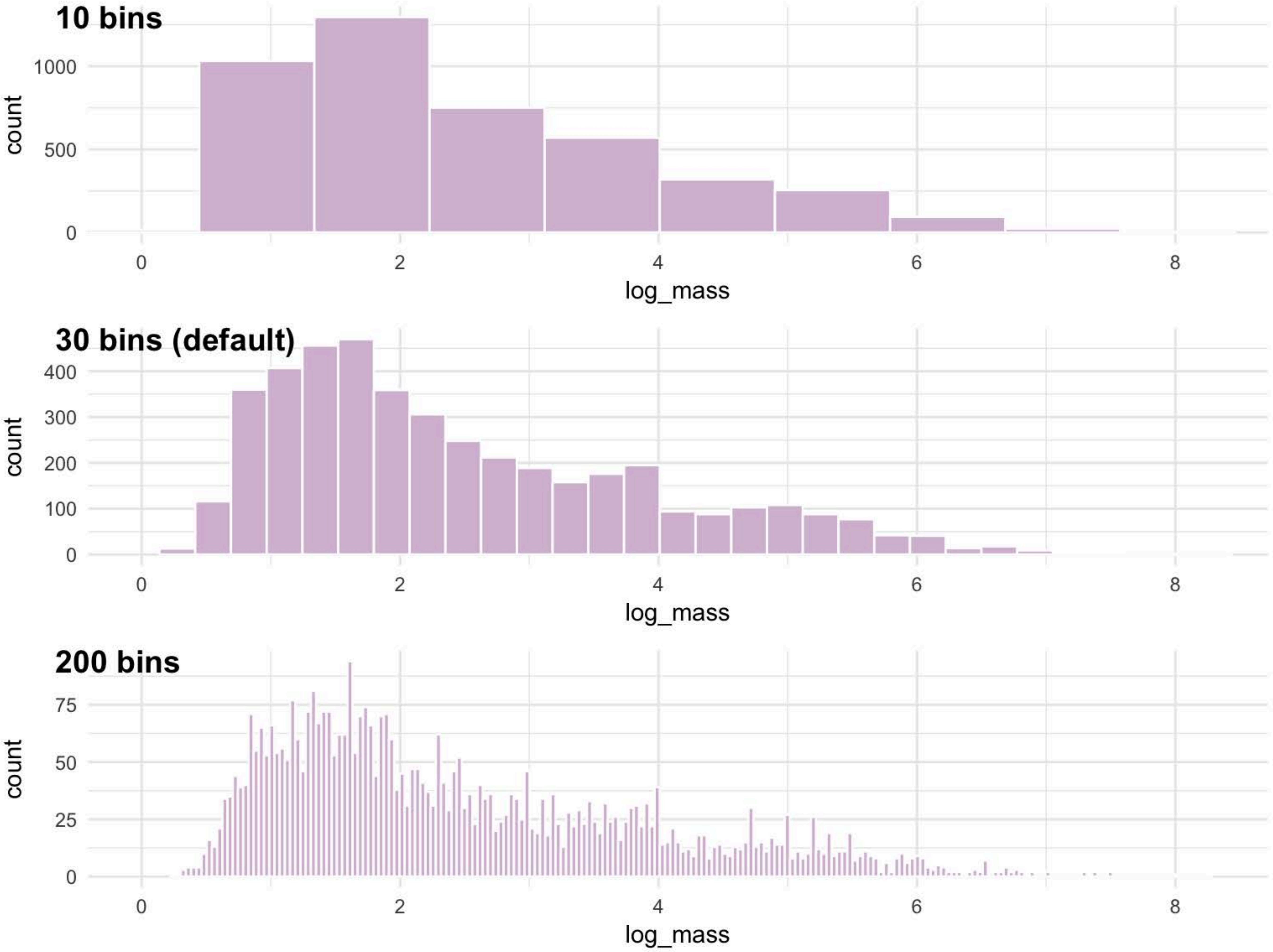
- *When making a histogram, always explore multiple bin widths.*

- Binning behavior can be modified using:

  - bins: `geom_histogram(bins = 30)`

  - binwidth: `geom_histogram(binwidth = 1)`

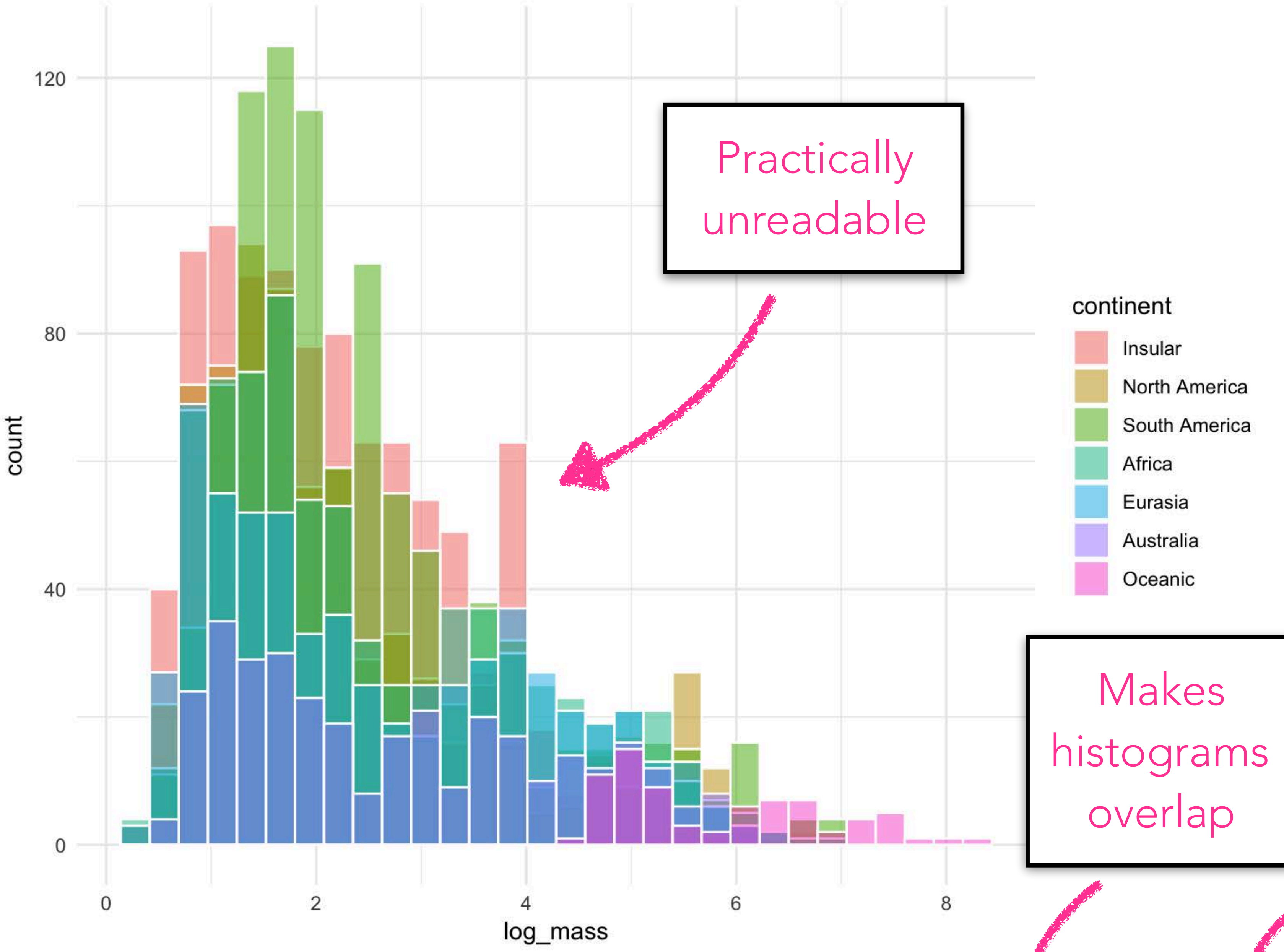
  - breaks: `geom_histogram(breaks = 0:10)`


10 bins
count
1000
500
0
0
2
4
6
8
log_mass
30 bins (default)
count
400
300
200
100
0
0
2
4
6
8
log_mass
200 bins
count
75
50
25
0
0
2
4
6
8
log_mass

HISTOGRAMS

- Histograms are generally bad at showing multiple distributions together
- For example…

YOUR TURN

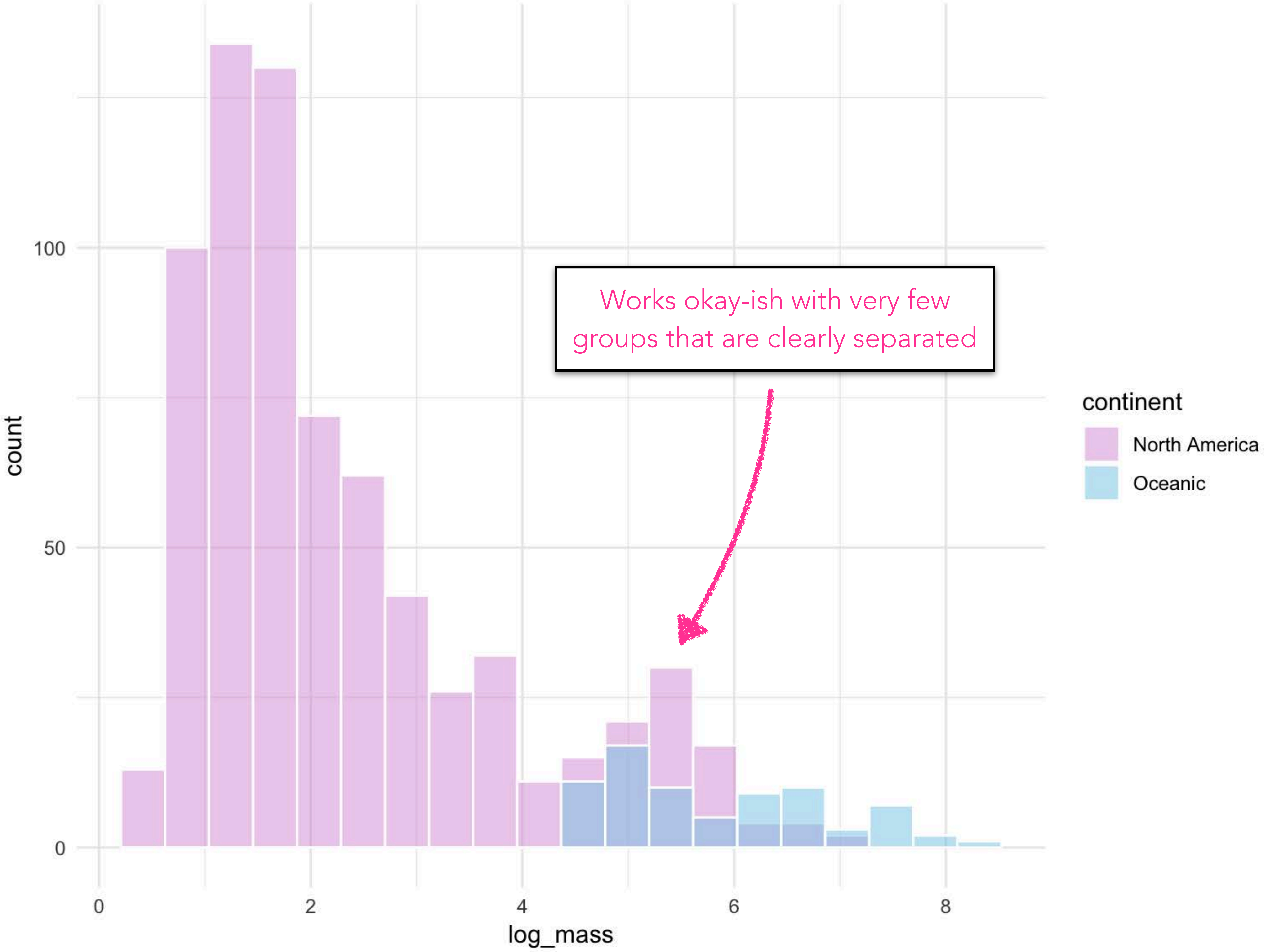


Practically unreadable

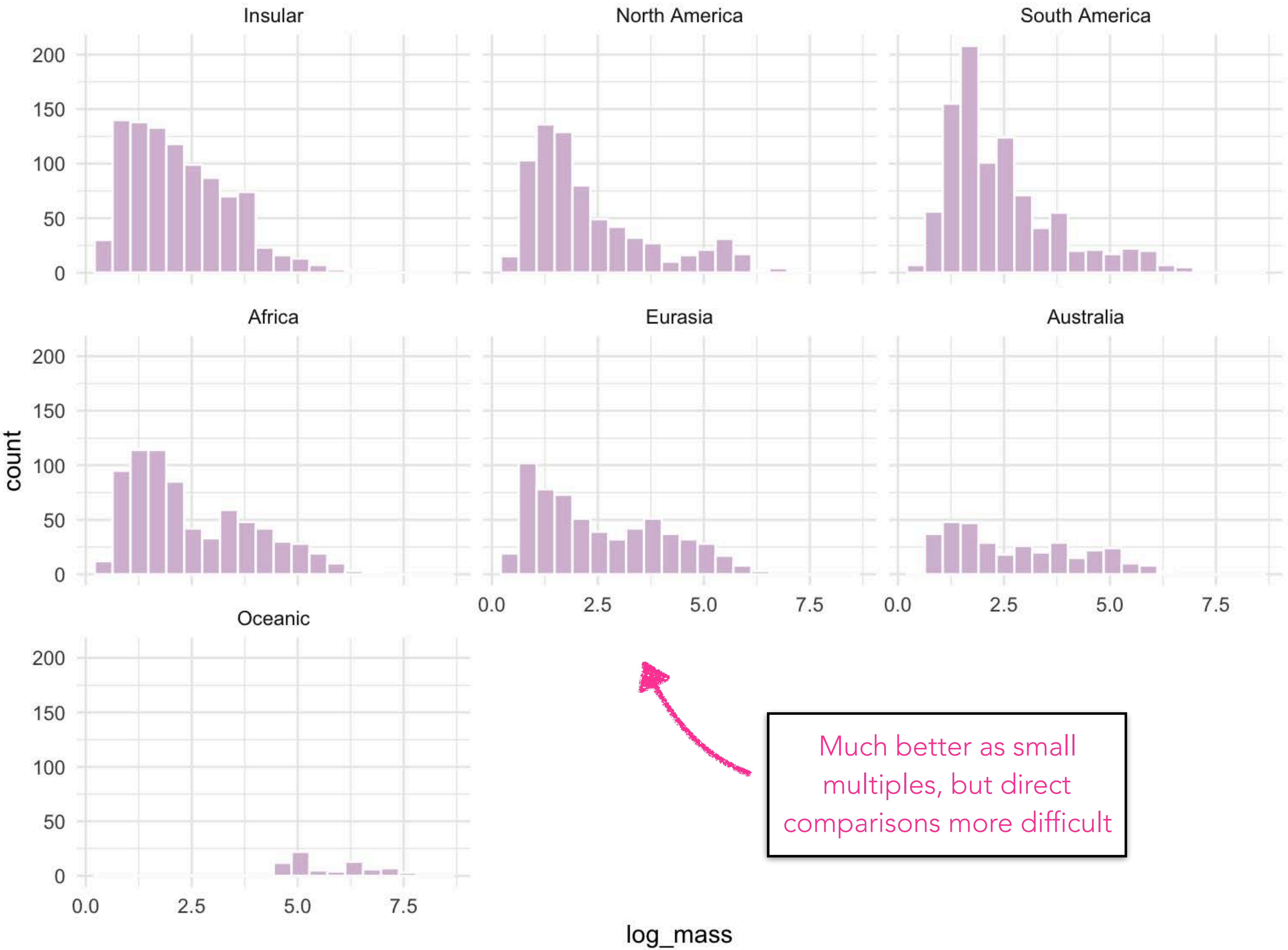
Makes histograms overlap

Makes them partially transparent

```
ggplot(lqm, aes(x = log_mass, fill = continent)) +
  geom_histogram(color = "white", position = "identity", alpha = 0.5)
```

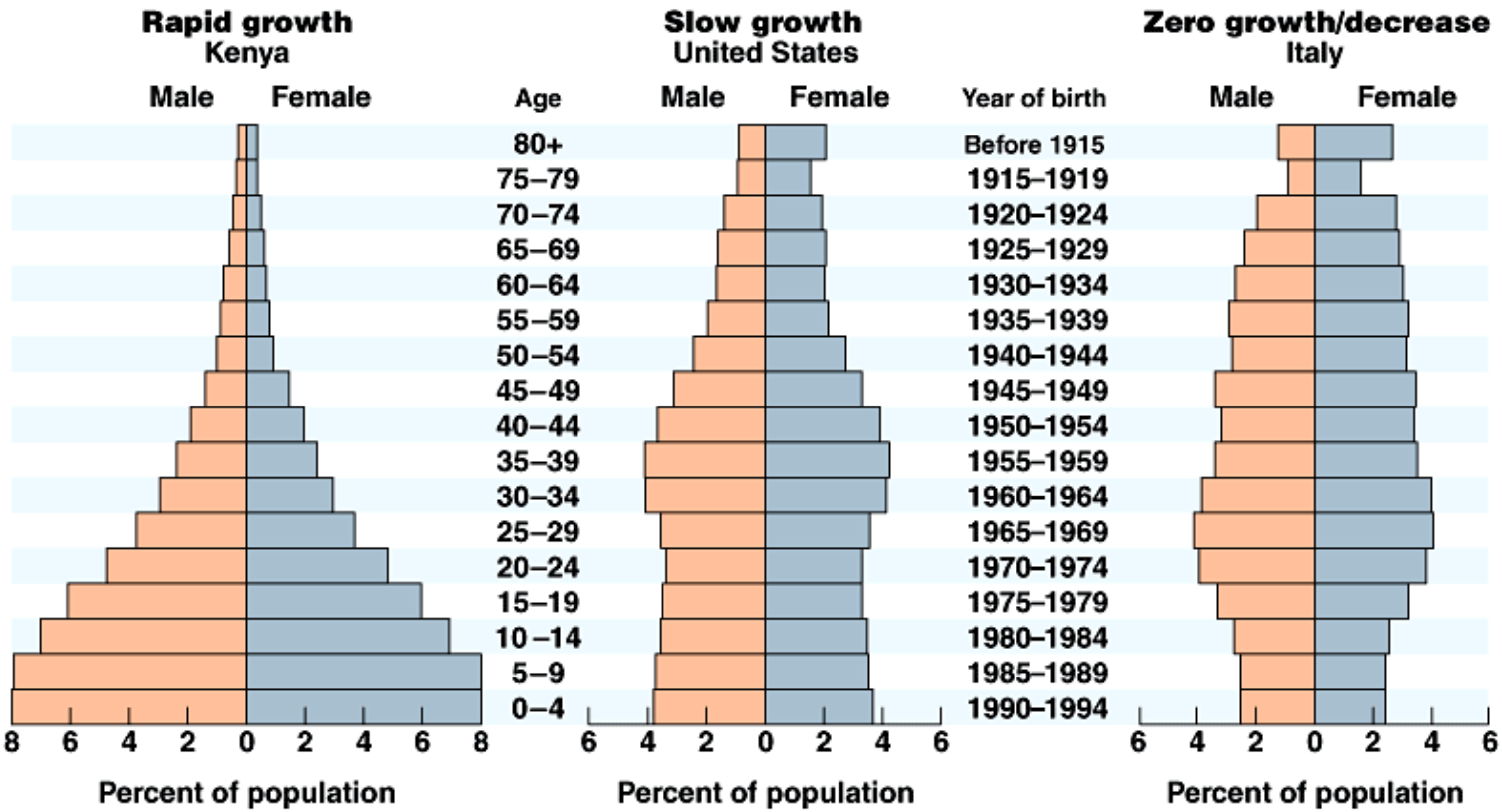


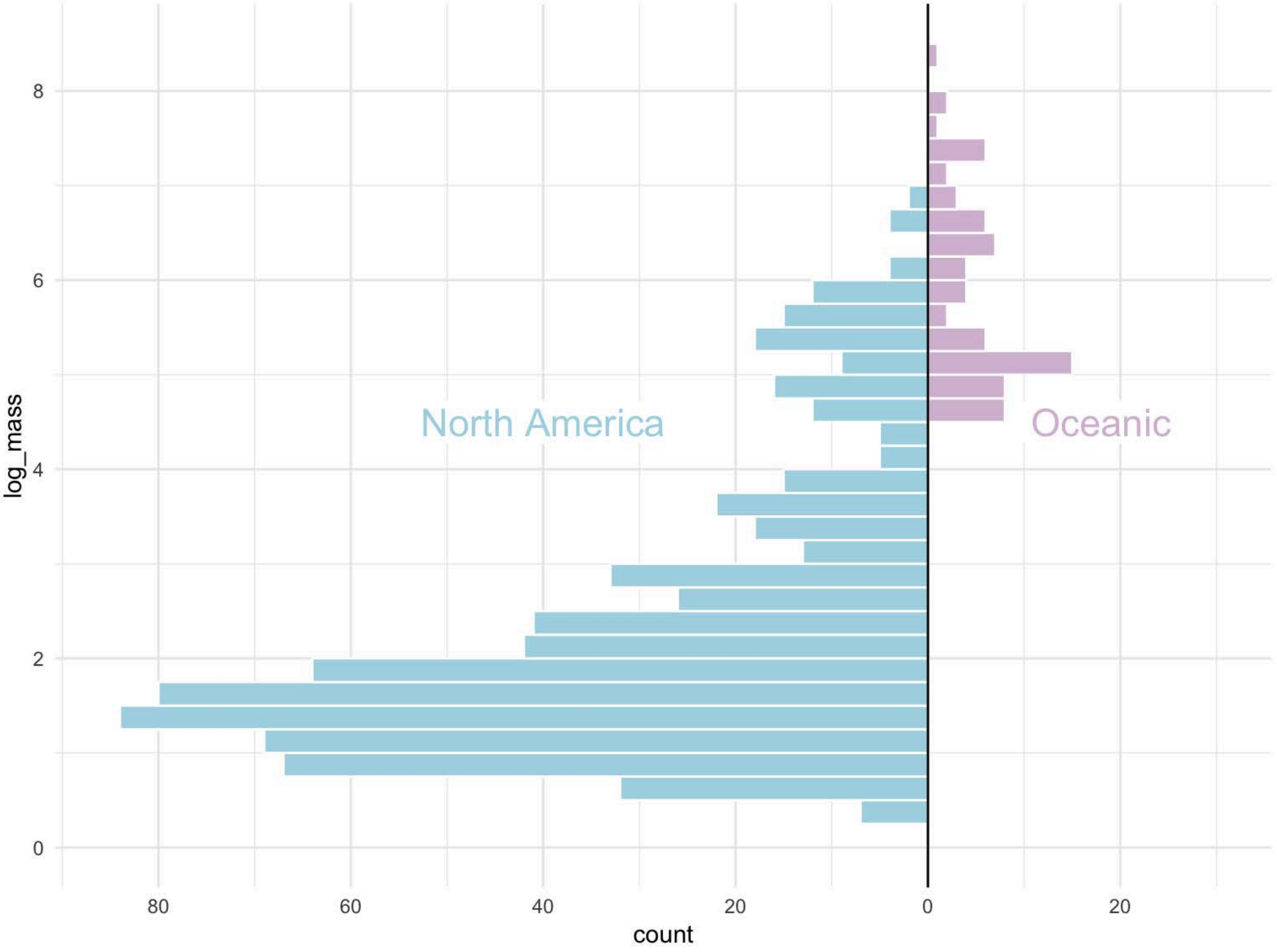
Works okay-ish with very few groups that are clearly separated

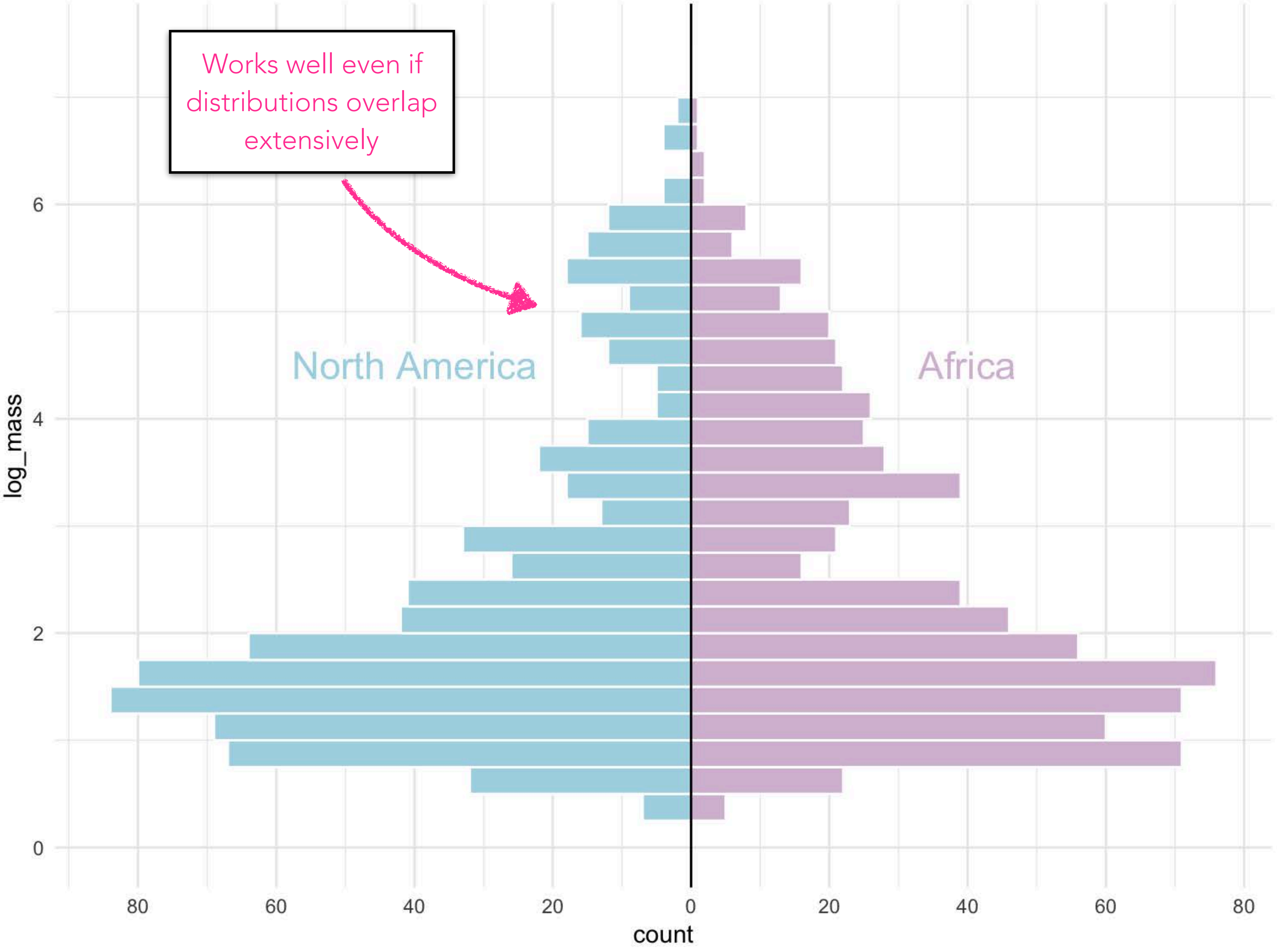


Much better as small multiples, but direct comparisons more difficult

“Age pyramid” style plot is ideal for comparing two (and only two) distributions

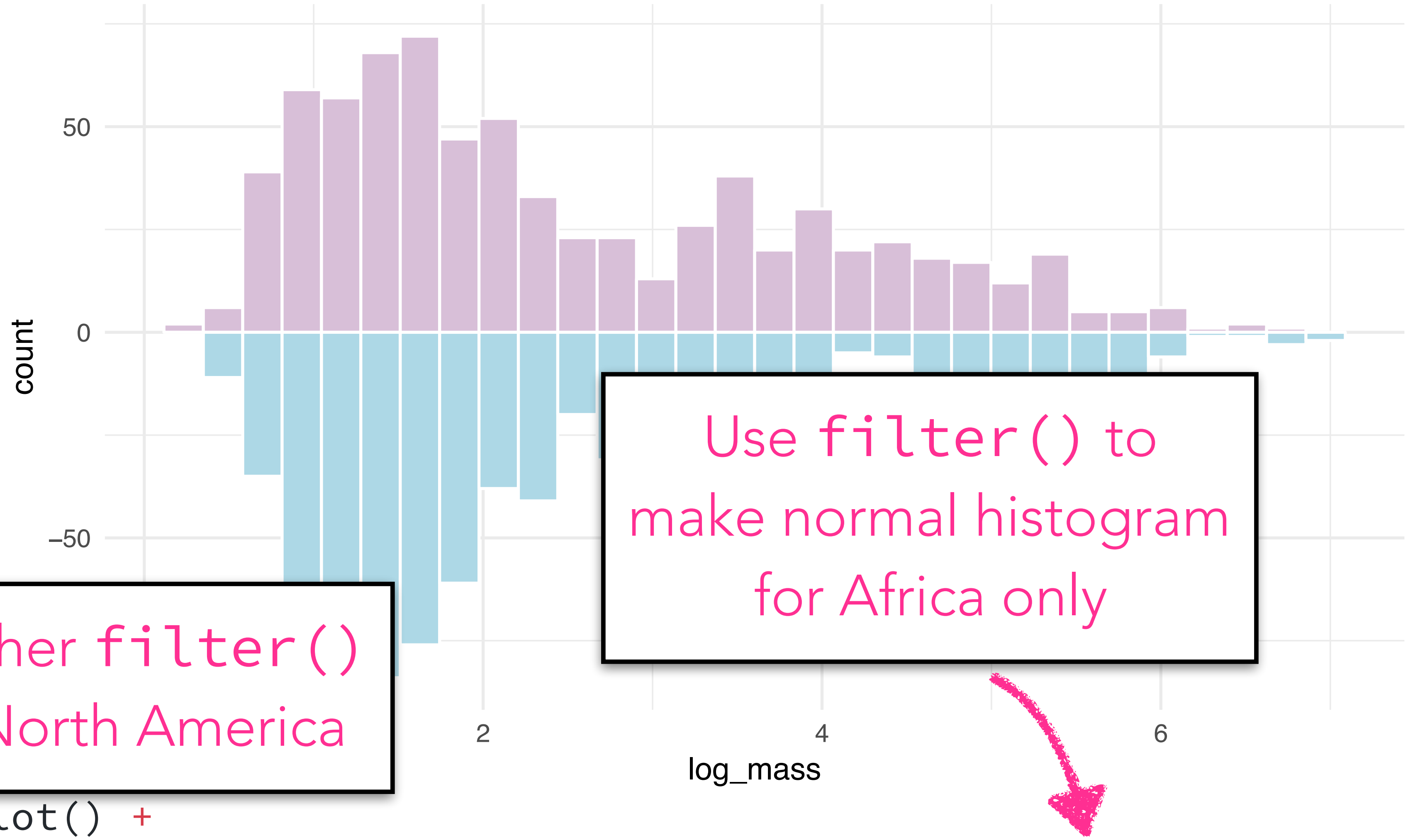



North America
Oceanic
log_mass
8
6
4
2
0
80
60
40
20
0
20
count



Works well even if distributions overlap extensively

A STARTER…



Use `filter()` to make normal histogram for Africa only

Another `filter()` for North America

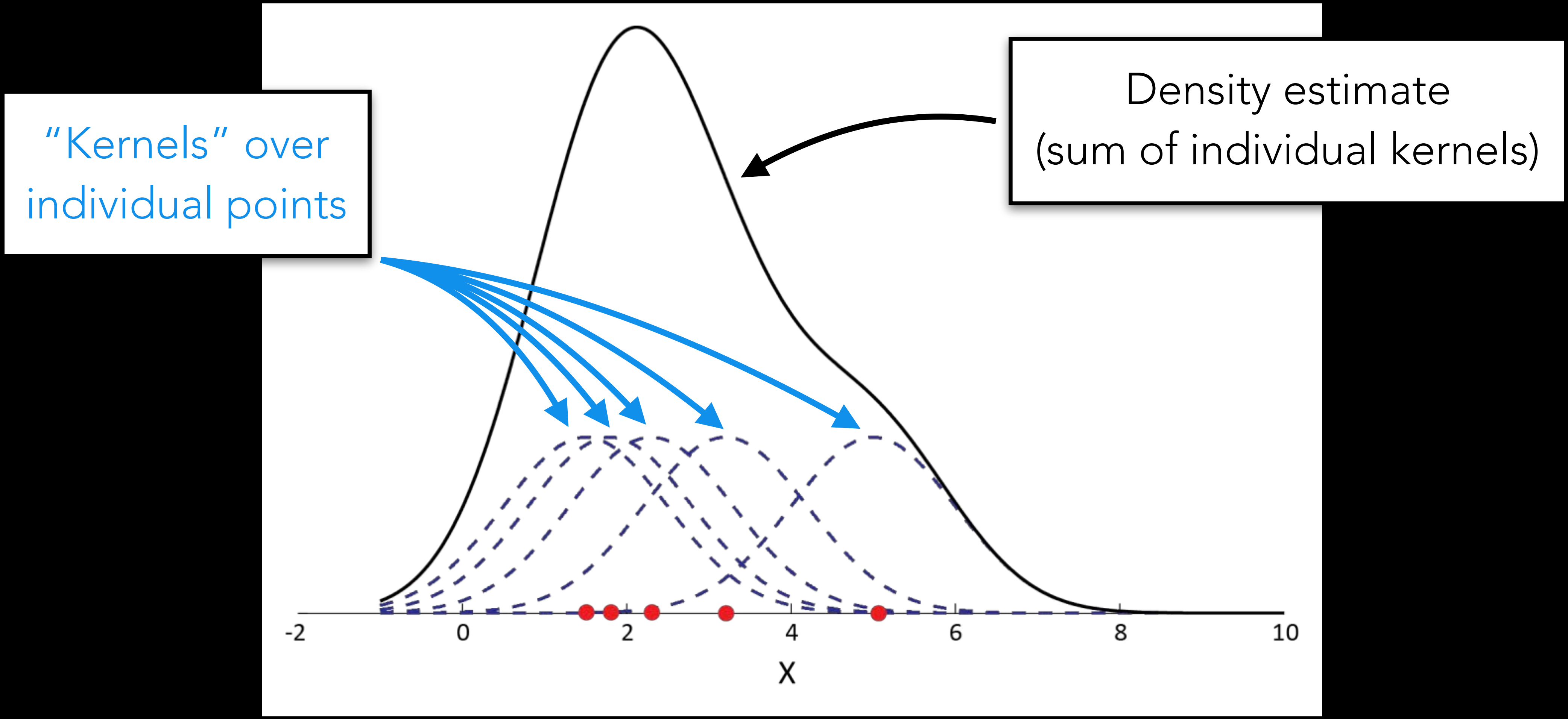
Use the `stat()` trick to get counts with flipped sign

```
ggplot() +
  geom_histogram(data = filter(lqm, continent == "Africa"),
                 aes(x = log_mass),
                 fill = "thistle", color = "white") +
  geom_histogram(data = filter(lqm, continent == "North America"),
                 aes(x = log_mass, y = -stat(count)),
                 fill = "lightblue", color = "white")
```

DENSITY PLOTS

DENSITY PLOTS

- Like histograms, density plots show the distribution of a single variable
- Uses smoothing to estimate the *density* of points over a *continuous* range of data values.



"Kernels" over individual points

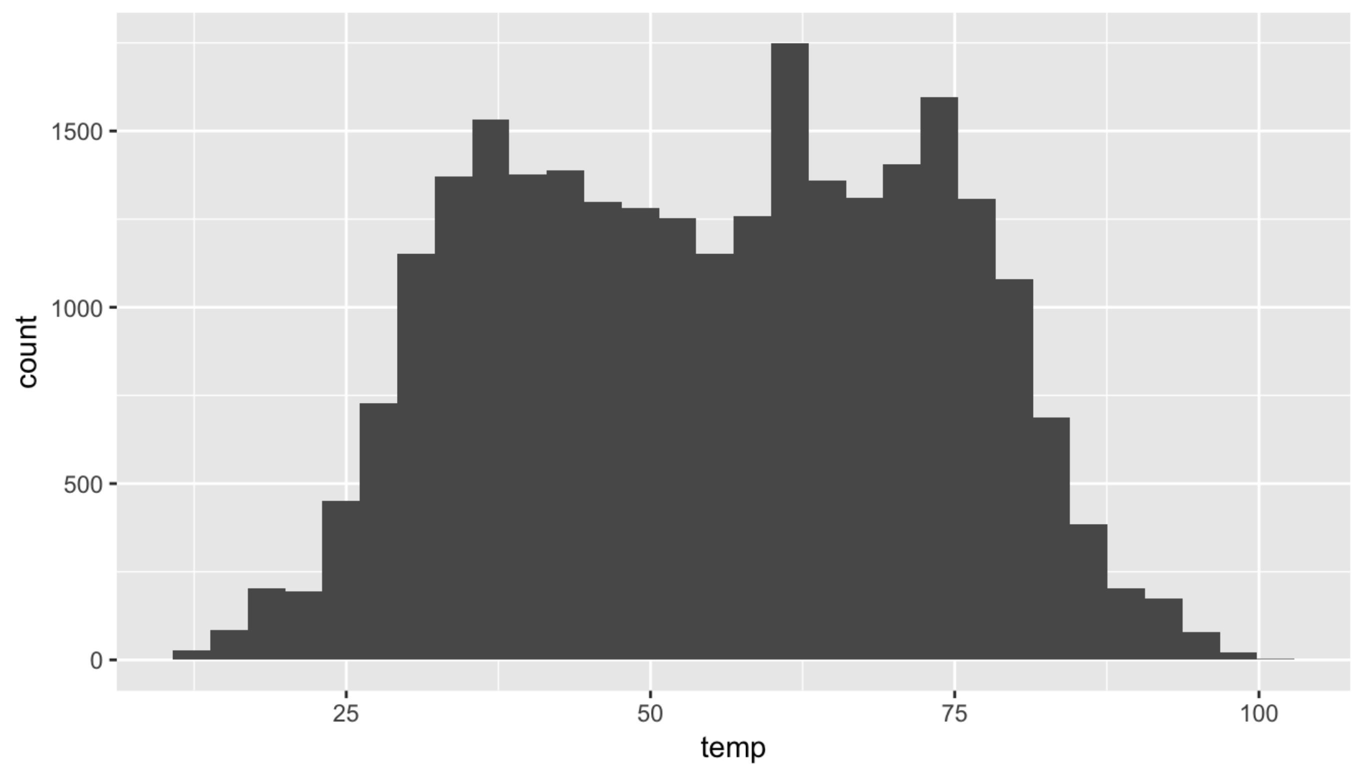
Density estimate (sum of individual kernels)

DENSITY PLOTS

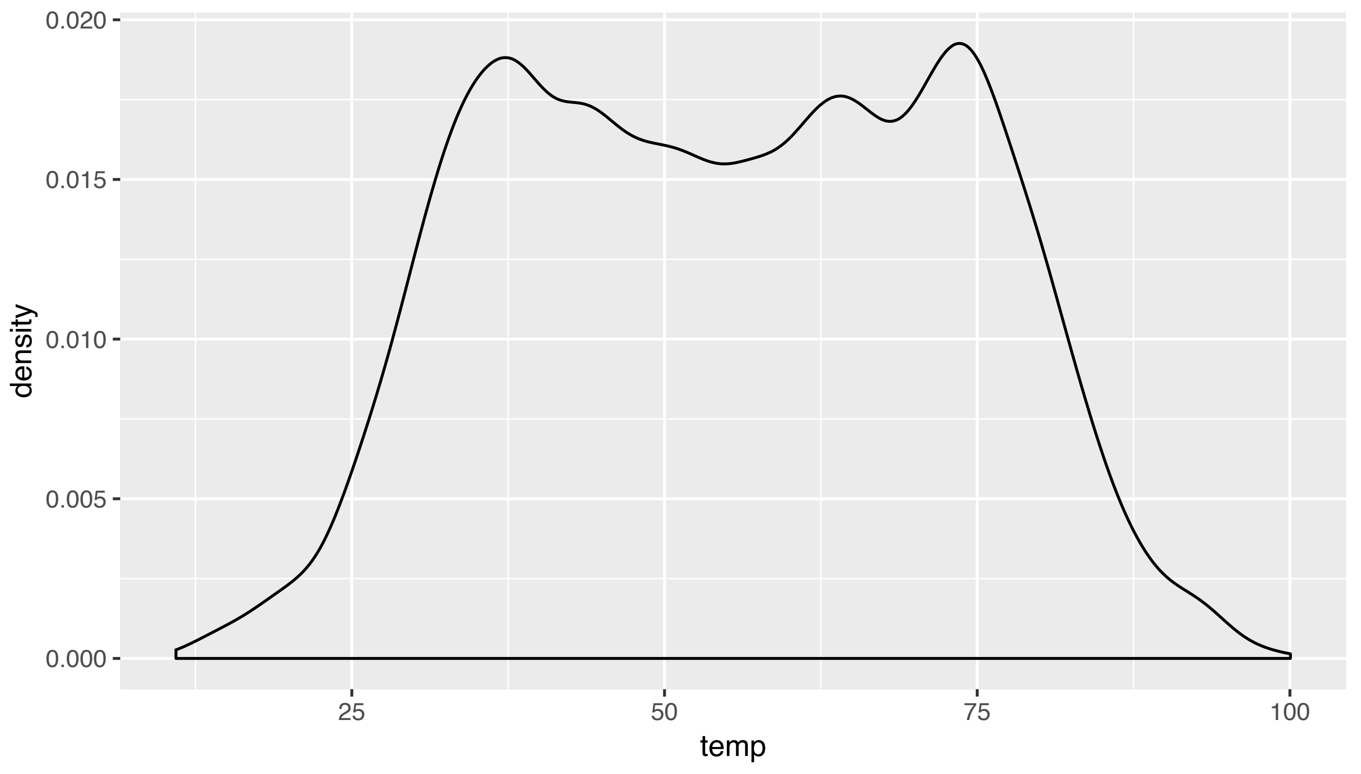
- Geometric object is `geom_density()`

- Aesthetic mappings are same as for histogram

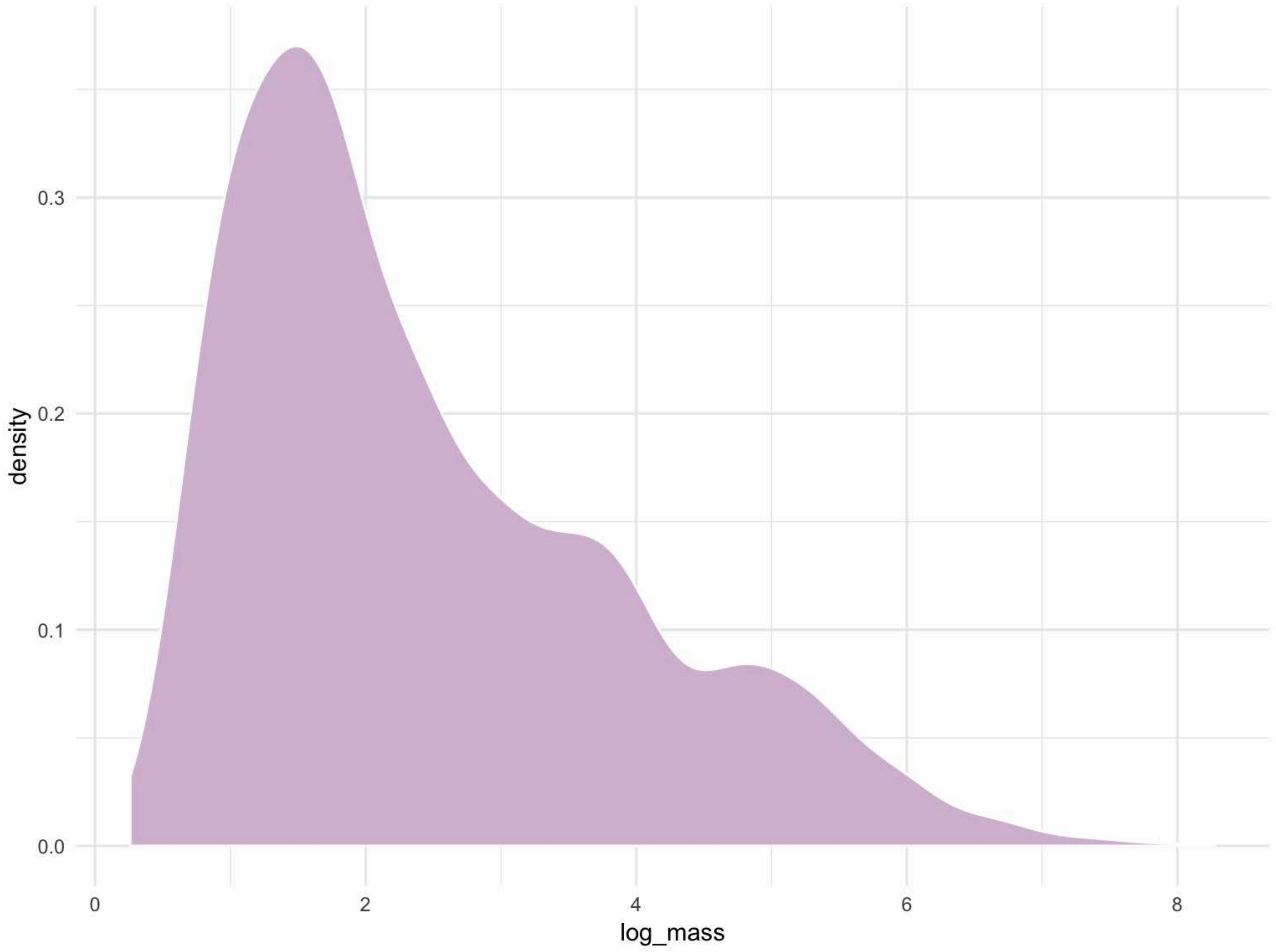
Histogram



Density plot



YOUR TURN



```
ggplot(lqm, aes(x = log_mass)) +
  geom_density(fill = "thistle", color = "white")
```

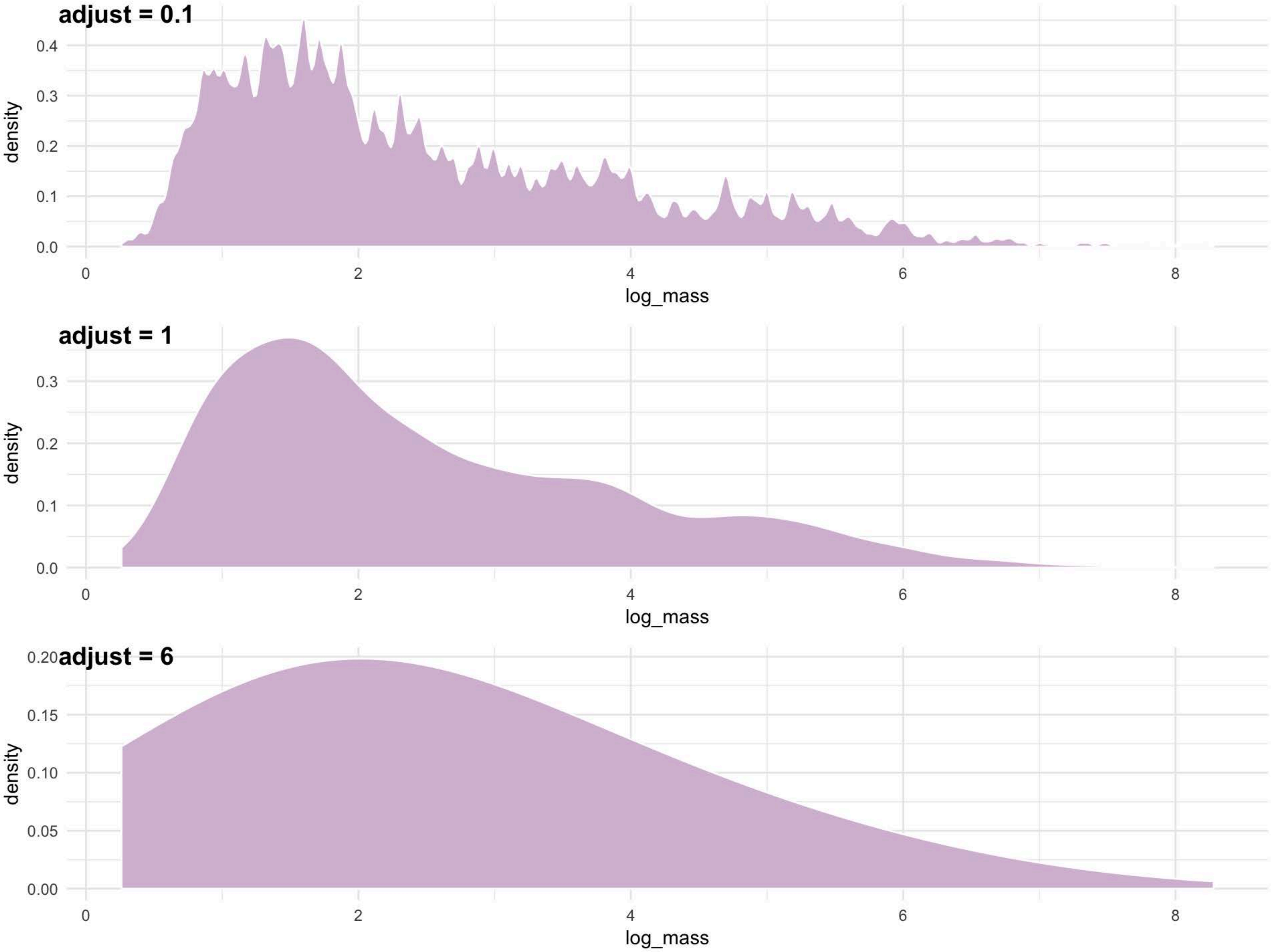
DENSITY PLOTS

- Like histograms, the visual appearance of density plots depends on the amount of smoothing.

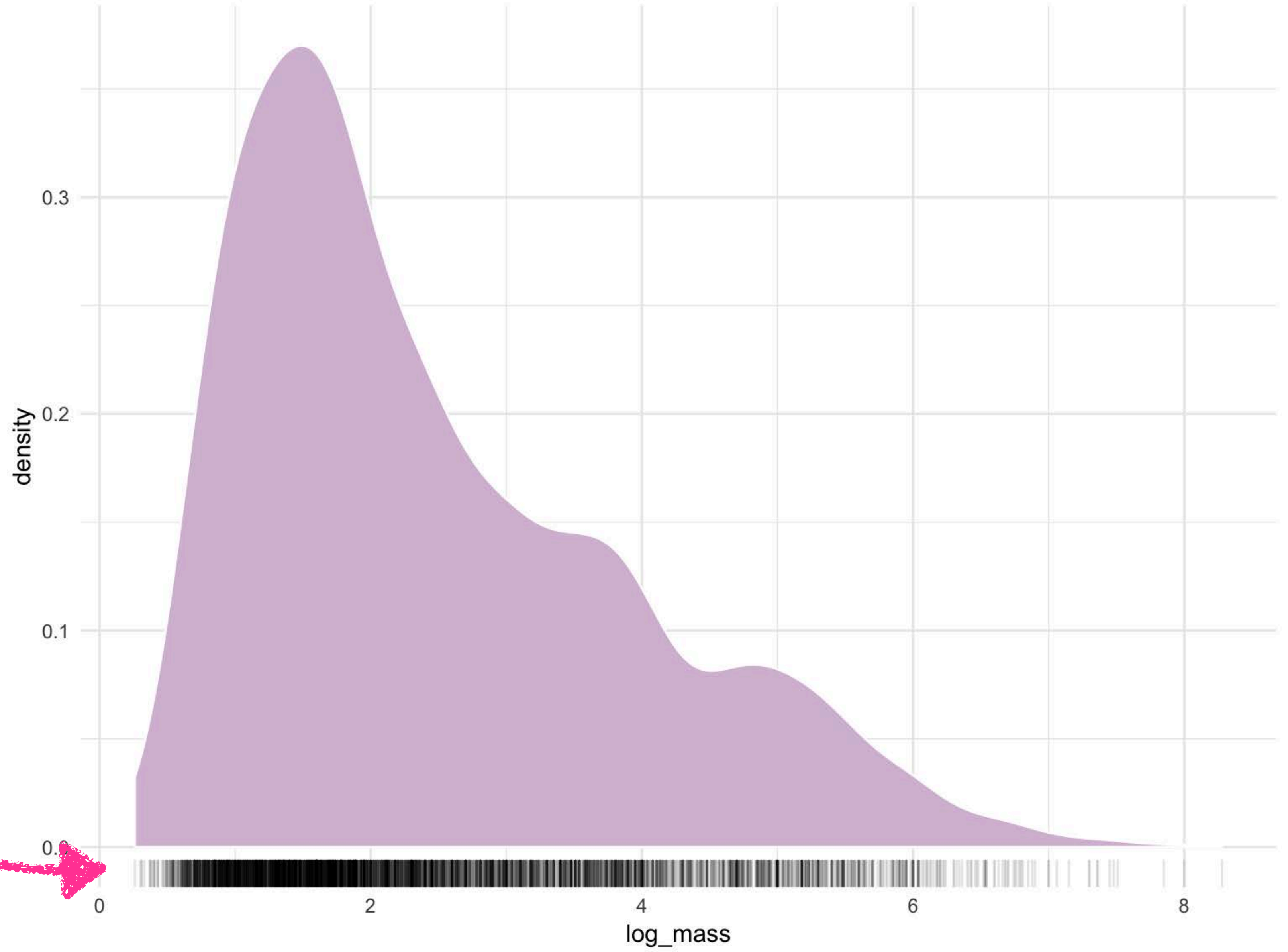
- Smoothing behavior can be adjusted in various complex ways, but the only one you're likely to use is:

  - adjust: `geom_density(adjust = 1)`

Multiplier for the amount of smoothing (1 = default)


adjust = 0.1
0.4
0.3
0.2
0.1
0.0
density
0
2
4
6
8
log_mass
adjust = 1
0.3
0.2
0.1
0.0
density
0
2
4
6
8
log_mass
adjust = 6
0.20
0.15
0.10
0.05
0.00
density
0
2
4
6
8
log_mass

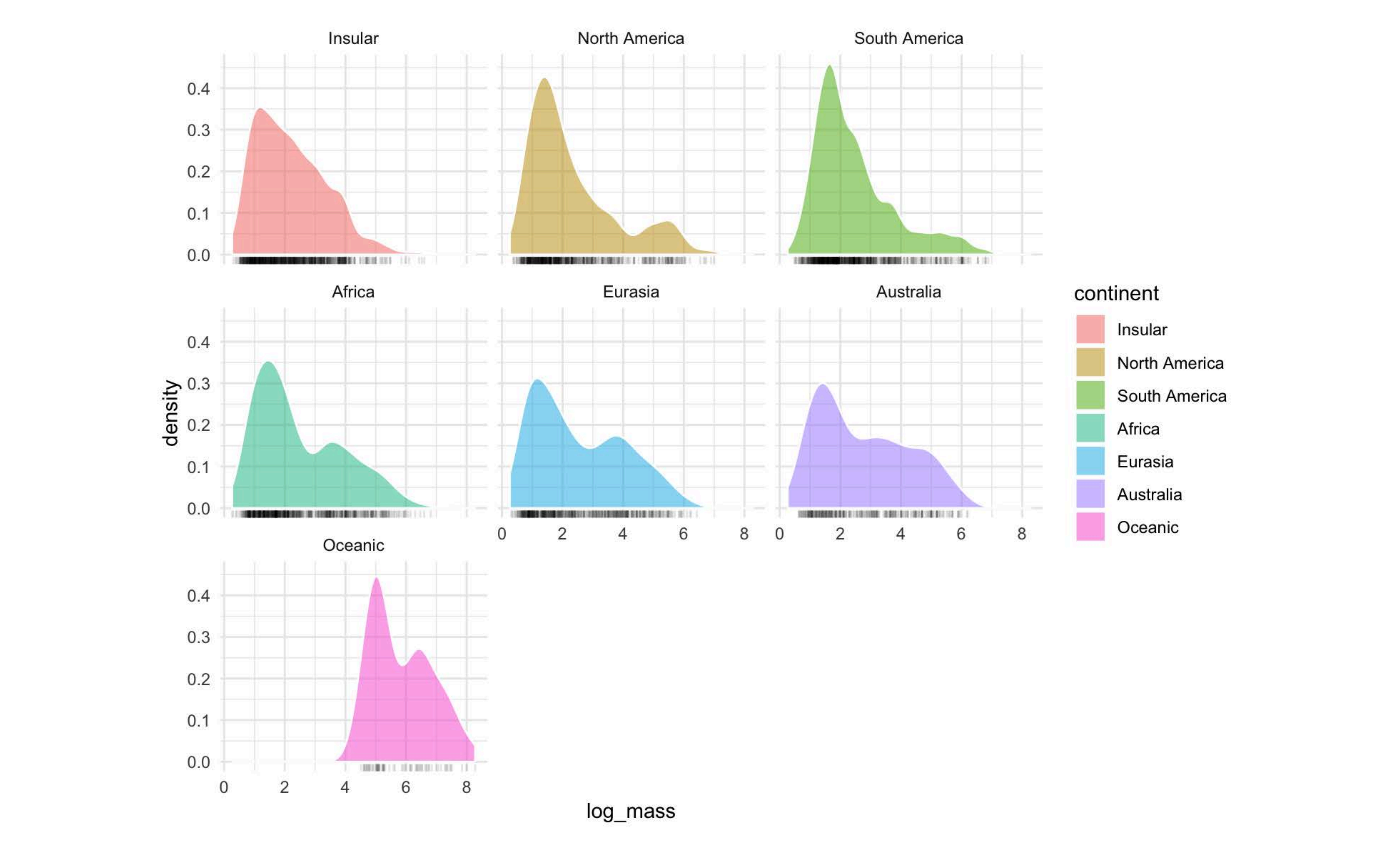
DENSITY PLOTS



Rug plot shows data points as small marks along axis

```
ggplot(lqm, aes(x = log_mass)) +
  geom_density(fill = "thistle", color = "white") +
  geom_rug(alpha = 0.1)
```

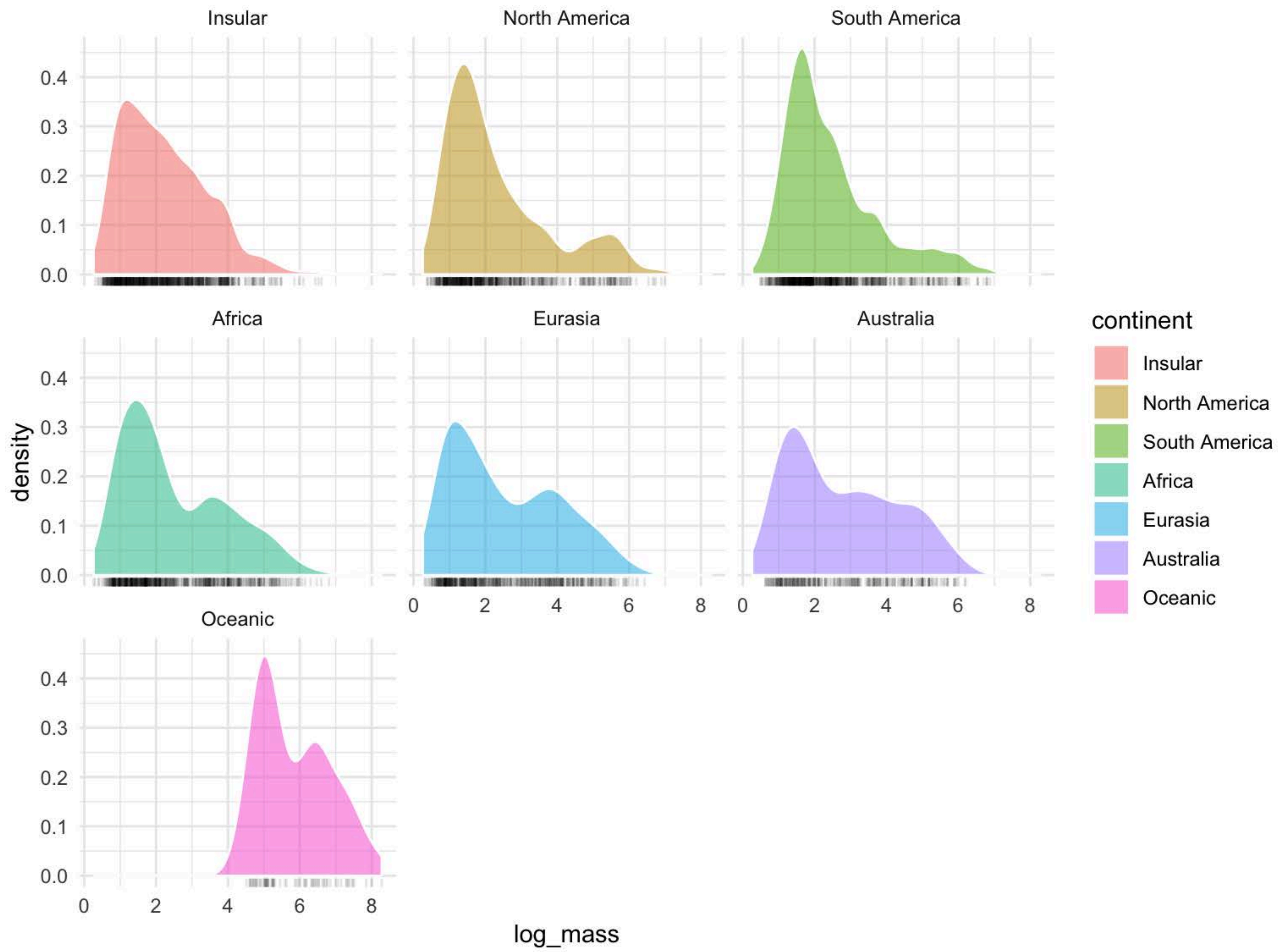
YOUR TURN



Try to reverse-engineer this plot.

5:00

YOUR TURN



```
ggplot(lqm, aes(x = log_mass, fill = continent)) +
  geom_density(color = "white", alpha = 0.5) +
  geom_rug(alpha = 0.1) +
  facet_wrap(~continent)
```

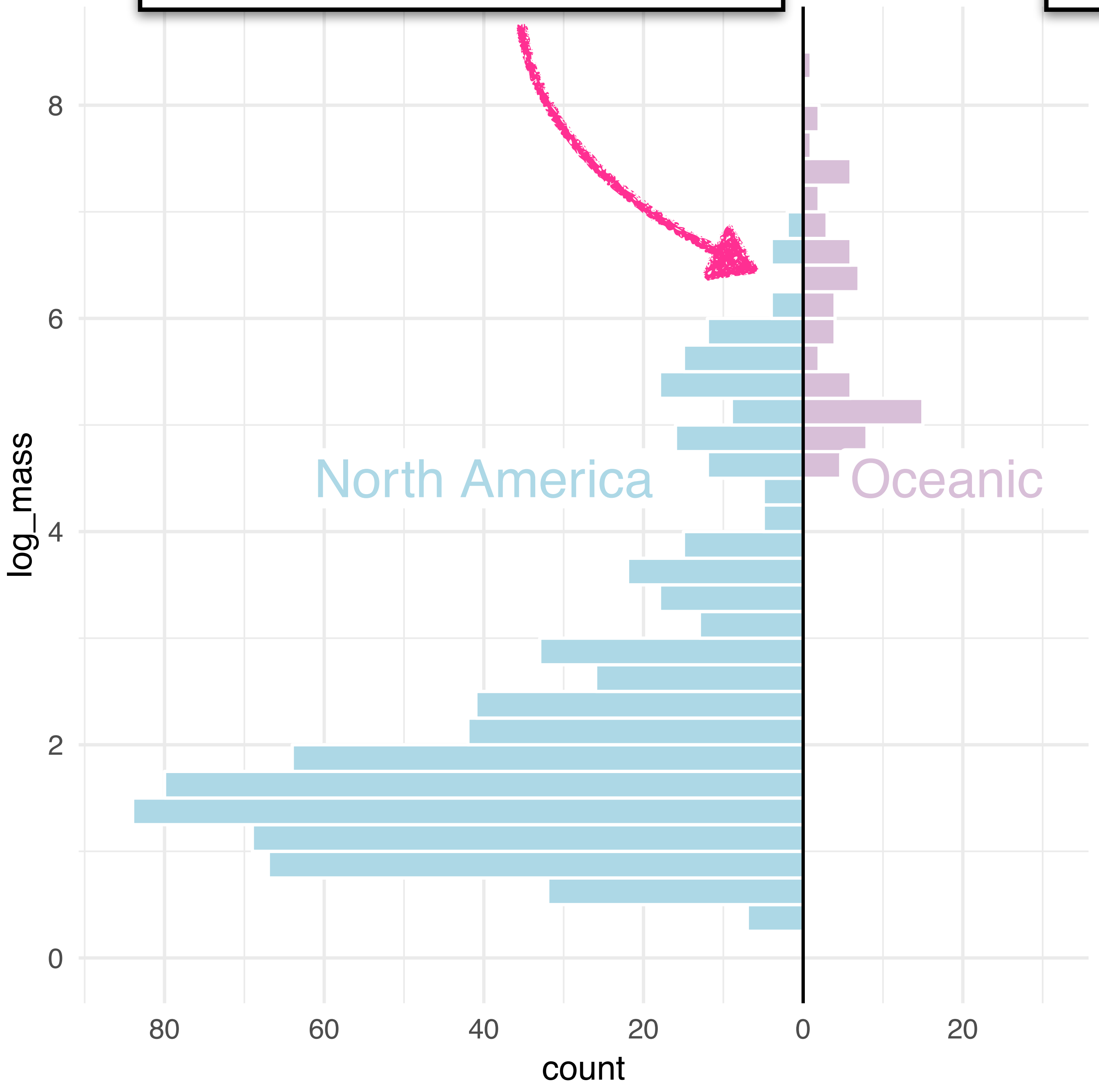
HISTOGRAMS VS. DENSITY PLOTS

- Histograms are much more common, and are therefore more familiar to a wide audience

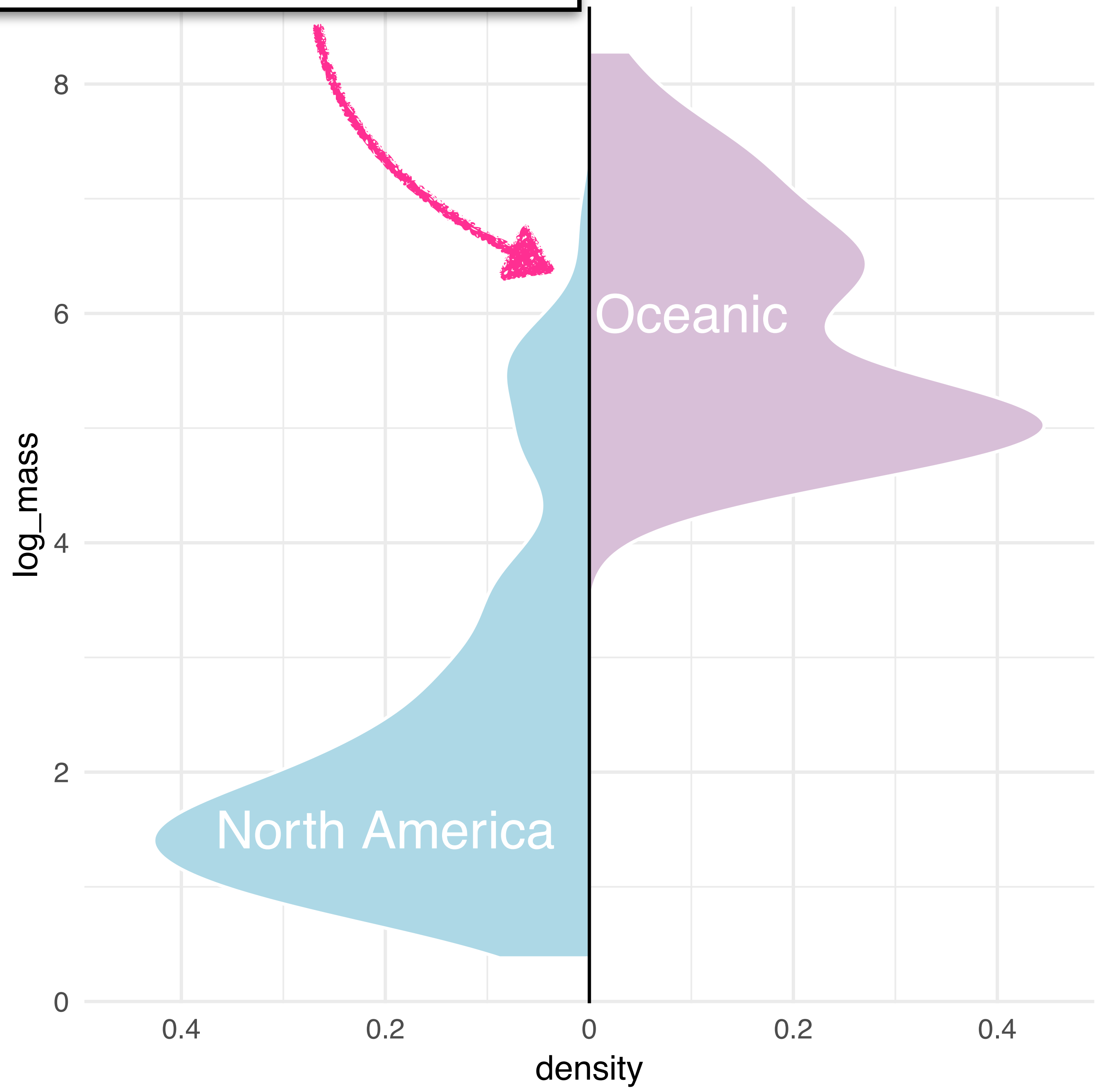
- Histograms use a more intuitive stat (count vs. density)

- Whereas the appearance of a histograms depends on sample size, densities are *scaled*

Smaller sample size for "Oceanic" means bars are small

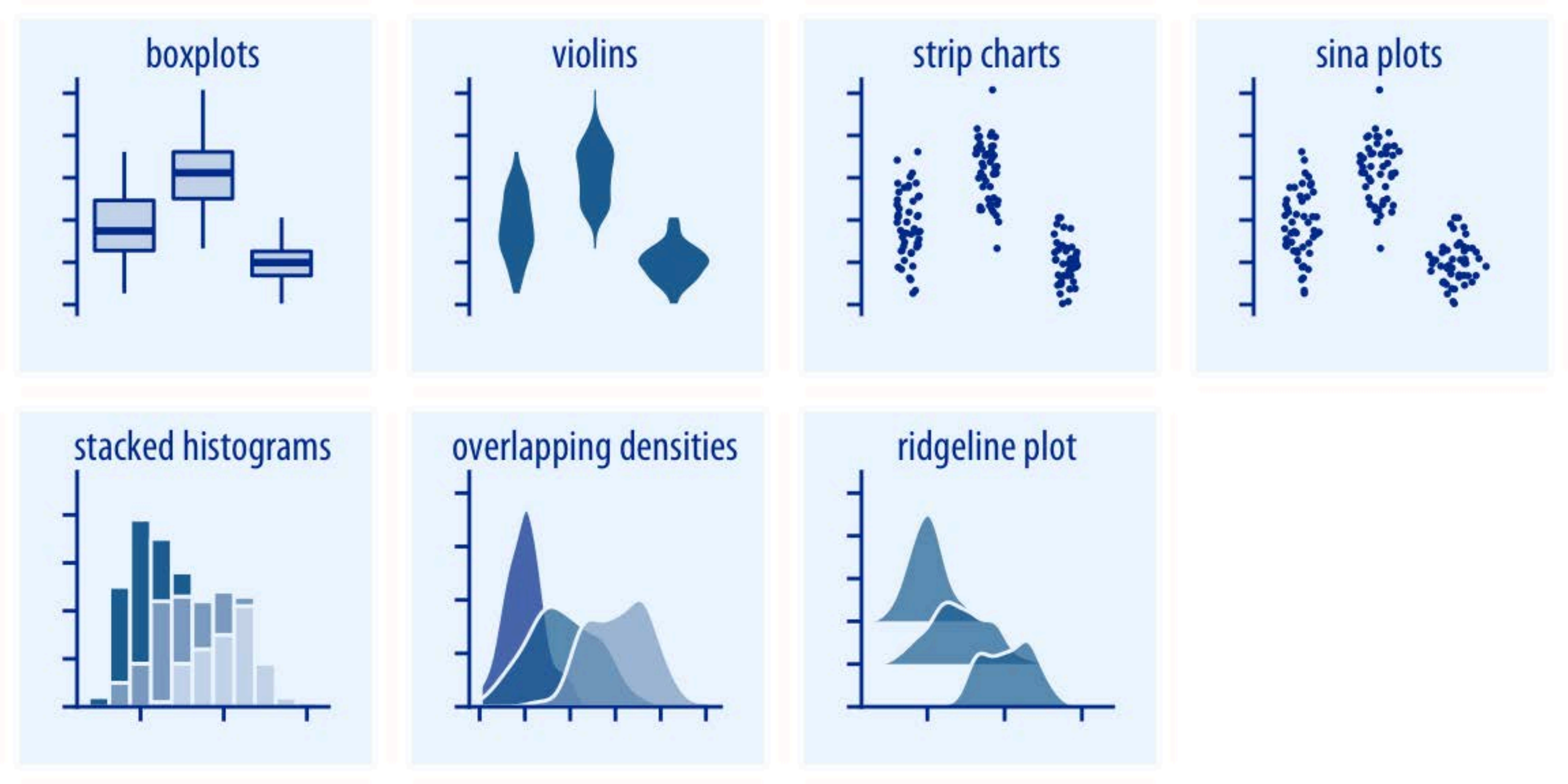


Densities have equal areas (area under the curve sums to 1)

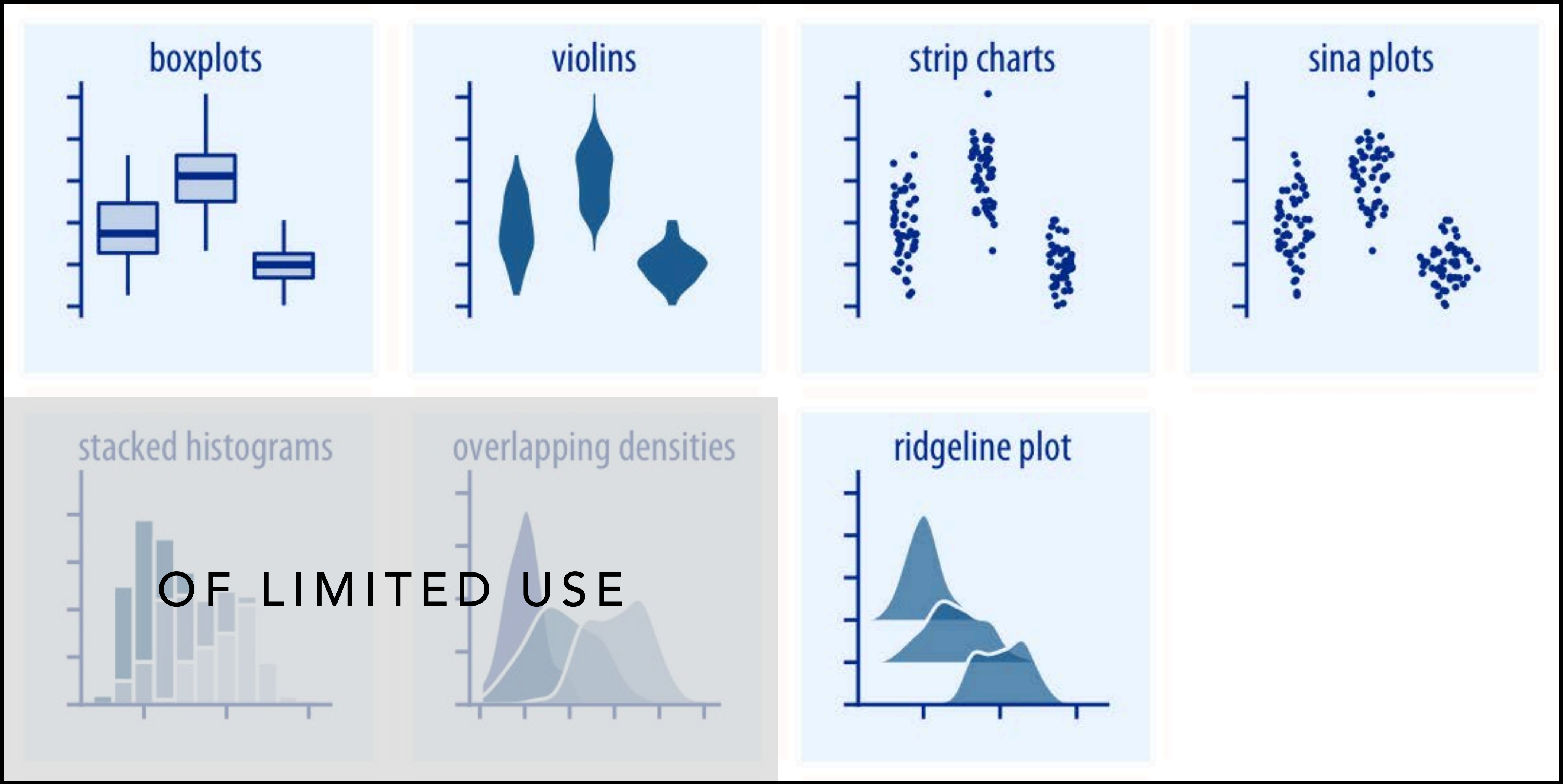


WHAT IF YOU NEED TO SHOW MULTIPLE DISTRIBUTIONS TOGETHER?

WHAT IF YOU NEED TO SHOW MULTIPLE DISTRIBUTIONS TOGETHER?

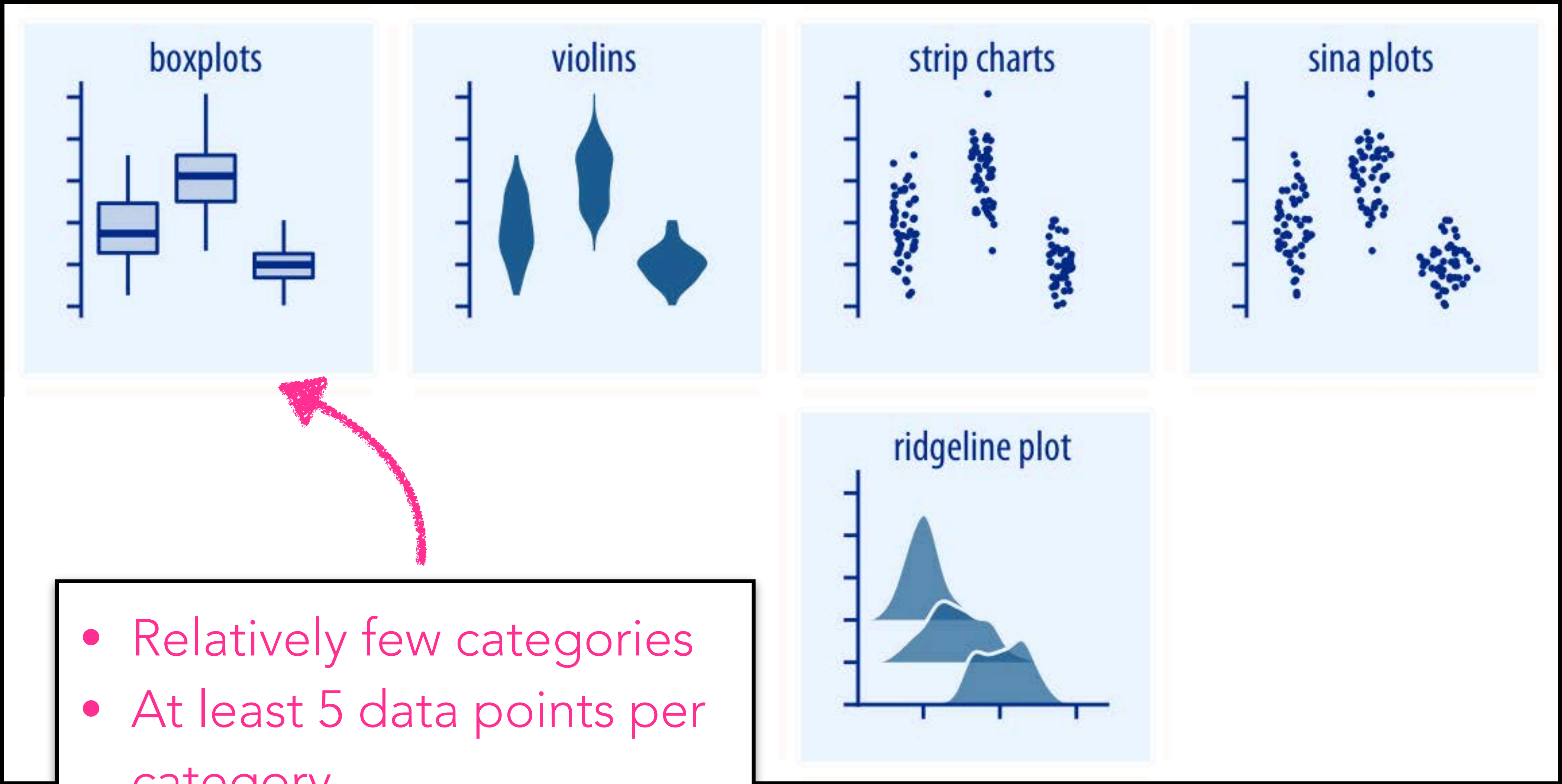


WHAT IF YOU NEED TO SHOW MULTIPLE DISTRIBUTIONS TOGETHER?



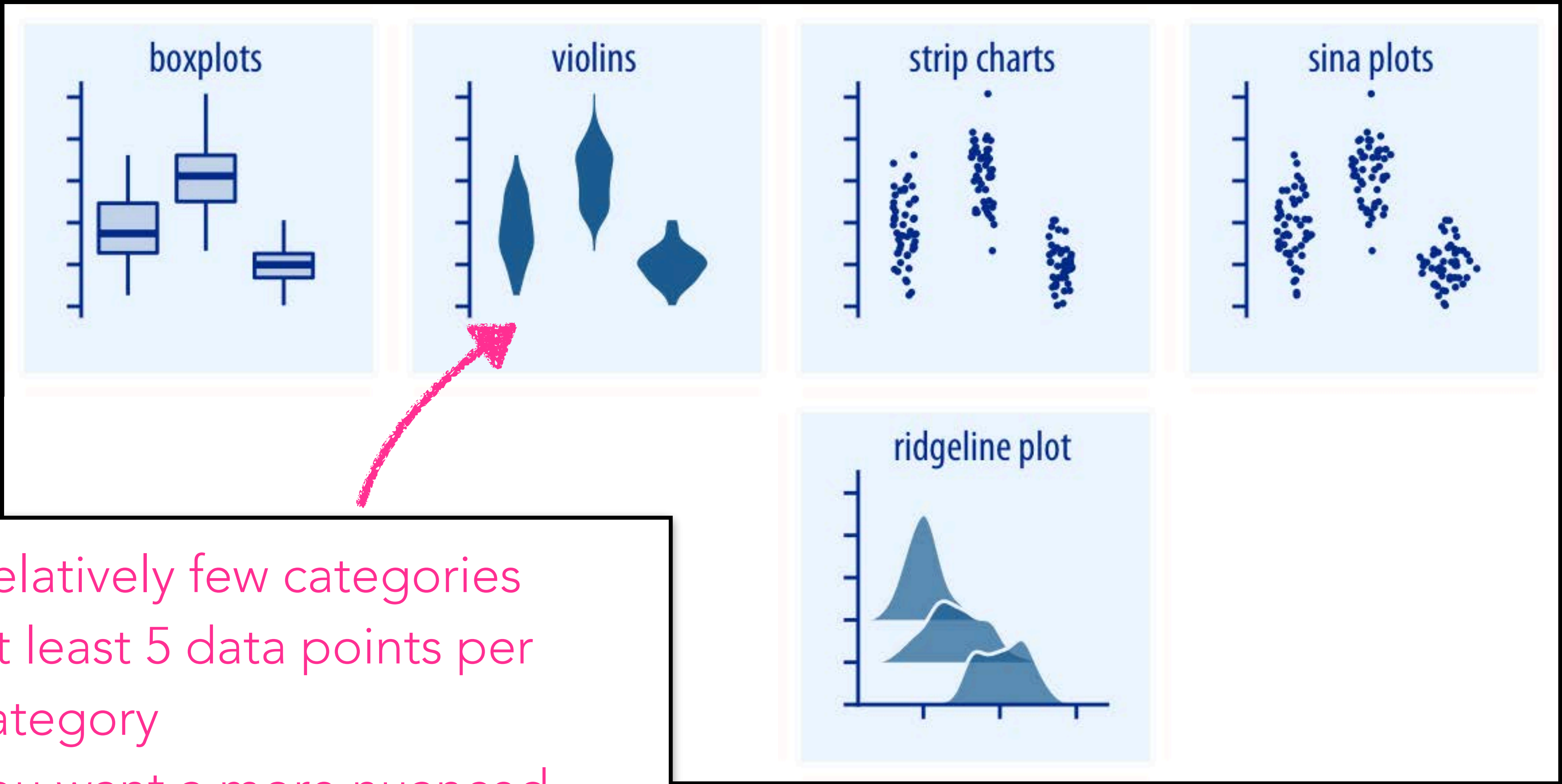
OF LIMITED USE

USAGE GUIDELINES



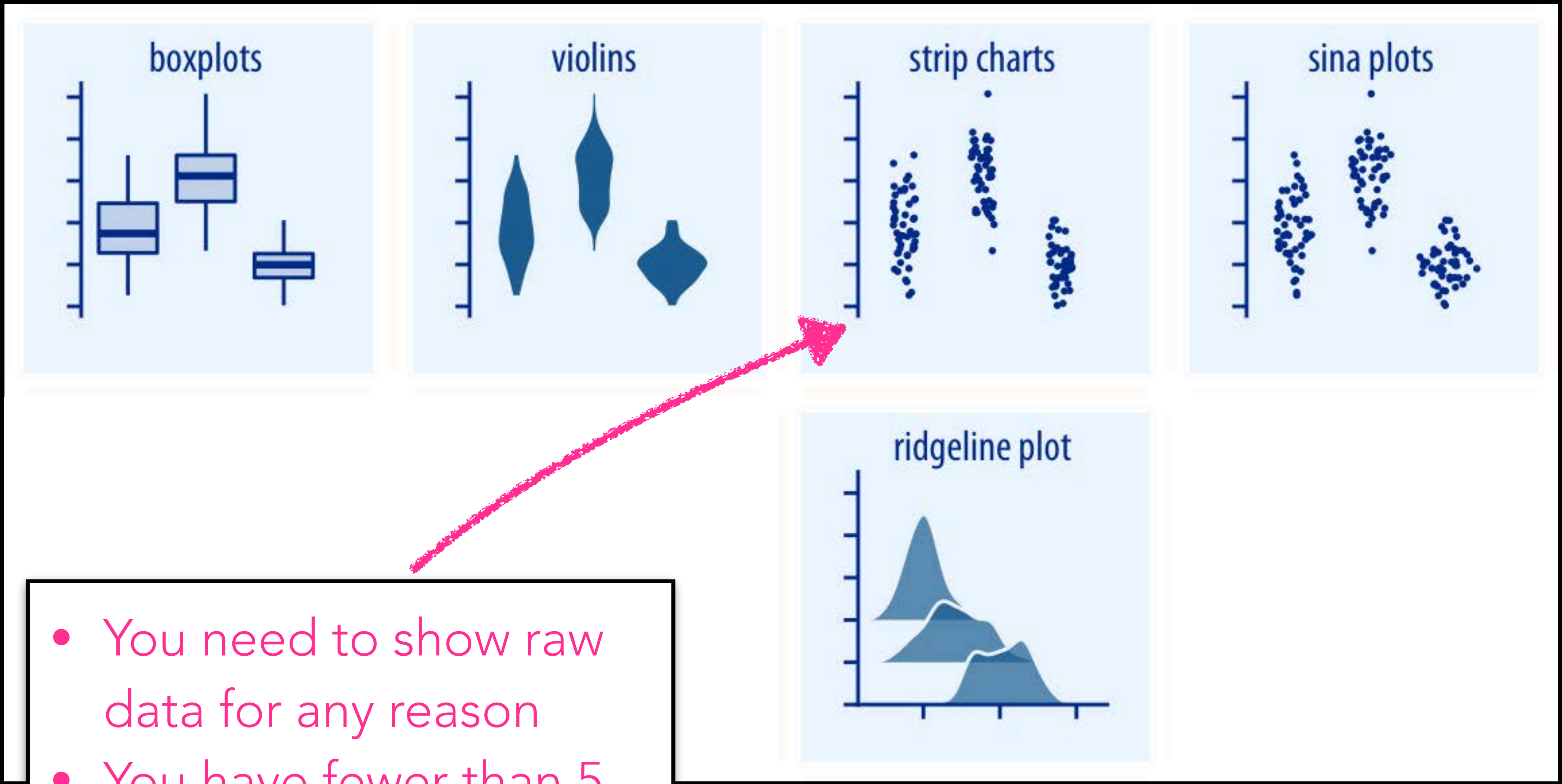
- Relatively few categories
- At least 5 data points per category
- Statistical comparisons are important

USAGE GUIDELINES



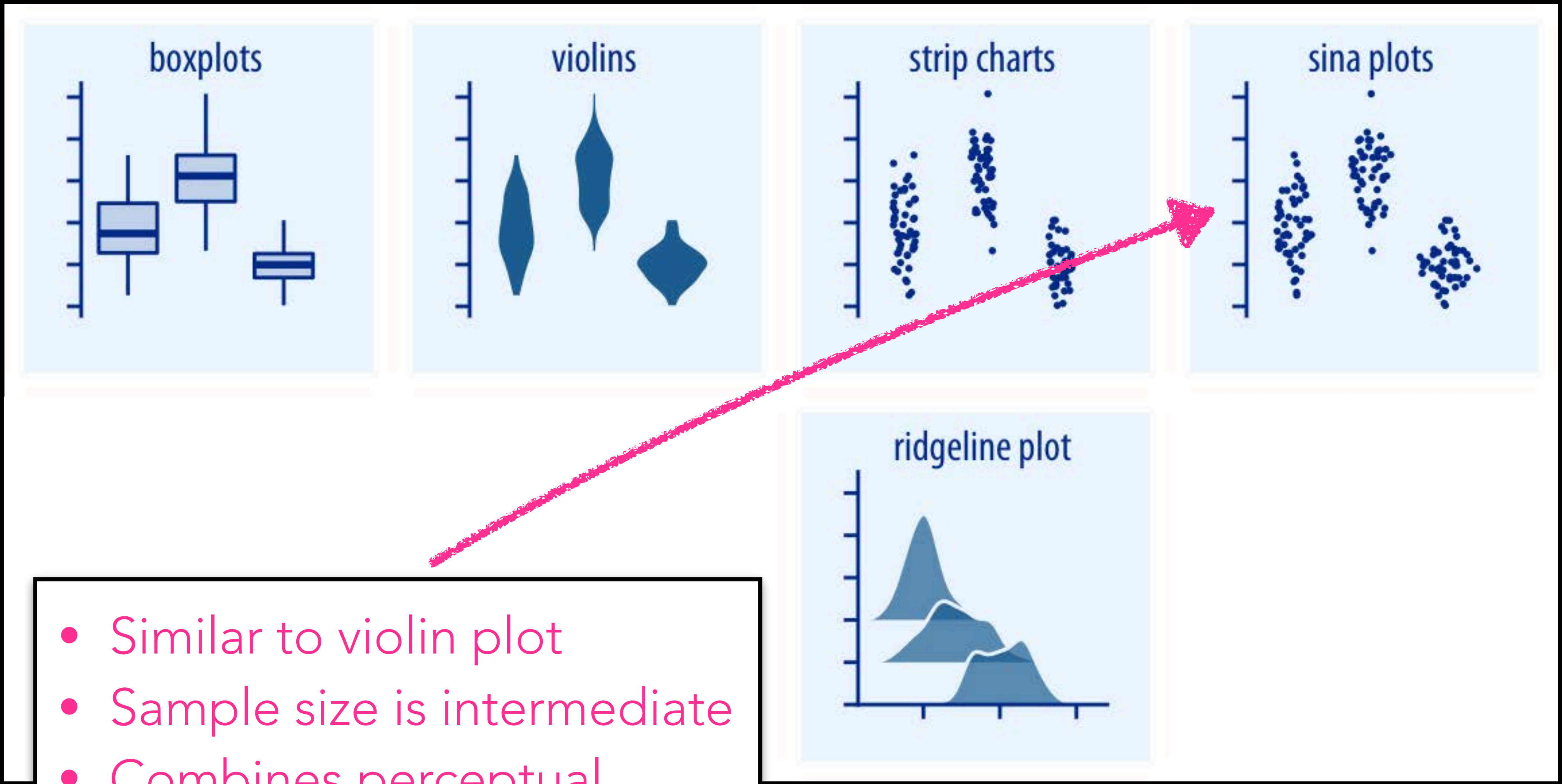
- Relatively few categories
- At least 5 data points per category
- You want a more nuanced comparison of the *shapes* of the distributions rather than statistical values

USAGE GUIDELINES



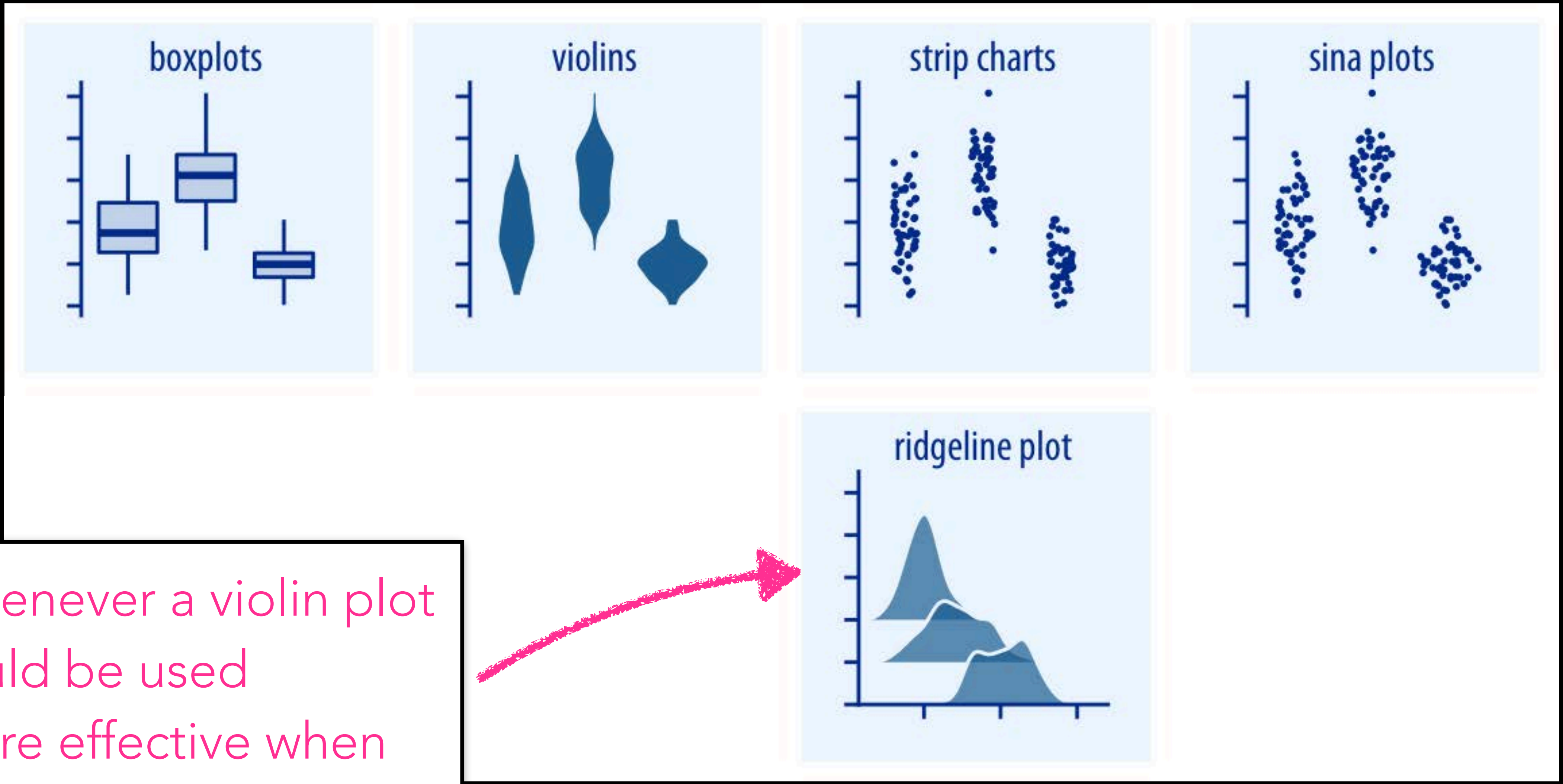
- You need to show raw data for any reason
- You have fewer than 5 data points in some categories

USAGE GUIDELINES



- Similar to violin plot
- Sample size is intermediate
- Combines perceptual properties of violins and strip charts

USAGE GUIDELINES

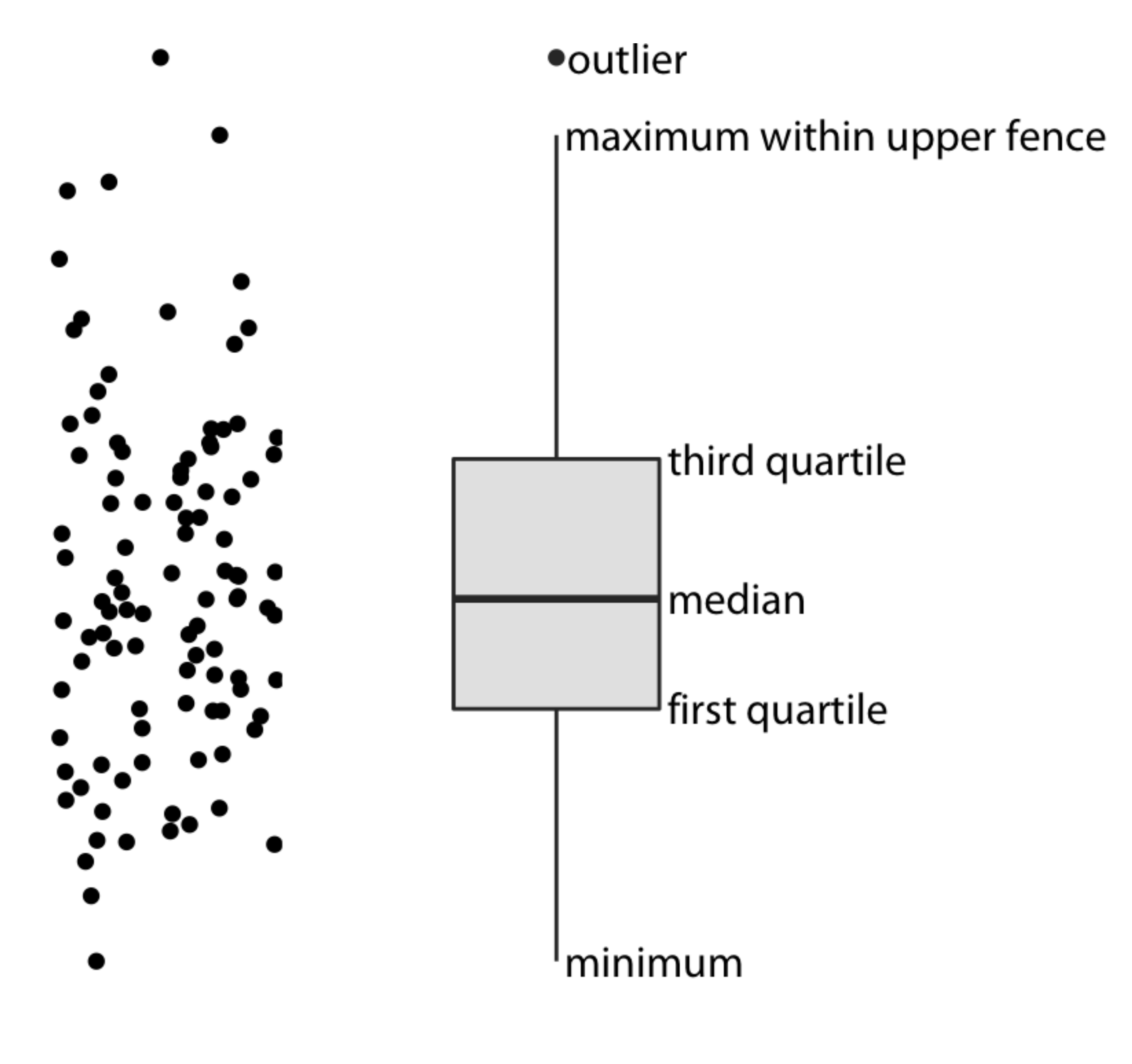


- Whenever a violin plot could be used
- More effective when there are *many* categories that you want to compare

BOXPLOTS

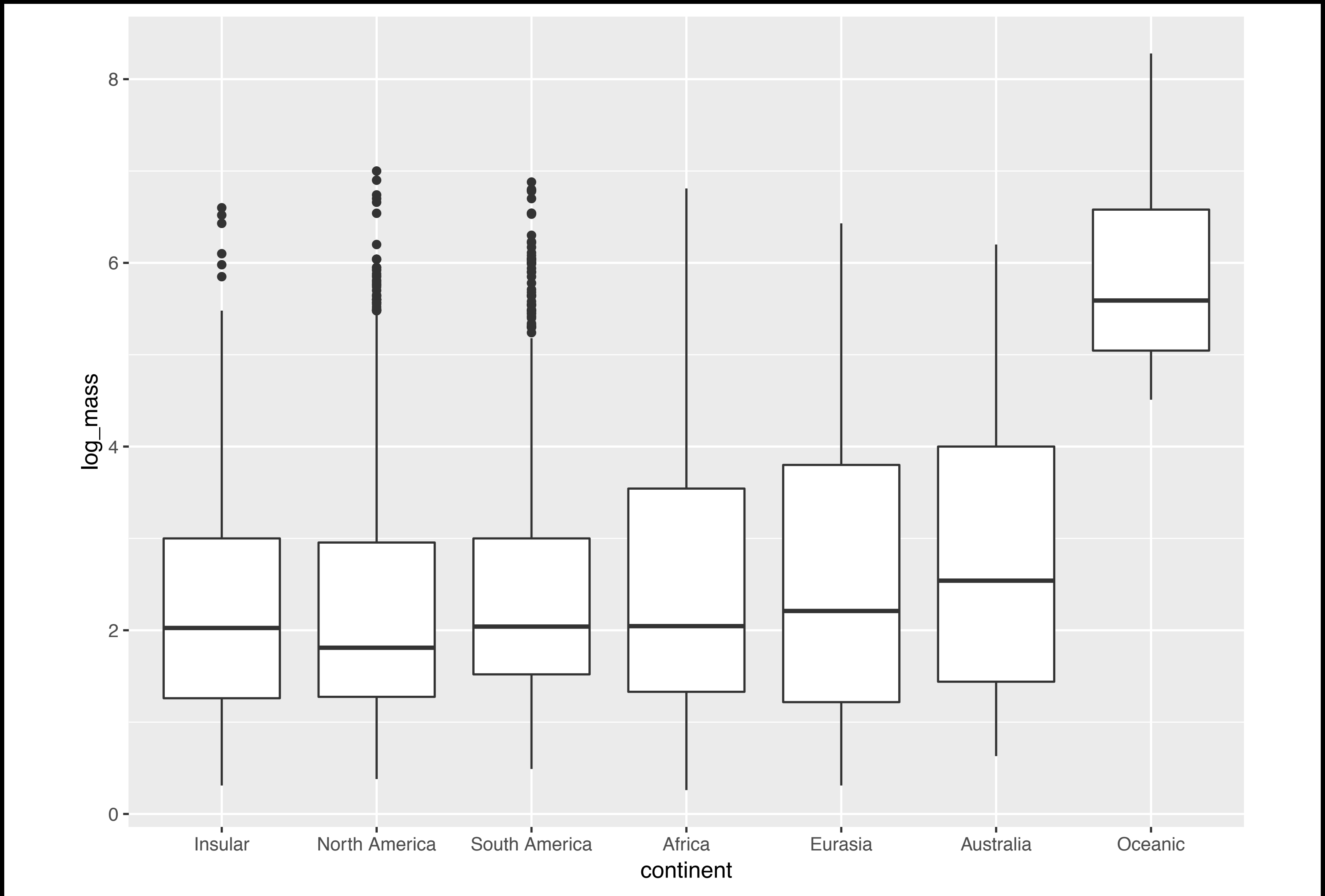
BOXPLOTS

- Each box plot shows a standardized set of 5 statistical summaries of the distribution, as well as individual outliers
- Very common in scientific lit
- But beware of using with lay audience (e.g., “what does the line sticking out of the top show exactly?”)



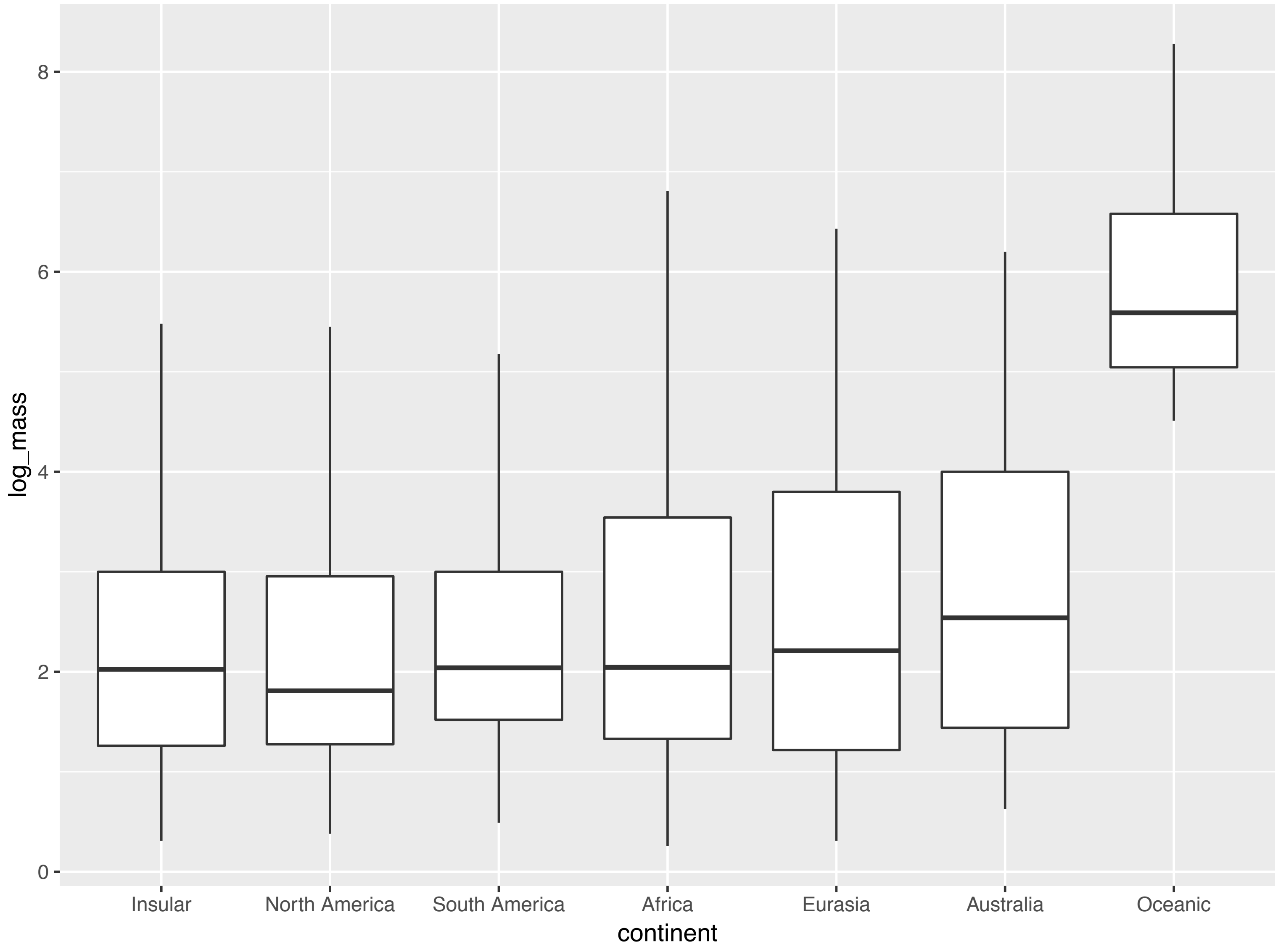
BOXPLOTS

- Geometric object is `geom_boxplot()`
- Categorical variable to mapped to x, and numeric variable mapped to y



BOXPLOTS

- Suppress outliers with `outlier.shape = NA`

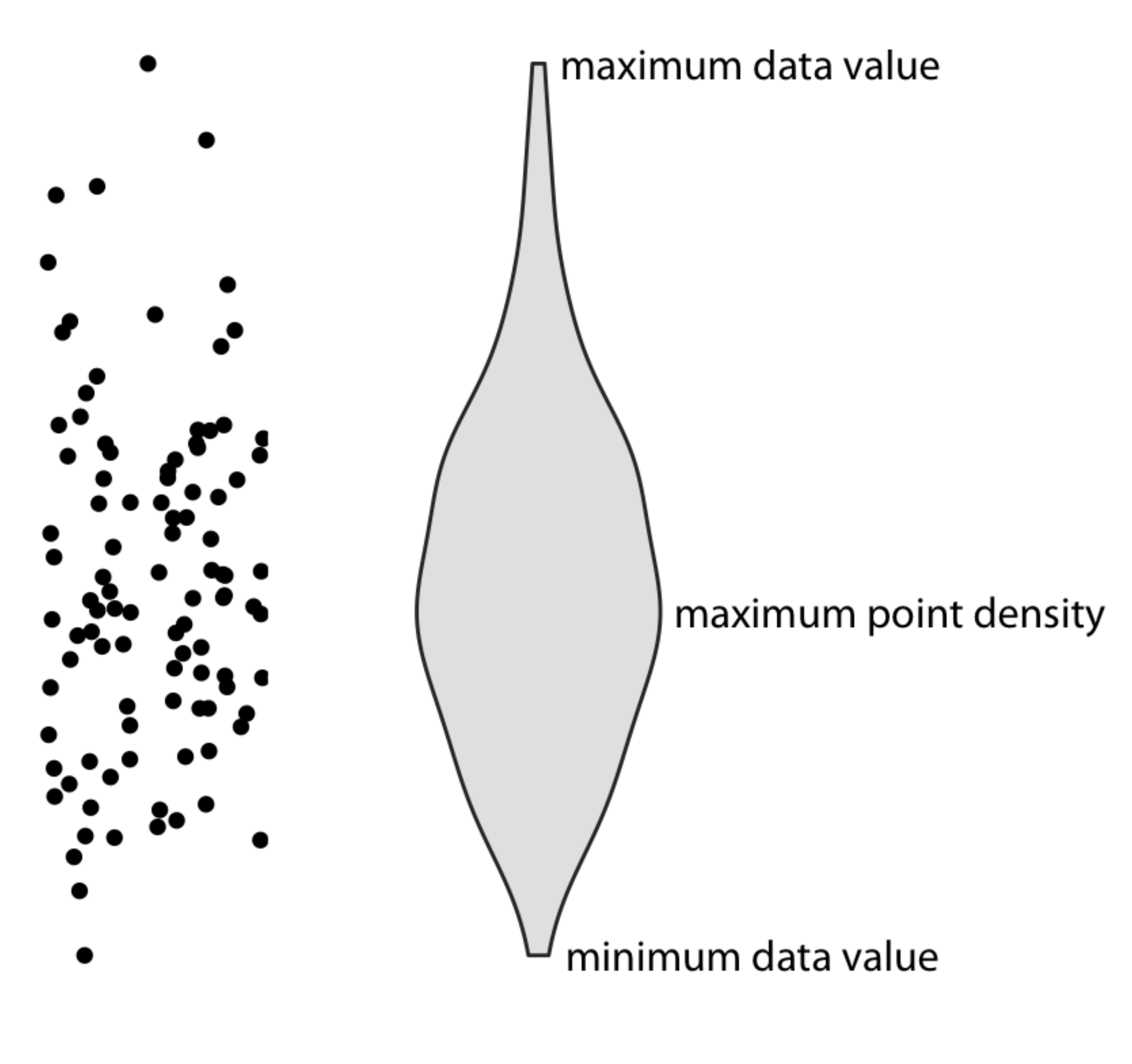


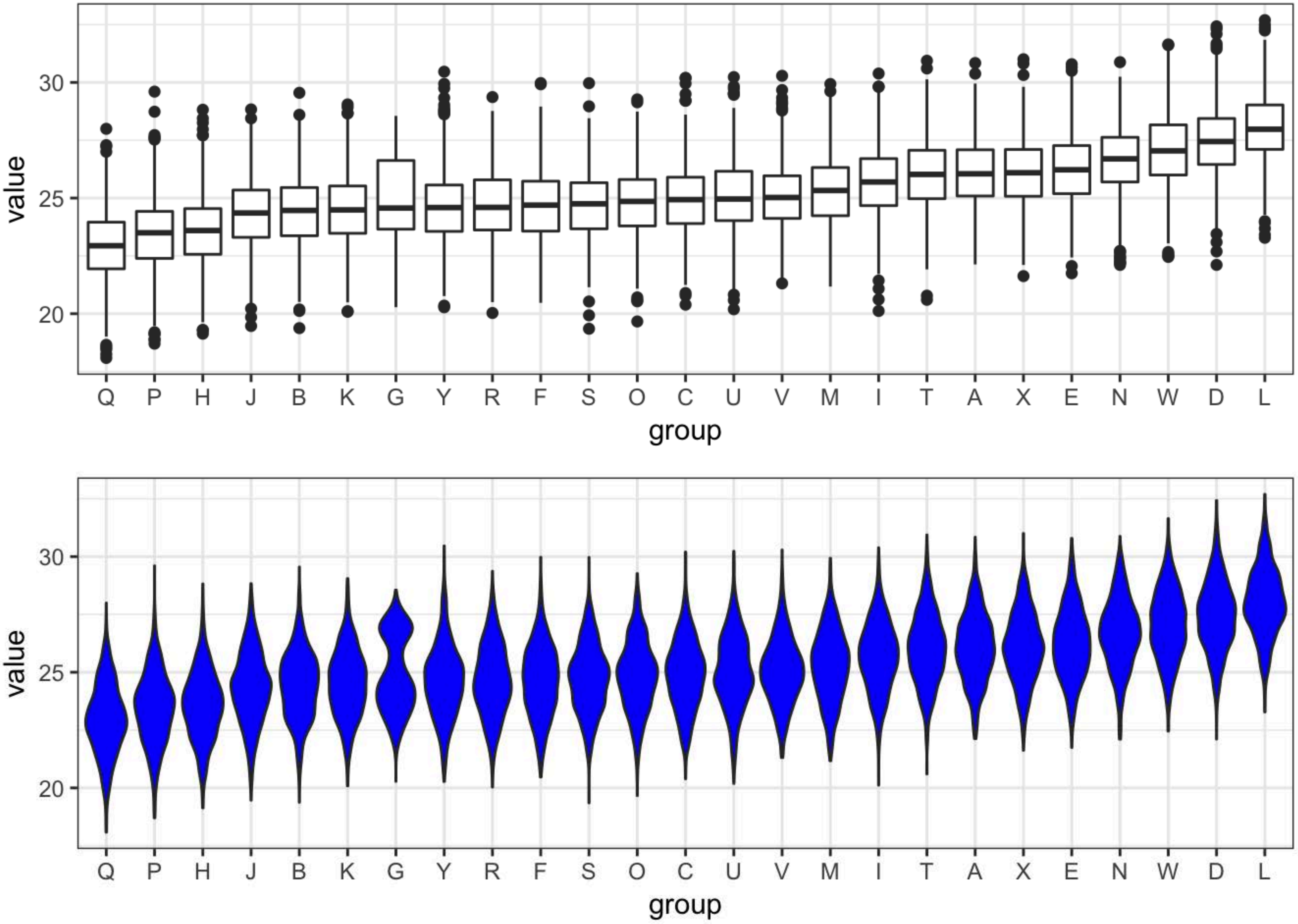
```
ggplot(lqm, aes(x = continent, y = log_mass)) +
  geom_boxplot(outlier.shape = NA)
```

VIOLIN PLOTS

VIOLIN PLOTS

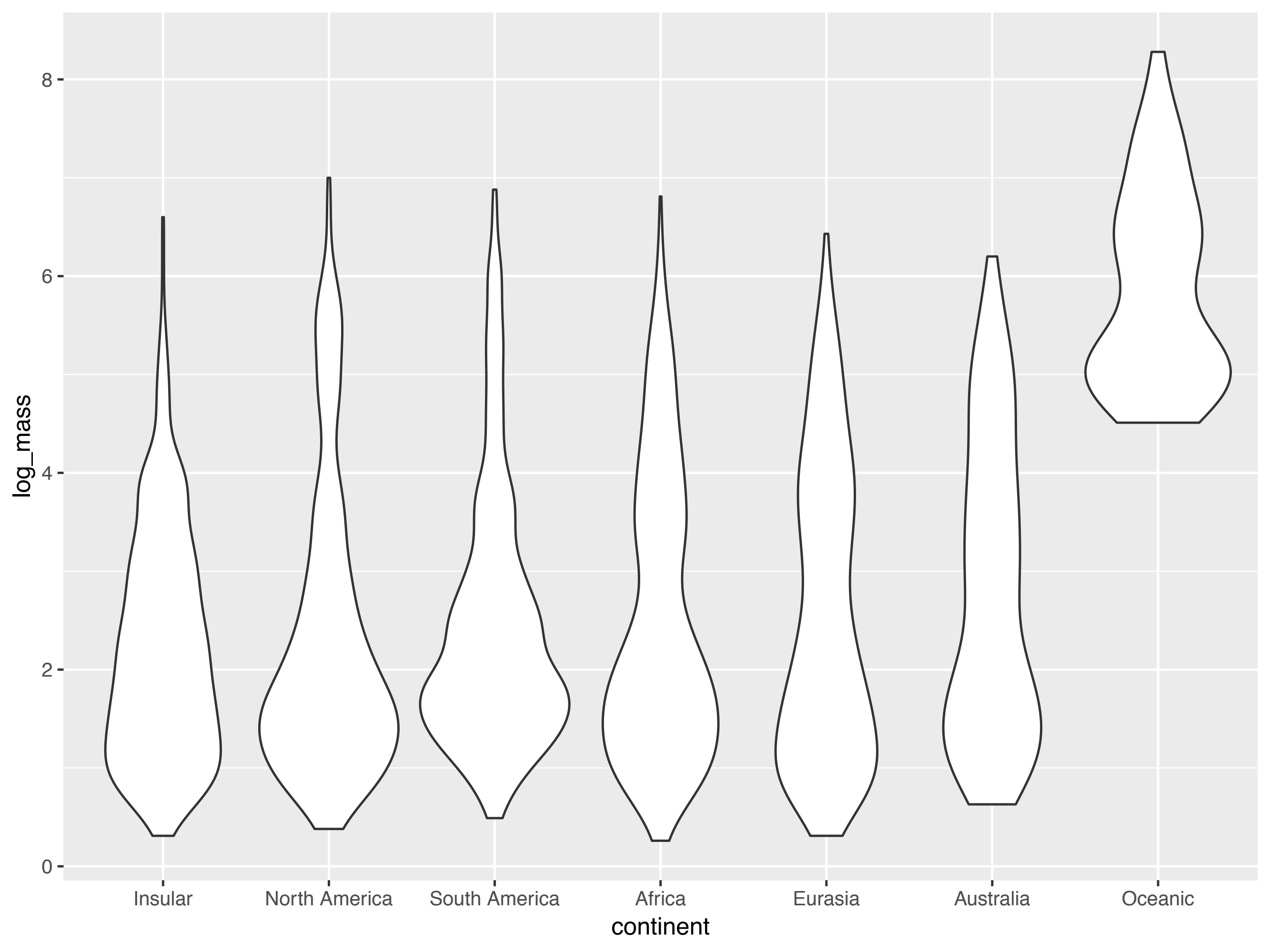
- Each violin is essentially a density plot rotated vertically and mirrored across the vertical
- Less standardized and precise, but also easier to intuit
- Better at showing multi-modal distributions



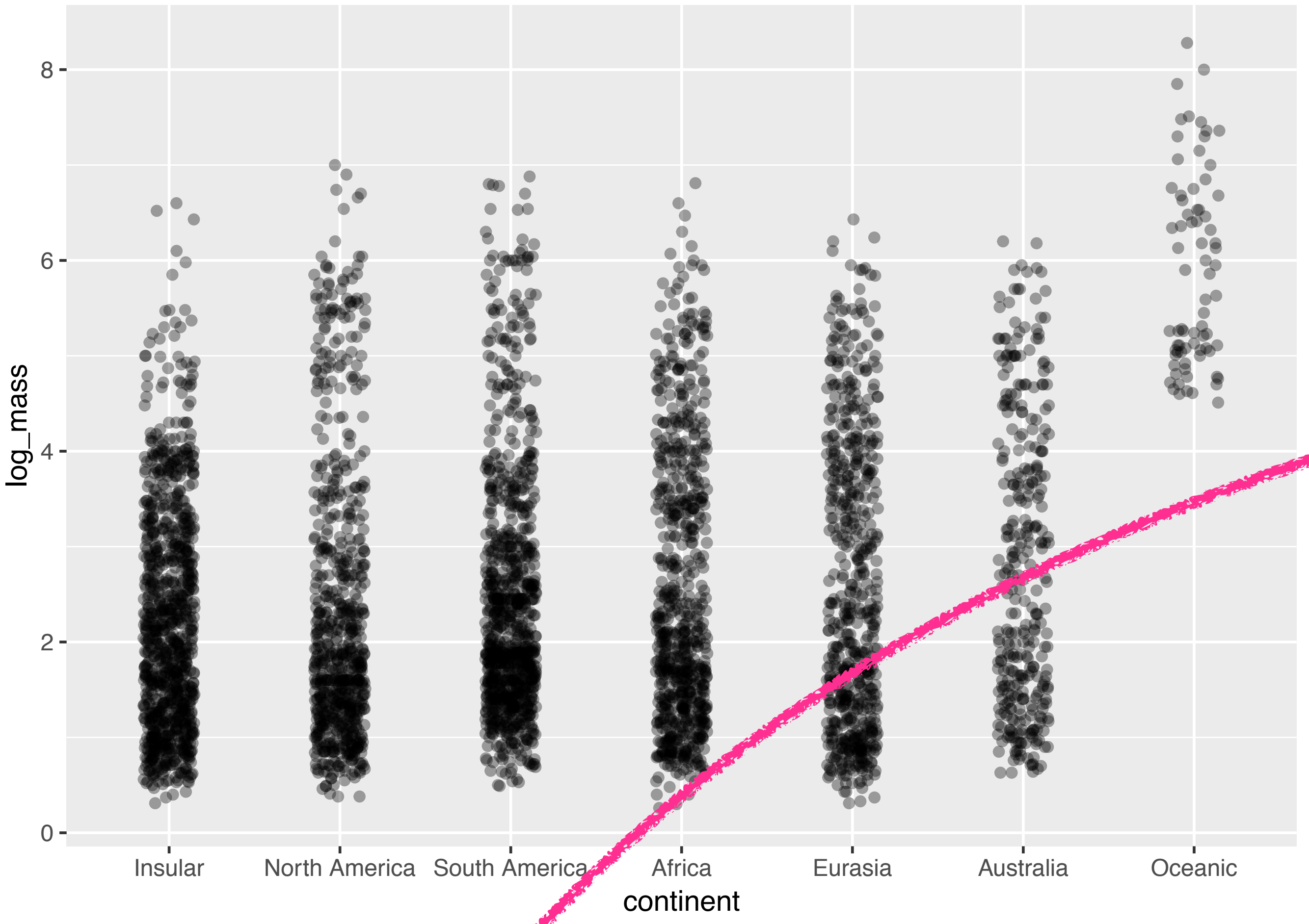

value
30
25
20
Q
P
H
J
B
K
G
Y
R
F
S
O
C
U
V
M
I
T
A
X
E
N
W
D
L
group
value
30
25
20
Q
P
H
J
B
K
G
Y
R
F
S
O
C
U
V
M
I
T
A
X
E
N
W
D
L
group

VIOLIN PLOTS

- Geometric object is `geom_violin()`
- Mappings same as boxplot



STRIP PLOTS



`geom_jitter()` makes randomly displaced points

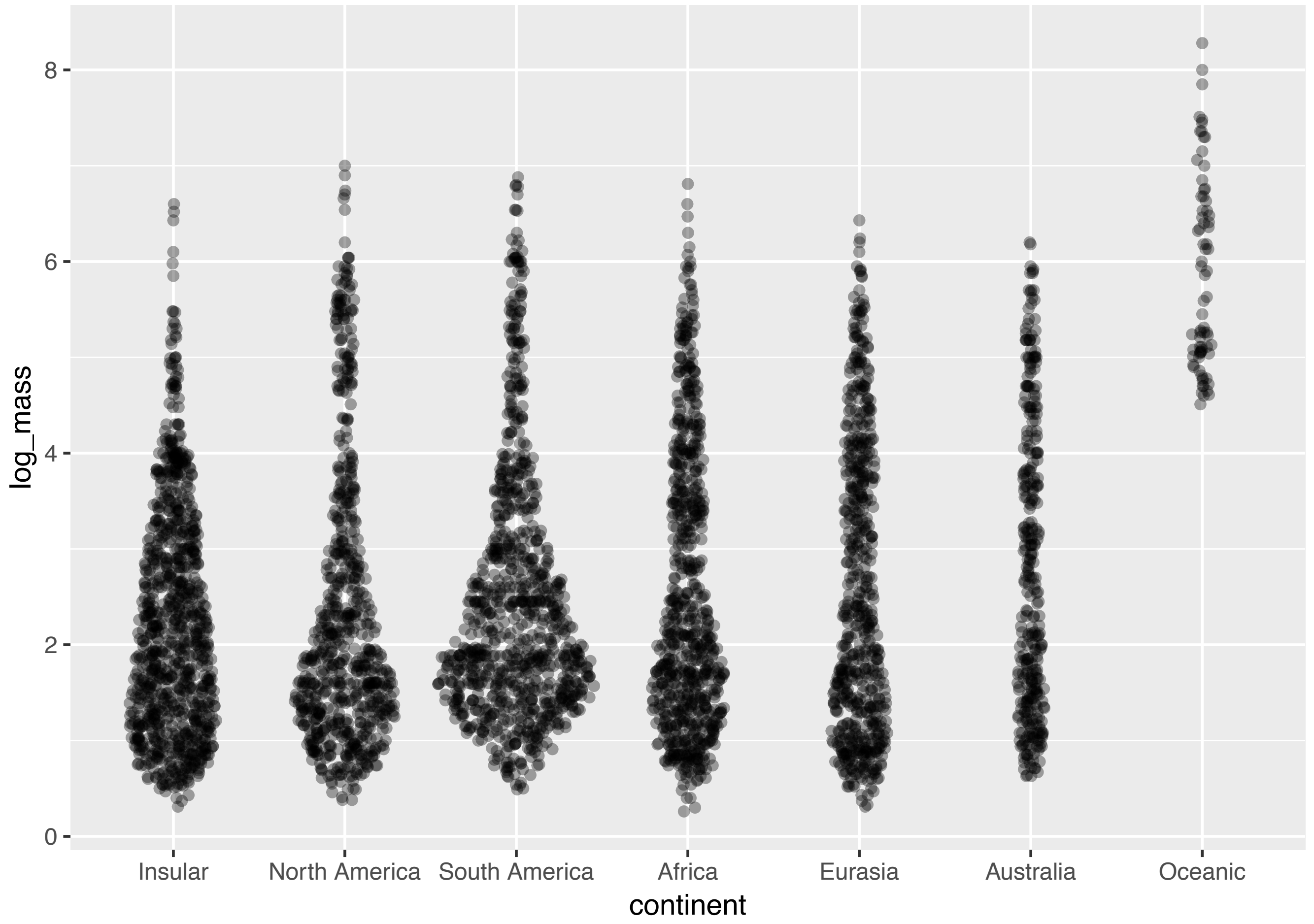
`height` and `width` parameters control amount of jitter

```
ggplot(lqm, aes(x = continent, y = log_mass)) +
  geom_jitter(width = 0.15, height = 0, alpha = 0.35)
```

SINA PLOTS

Not a default geom!

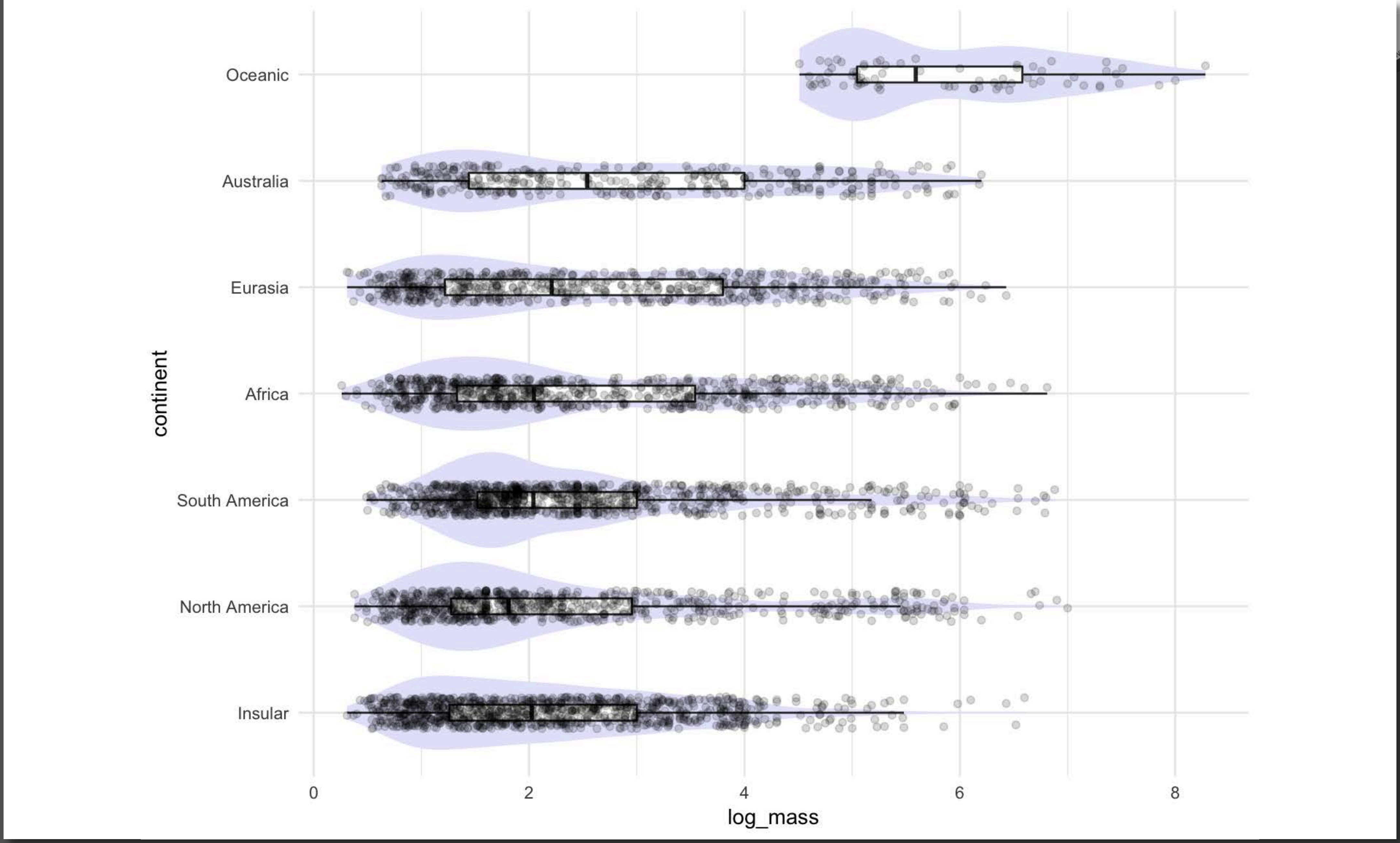
Provided by the "ggforce" package.



```
library("ggforce")

ggplot(lqm, aes(x = continent, y = log_mass)) +
  geom_sina(alpha = 0.15)
```

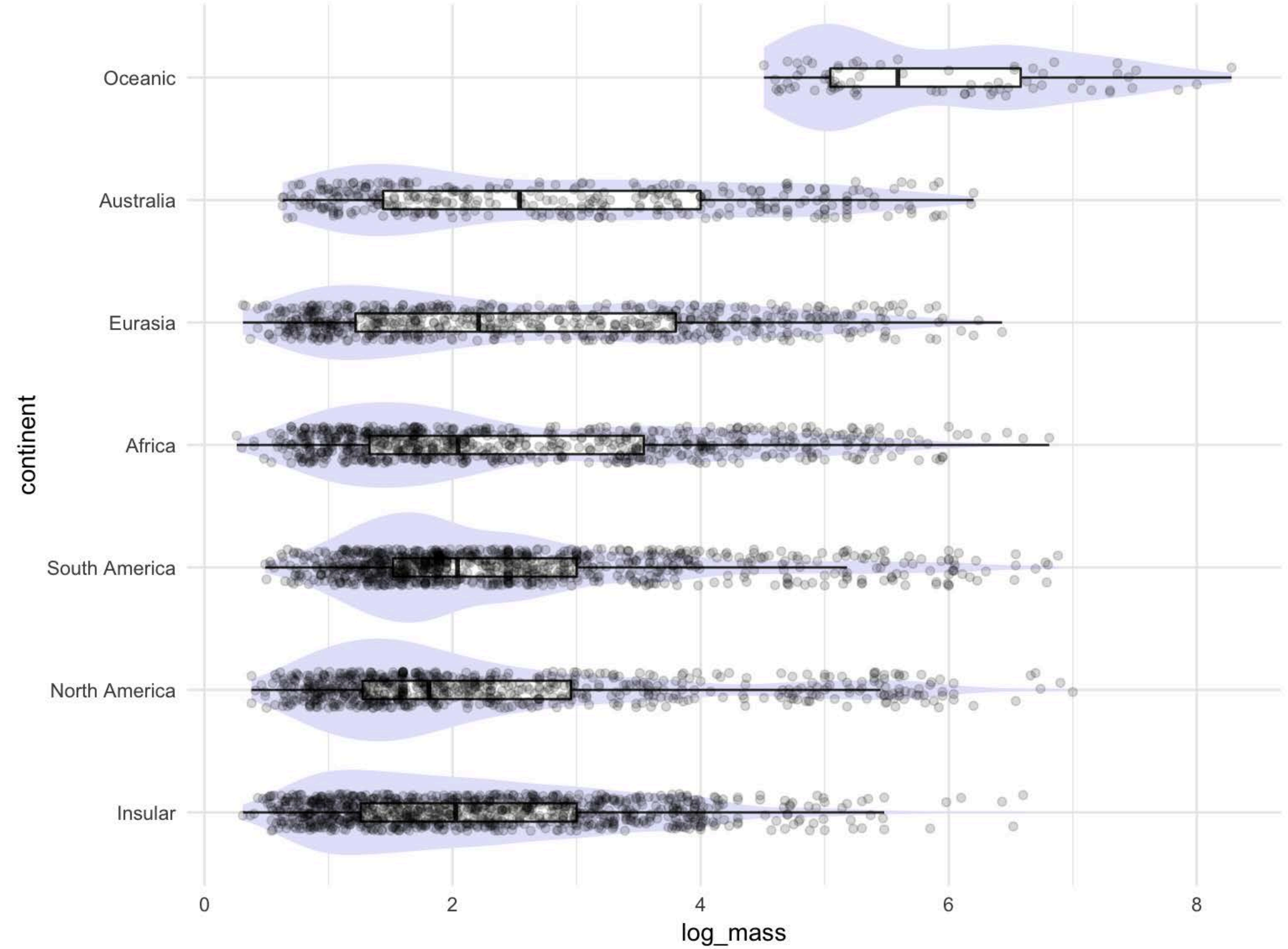
YOUR TURN



Try to reverse-engineer this (rather busy) plot.

5:00

YOUR TURN

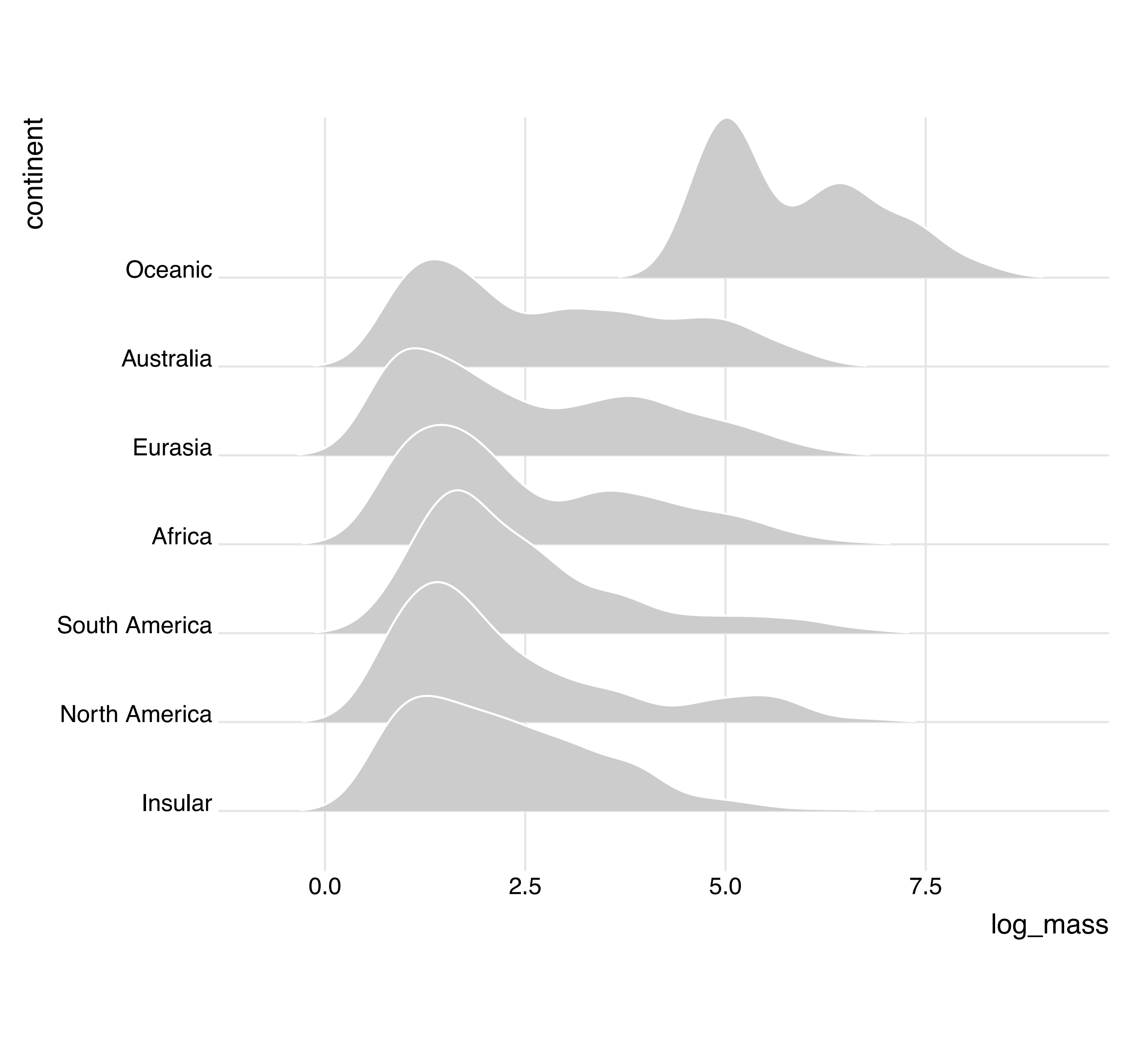


```
ggplot(lqm, aes(x = continent, y = log_mass)) +
  geom_violin(fill = "lavender", color = "white") +
  geom_boxplot(fill = "white", width = 0.15, outlier.shape = NA) +
  geom_jitter(width = 0.15, height = 0, alpha = 0.15) +
  coord_flip()
```

RIDGELINE PLOTS

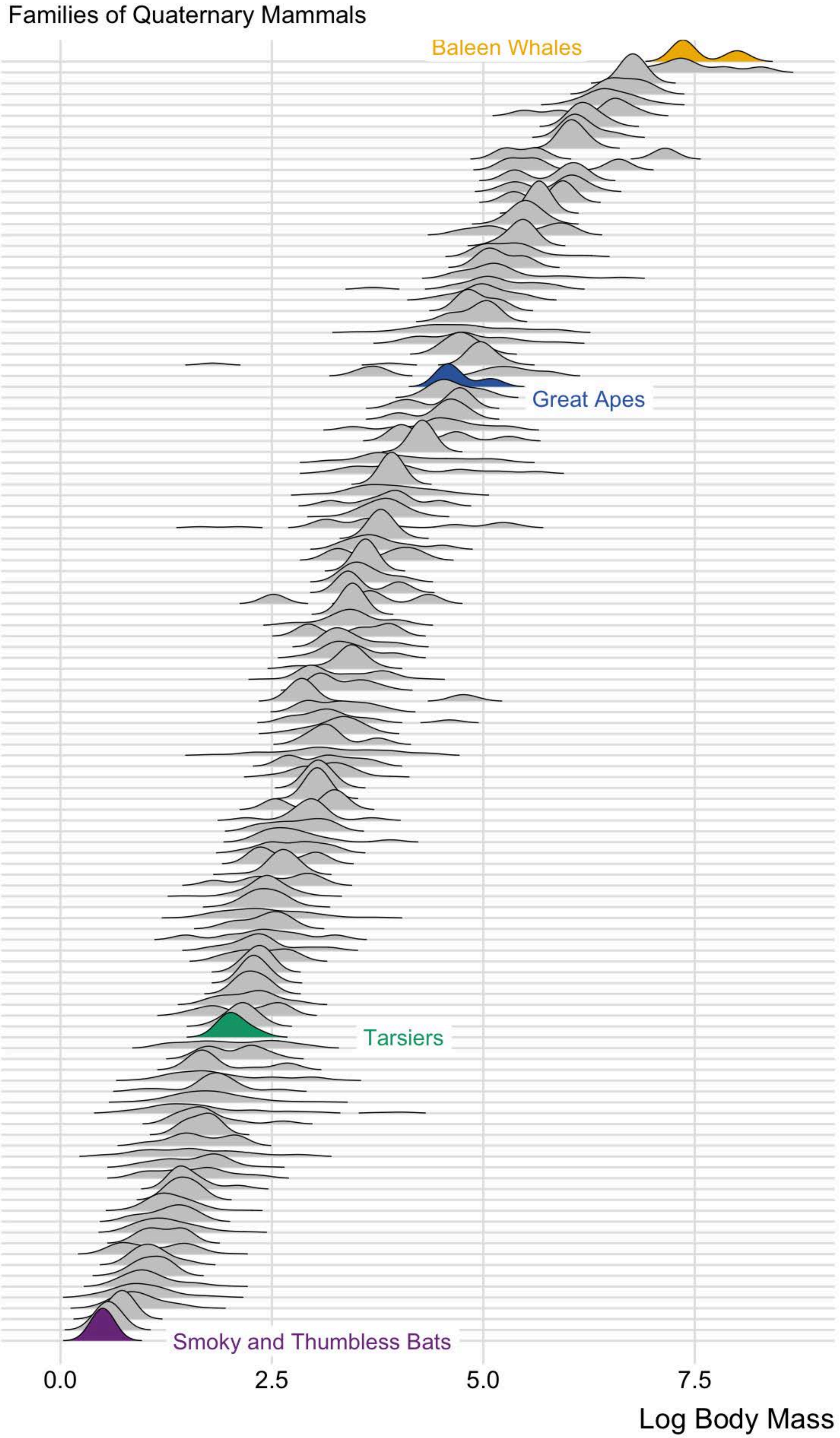
RIDGELINE PLOTS

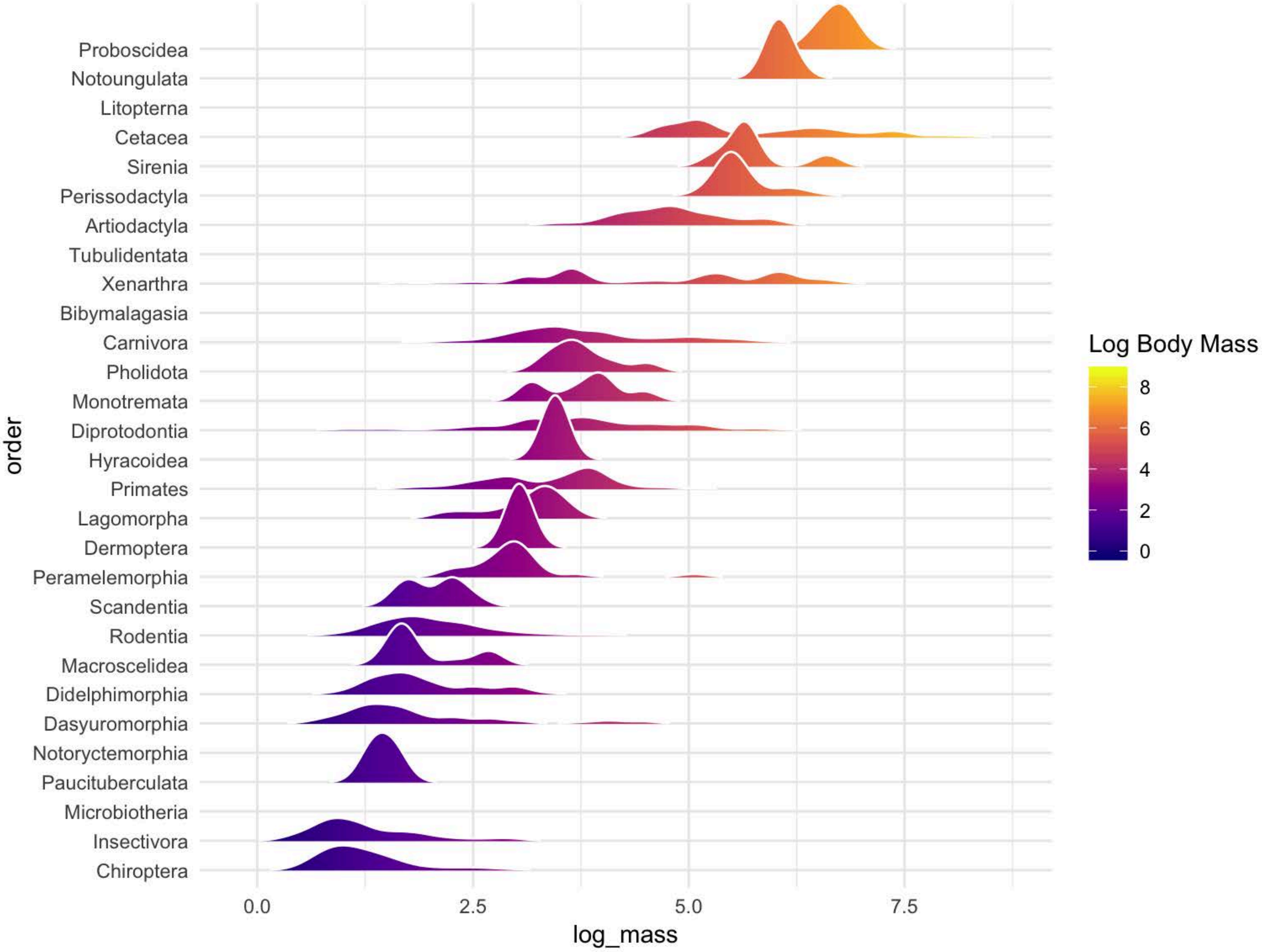
- Each ridge is a vertically offset density plot
- Rarely used, but easy to compare shapes and relative heights
- Not good if specific statistical values are of interest.



RIDGELINE PLOT

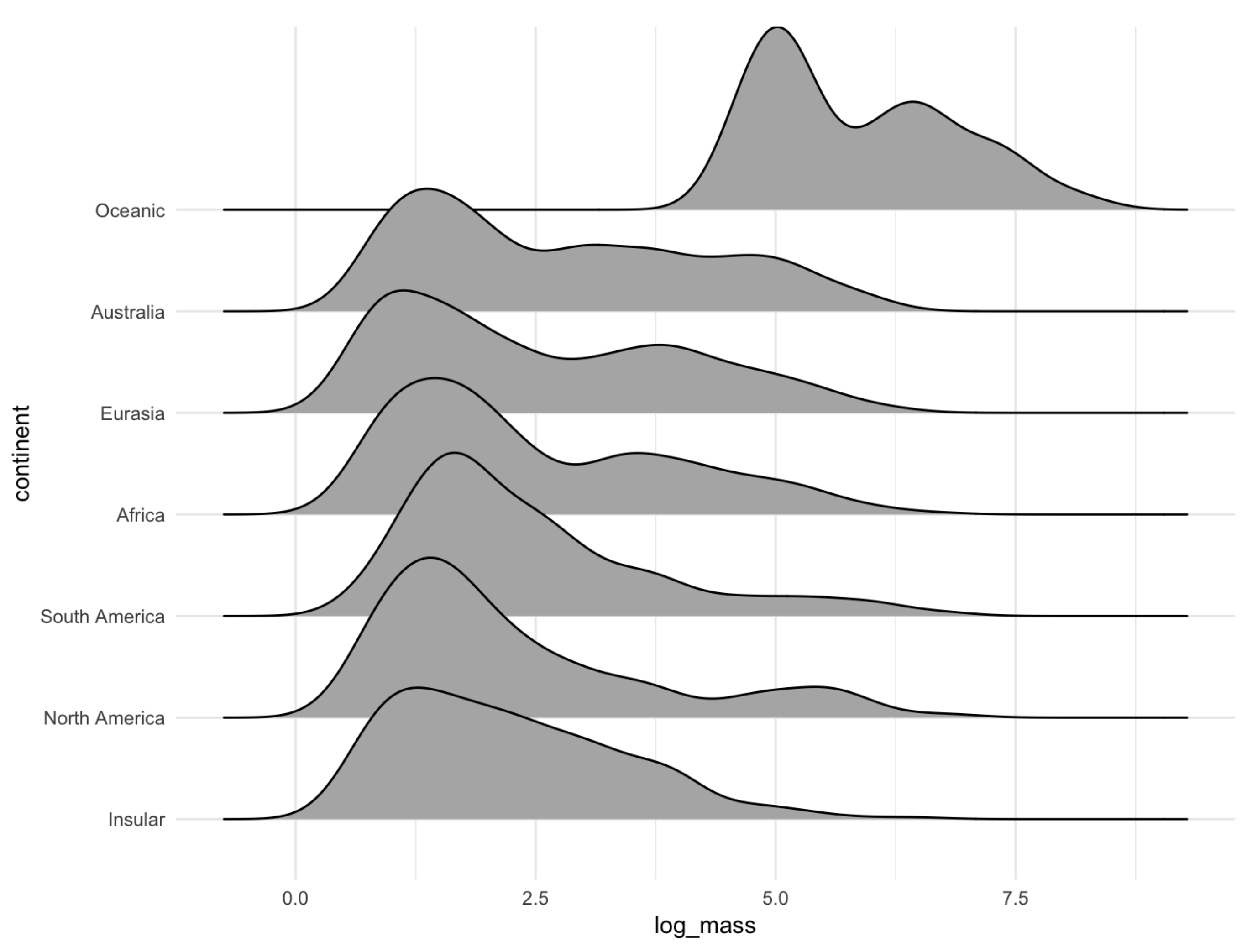
- Only method we have discussed that remains readable with large number of distributions




Proboscidea
Notoungulata
Litopterna
Cetacea
Sirenia
Perissodactyla
Artiodactyla
Tubulidentata
Xenarthra
Bibymalagasia
Carnivora
Pholidota
Monotremata
Diprotodontia
Hyracoidea
Primates
Lagomorpha
Dermoptera
Peramelemorphia
Scandentia
Rodentia
Macroscelidea
Didelphimorphia
Dasyuromorphia
Notoryctemorphia
Paucituberculata
Microbiotheria
Insectivora
Chiroptera
order
0.0
2.5
5.0
7.5
log_mass
Log Body Mass
8
6
4
2
0

RIDGELINE PLOTS

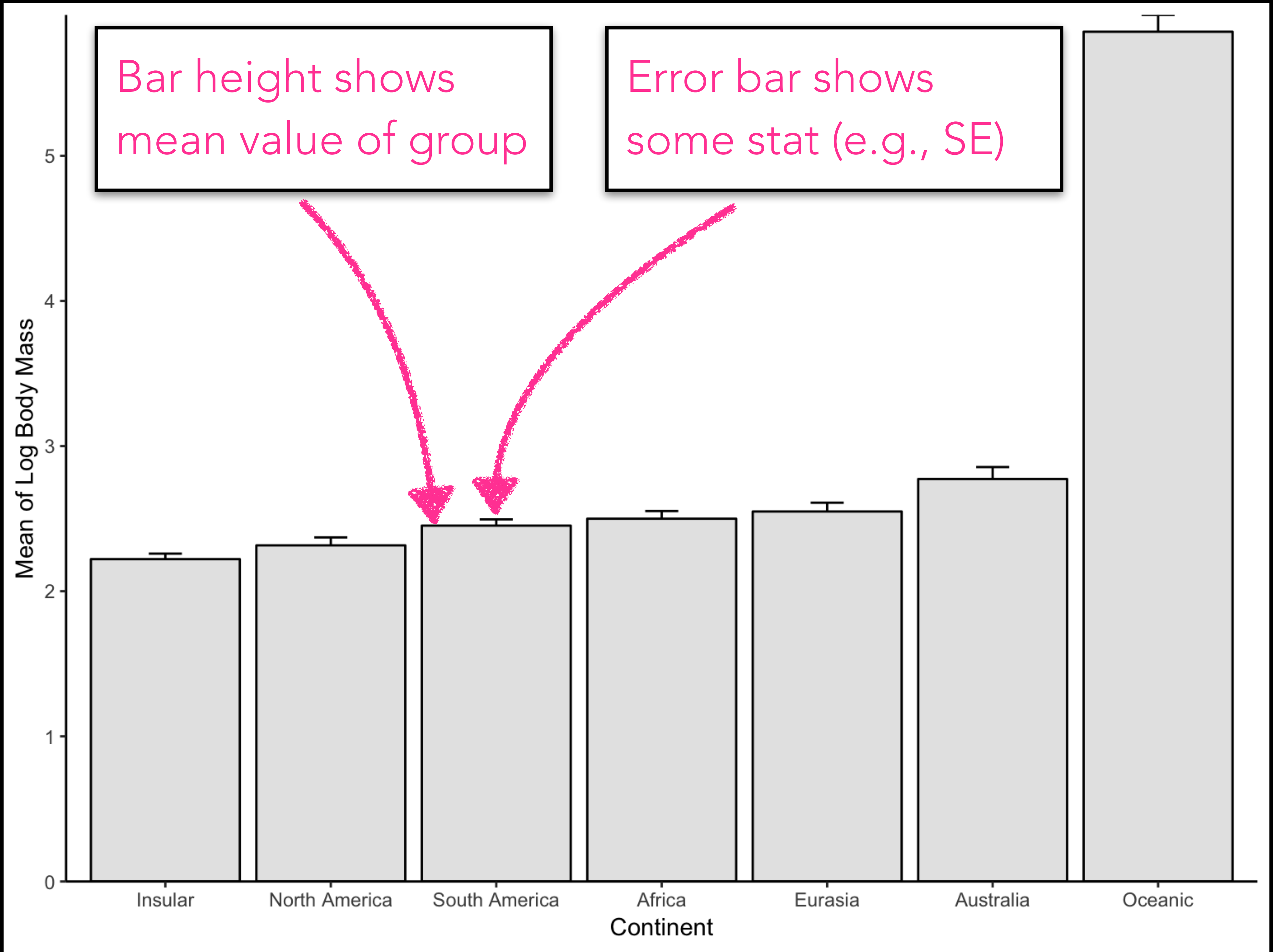
- Not a default geom—provided by "ggridges" package
- Basic geometric object is `geom_density_ridges()`
- Like density plot, numeric variable mapped to x but categorical variable to mapped to y



```
library("ggridges")

ggplot(lqm, aes(x = log_mass, y = continent)) +
  geom_density_ridges()
```

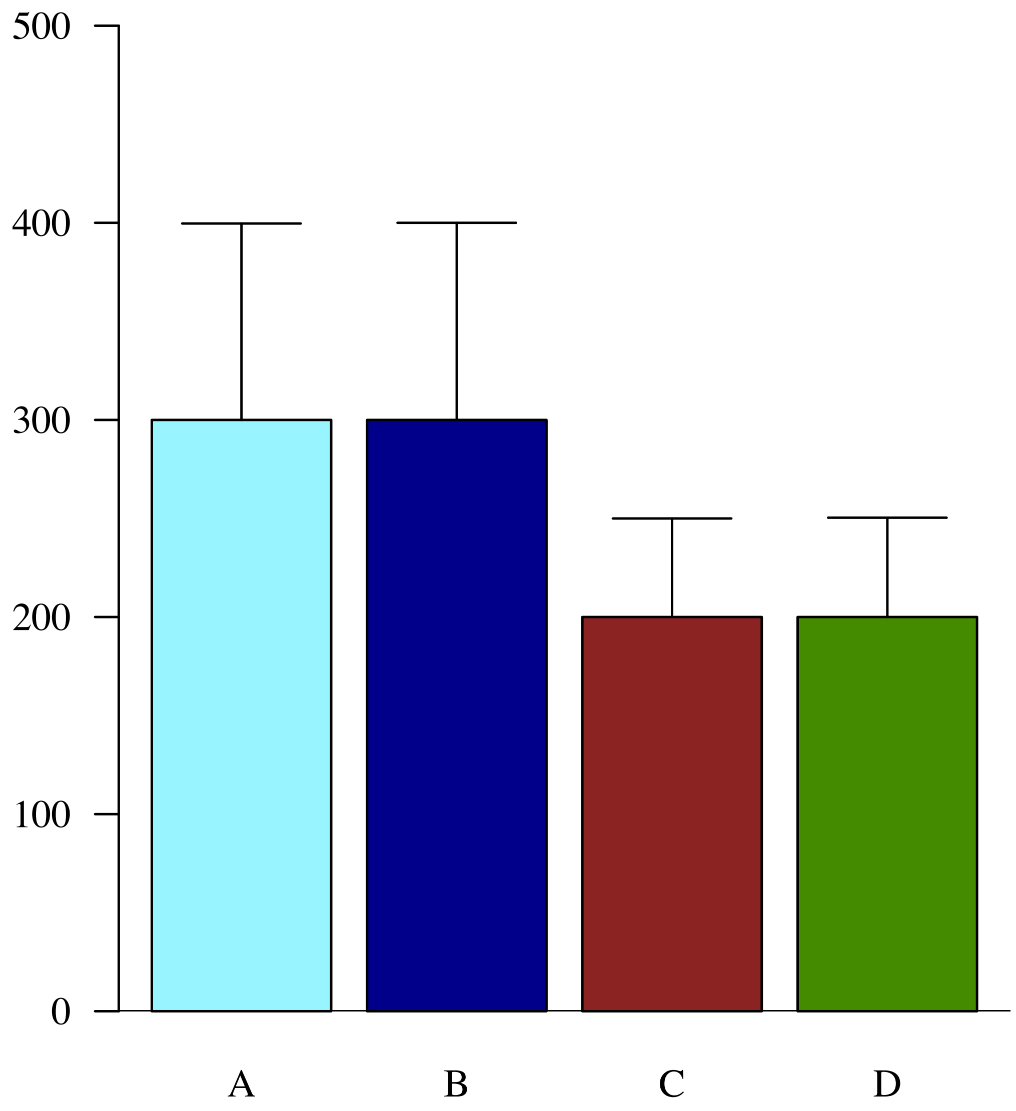
WHAT ABOUT A "DYNAMITE PLOT"?

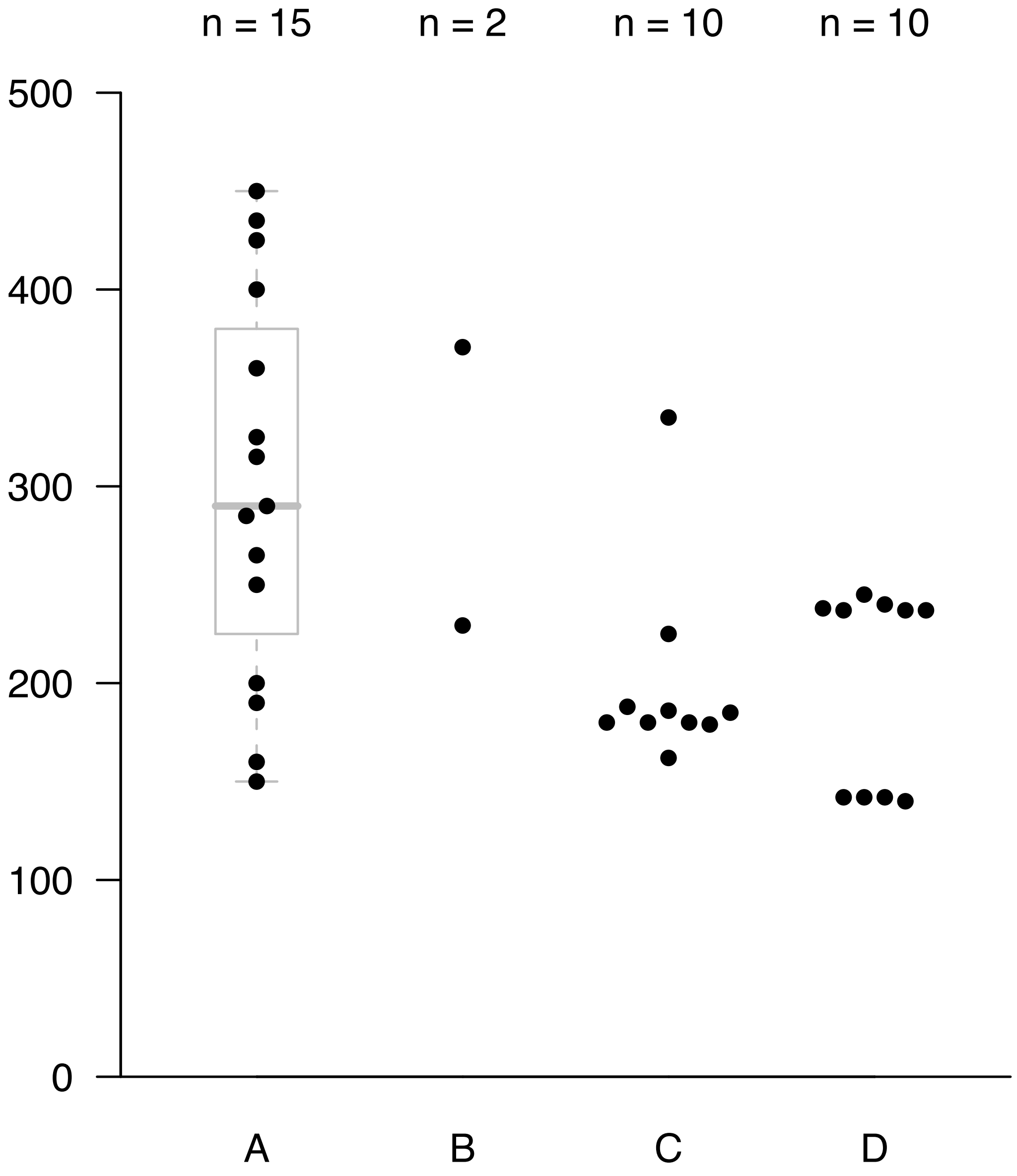


Bar height shows mean value of group

Error bar shows some stat (e.g., SE)

DON'T USE THEM FOR DISTRIBUTIONS





http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TatsukiRcode/Poster3.pdf

PROBLEMS WITH DYNAMITE PLOTS

- Hides the data for no particular reason
- Does not show possible skew or multi-modality
- Anchors data at zero, *which might not make sense for the distribution*
    - Data in small range: : part of the range covered by the bar might have never been observed in the sample
    - Log scale (log(0) is undefined)
- *Bar plots are good for showing counts and single amounts of categories, but not for summarizing continuous data*