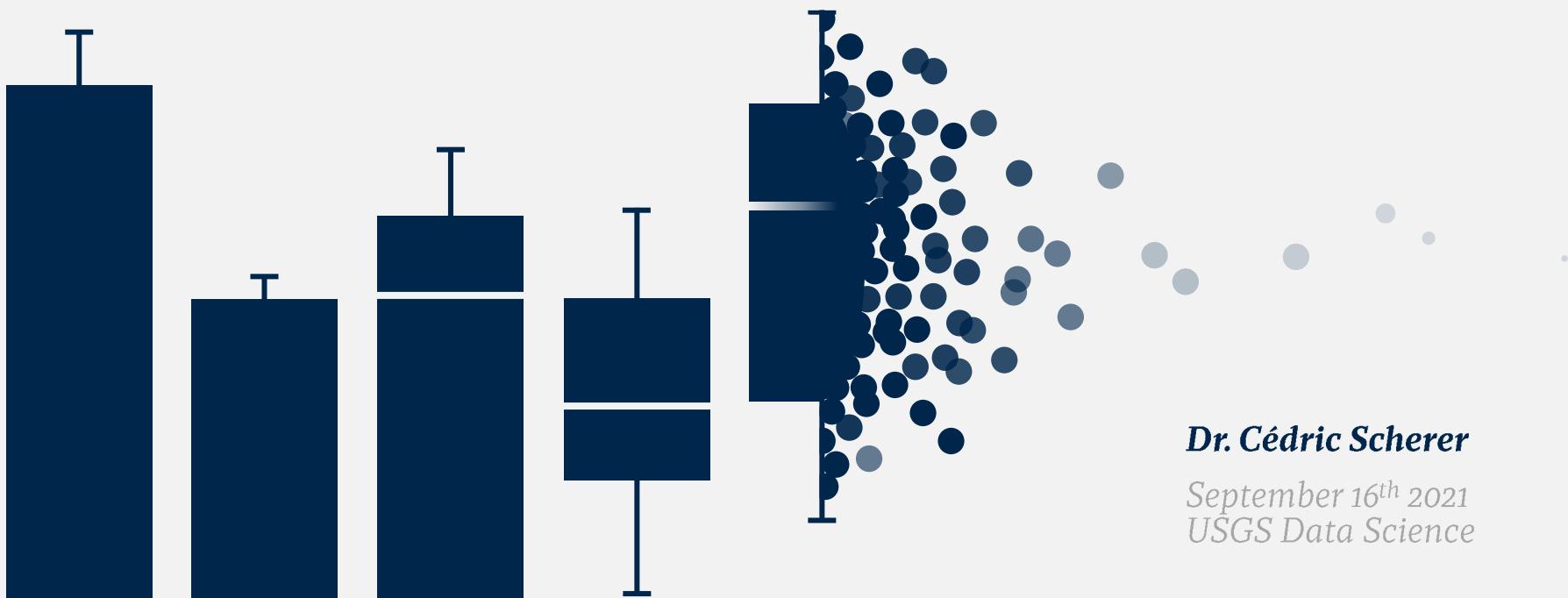


BEYOND BAR AND BOX PLOTS

Chart alternatives and how to design them with ggplot2

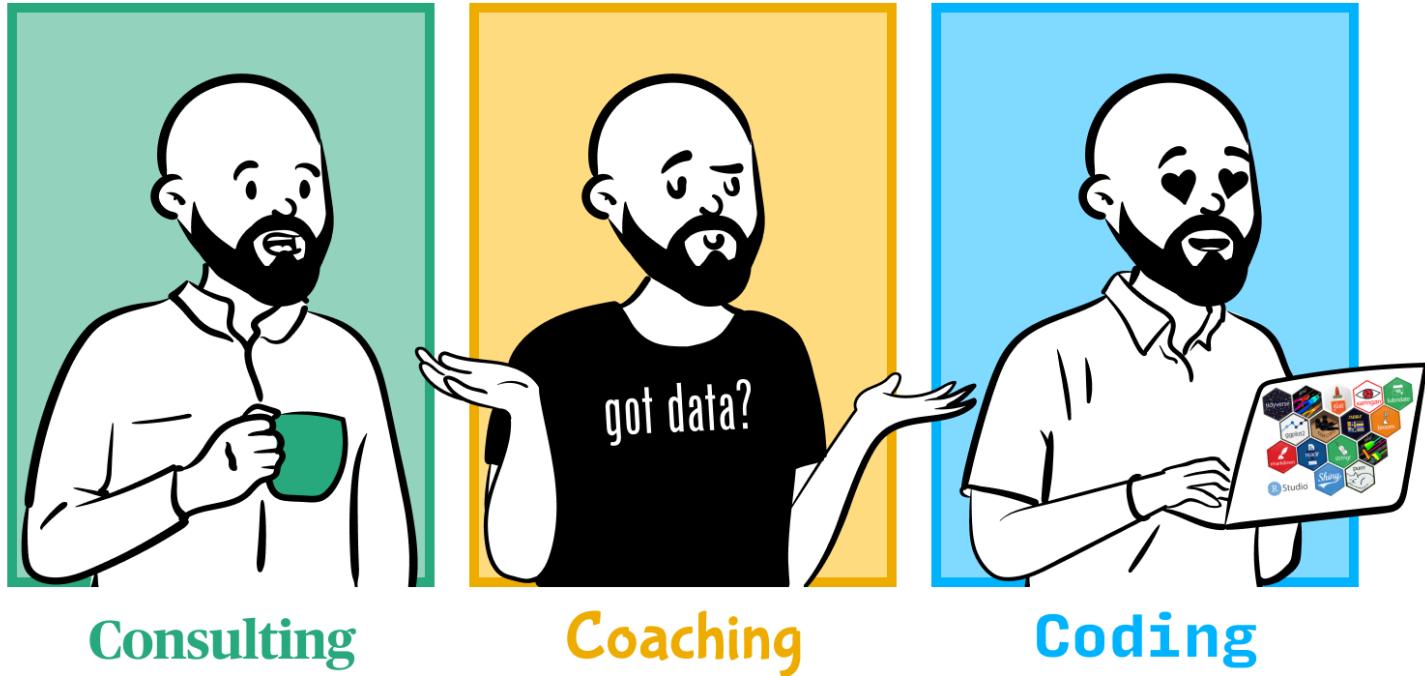


Dr. Cédric Scherer

September 16th 2021
USGS Data Science

Cédric Scherer

Freelance Data Visualization Specialist
Computational Ecologist at IZW Berlin



cedricscherer.com



@CedScherer



@z3tt



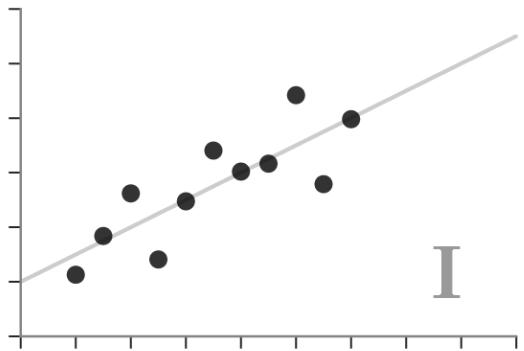
@cedscherer

*...make both calculations and graphs.
Both sorts of output should be studied;
each will contribute to understanding.*

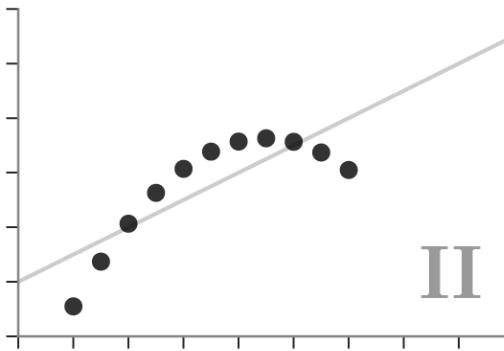
F. J. Anscombe (1973)

Anscombe's Quartet

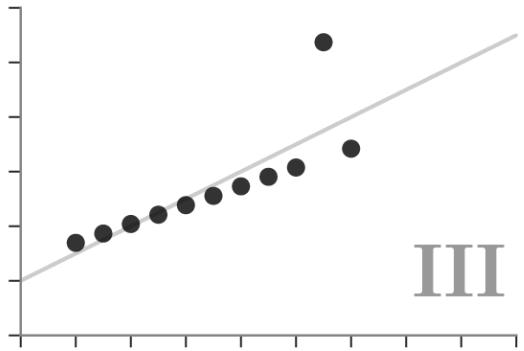
Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



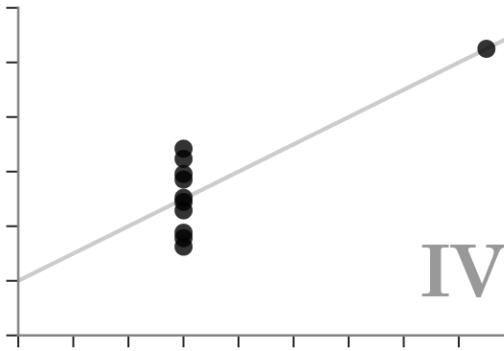
I



II



III



IV

Matejka & Fitzmaurice (2017)



cedricscherer.com



@CedScherer

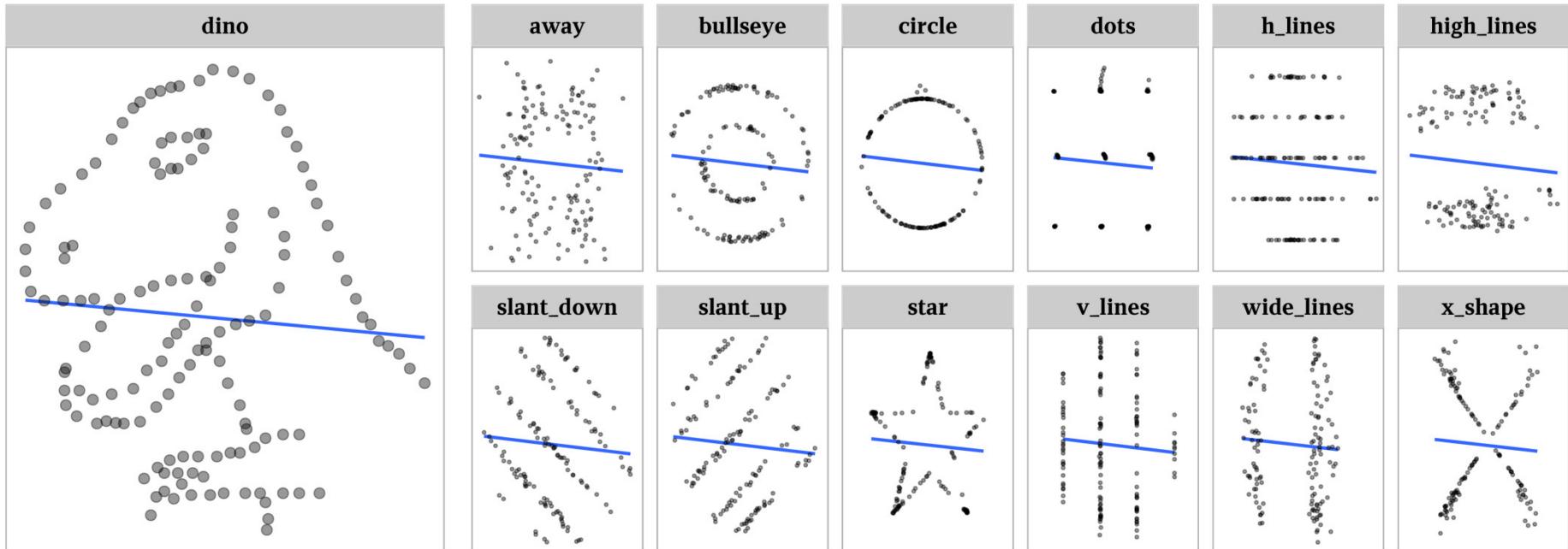


@z3tt



@cdscherer

The Datasaurus by Alberto Cairo shows us why visualisation is important, not just summary statistics.



PERSPECTIVE

Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

Tracey L. Weissgerber^{1*}, Natasa M. Milic^{1,2}, Stacey J. Winham³, Vesna D. Garovic¹

1 Division of Nephrology & Hypertension, Mayo Clinic, Rochester, Minnesota, United States of America,

2 Department of Biostatistics, Medical Faculty, University of Belgrade, Belgrade, Serbia, **3** Division of Biomedical Statistic and Informatics, Mayo Clinic, Rochester, Minnesota, United States of America

* weissgerber.tracey@mayo.edu

Weissgerber et al. (2015) PLoS Biol 13:e1002128 ([10.1371/journal.pbio.1002128](https://doi.org/10.1371/journal.pbio.1002128))



cedricscherer.com



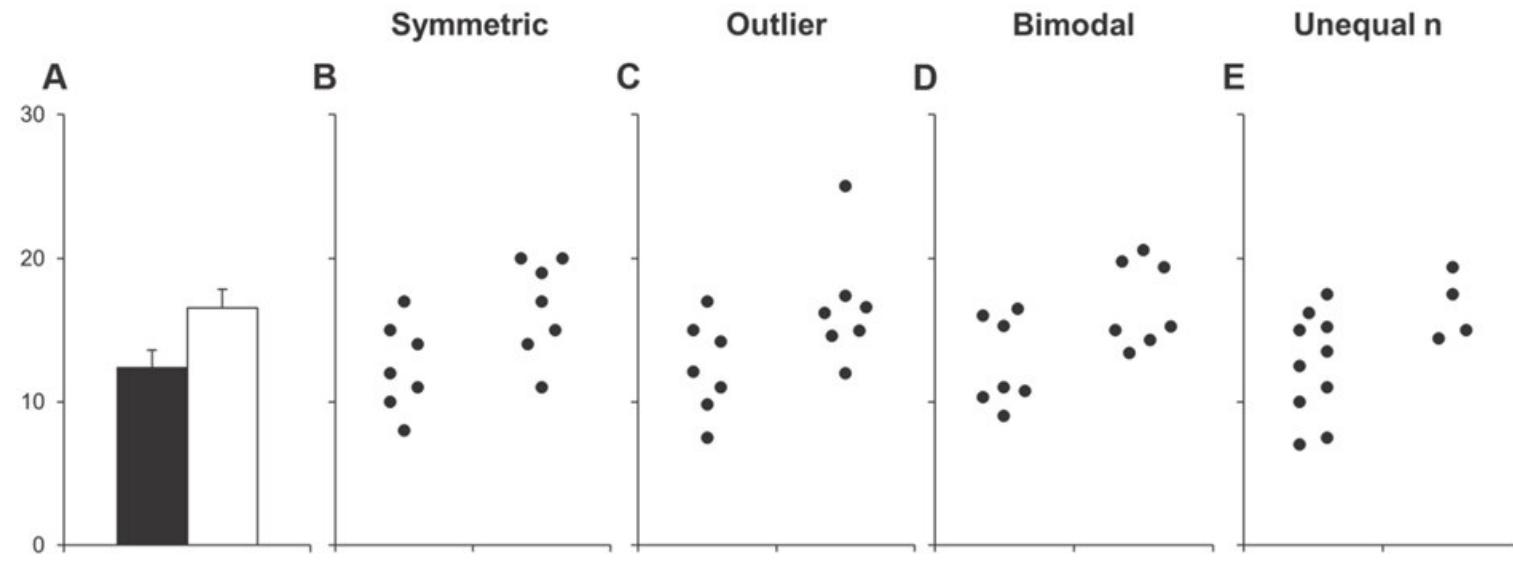
@CedScherer



@z3tt



@cedscherer



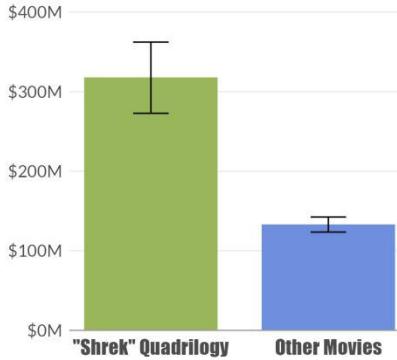
| Test | p value | | | |
|----------------------|---------|-------|-------|-------|
| T-test: Equal var. | 0.035 | 0.050 | 0.026 | 0.063 |
| T-test: Unequal var. | 0.035 | 0.050 | 0.026 | 0.035 |
| Wilcoxon | 0.054 | 0.073 | 0.128 | 0.103 |

Weissgerber et al. (2015) PLoS Biol 13:e1002128 ([10.1371/journal.pbio.1002128](https://doi.org/10.1371/journal.pbio.1002128))



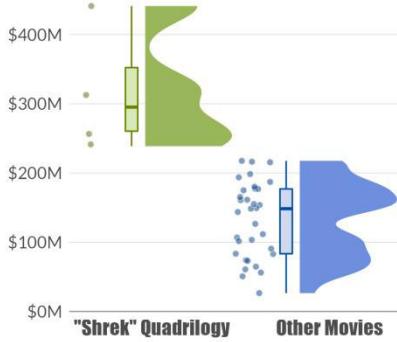


Domestic Box Office of DreamWorks Movies



© Dreamworks Animation

Domestic Box Office of DreamWorks Movies



Why Dynamite Plots Are Terrible—and Why You Should Use Something Else | Cédric Scherer | #30DayChartChallenge 2021 | Day 27: Educational



cedricscherer.com



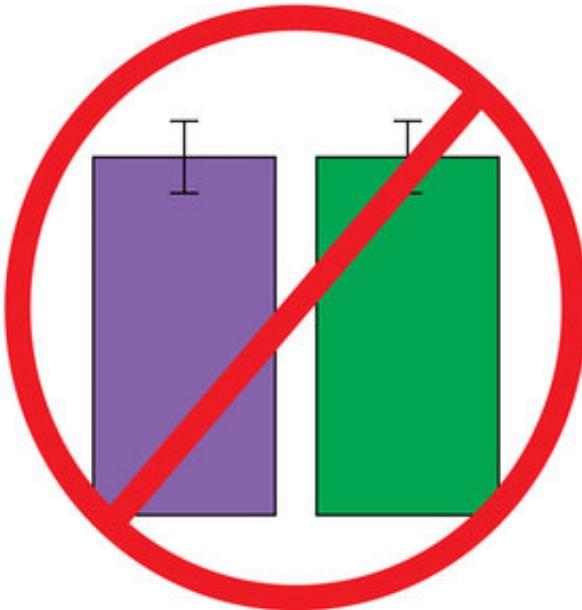
@CedScherer



@z3tt



@cedscherer



#barbarplots

barbarplots.github.io



cedricscherer.com



@CedScherer



@z3tt



@cedscherer

Welcome!

We are a group of young scientists interested in improving scientific communication. Learn more about our campaign in the video below.

To contact us with any questions, please email barbarplots@gmail.com.



Copyright © #barbarplots 2016

barbarplots.github.io
youtu.be/LFDbqw2xPhQ



cedricscherer.com



@CedScherer



@z3tt



@cedscherer

Dear Editor

Data visualization is a complex topic in the experimental sciences. While there are many ways to display data, many researchers choose to use bar plots. Generally, these plots only depict a group mean and standard error (or deviation). Unfortunately, most data are not as clean as bar plots make them seem, and since bar plots reveal very little about the distribution of the data, this kind of visualization can be misleading [1,2,3]. A further issue is that of the bar itself, which implies that the base of the y-axis is meaningful, which is not necessarily the case. The bar can then mislead readers [3].

We are a group of young scientists who recently started an initiative called "#barbarplots" aimed at raising awareness and improving scientific communication. As part of this campaign, 169 individuals pledged over 3,400€ in a Kickstarter project to send this package to editors of top journals. Journal editors like you have the power to greatly influence trends in the field. We would like to suggest that you begin a conversation with your editorial board about your journal's stance on data visualization, and whether you wish to encourage authors to use more informative techniques to plot distributions of data. Many high impact journals, such as *Nature Neuroscience*, the *Journal of Neuroscience*, and *PLoS Biology*, have already taken the step to #barbarplots in an effort to make articles more transparent.

To better spread the word, we wish to encourage your participation in our social media campaign. A t-shirt is enclosed in this package: put it on, take a selfie, and share it on Twitter or Facebook with the hashtag #barbarplots.

We hope that this initiative will encourage discussions between colleagues on the merits of various data visualization techniques and on transparency in science in general.

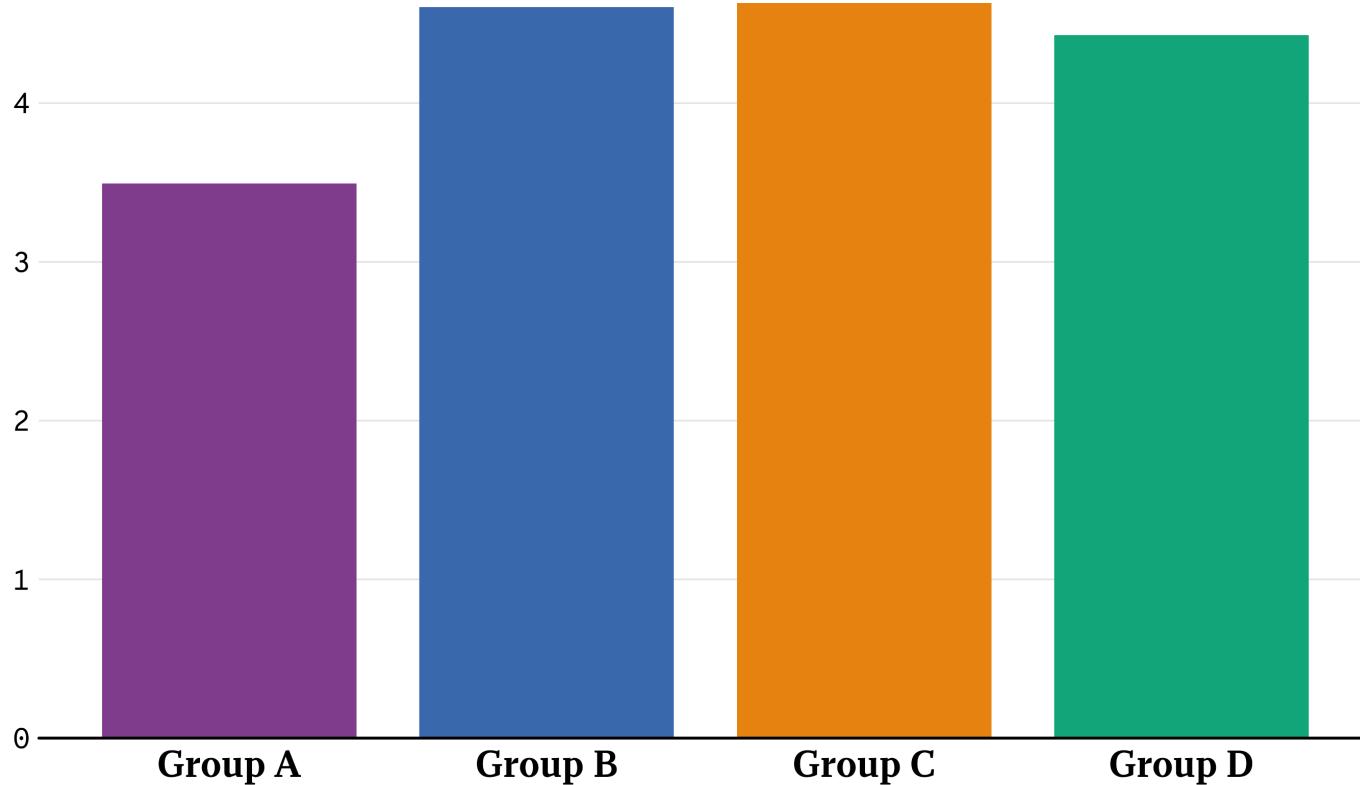
References

1. Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLoS Biology*, 13(4), 1–10. doi:10.1371/journal.pbio.1002128
2. Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., & Milic, N. M. (2016). From Static to Interactive: Transforming Data Visualization to Improve Transparency. *PLoS Biology*, 14(6), e1002484. doi:10.1371/journal.pbio.1002484
3. Saxon, E. (2015). Beyond bar charts. *BMC Biology*, 13(1), 1–2. doi:10.1186/s12915-015-0169-6



BAR PLOT

geom_bar(fun = "summary")



CODES

z3tt.github.io/beyond-bar-and-box-plots

GGPLOT2 SETUP

```
## general theme
theme_set(theme_void(base_family = "Roboto"))

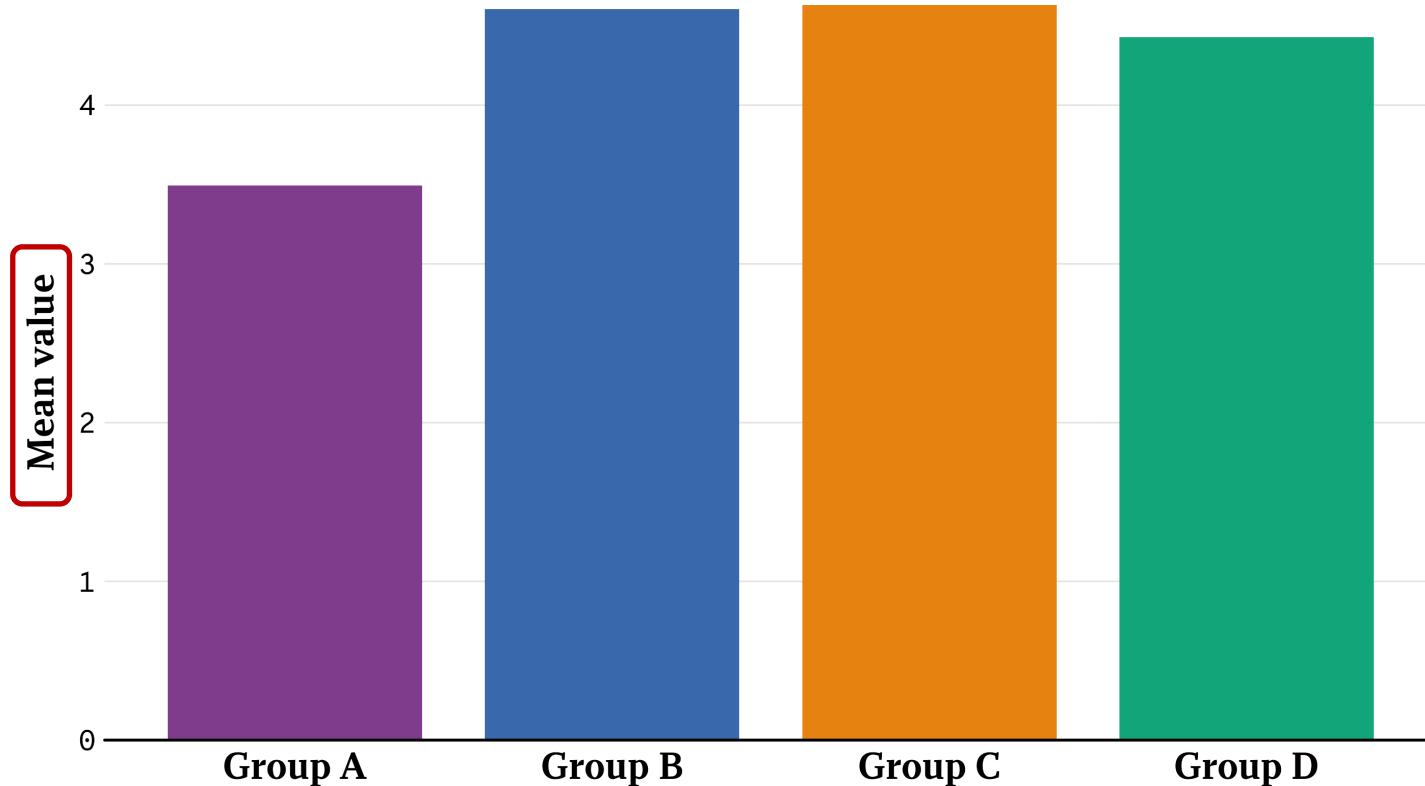
theme_update(
  axis.text.x = element_text(color = "black", face = "bold", size = 26,
                               margin = margin(t = 6)),
  axis.text.y = element_text(color = "black", size = 22, hjust = 1,
                               margin = margin(r = 6), family = "Roboto Mono"),
  panel.grid.major.y = element_line(color = "grey90", size = .6),
  axis.line.x = element_line(color = "black", size = 1),
  plot.margin = margin(rep(20, 4))
)

## custom colors
my_pal <- rcartocolor::carto_pal(n = 8, name = "Bold")[c(1, 3, 7, 2)]
```



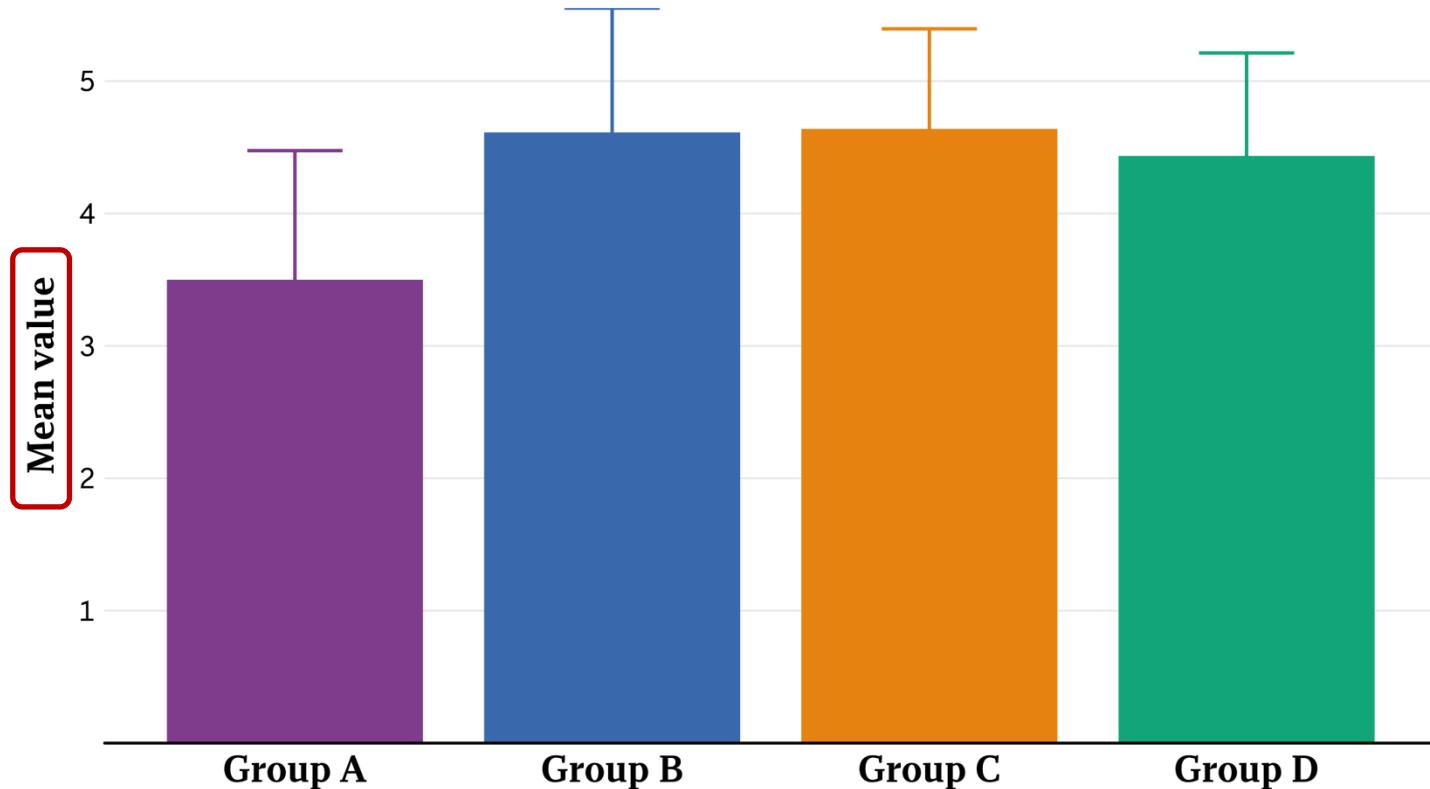
~~BAR PLOT~~

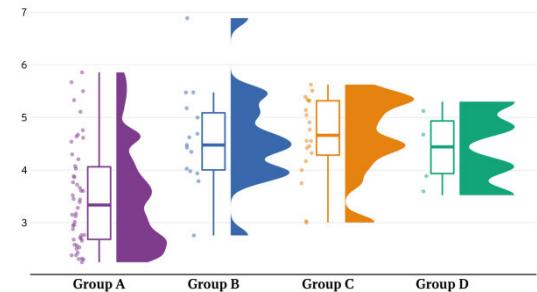
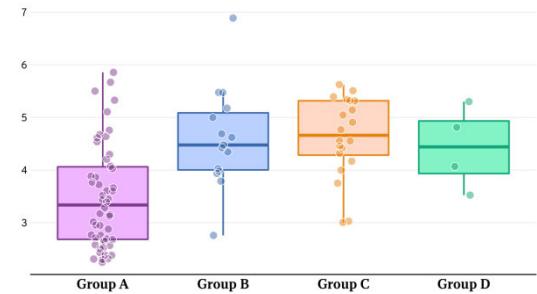
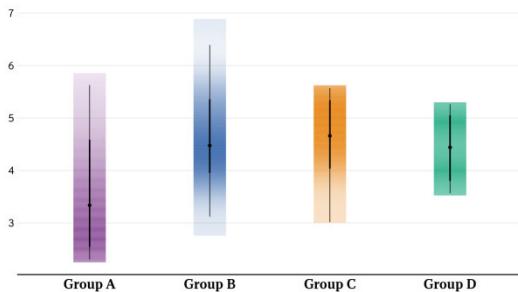
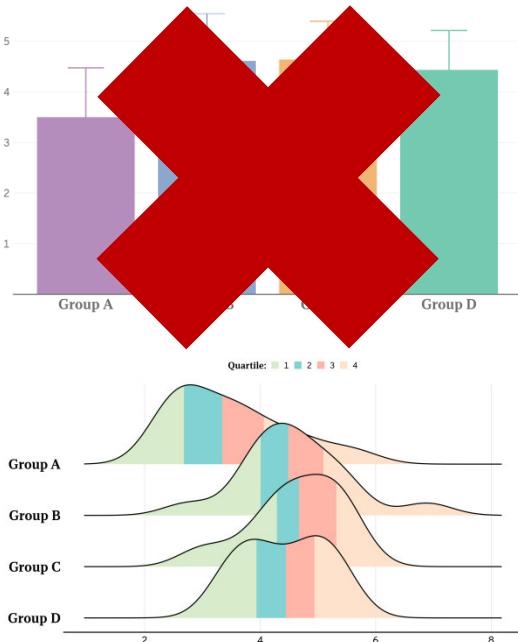
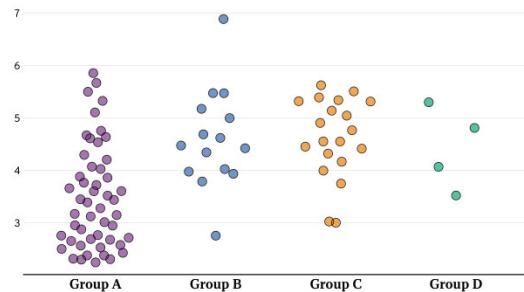
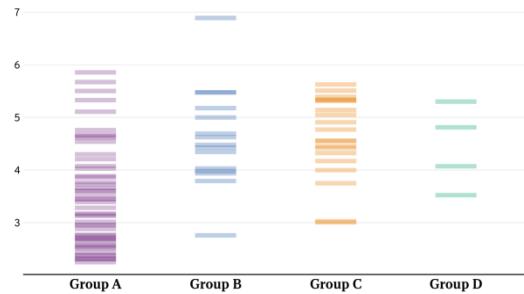
geom_bar(stat = "summary")



~~DYNAMITE PLOT~~

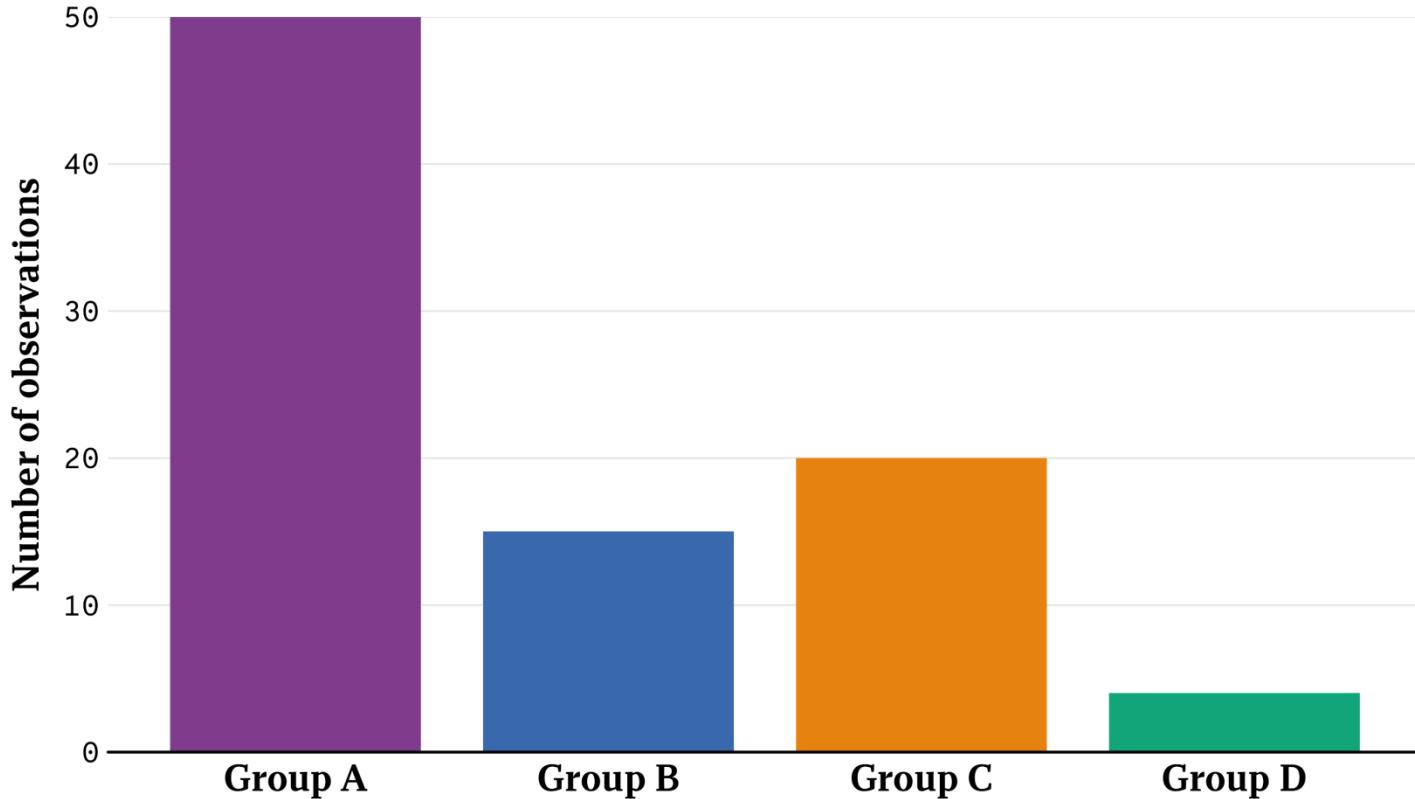
`geom_bar(stat = "summary") + stat_summary(geom = "errorbar")`



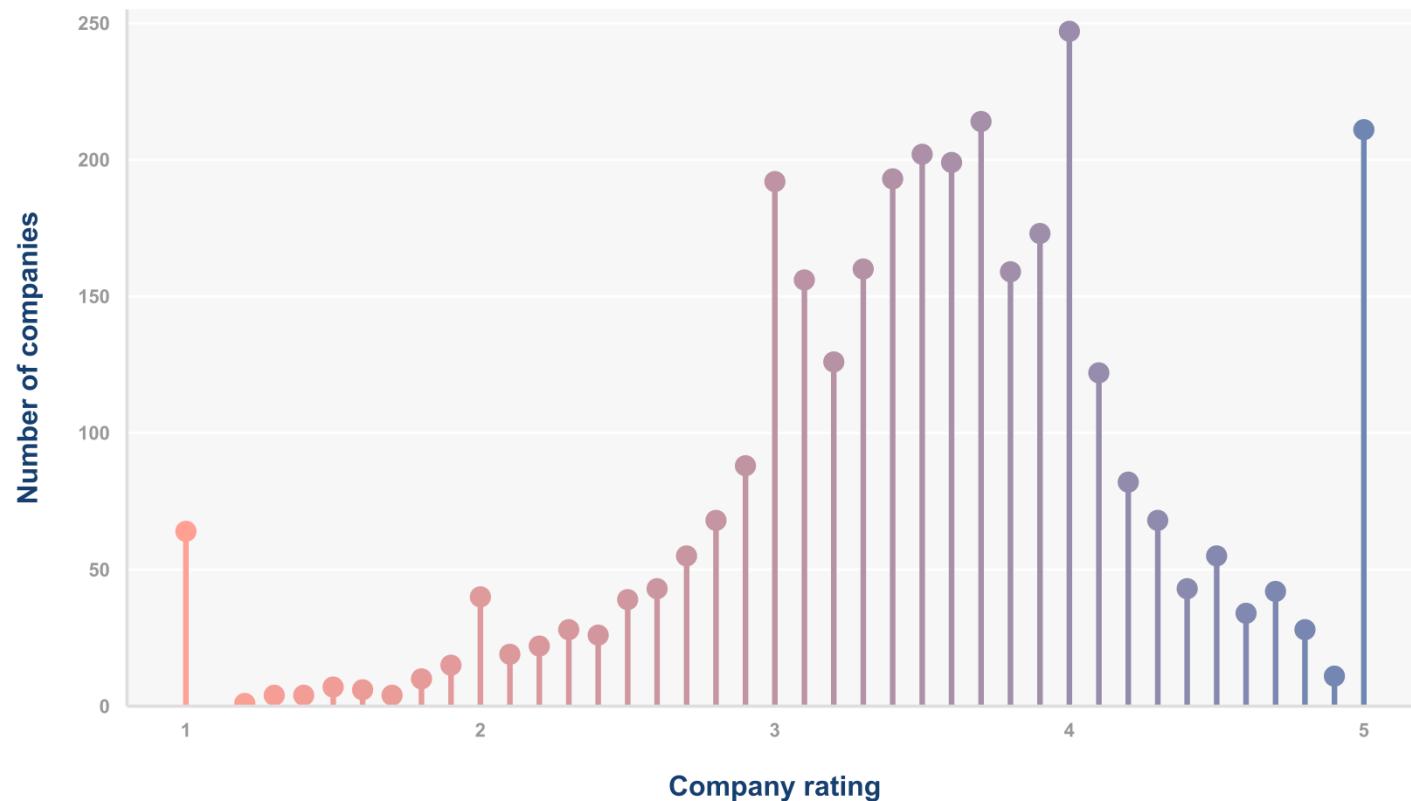


BAR PLOT ✓

geom_bar(stat = "count")

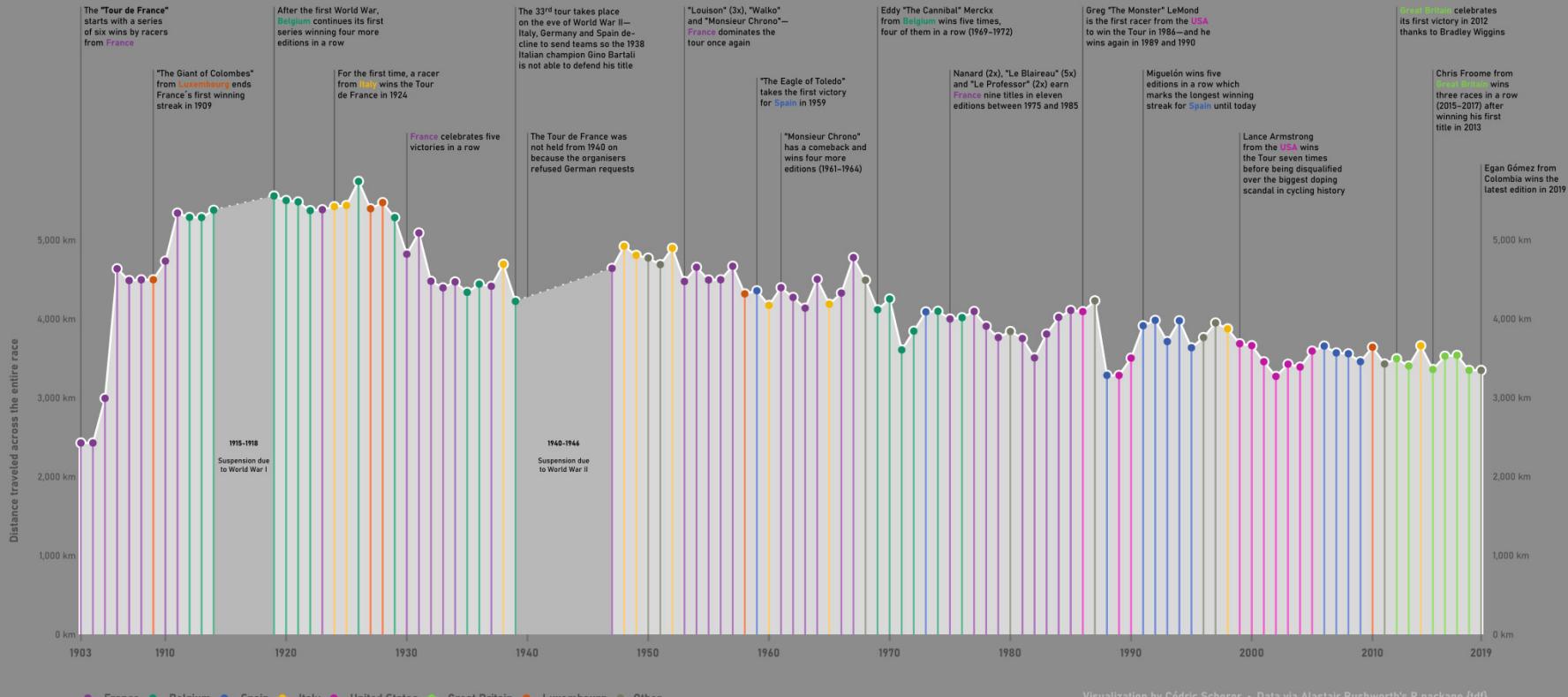


Do more low- or high-rated companies post e-commerce jobs?



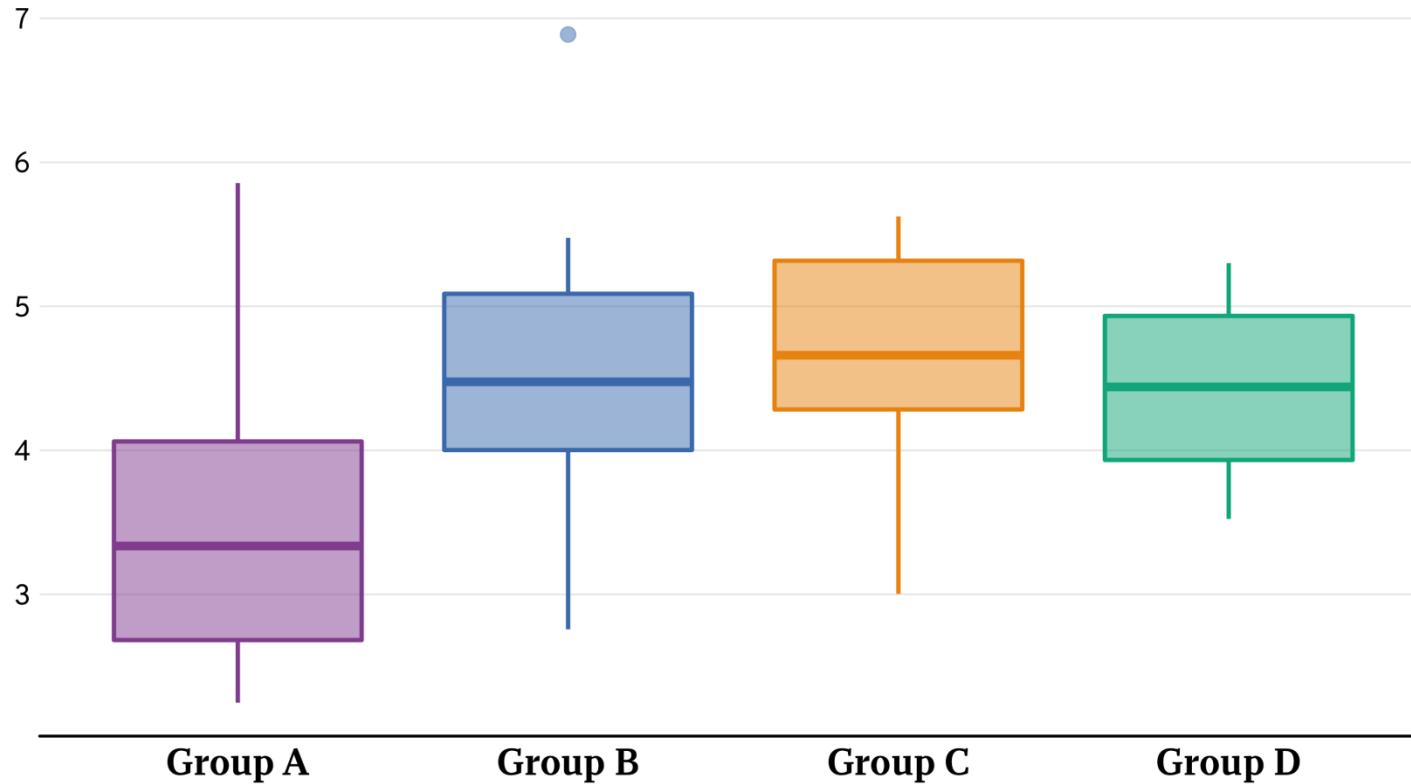
Source: Clerk.io

L'histoire du Tour de France de 1903 à 2019



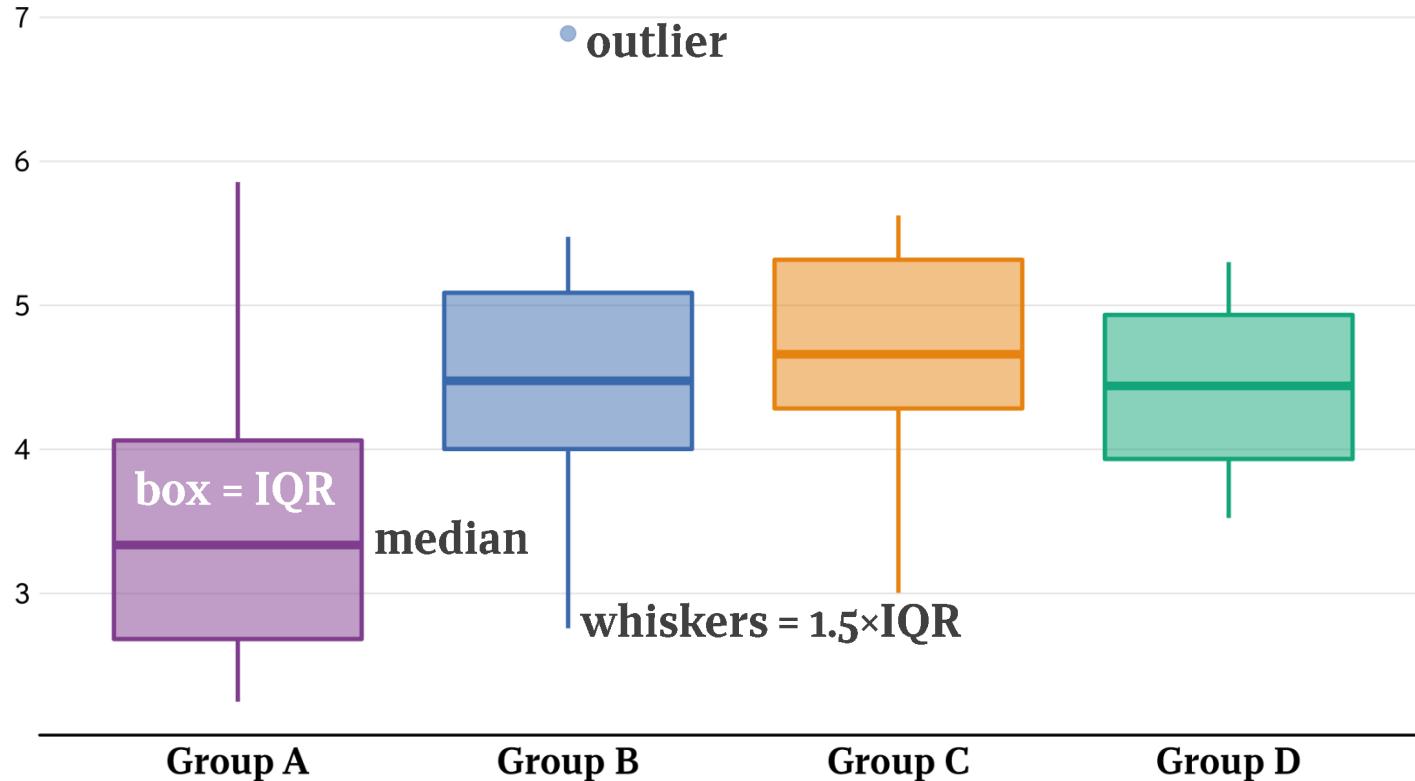
BOX-AND-WHISKERS PLOT

geom_boxplot()



BOX-AND-WHISKERS PLOT

geom_boxplot()



BOX-AND-WHISKERS PLOT

geom_boxplot()

```
g <- ggplot(data, aes(x = group, y = value, color = group, fill = group)) +  
  scale_y_continuous(breaks = 1:9) +  
  scale_color_manual(values = my_pal, guide = "none") +  
  scale_fill_manual(values = my_pal, guide = "none")
```



BOX-AND-WHISKERS PLOT

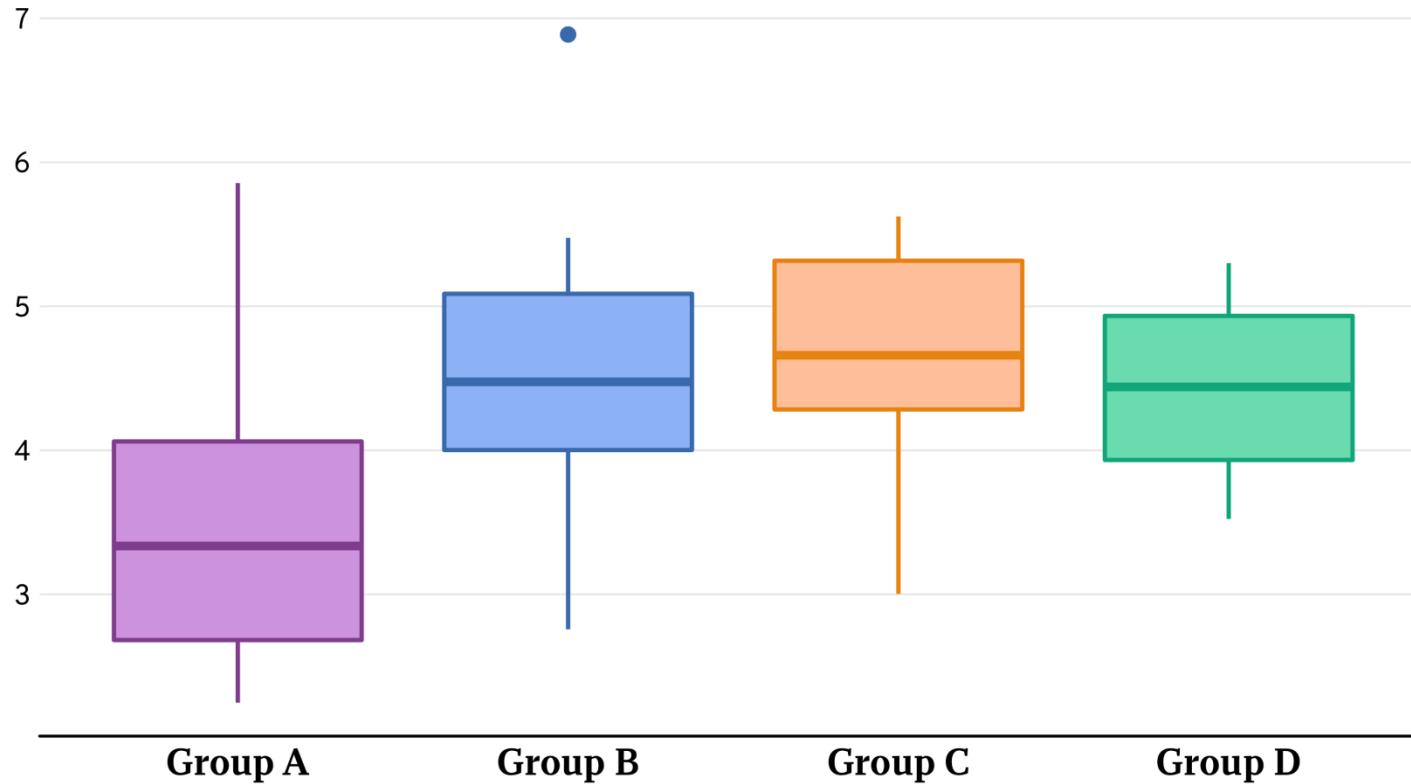
geom_boxplot()

```
g <- ggplot(data, aes(x = group, y = value, color = group, fill = group)) +  
  scale_y_continuous(breaks = 1:9) +  
  scale_color_manual(values = my_pal, guide = "none") +  
  scale_fill_manual(values = my_pal, guide = "none")  
  
g +  
  geom_boxplot(  
    aes(fill = group), alpha = .5,  
    size = 1.5, outlier.size = 5  
  )
```



BOX-AND-WHISKERS PLOT

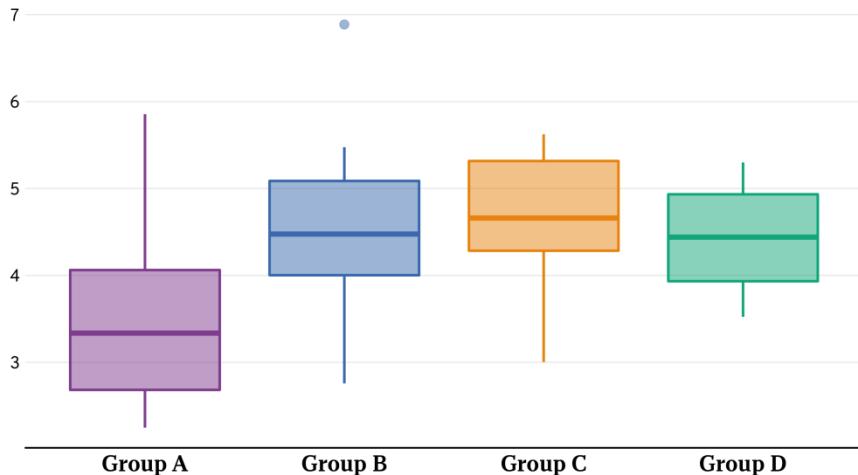
geom_boxplot()



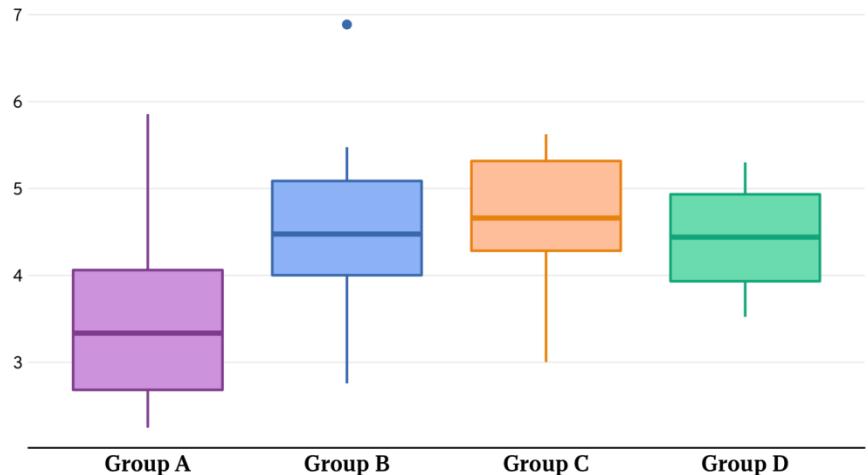
BOX-AND-WHISKERS PLOT

geom_boxplot()

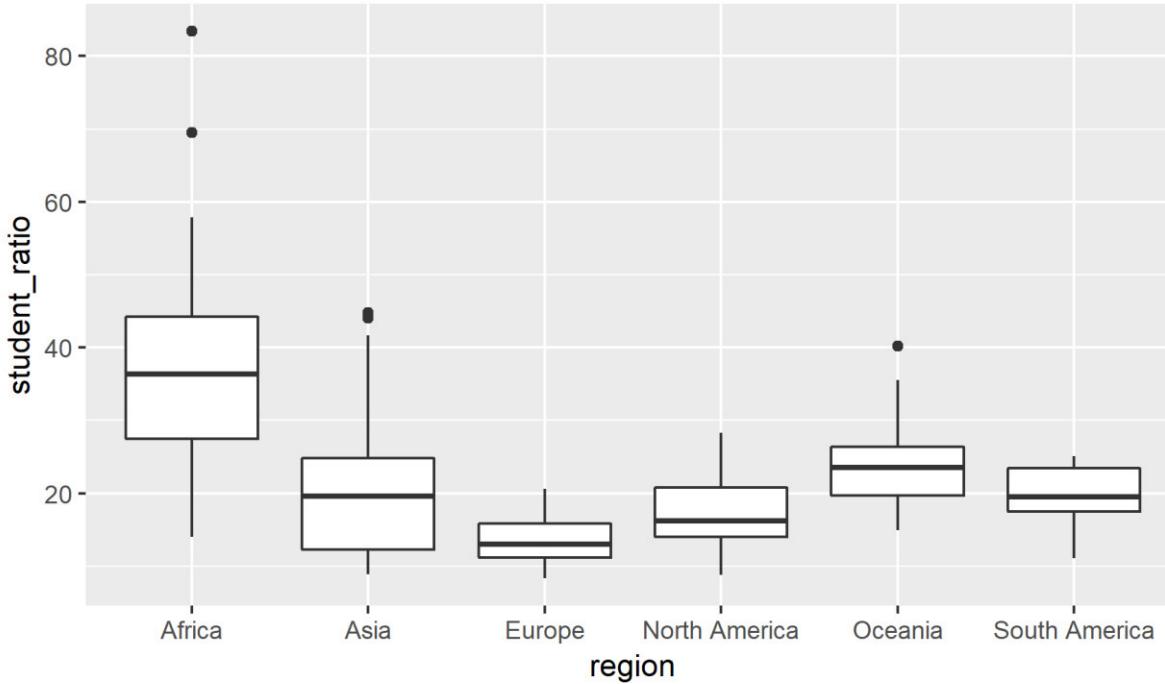
```
aes(fill = group),  
alpha = .5
```



```
aes(fill = group,  
fill = after_scale(colorspace::lighten(fill, .5)))
```



The Evolution of a ggplot



Data: UNESCO Institute for Statistics
Visualization by Cédric Scherer

["The Evolution of a ggplot \(Ep. 1\)"](#)



cedricscherer.com



@CedScherer



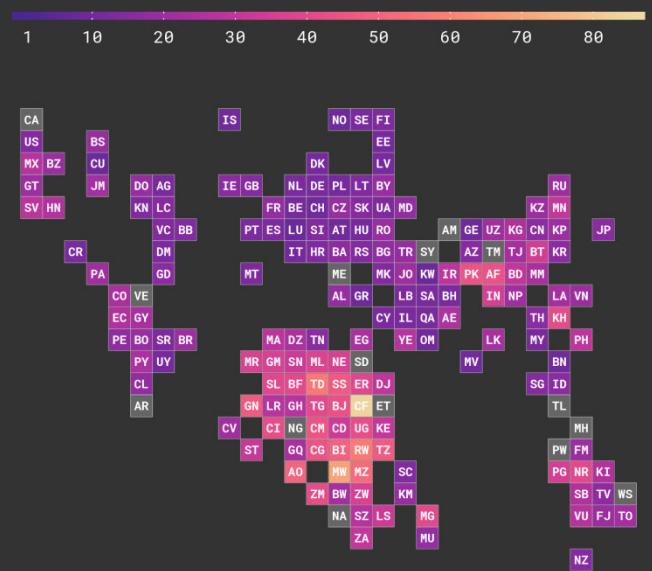
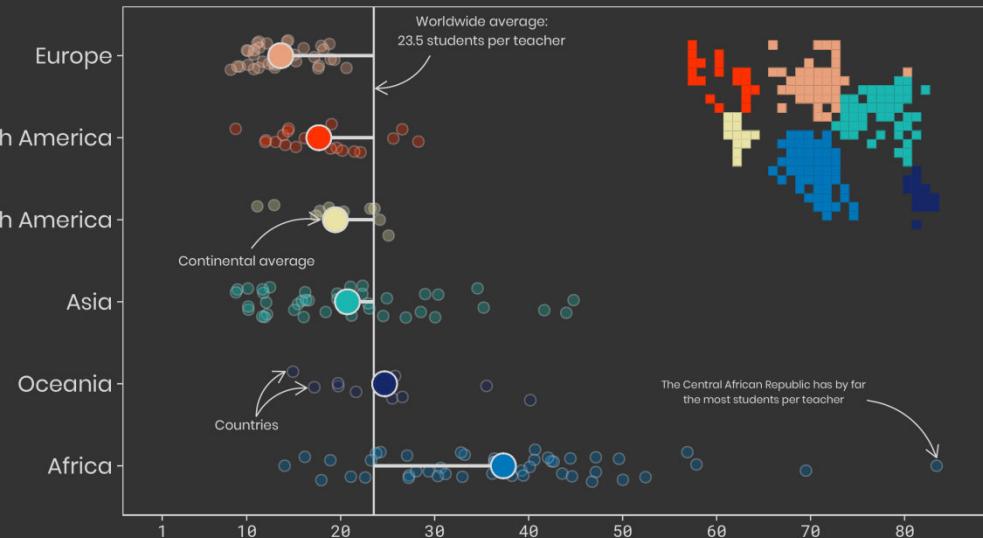
@z3tt



@cedscherer

Global student to teacher ratios in primary education

Latest reported student to teacher ratio per country and continent (2012–2018)

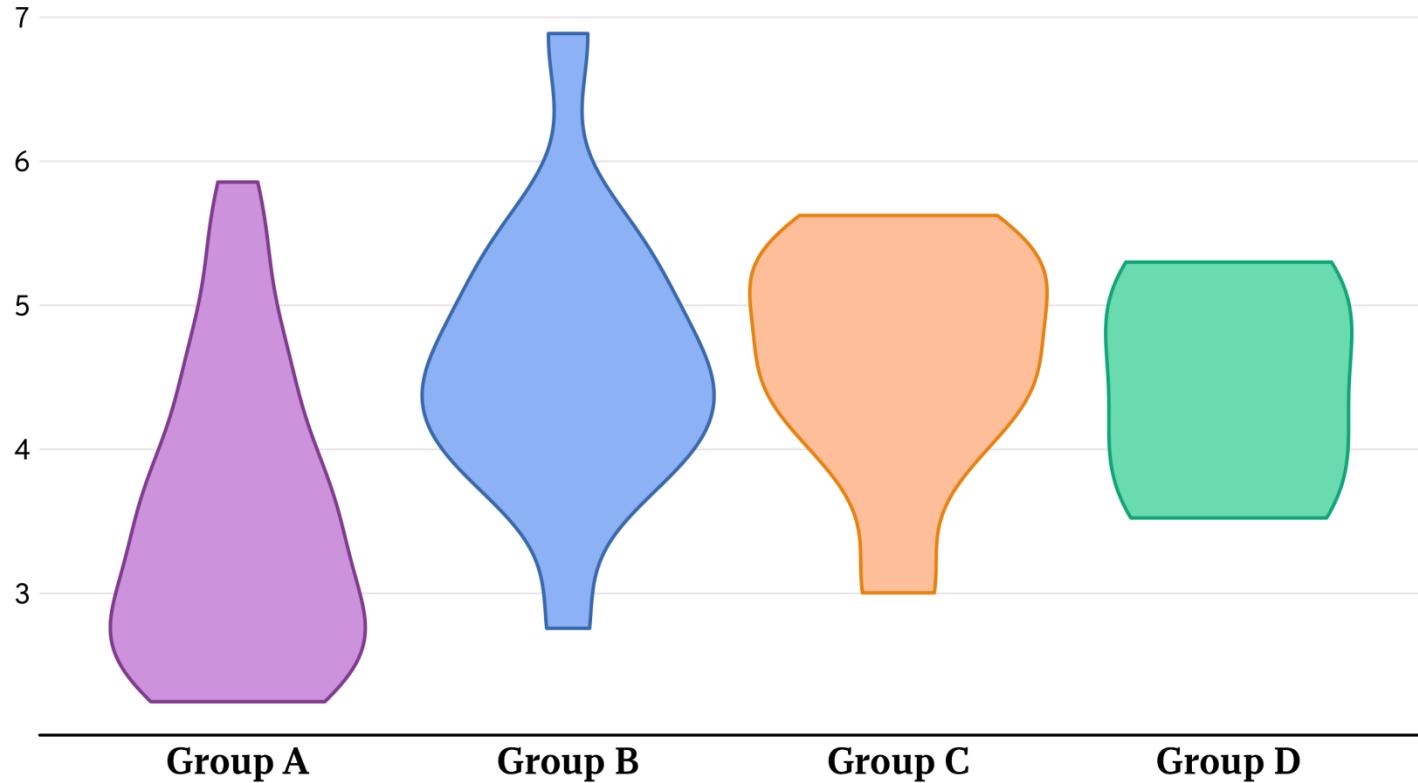


Visualization by Cédric Scherer | Data: "eAtlas of Teachers" by UNESCO

ALTERNATIVE CHART TYPES TO VISUALIZE *Distributions*

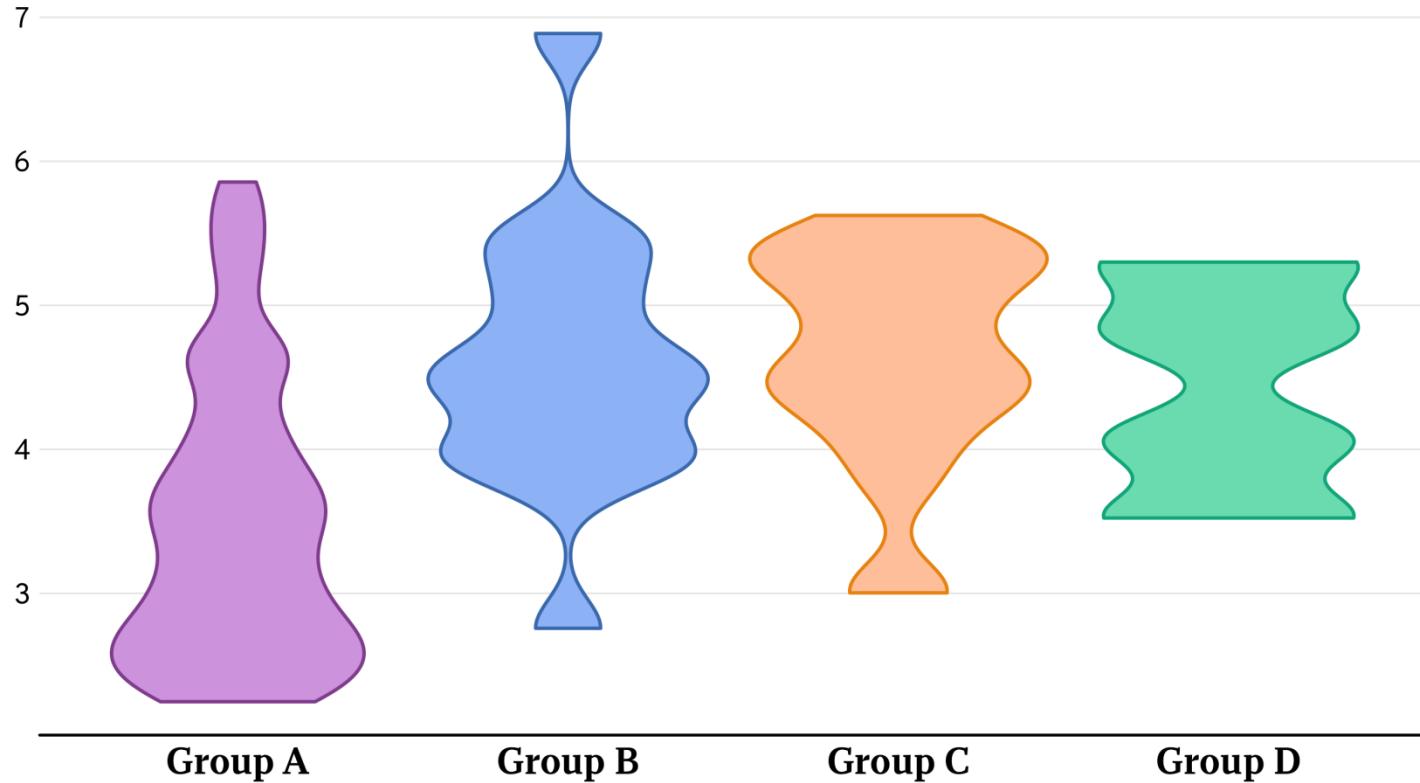
VIOLIN PLOTS

geom_violin()



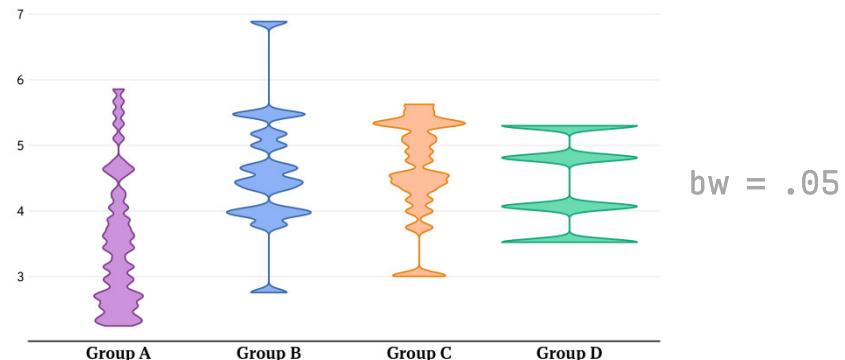
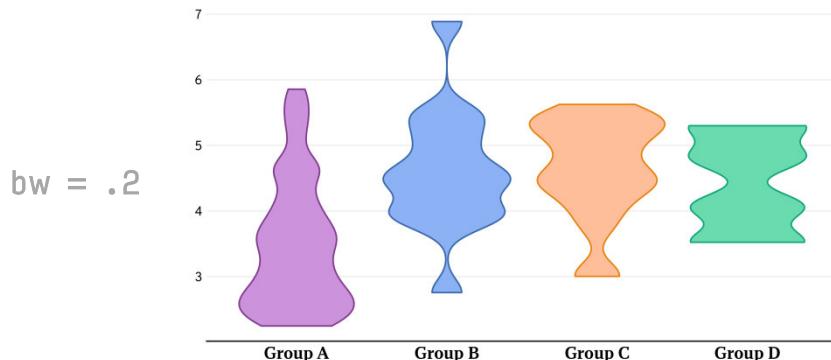
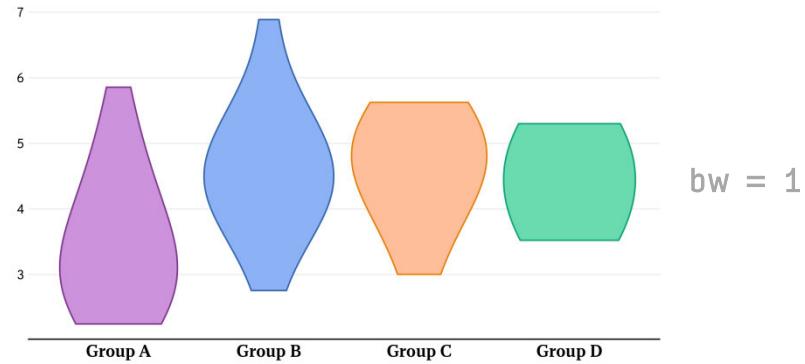
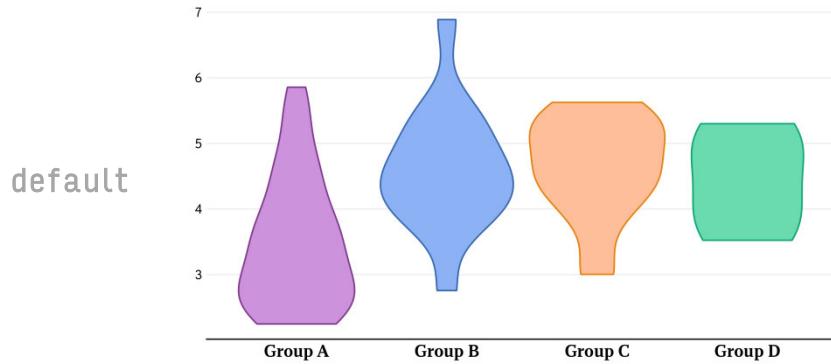
VIOLIN PLOTS

geom_violin(bw = .2)



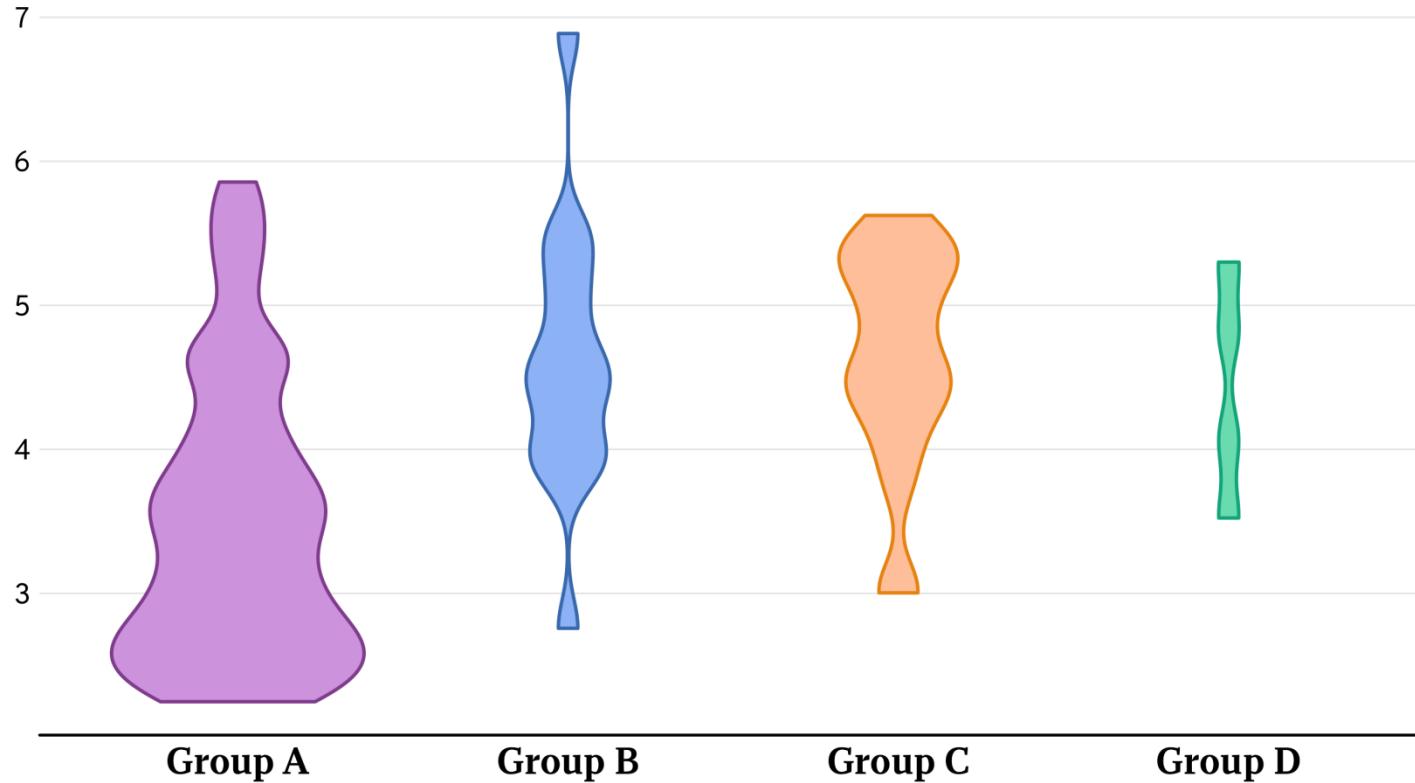
VIOLIN PLOTS

geom_violin()

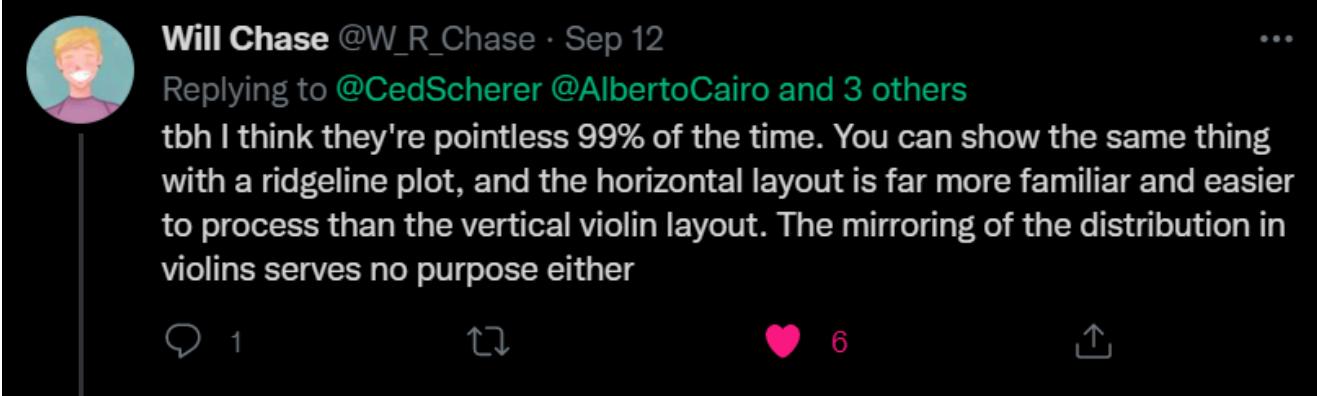


VIOLIN PLOT (SCALED BY COUNTS)

geom_violin(scale = "count")



VIOLIN PLOTS IN THE WILD?



Will Chase @W_R_Chase · Sep 12

Replying to @CedScherer @AlbertoCairo and 3 others

tbh I think they're pointless 99% of the time. You can show the same thing with a ridgeline plot, and the horizontal layout is far more familiar and easier to process than the vertical violin layout. The mirroring of the distribution in violins serves no purpose either

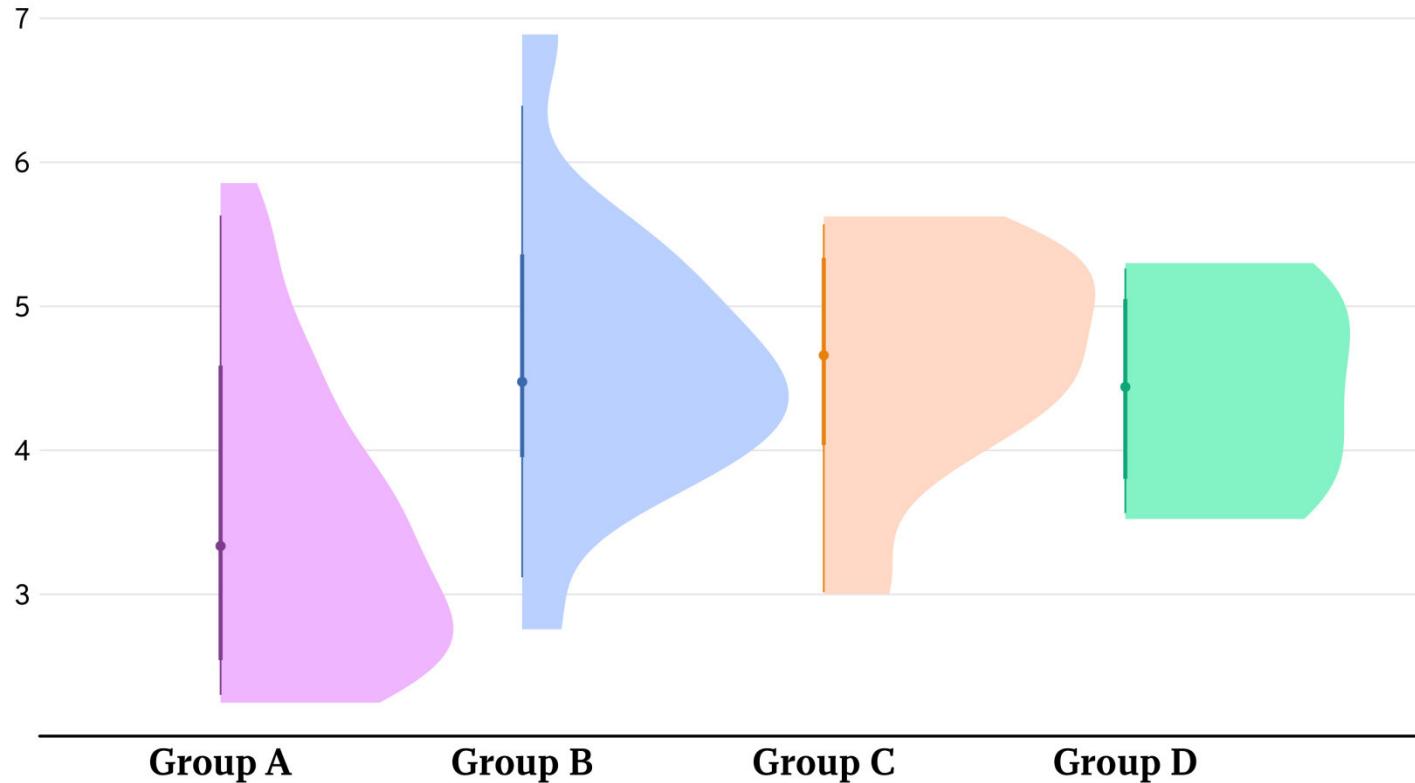
1

6



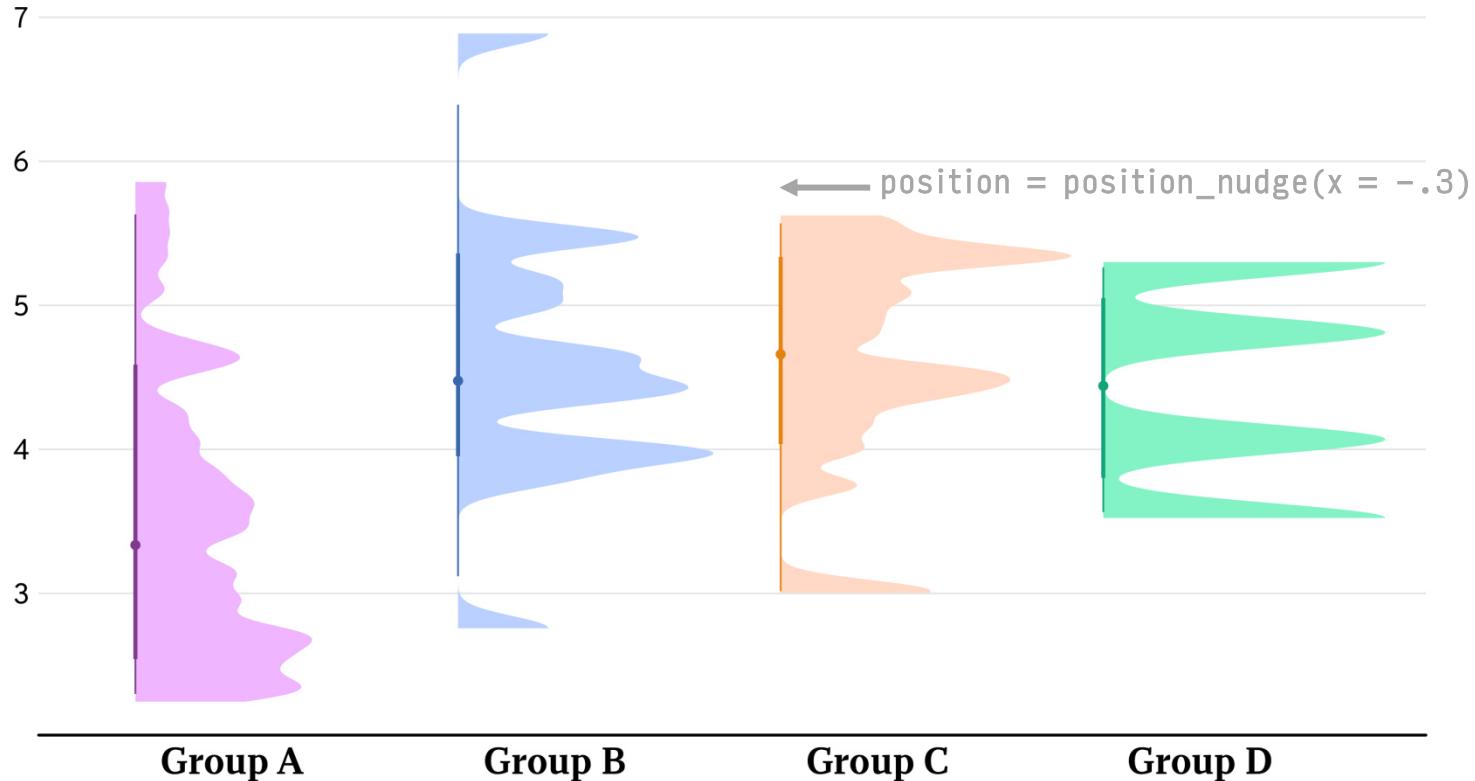
VIOLIN PLOT (HALF-EYES)

ggdist::stat_halfeye()



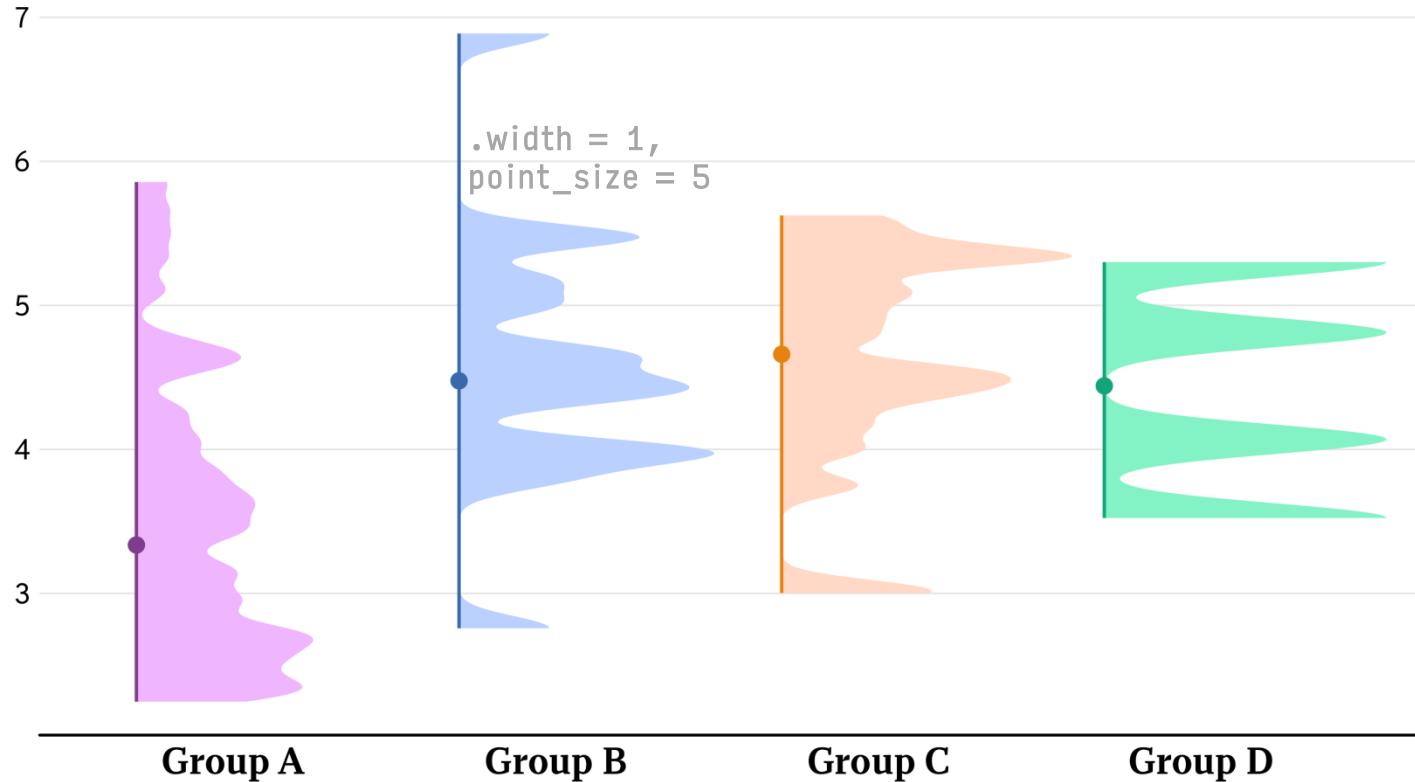
VIOLIN PLOT (HALF-EYES)

`ggdist::stat_halfeye(adjust = .2)`



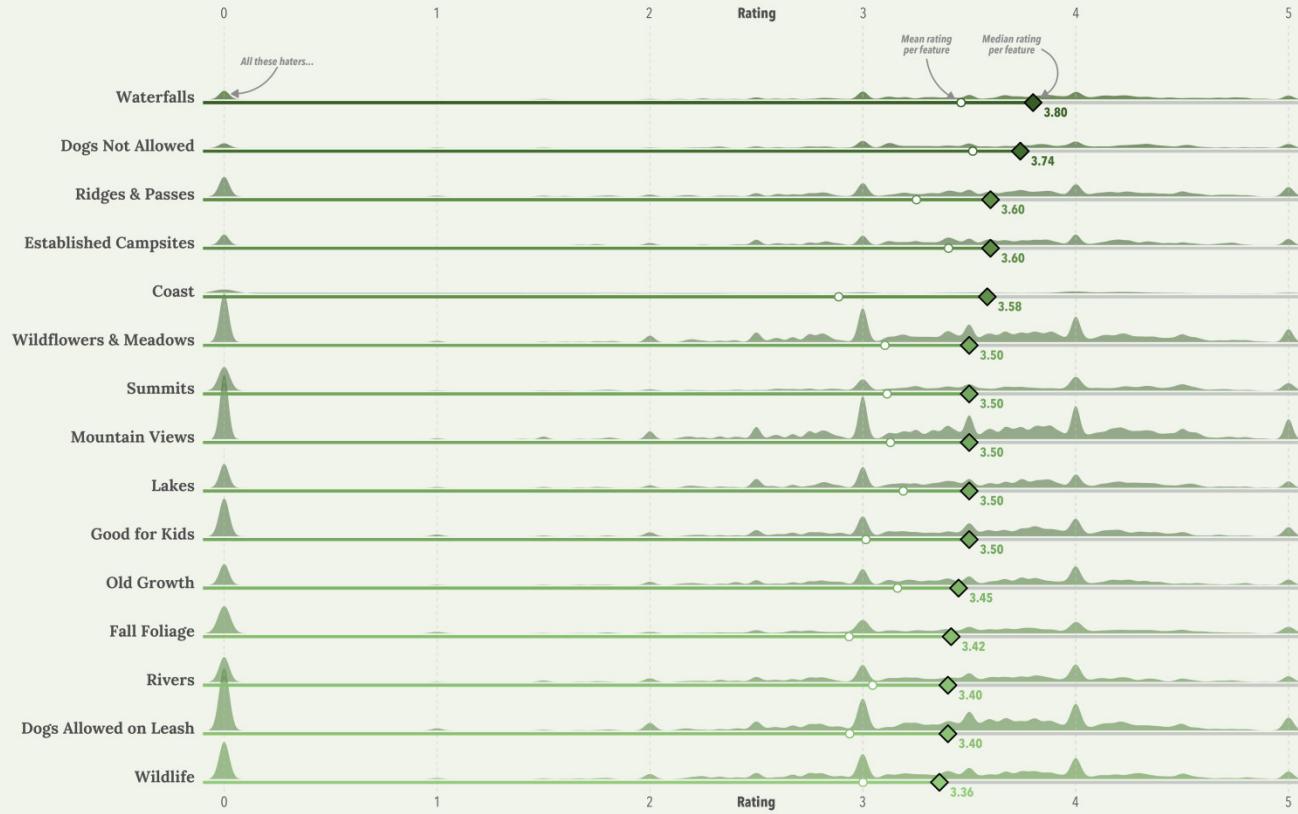
VIOLIN PLOT (HALF-EYES)

ggdist::stat_halfeye(adjust = .2)

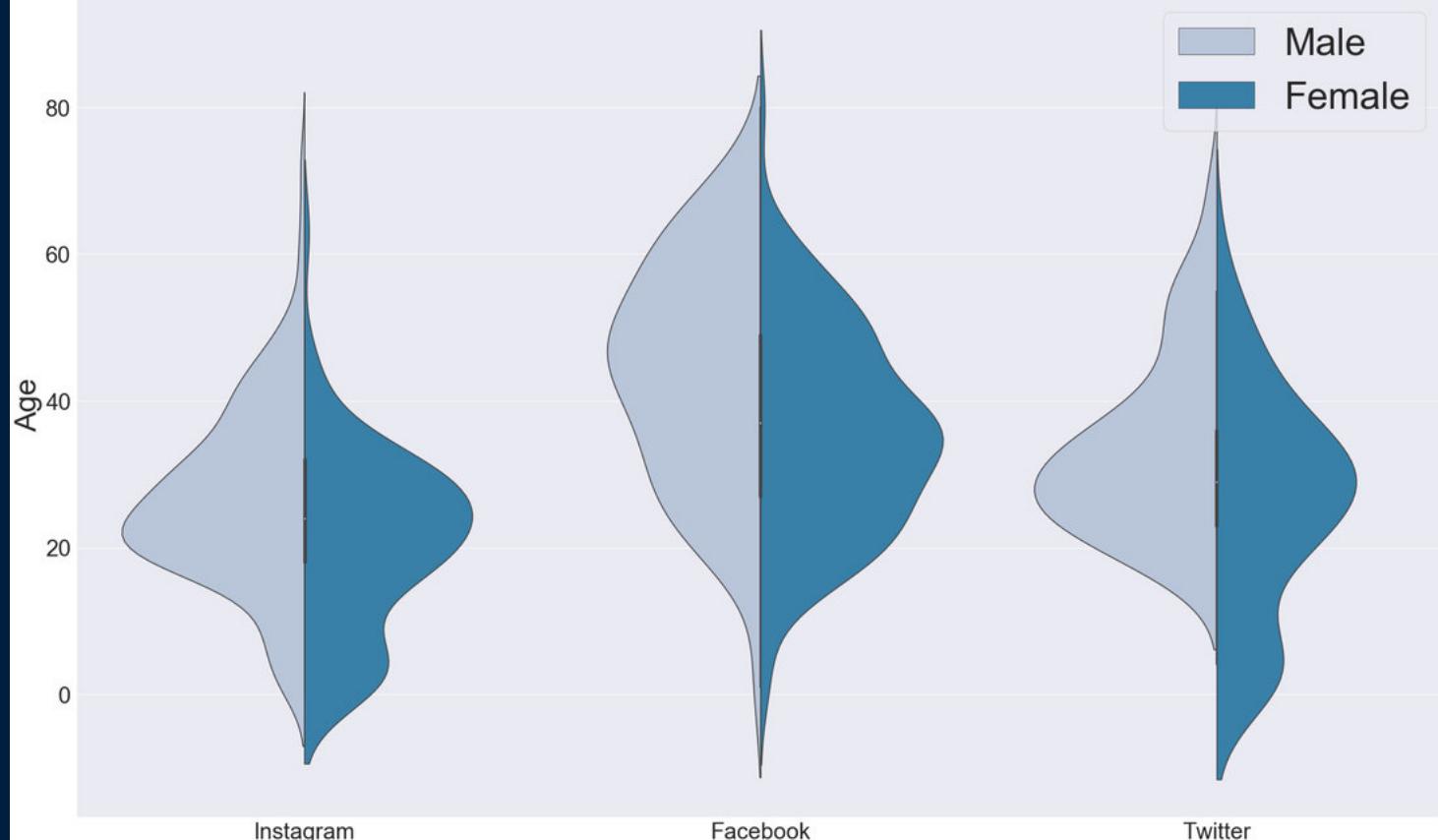


Hiking in Washington: Waterfalls are great—Wildlife and dogs rather not

The Washington Trail Association (WTA) offers a hiking guide that claims top be “*the most comprehensive database of hikes in Washington*”. It comprises content written by local hiking experts and user submitted information such as ratings for each trail and specification of certain features. The plot shows the overall rating of 1,924 Washington trails listed in the WTA database grouped by its features and ordered by the highest median rating.



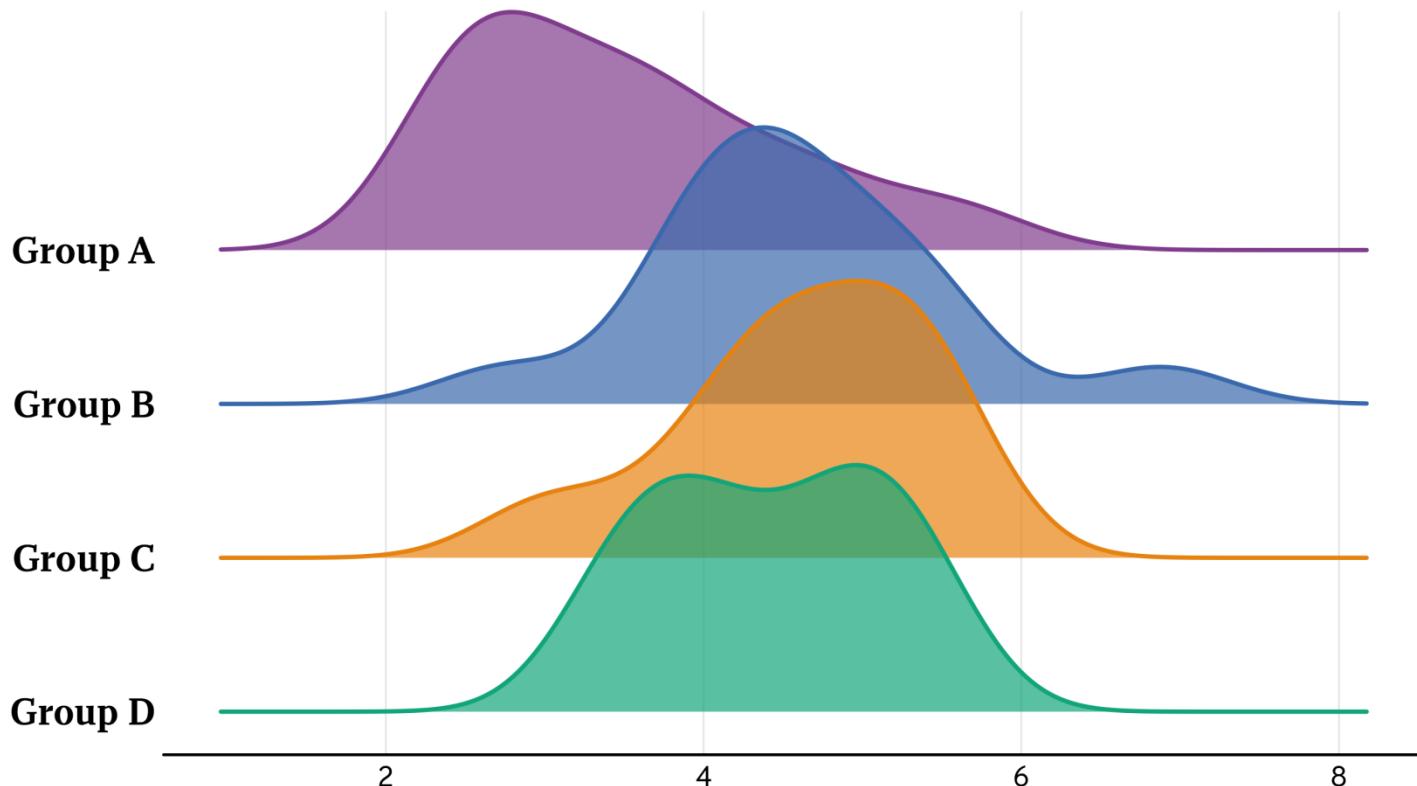
AGE DISTRIBUTION OF SOCIAL MEDIA FOLLOWERS



[“Violin Plots” by LondonSoda](#)

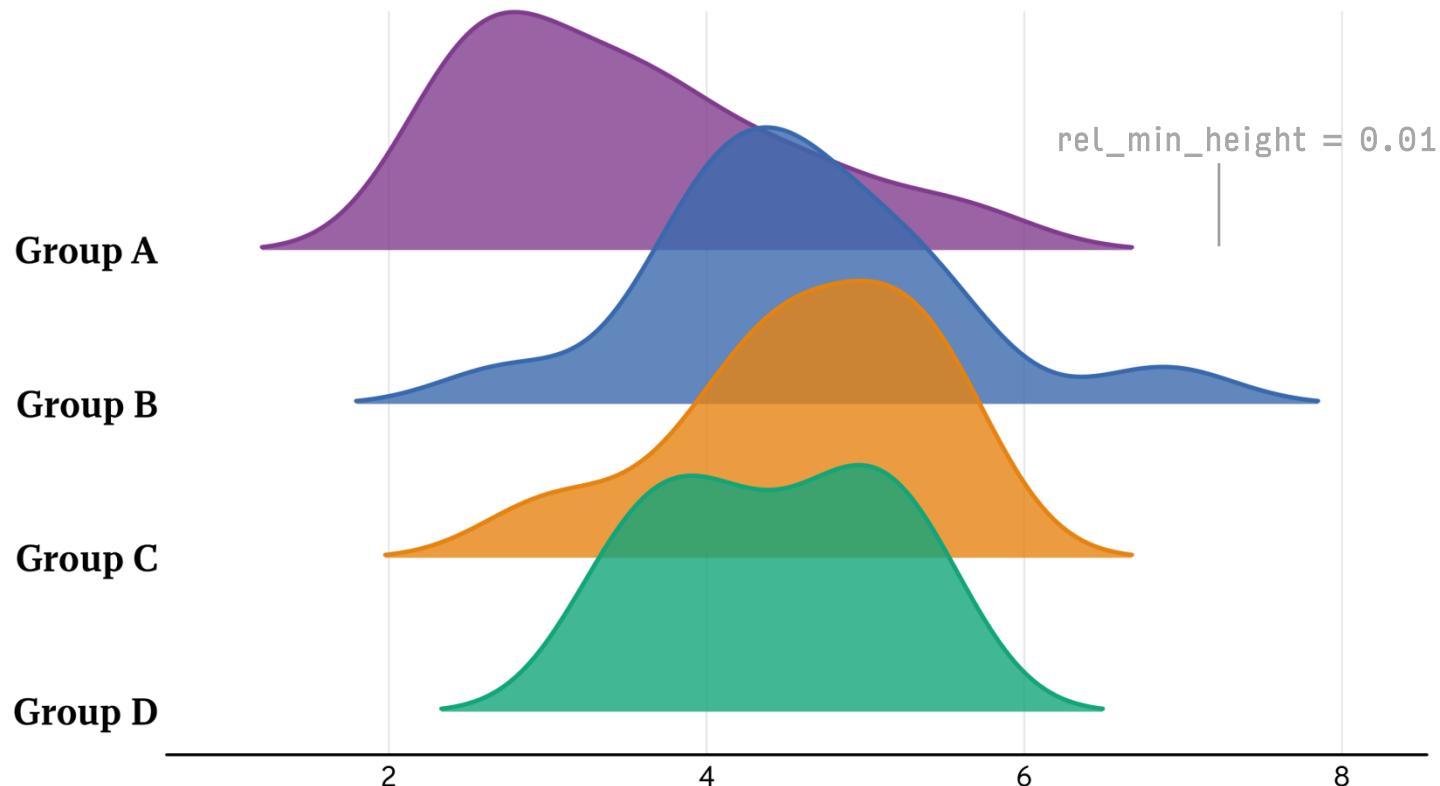
RIDGE LINE PLOT

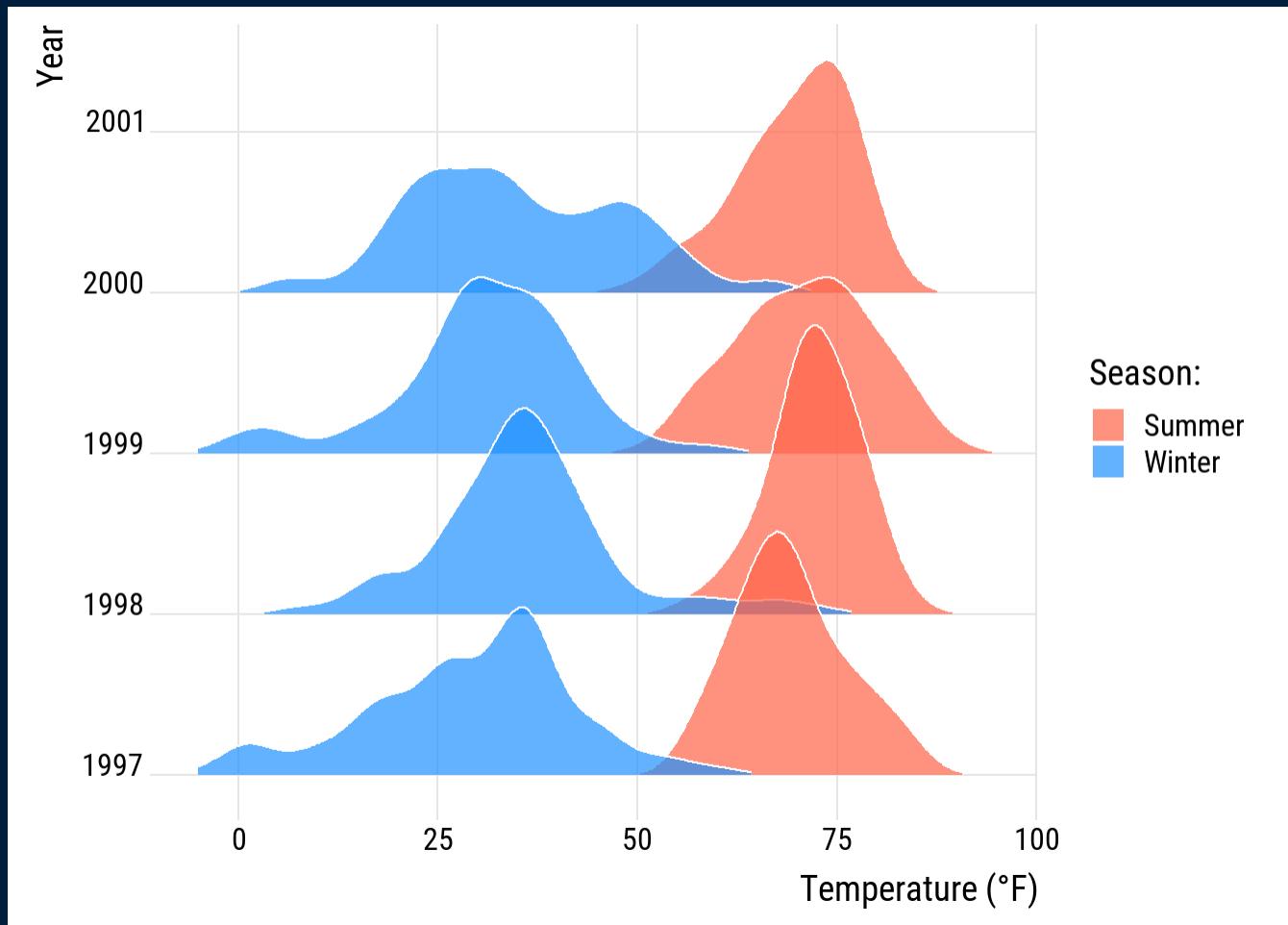
ggridges::geom_density_ridges()



RIDGE LINE PLOT (CUTTED)

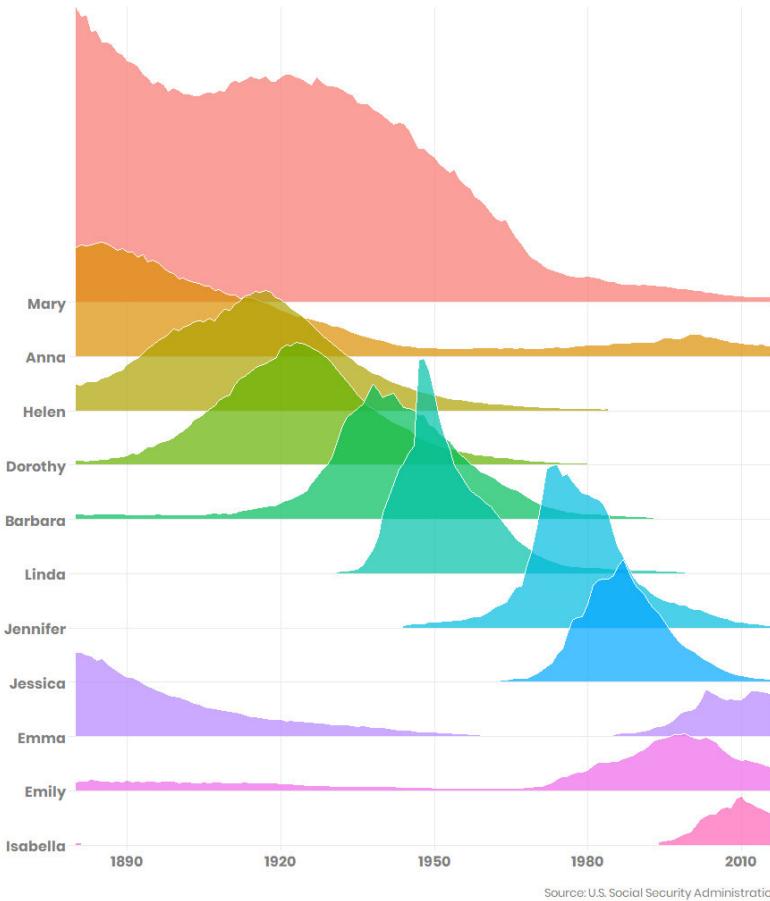
`ggridges::geom_density_ridges()`





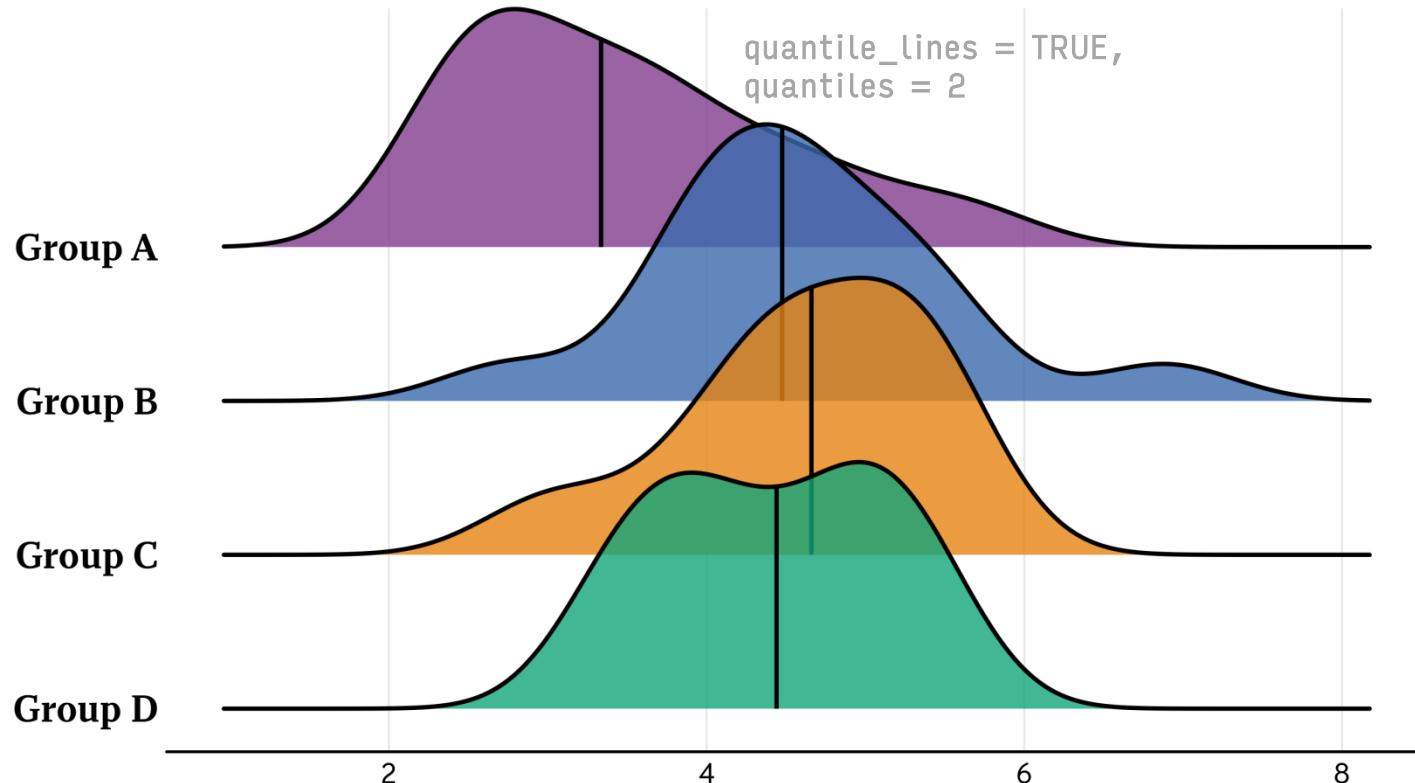
Most popular girl names in the U.S.

Top 2 names with the highest mean and/or maximum per quarter are shown.



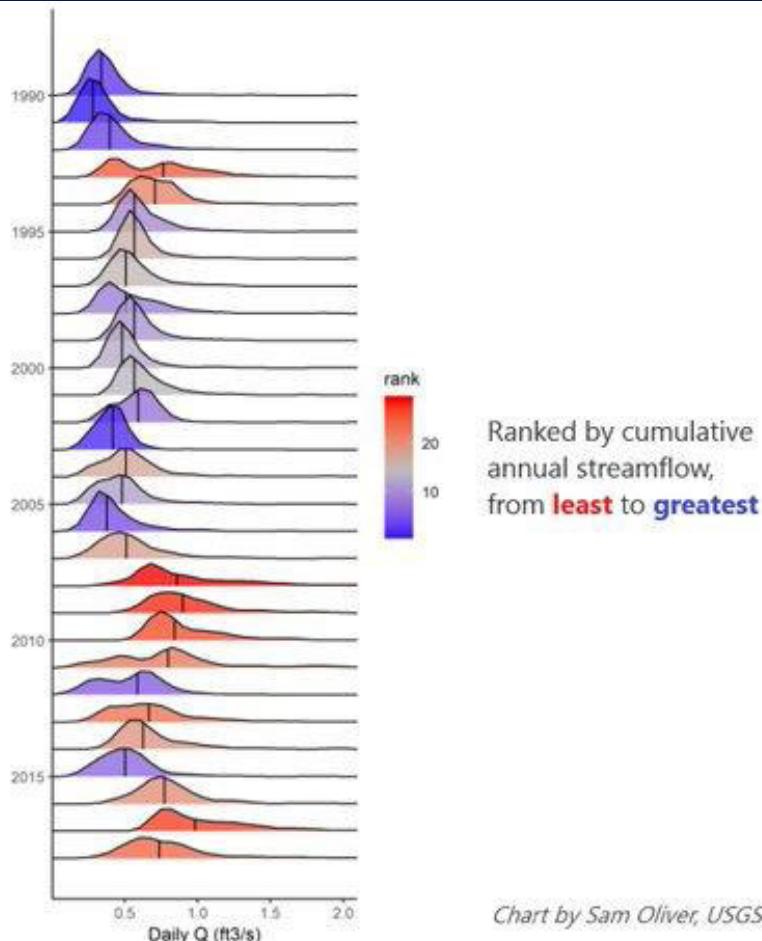
RIDGE LINE PLOT WITH MEDIAN

`ggridges:::stat_density_ridges()`



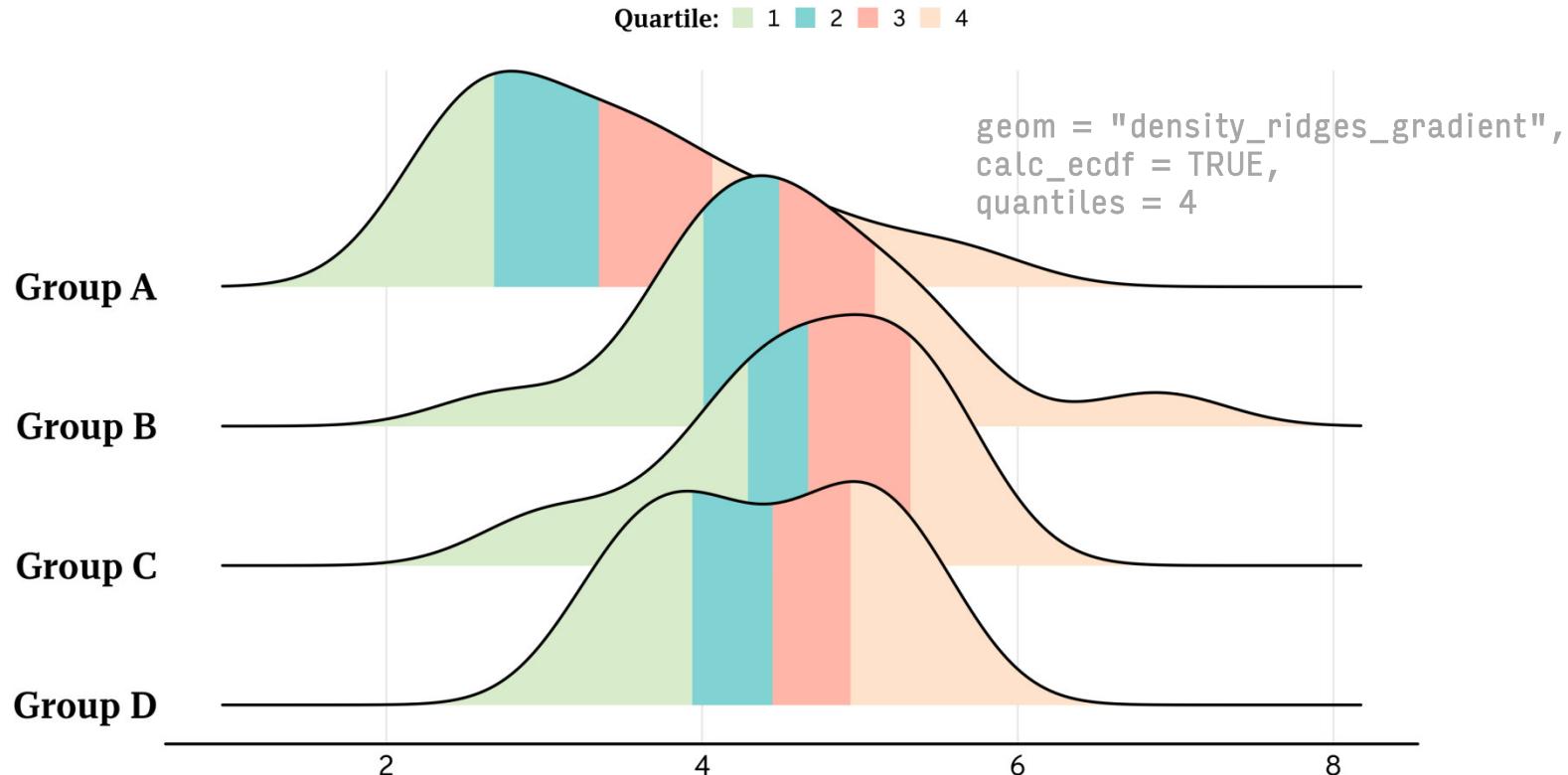
Distributions of daily streamflow on the Yahara River by year

1990 - 2018



RIDGE LINE PLOT WITH QUARTILES

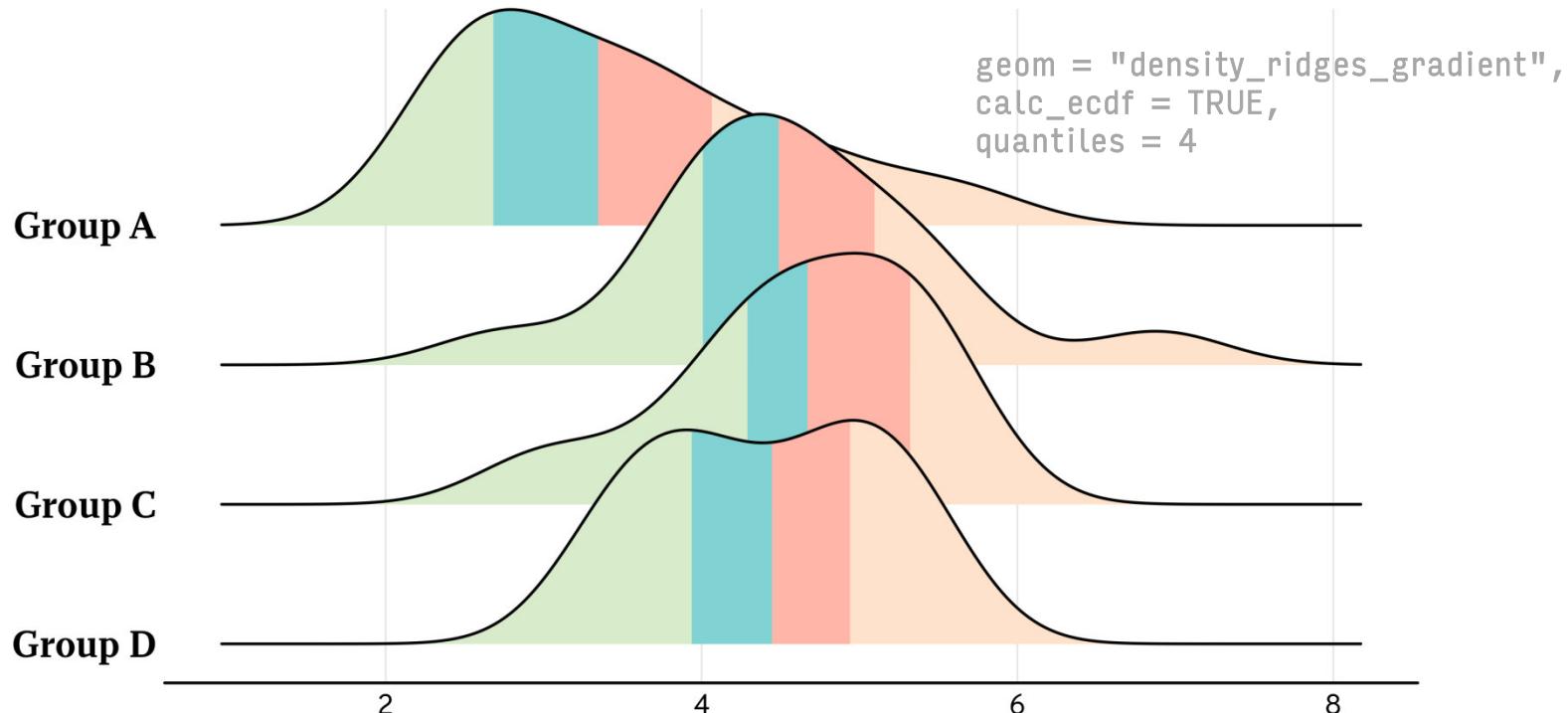
ggridges:::stat_density_ridges()



RIDGE LINE PLOT WITH QUARTILES

ggridges:::stat_density_ridges()

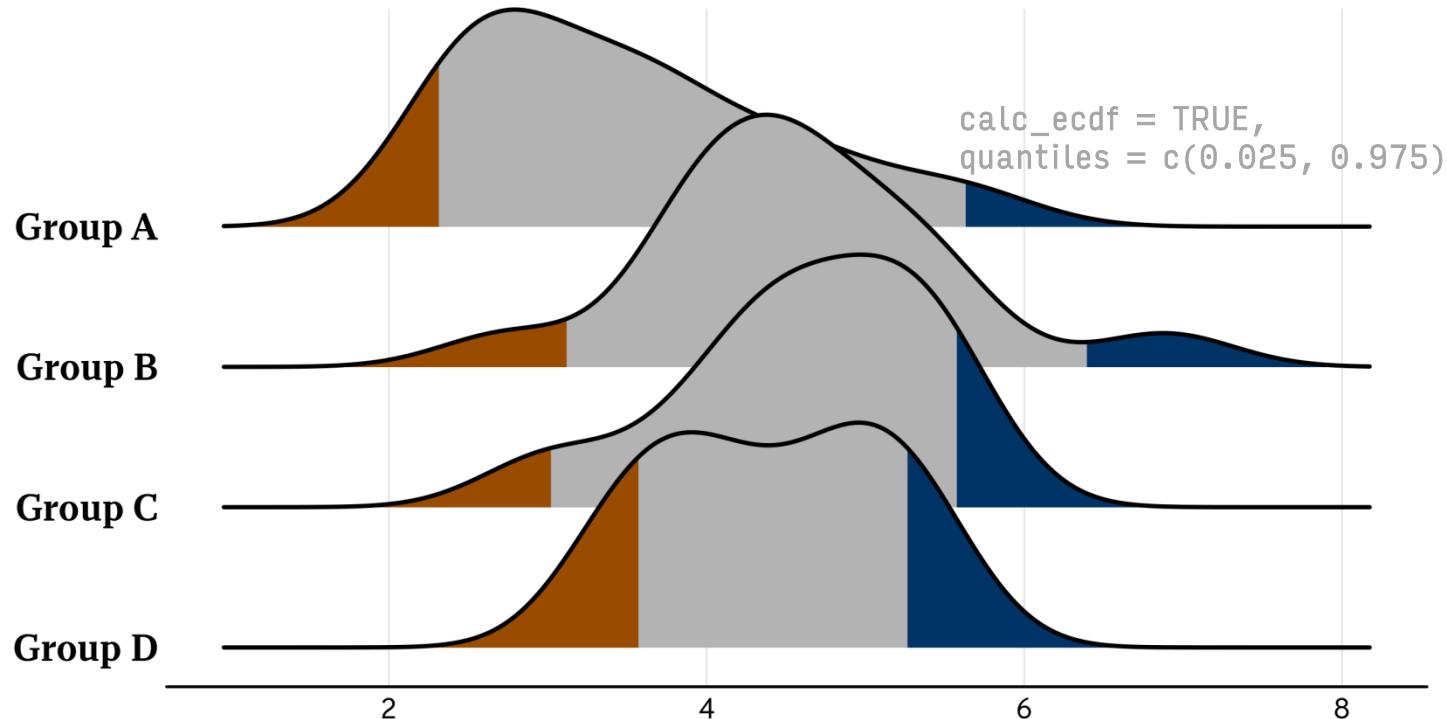
aes(fill = factor(stat(quantile))) Quartile: ■ 1 ■ 2 ■ 3 ■ 4



RIDGE LINE PLOT WITH MEDIAN

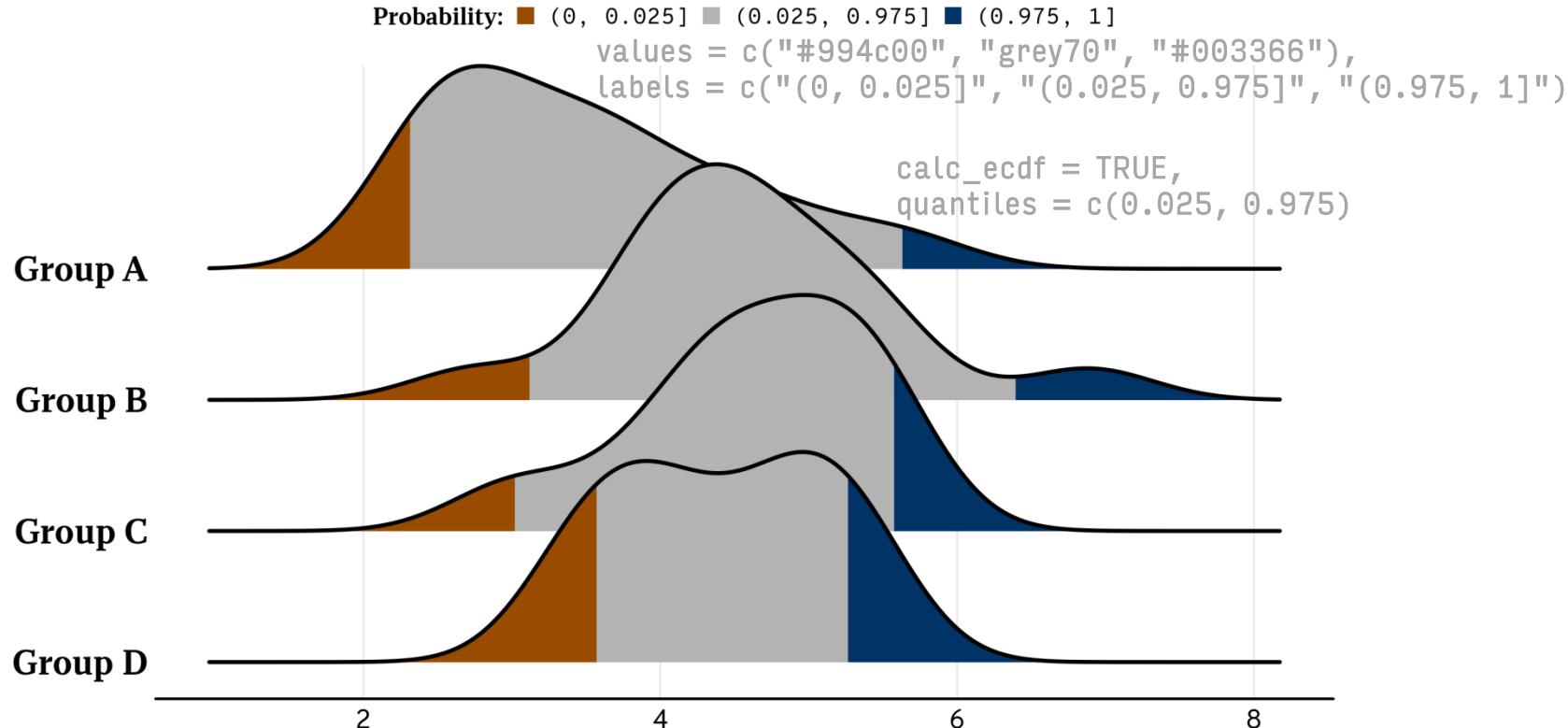
ggridges:::stat_density_ridges()

Probability: ■ (0, 0.025] ■ (0.025, 0.975] ■ (0.975, 1]



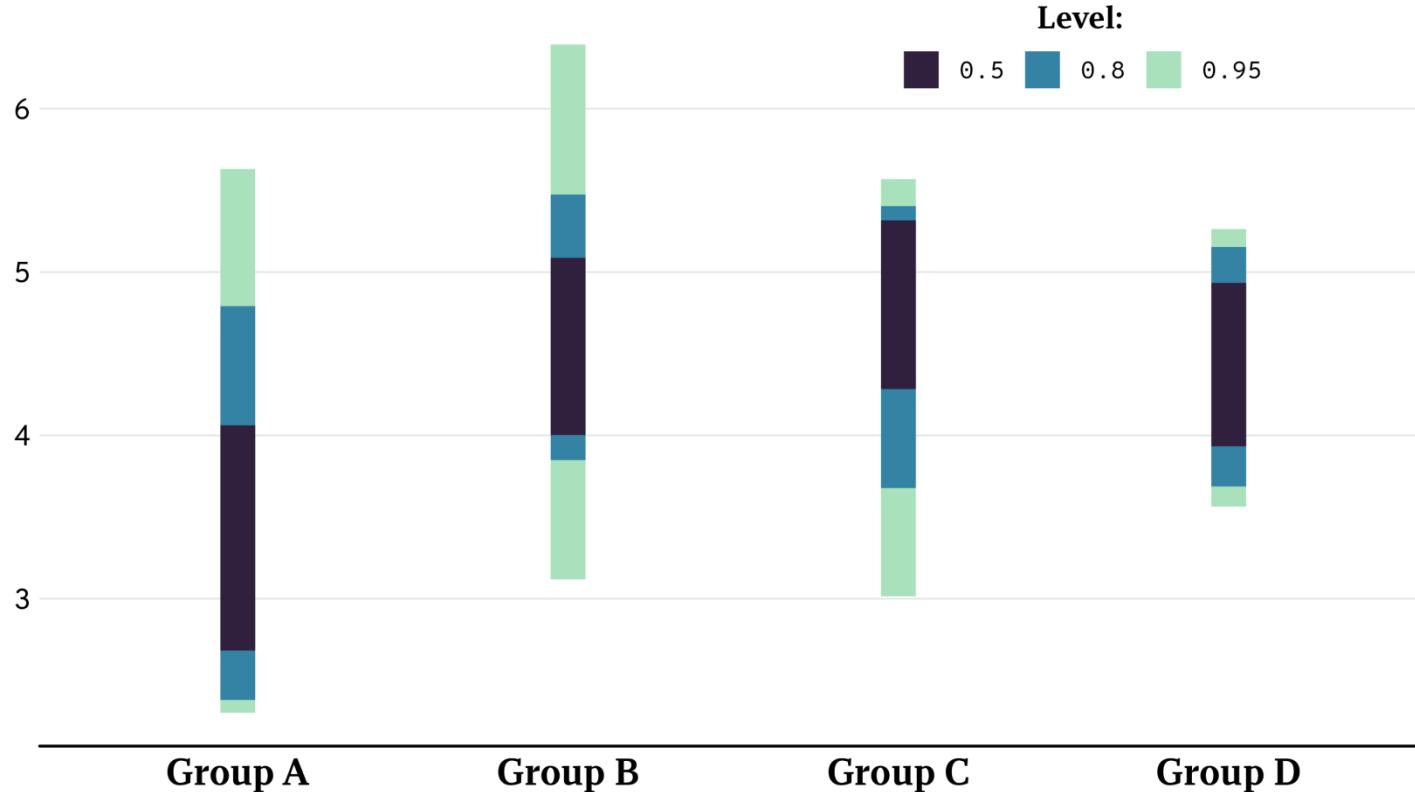
RIDGE LINE PLOT WITH MEDIAN

ggridges:::stat_density_ridges()



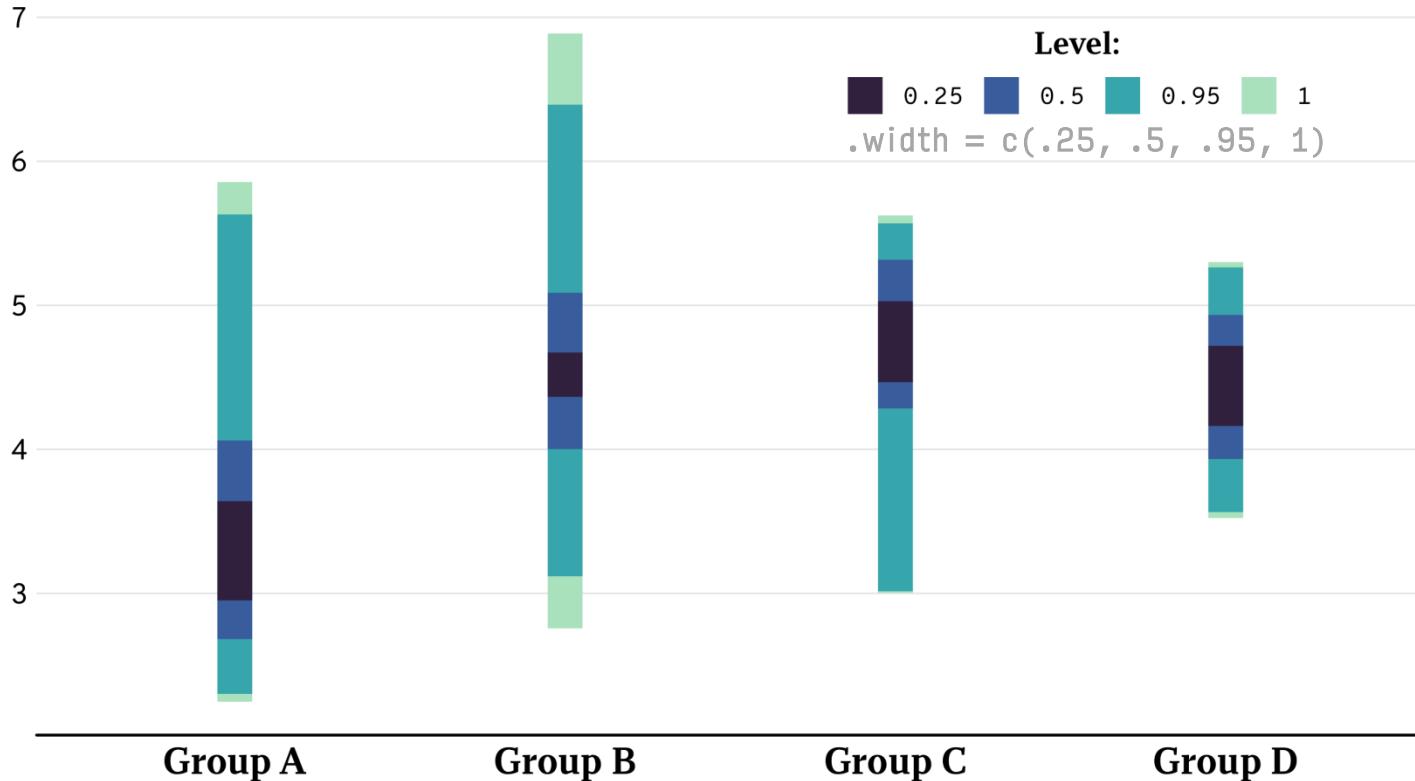
INTERVAL STRIPS

ggdist::stat_interval()



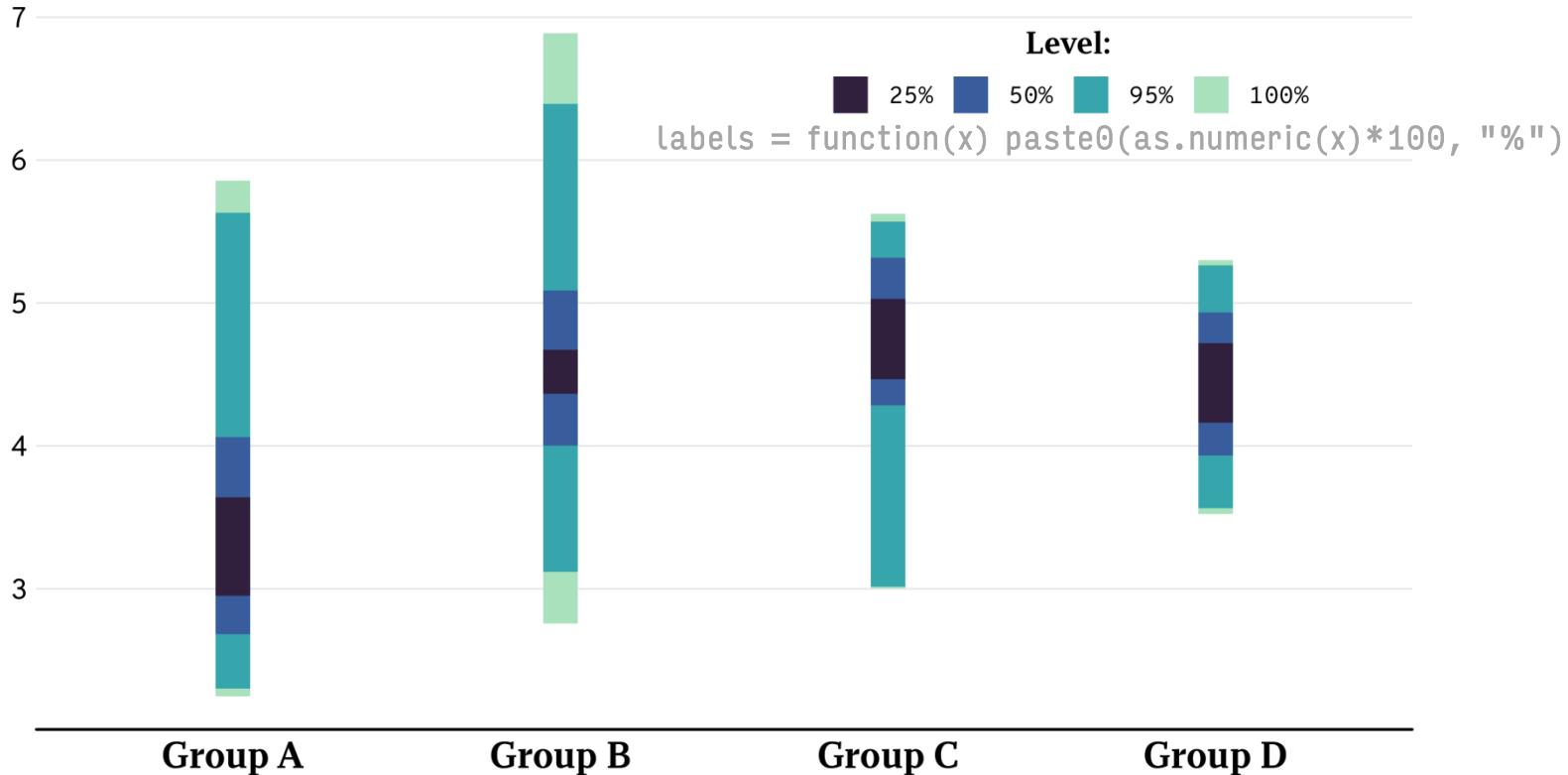
INTERVAL STRIPS

ggdist::stat_interval()



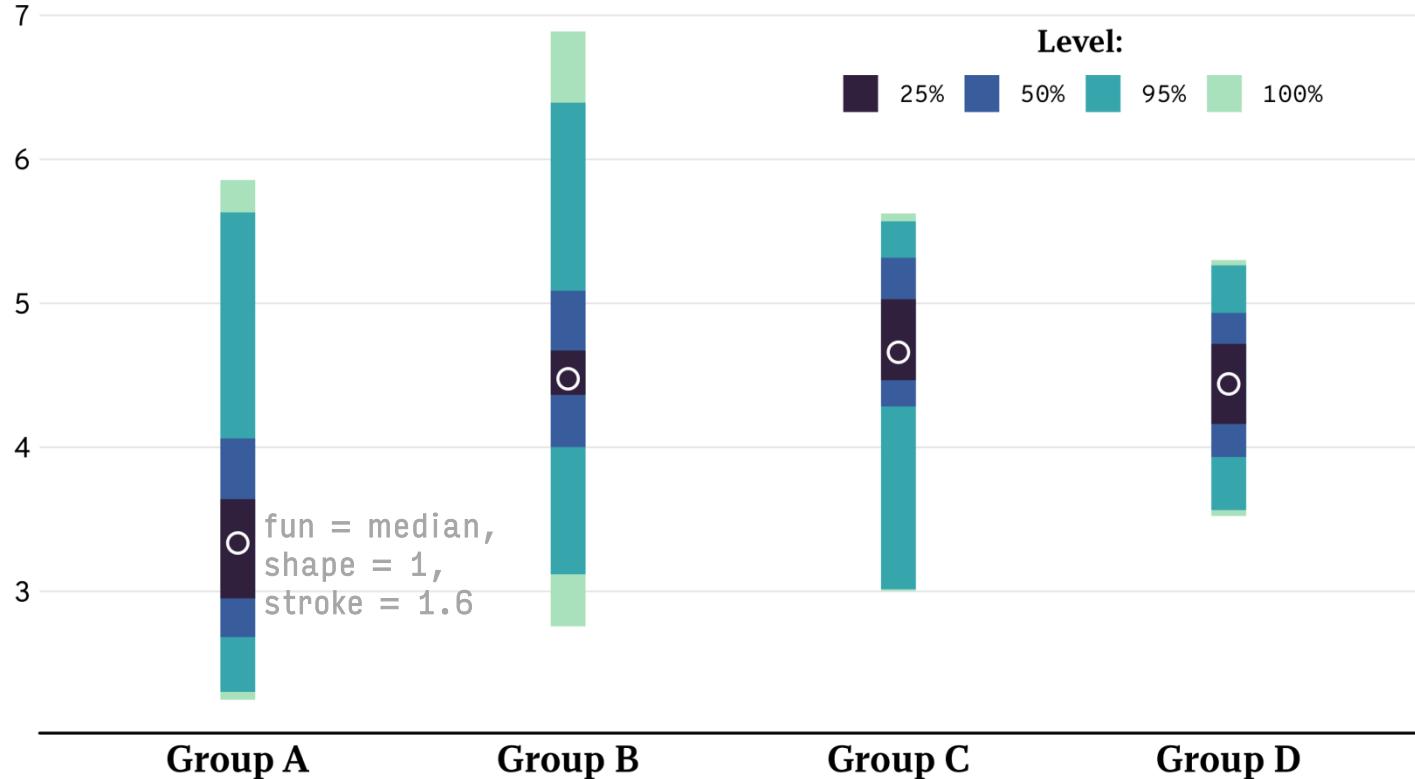
INTERVAL STRIPS

ggdist::stat_interval()



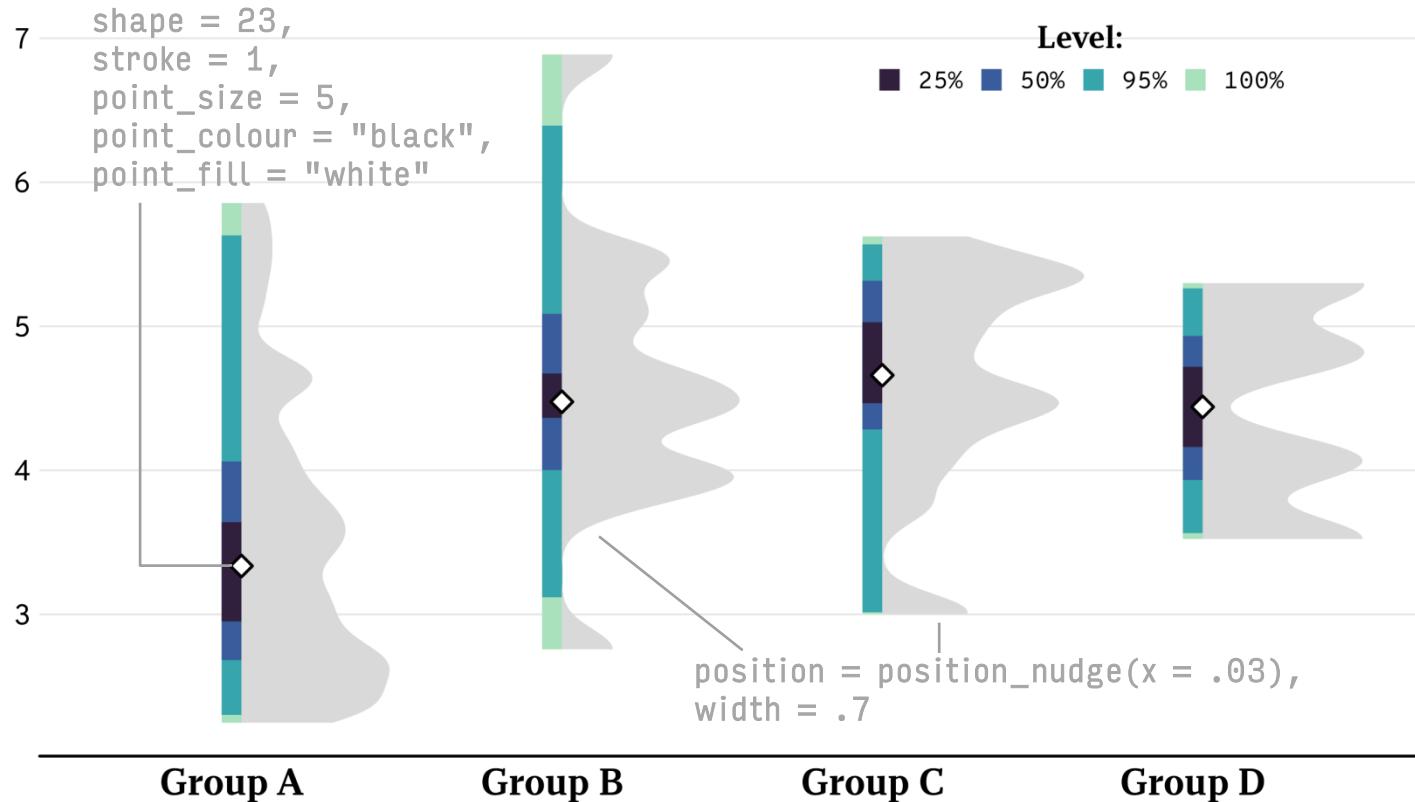
INTERVAL STRIPS

ggdist::stat_interval() + stat_summary(geom = "point")



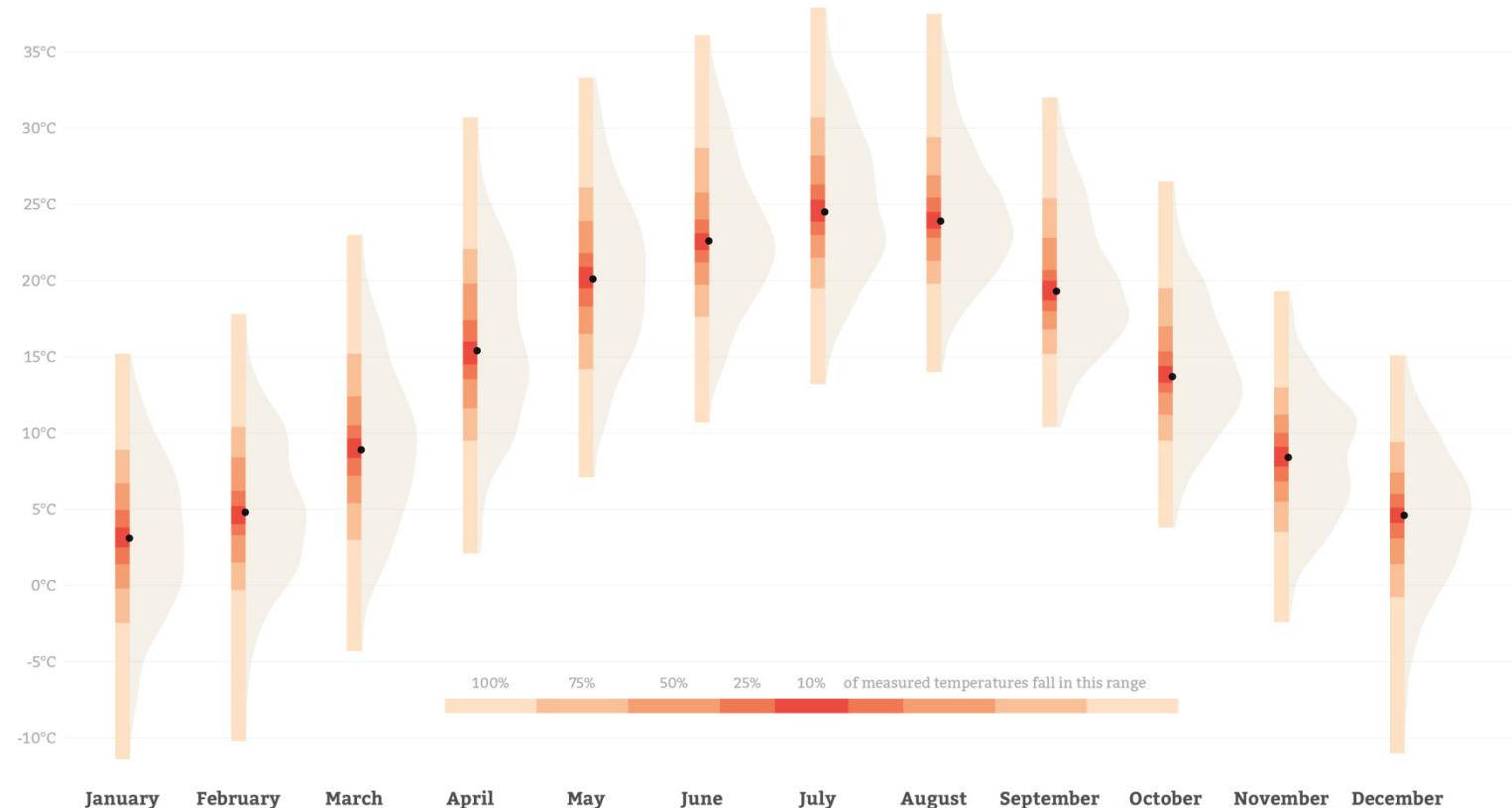
INTERVAL STRIPS X HALF-EYE

ggdist::stat_interval() + ggdist::stat_halfeye()



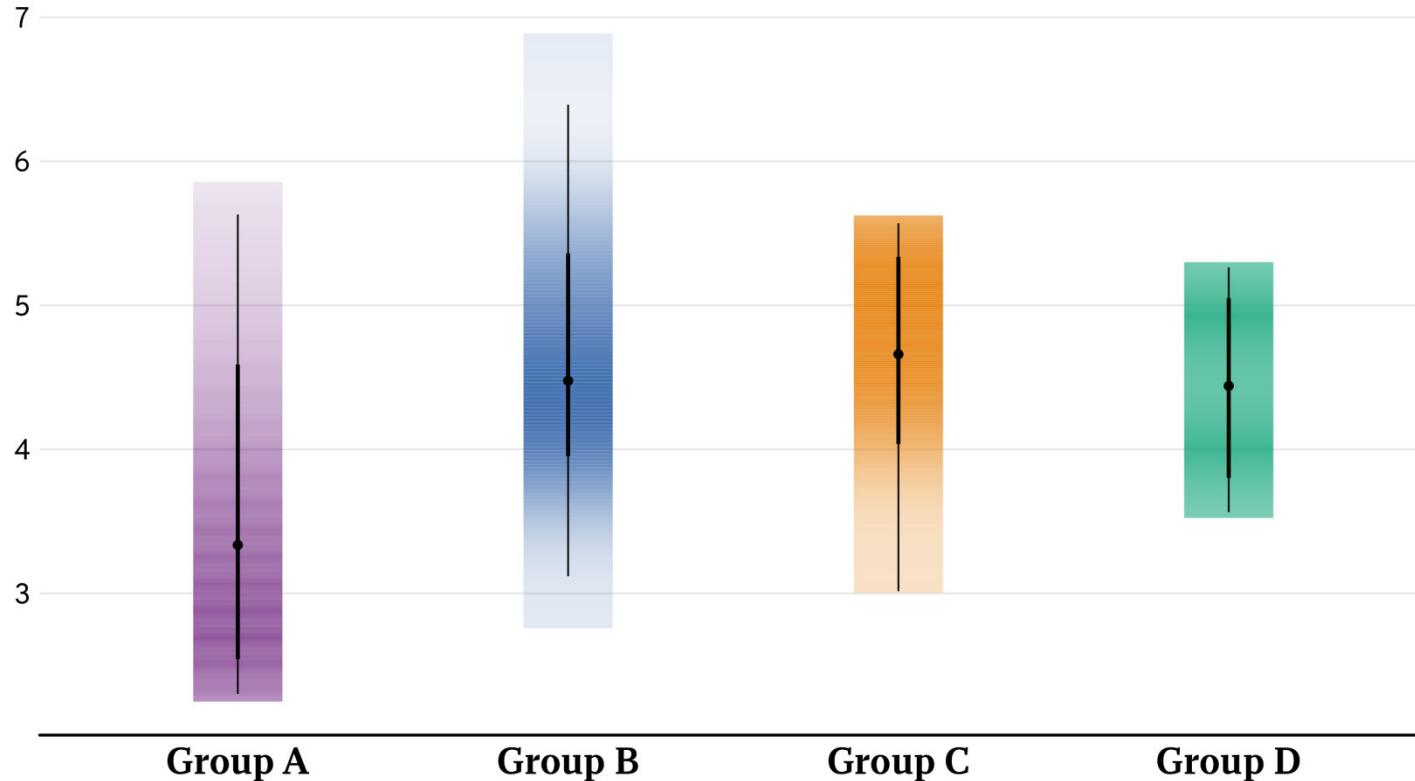
Daily Temperatures in Berlin, Germany

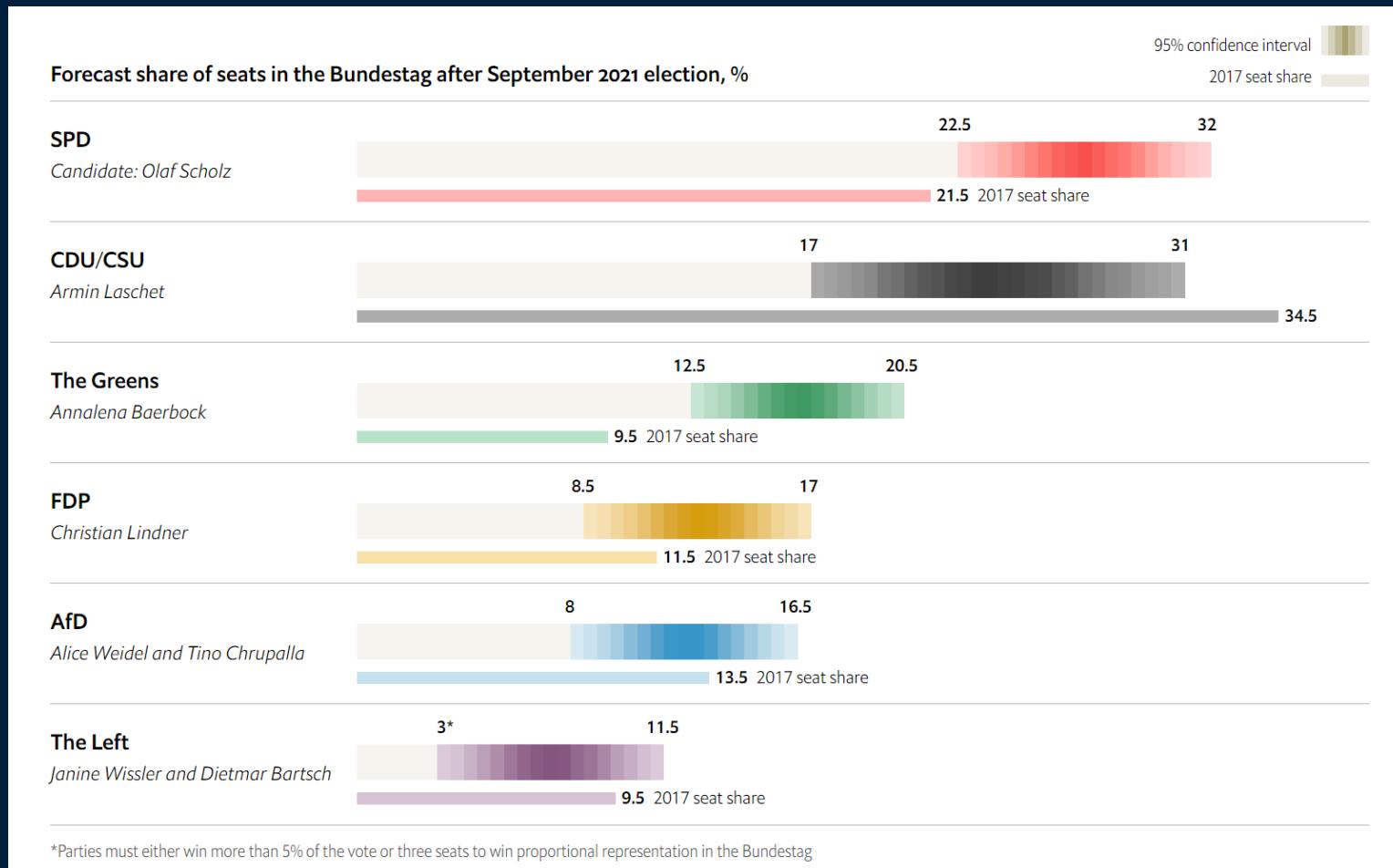
Range and distribution of maximum daily temperatures in Celsius per month from 2000 to 2018 measured in Berlin-Dahlem, Germany



GRADIENT INTERVAL STRIPS

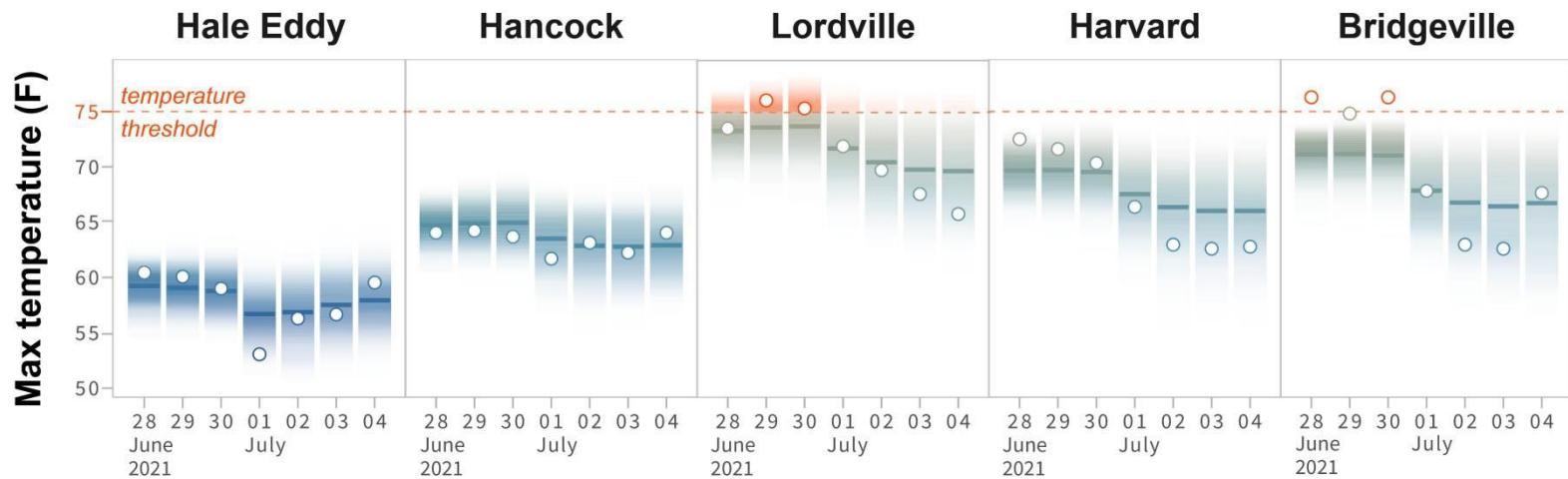
ggdist::stat_gradientinterval()





7-day-ahead forecasts of maximum stream temperature at 5 sites in the Delaware River Basin

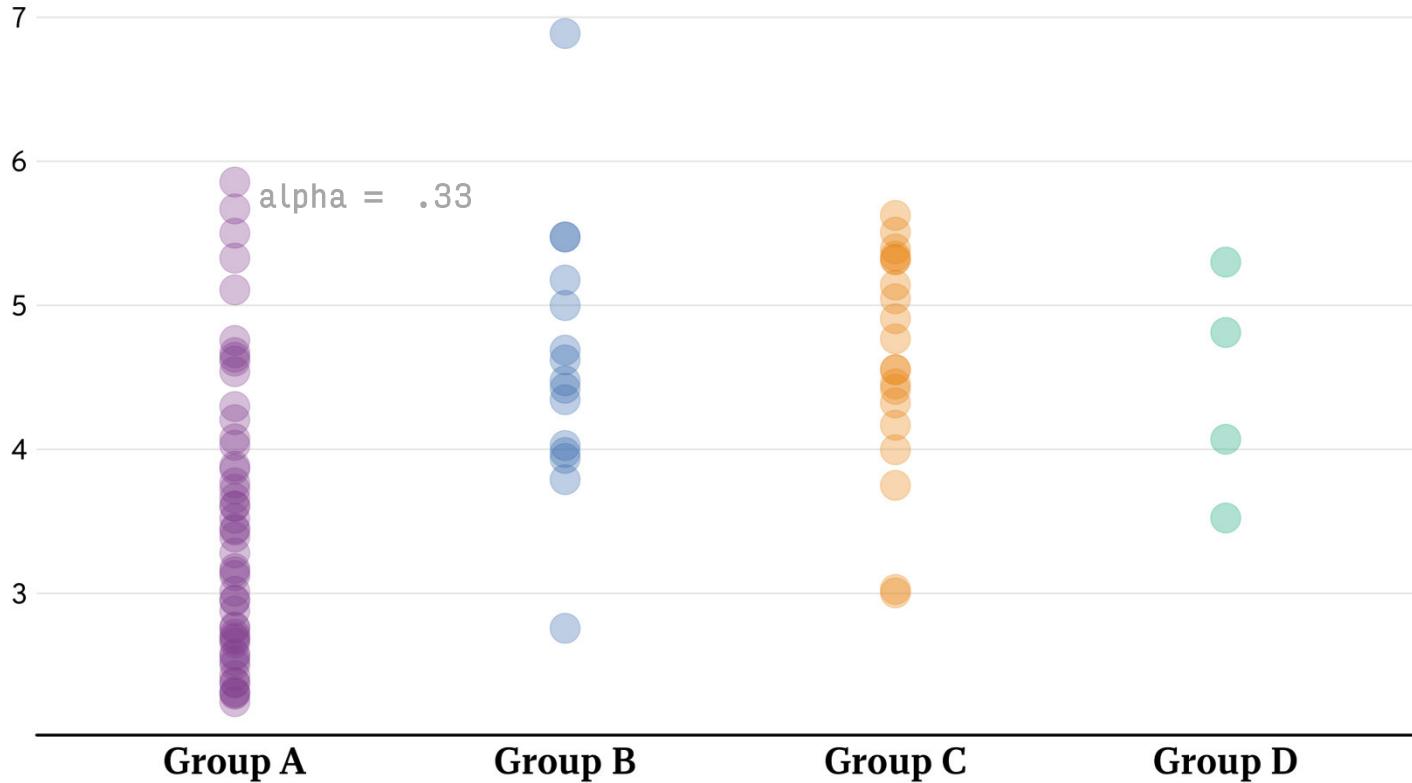
observed max temp
mean prediction
90% CI



SHOW THE *Raw* DATA

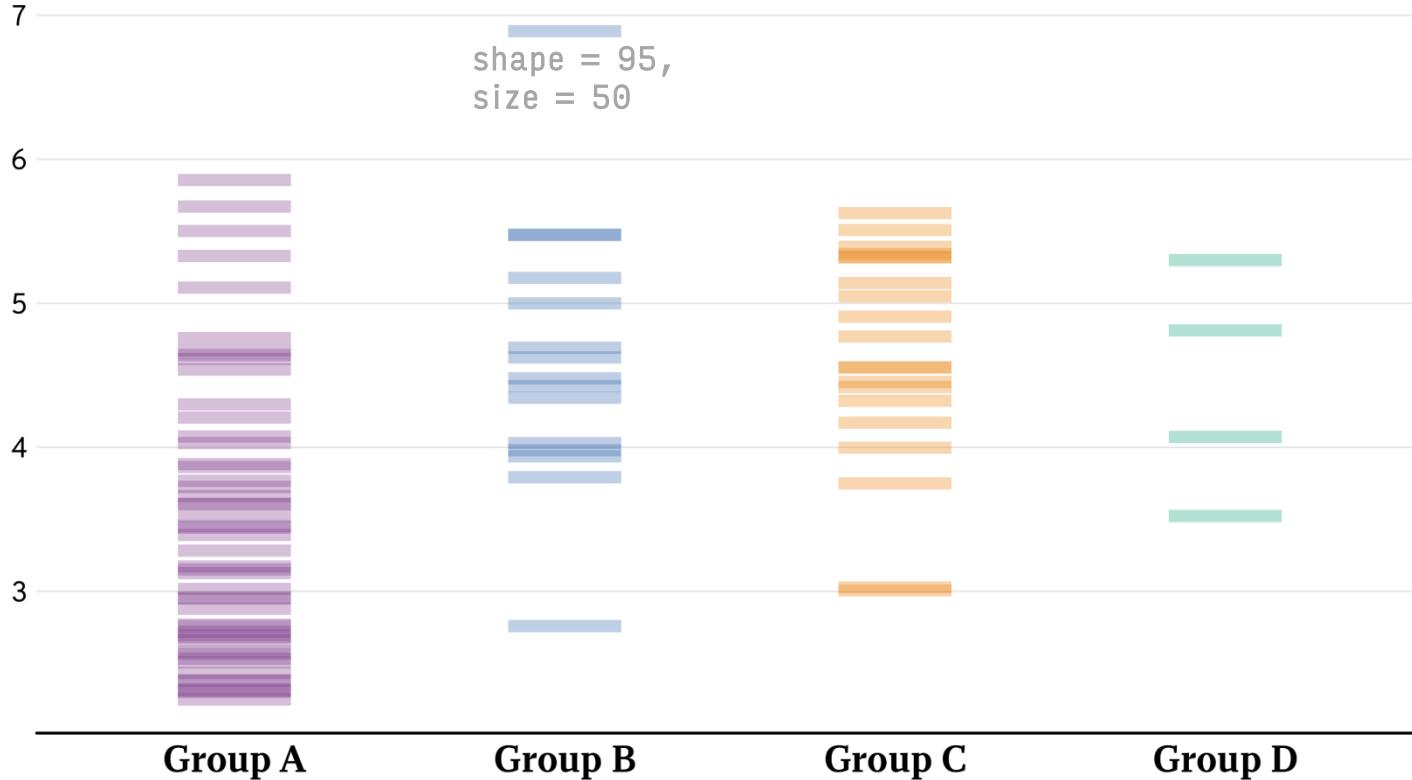
SCATTERPLOT (STRIP CHART)

geom_point()



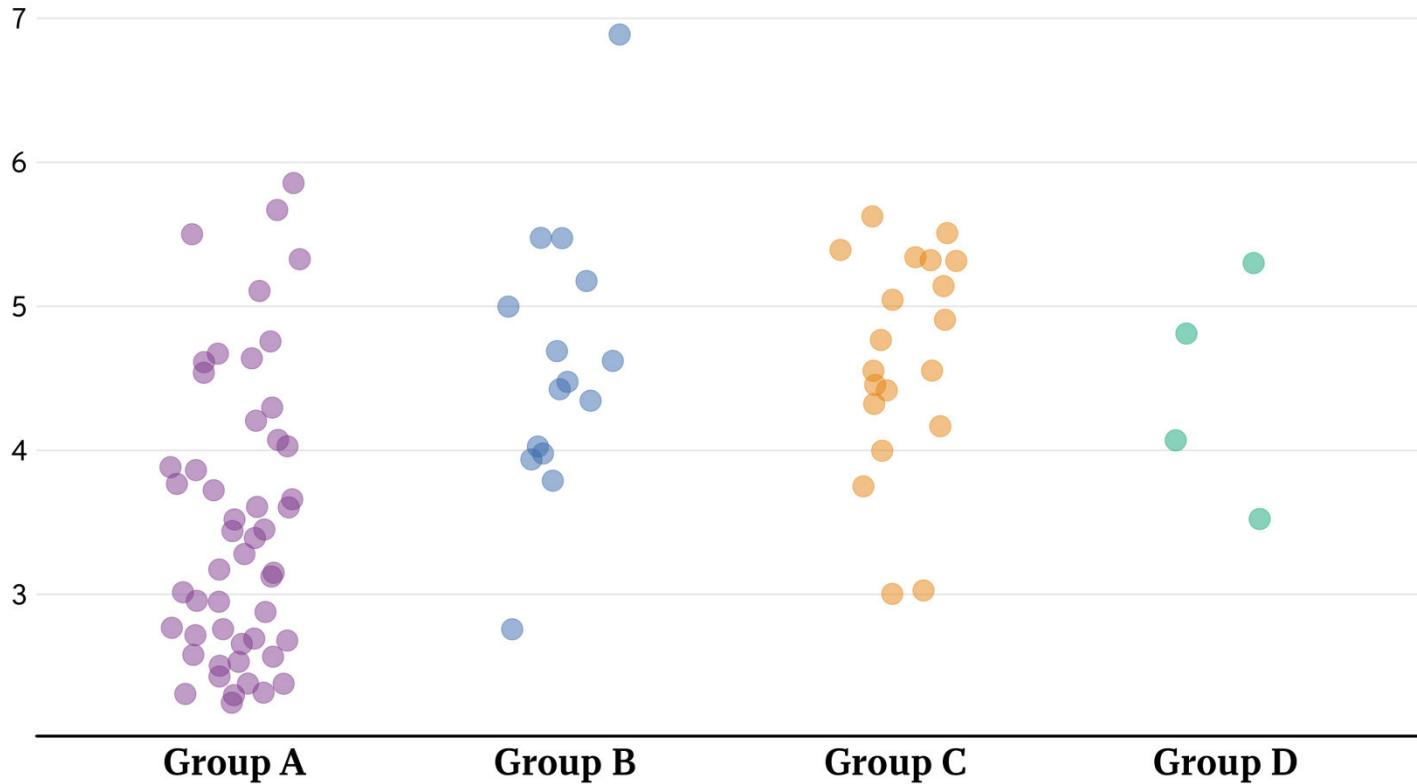
BARCODE PLOT (STRIP CHART)

geom_point()



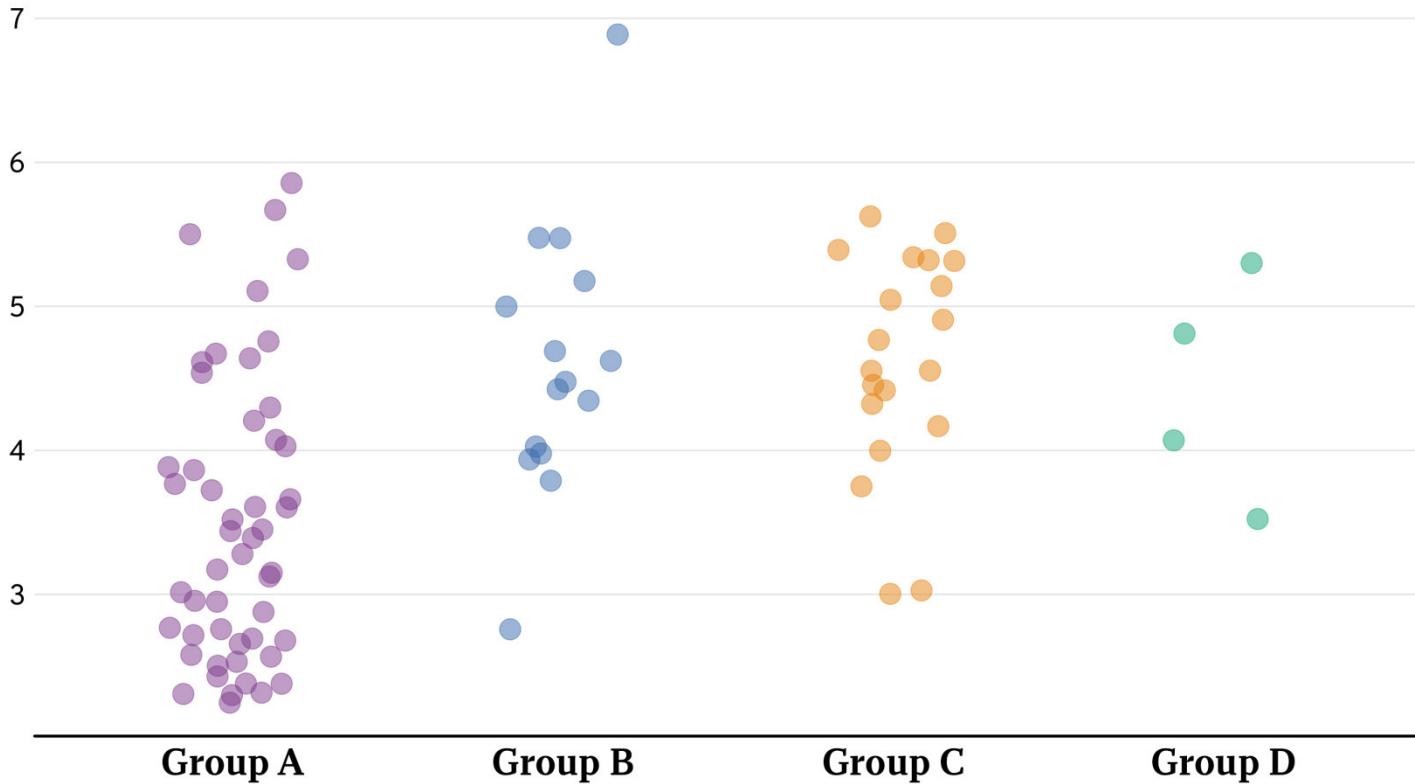
JITTER STRIP CHART

`geom_jitter()`



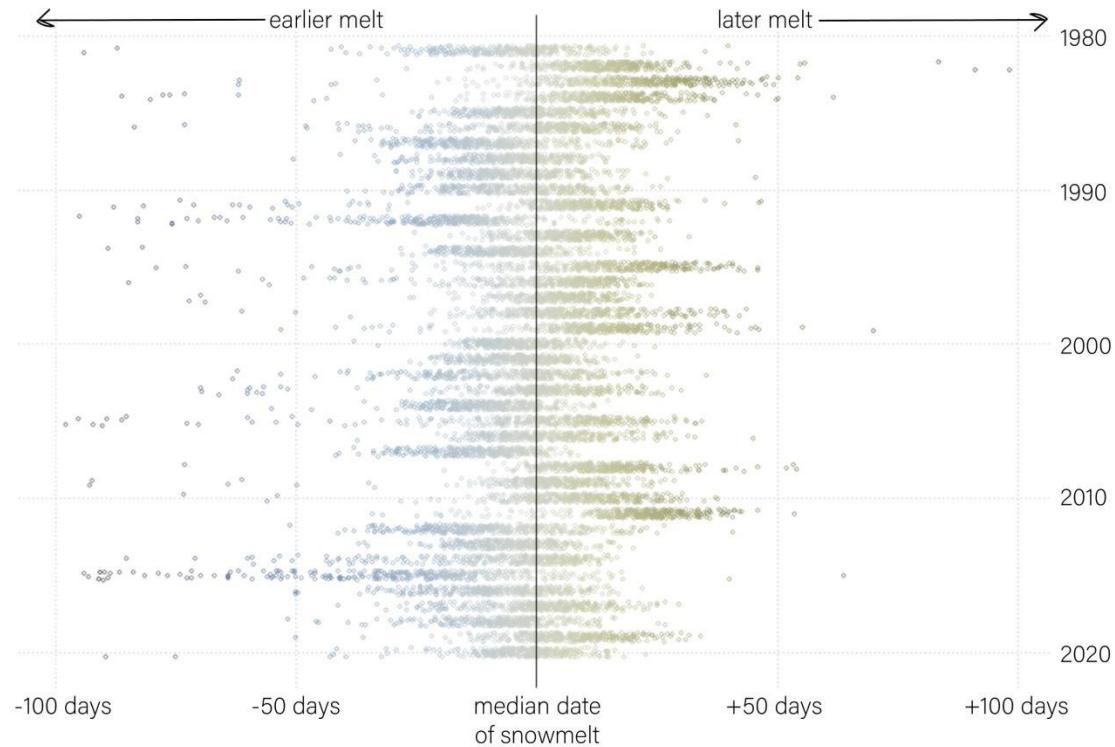
JITTER STRIP CHART

geom_point(position = position_jitter())



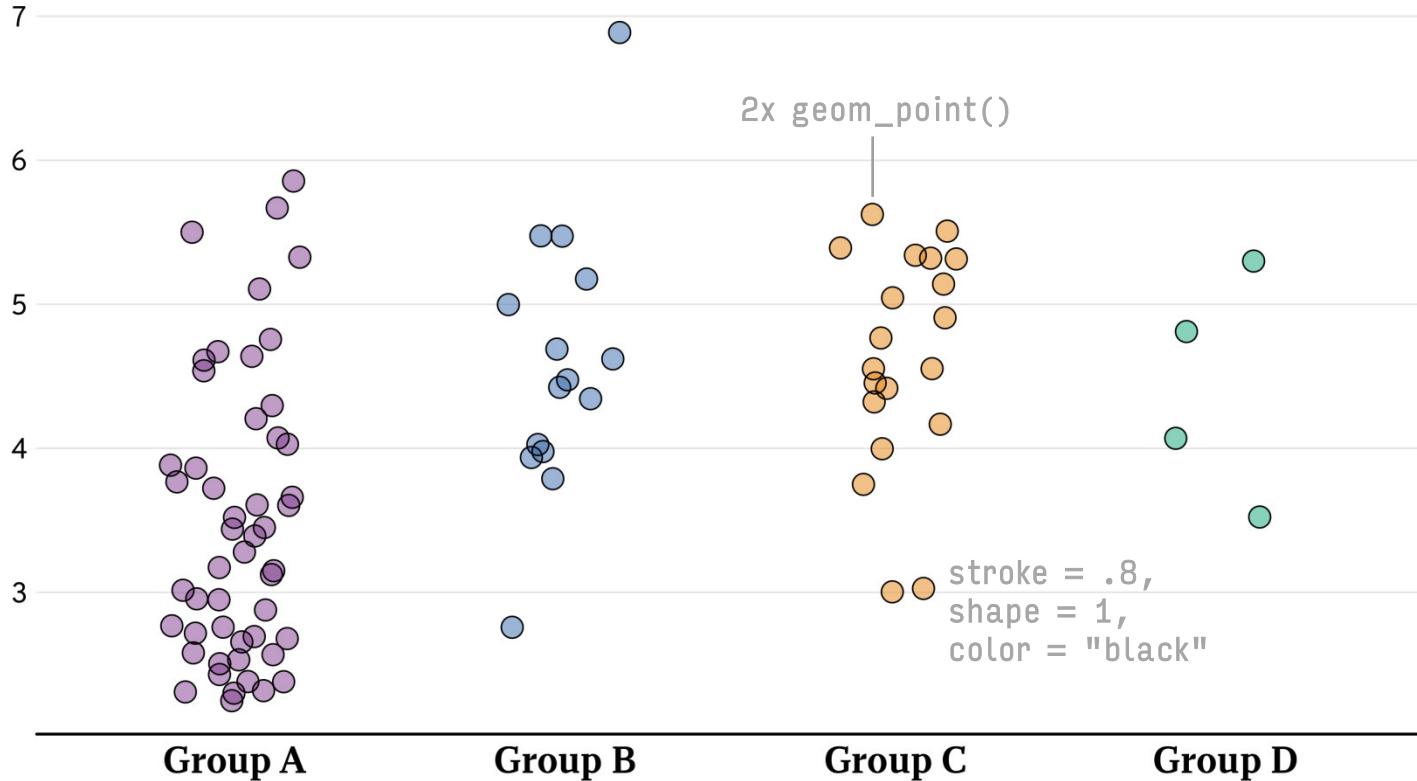
Snowmelt timing (1981-2020)

the difference in days from median snowmelt date at USDA-NRCS snow telemetry (SNOTEL) sites



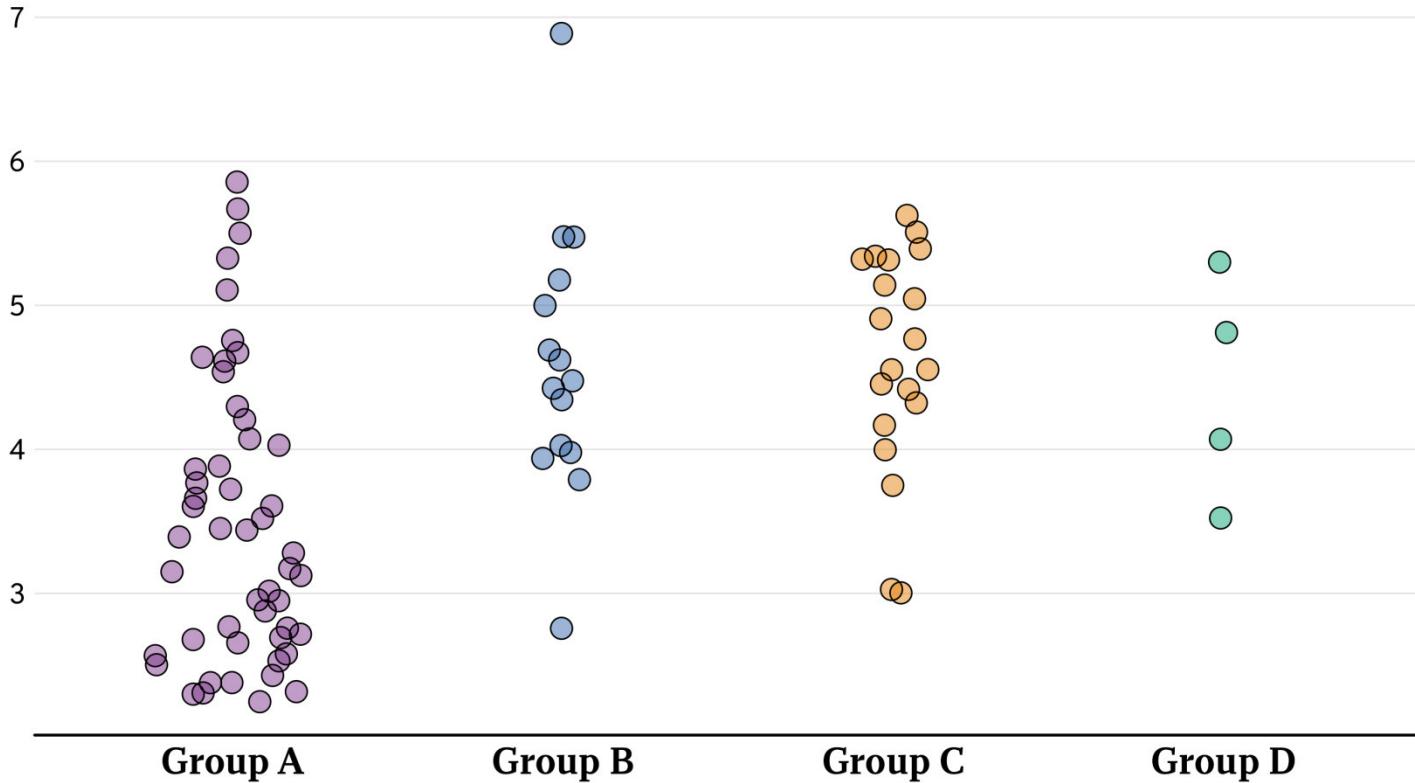
JITTER STRIP CHART

`geom_point(position = position_jitter(seed = 0))`



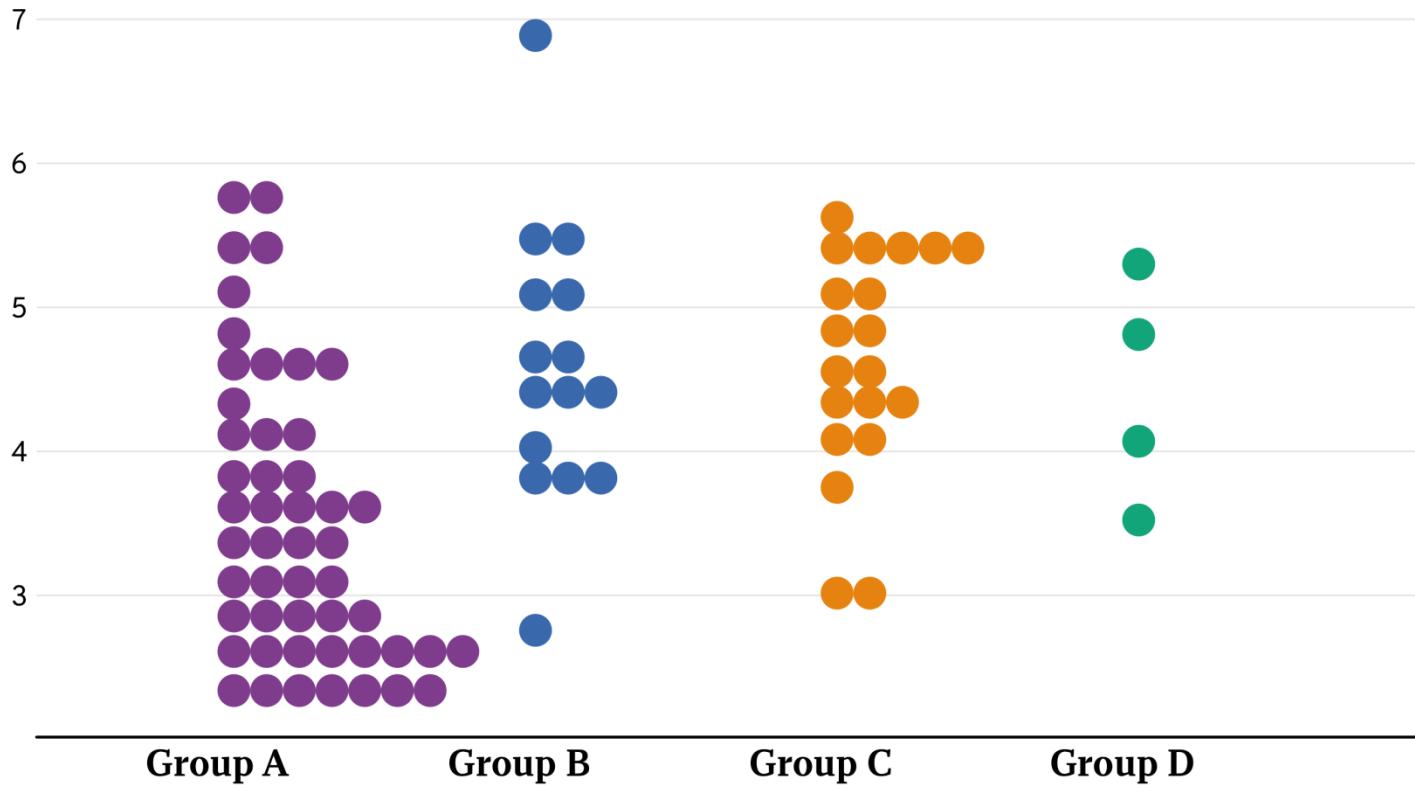
SINA PLOT

ggforce::geom_sina(seed = 0)



DOT PLOT

ggdist::stat_dot()



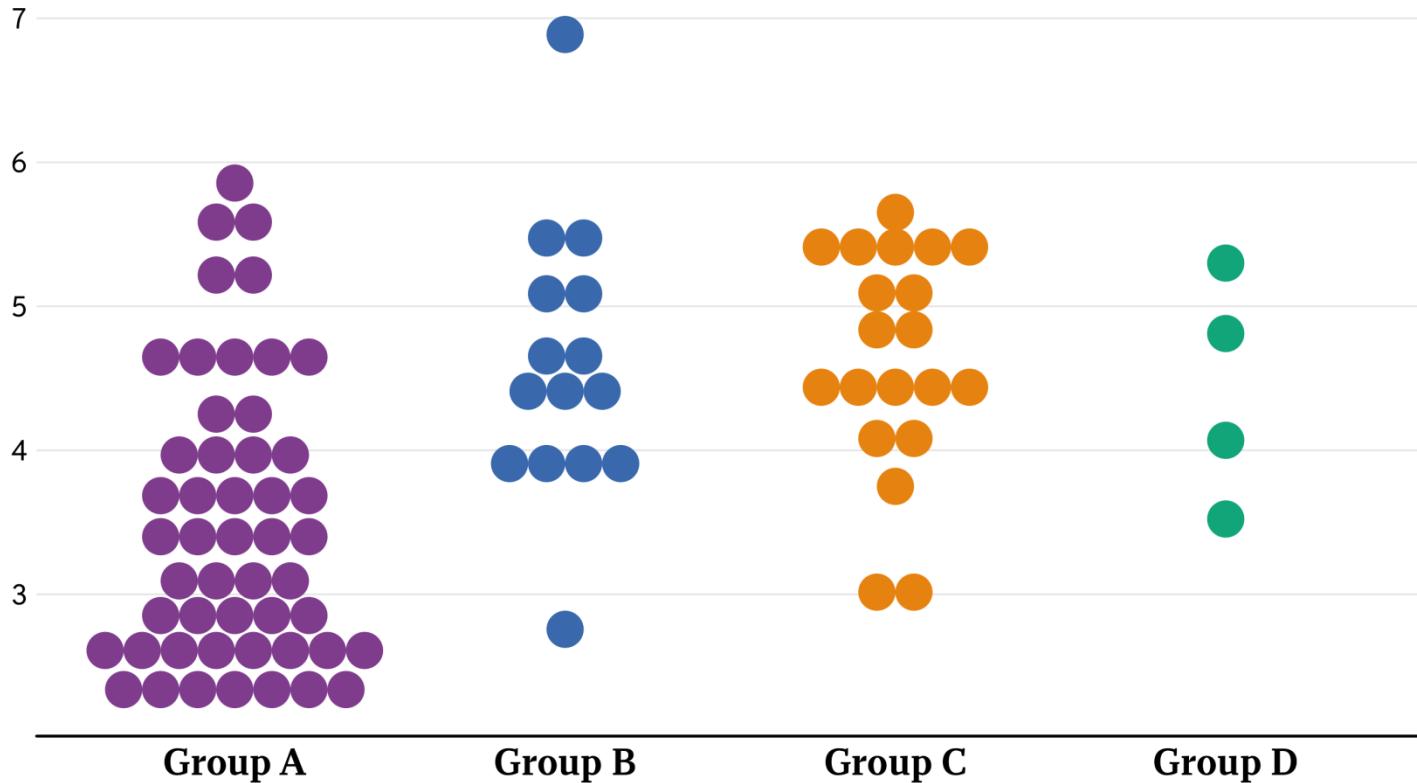
DOT PLOT

ggdist::stat_dots()



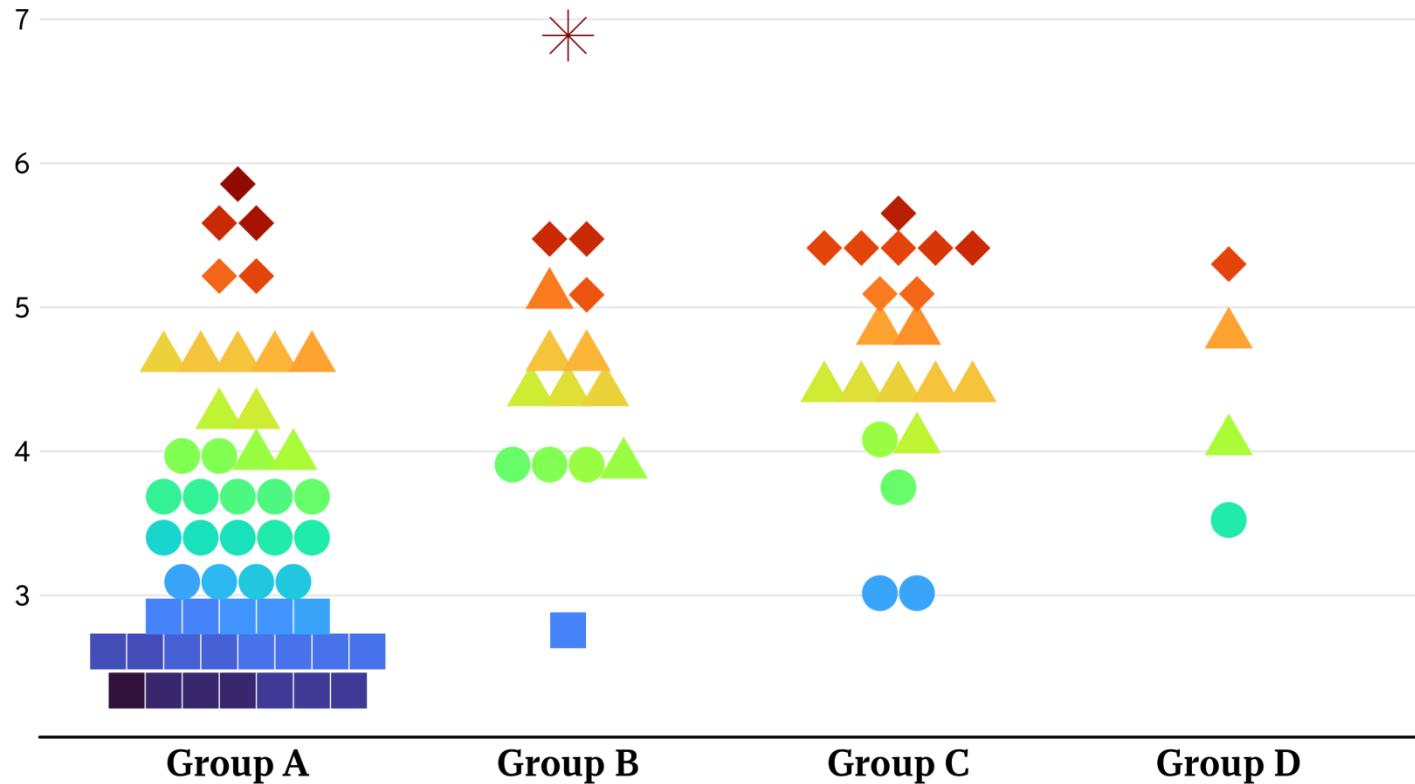
DOT PLOT

ggdist::stat_dots(side = "both")



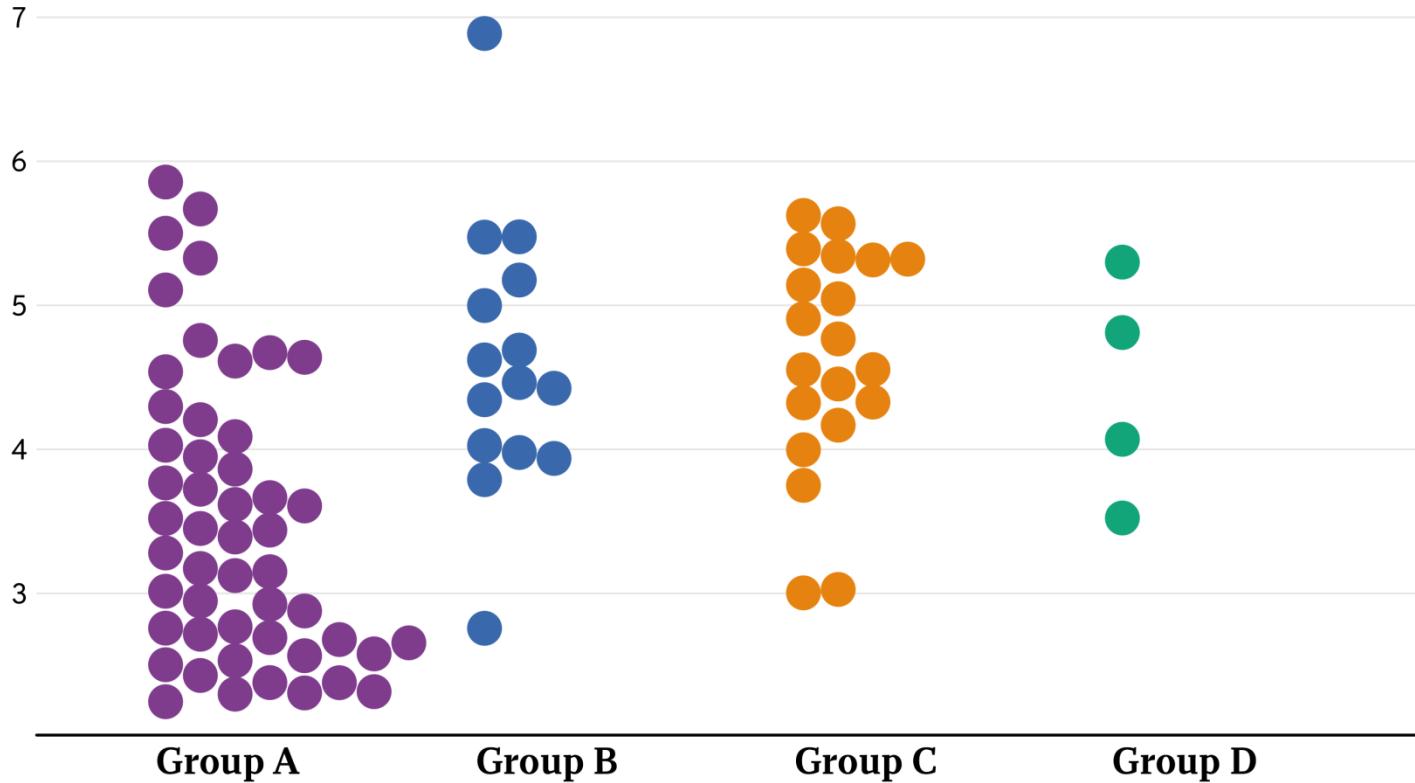
DOT PLOT

ggdist::stat_dots(side = "both")



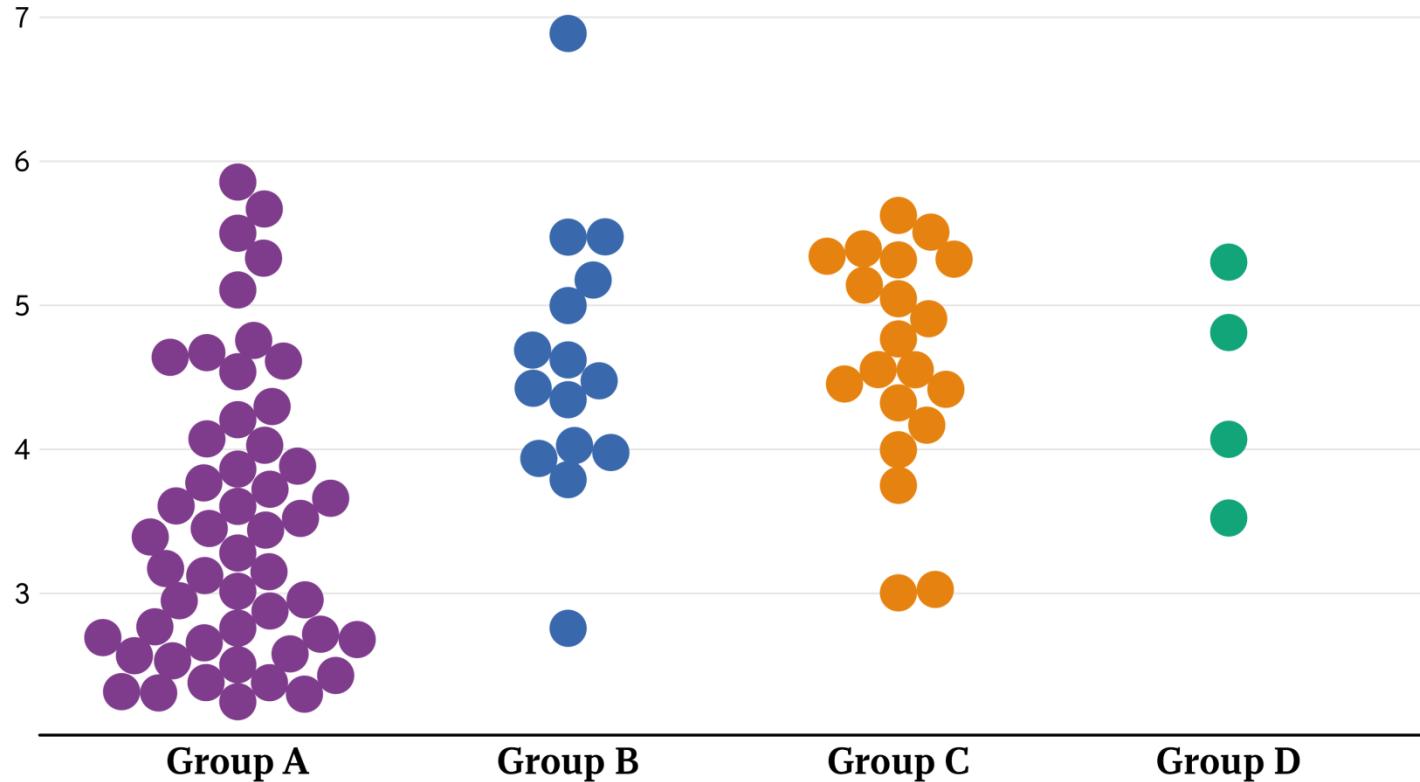
DOT PLOT

ggdist::stat_dots(layout = "weave")



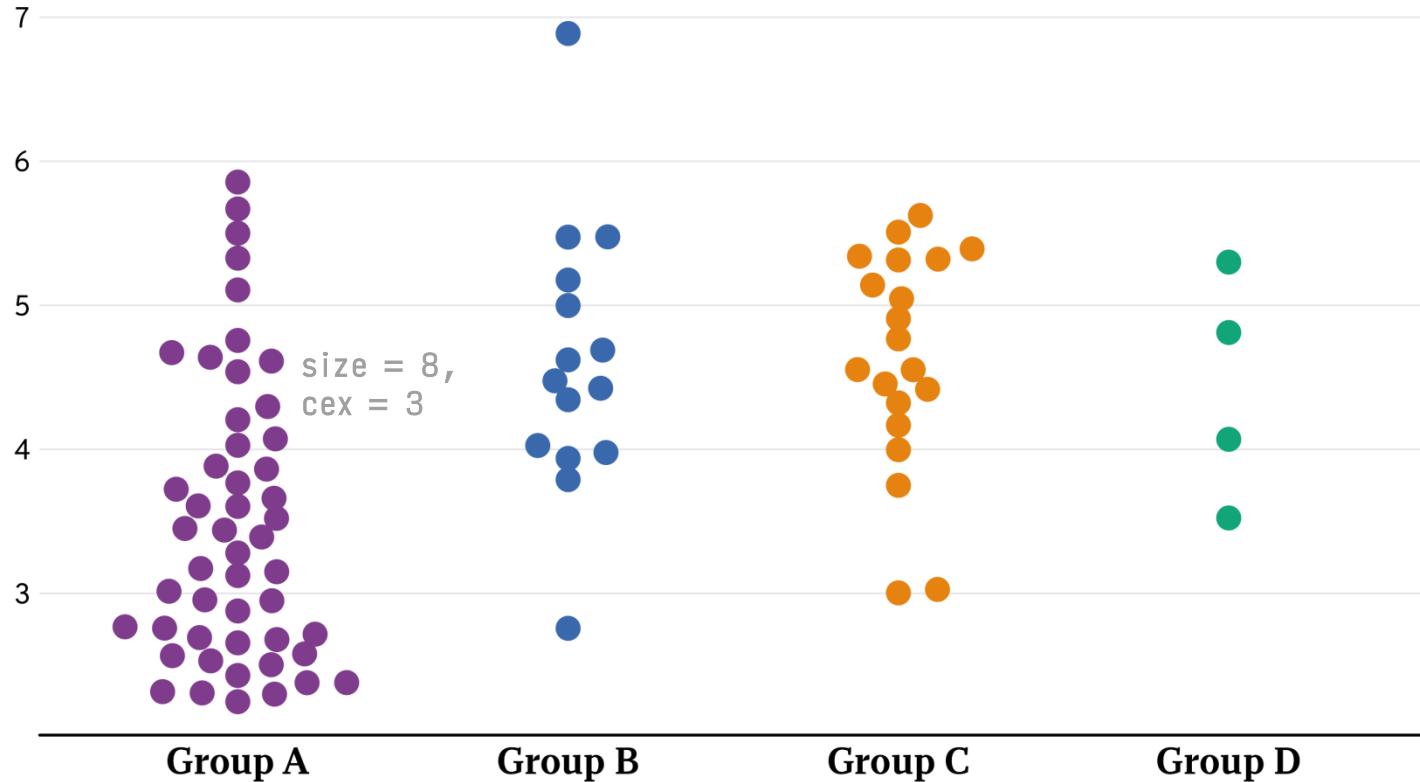
BEESWARM PLOT

`ggdist::stat_dot(layout = "swarm", side = "both")`



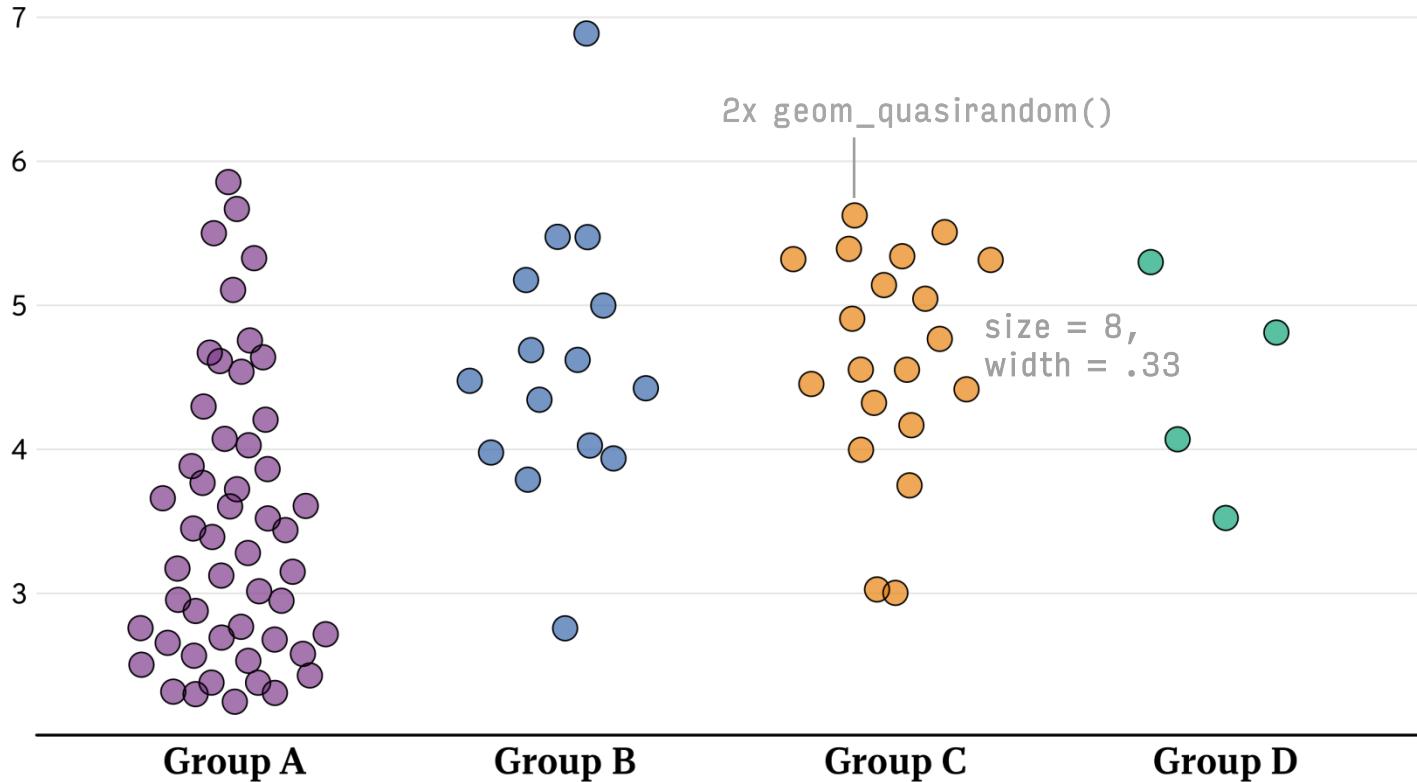
BEESWARM PLOT

ggbeeswarm::geom_beeswarm()

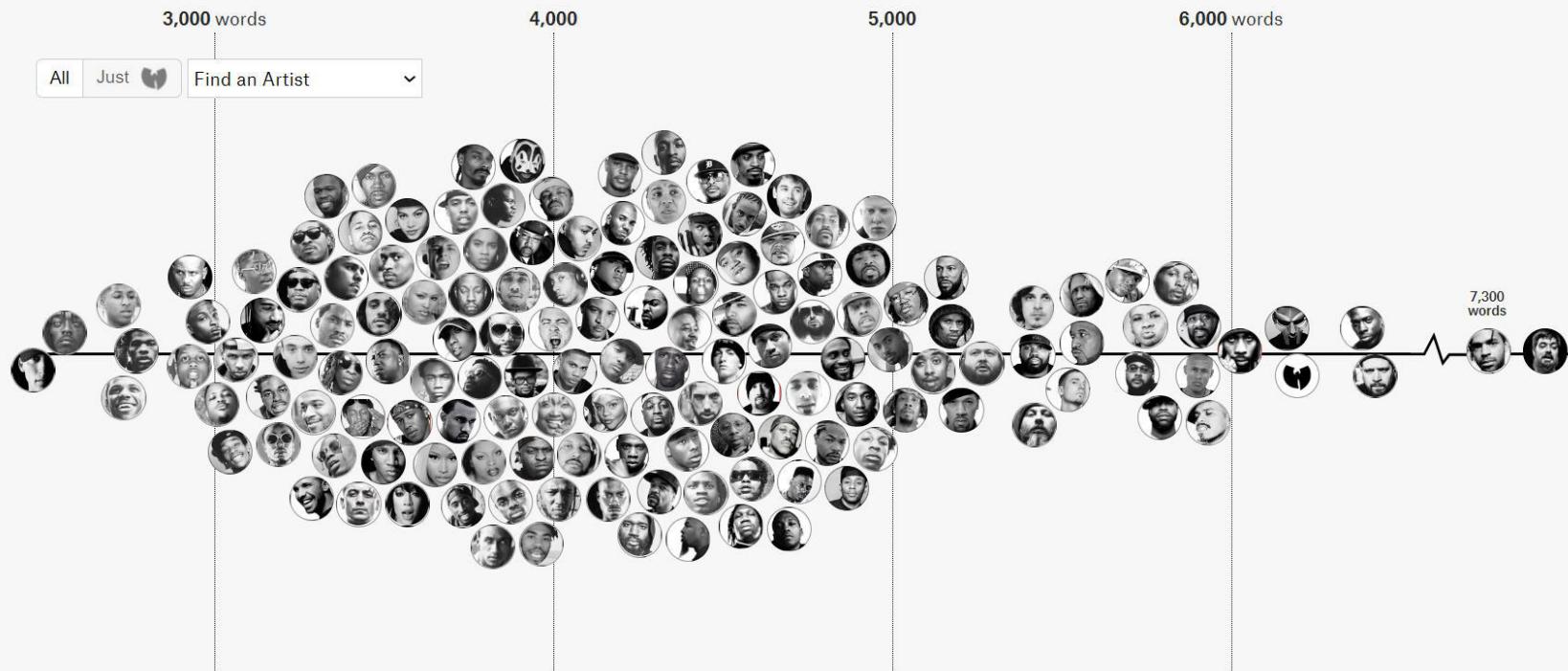


BEESWARM PLOT

ggbeeswarm::geom_quasirandom()



of Unique Words Used Within Artist's First 35,000 Lyrics



"[The Largest Vocabulary In Hip Hop](#)" by Matt Daniels (*The Pudding*)

YouTube DE

Search

Beeswarm Charts

One Chart at a Time

Beeswarm Charts with Cédric Scherer (Ep. 32)

473 views • 23 Feb 2021

14 ▾ 0 ▾ SHARE ▾ SAVE ▾

All Listenable Related From Jon Schwabish

Jon Schwabish 2.15K subscribers

SUBSCRIBE

Effective Data Visualization
How to design impactful and aesthetically pleasing charts

Effective Data Visualization
with Cédric Scherer & Matthe...

Heureka Labs 1.8K views • 11 months ago

1:34:38

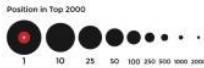
youtu.be/kTtV2GRckj0

TOP 2000 ❤️ 70's & 80's

Since 1999 the 2000 most popular songs of all time, as voted by the show's audience, are played on Dutch national Radio 2 in a yearly marathon. The 2000 songs are on the air between noon on December 25th until New Year's Eve and over half of the Dutch population listens to the Top 2000 each year.

Each ● to the right represents a song in the Top 2000. It is placed according to its year of release. In the legend below you can see what the size and color of a song means.

The bulk of the songs and most of the top 10 are from the 70's & 80's.



The oldest song on the list, Billy Holiday's "Strange Fruit," is from 1939. It's 17 years older than the second-oldest song; if it will still be barely in now, on position 1989.



Newly discovered
Although already released in 1972, 'Stamen' from David Bowie is the highest new song in the list. It never appeared in the previous 17 editions of the Top 2000 and entered in 2016 on position 270.

Prince
Fernando Lopes

Prince
Another legend who passed away in 2016 (on April 21st). It seems that new people discovered his works, with all 9 songs that were in 2015's list rising significantly and 8 more songs joining in 2016.

Archives
Baudouin

AVONIO
Boudewijn de Groot · 1997

High riser

High User
Adela's When we were young from 2015 apparently needed some time to become fully appreciated. It is the song with the highest increase in the list, shooting 1599 places from position 1743 to 164.

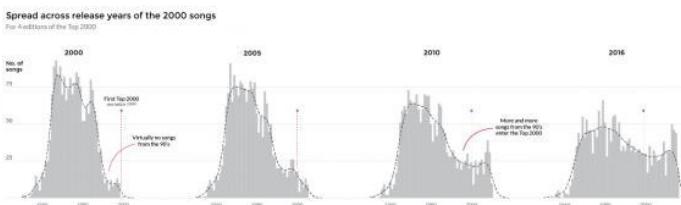
2016's most popular

non
the list in 2015 due to a
no-campaign, nobody can
expect that Pokorný had on
Milek's daily lives in 2055.
— Venit by Jason Pölzer rises
to no results 2024

David Bowie
Passing away only days after the release of his new album *Blackstar* on January 10th 2016. His legend remains strong with 26 songs in the Top 2000, 148 most popular song

The Beatles

No other artist or band has more songs in the Top 2000 as the Beatles. With 38 songs they are responsible for 14% of all titles before 1970. Nonetheless, only 5 years ago they still had 50 songs in the list.



The charts on the right represent all 2000 songs from 2part software of the Top 2000 field in 2000, 2005, 2010 and the most recent 2014 edition.

The songs are stacked according to their year of release. The higher a rectangle, the more songs that are in the top 2000 list from that release year.

The black chart below represents a smoothed curve over all 2000 songs. This makes the distribution of chart positions more transparent.

Downloaded from <https://academic.oup.com/imrn/article/2020/10/3333> by the University of Texas at Austin on 11 October 2020

Journal of Clinical Anesthesia 2004;16:949-953. First published online as an Author Manuscript, and later made available in the journal issue. © 2004 Elsevier Inc. All rights reserved.

**But they're losing tracks
to the new Millennium.**

It makes sense that the Top 2000 will be more spread out for each new edition, since there are more songs to choose from. However, if we compare the distributions of the Top 2000 songs over 4 editions, we see that especially the 90's has been gaining a lot of popularity.

Even though all songs from the 90's were out in the 2000 edition, only a few songs from that decade were chosen. Whereas in the 2016 edition the number of songs from the 90's has risen significantly. This could be due to a new generation who has grown up during the 90's taking over from those who voted in the early 2000's (who apparently didn't appreciate the music much).

How the growing racial diversity in the US is reflected in schools

Comparison of the Simpson diversity index, a quantitative measure that reflects the racial diversity, during the school years **1994–1995** and **2016–2017** for schools with the racial diversity being lower than (● or ▲) or higher than/equal to (● or ▲) the state's median.

Dotted lines indicate the US median diversity and numbers the number of schools with only one ethnicity (dots removed for better readability).







Leland Wilkinson

May 28, 2016 at 3:43 pm

People interested in these plots need to read my paper

http://moderngraphics11.pbworks.com/f/wilkinson_1999.DotPlots.pdf

before programming dot plots. Some of the implementations above are useful. One, however, is misleading and should be avoided. That's the "beeswarm" plot (why did we need a new name for this 100 year old plot?). In the vertical version of the "beeswarm" plot, the Y values are placed at their proper locations but the X values are arbitrarily ordered by the Y values. This creates a visual artifact of U-shaped dot stacks that misrepresent the structure of the data. There are also other examples in the "beeswarm" R program that allow the dots to be asymmetric around a vertical center line. This, too, induces a visual artifact. Dot plots need to be a faithful representation of a density (this is a well-defined statistical concept) and need to converge to a population density as sample size increases.

The main point of my paper was not to devise an algorithm for producing dot plots, but to show that an admissible dot plot algorithm (there are several) needs to be evaluated on its Integrated Mean Square Error (IMSE). I also showed that when this is properly defined, the dot plot loss function resembles the ones used for histograms and kernel density estimates.

As I tried to explain in The Grammar of Graphics, valid visualizations need to pay attention to the underlying mathematical and statistical models on which they are based. It is not enough to draw unstructured pictures of "data," pretty as they may be.



cedricscherer.com



@CedScherer



@z3tt

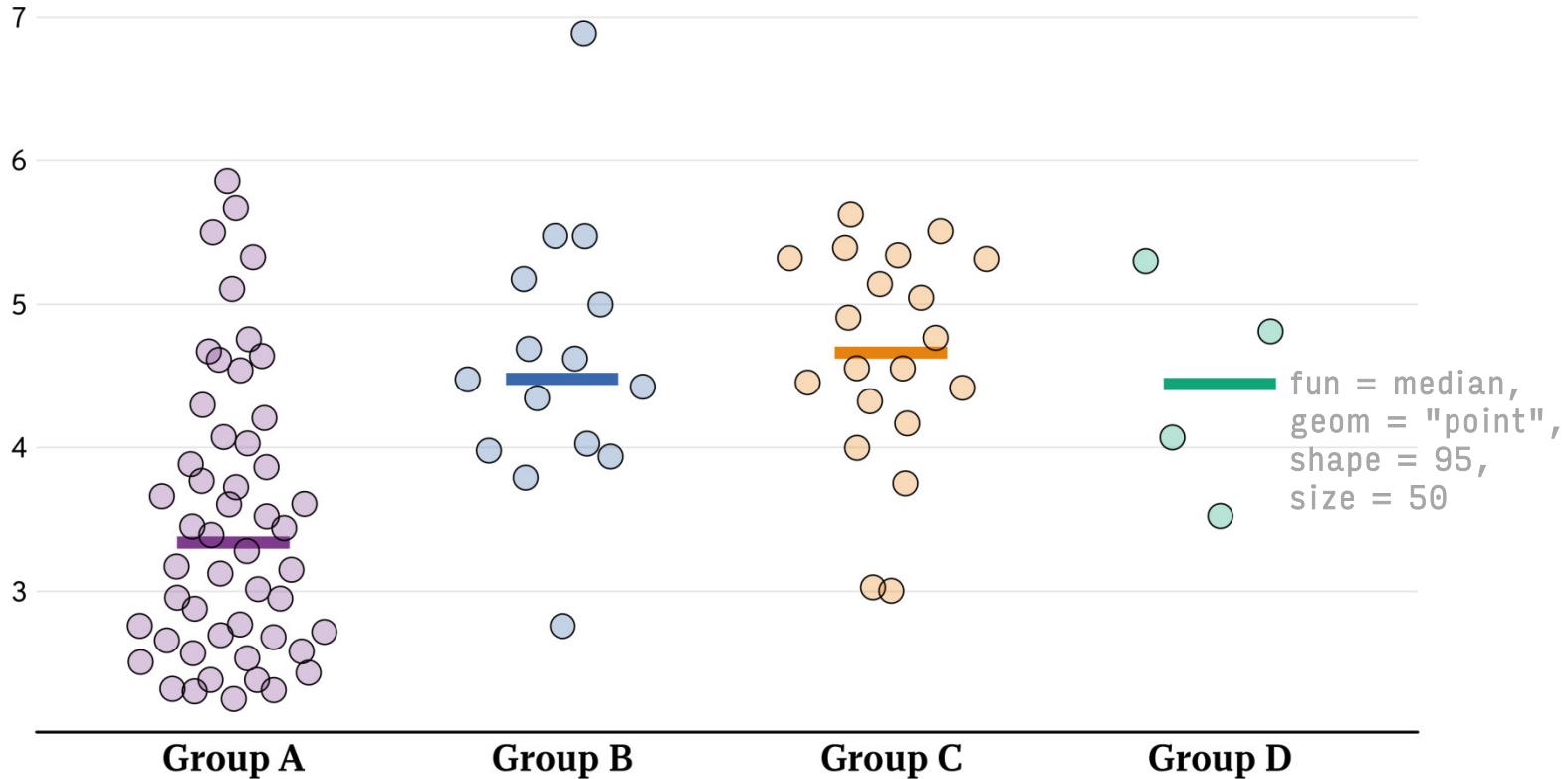


@cedscherer

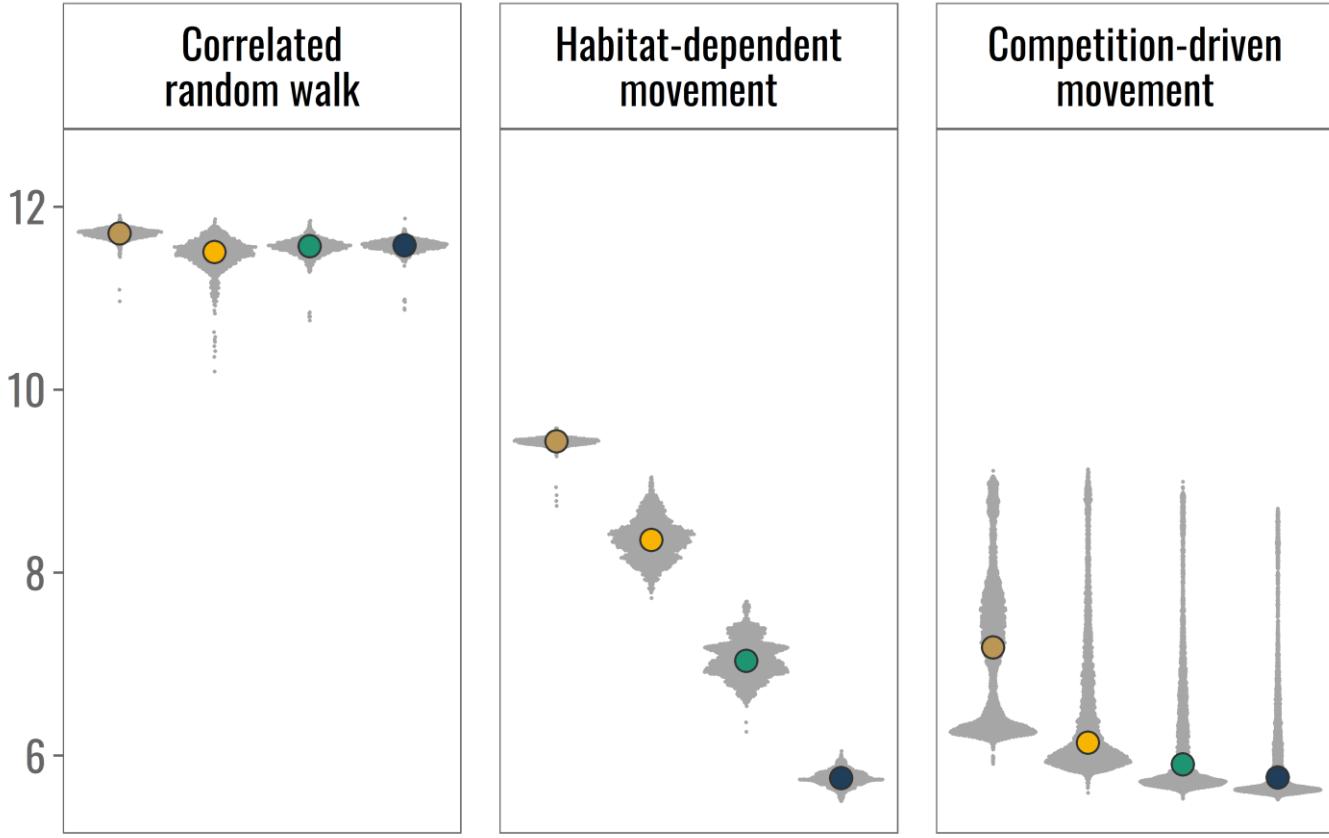
THE BEST OF BOTH WORLDS: *Hybrid Charts*

BEESWARM PLOT WITH MEDIAN

ggbeeswarm::geom_beeswarm() + stat_summary()



Visited cells per week

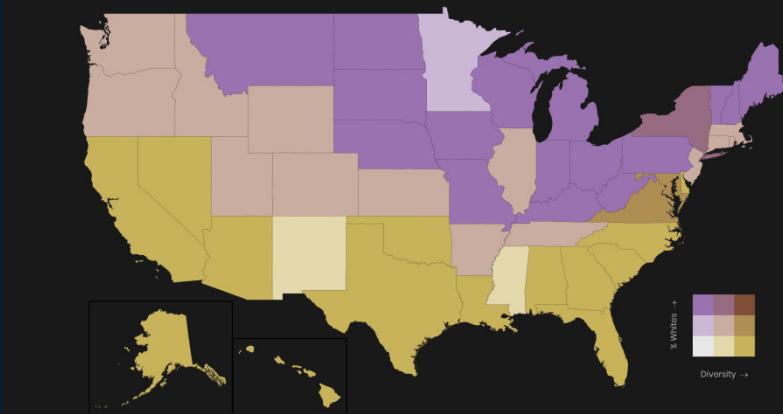


Landscape scenario

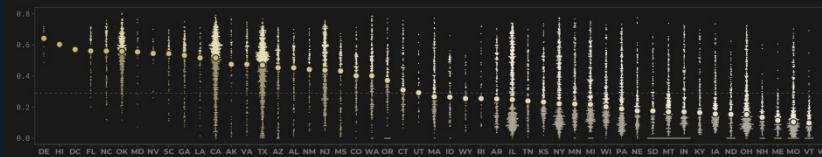
- Homogeneous
- Large clusters
- Small clusters
- Random

How diverse are schools in the US?

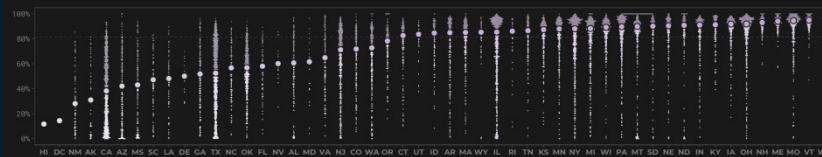
Bivariate map showing the combination of racial diversity measured as [Simpson index](#), a quantitative diversity measure, and the proportion of students with white ethnicity during the school year 2016–2017.



Simpson diversity index for all schools grouped per state, ranked from states with **high diversity** to those with **low diversity**.
Larger dots represent each state's median diversity with darker colored points laying below and lighter colored points laying above this value.

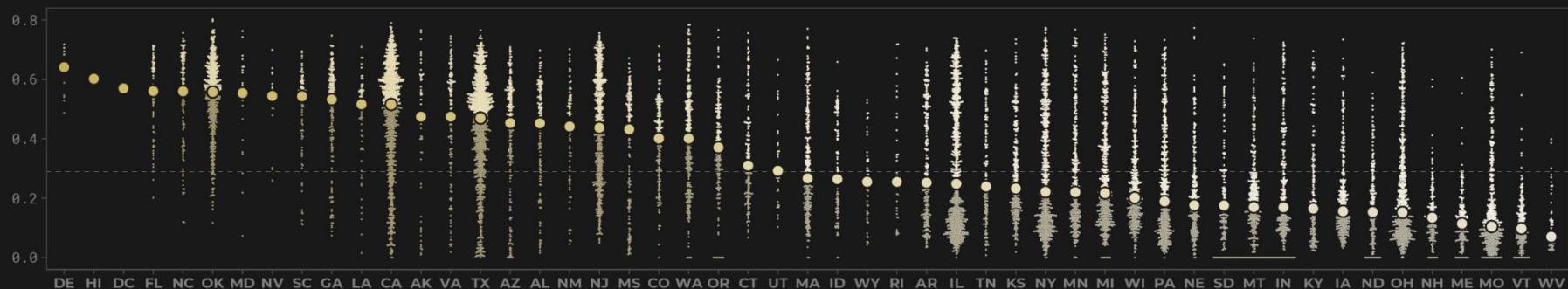


Proportion of white students for all schools grouped per state, ranked from states with **few white students** to those with **many white students**.
Larger dots represent each state's median proportion with lighter colored points laying below and darker colored points laying above this value.



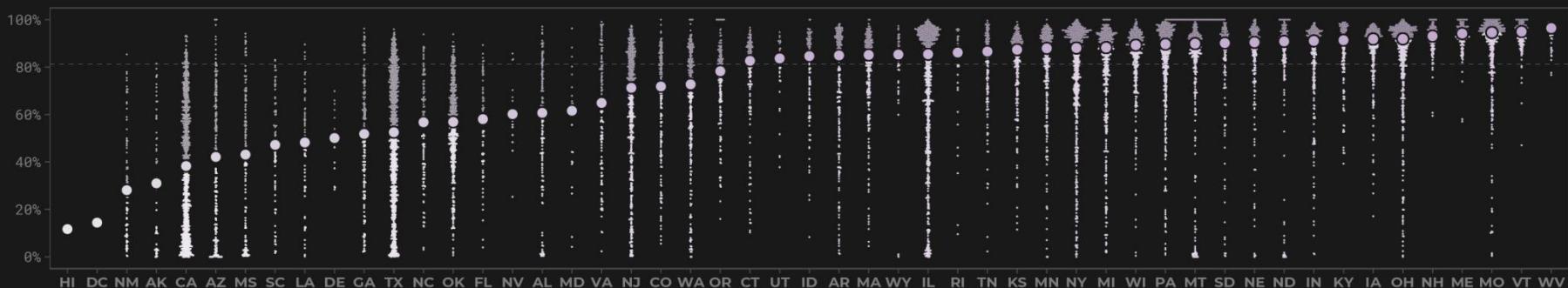
Simpson diversity index for all schools grouped per state, ranked from states with **high diversity** to those with **low diversity**.

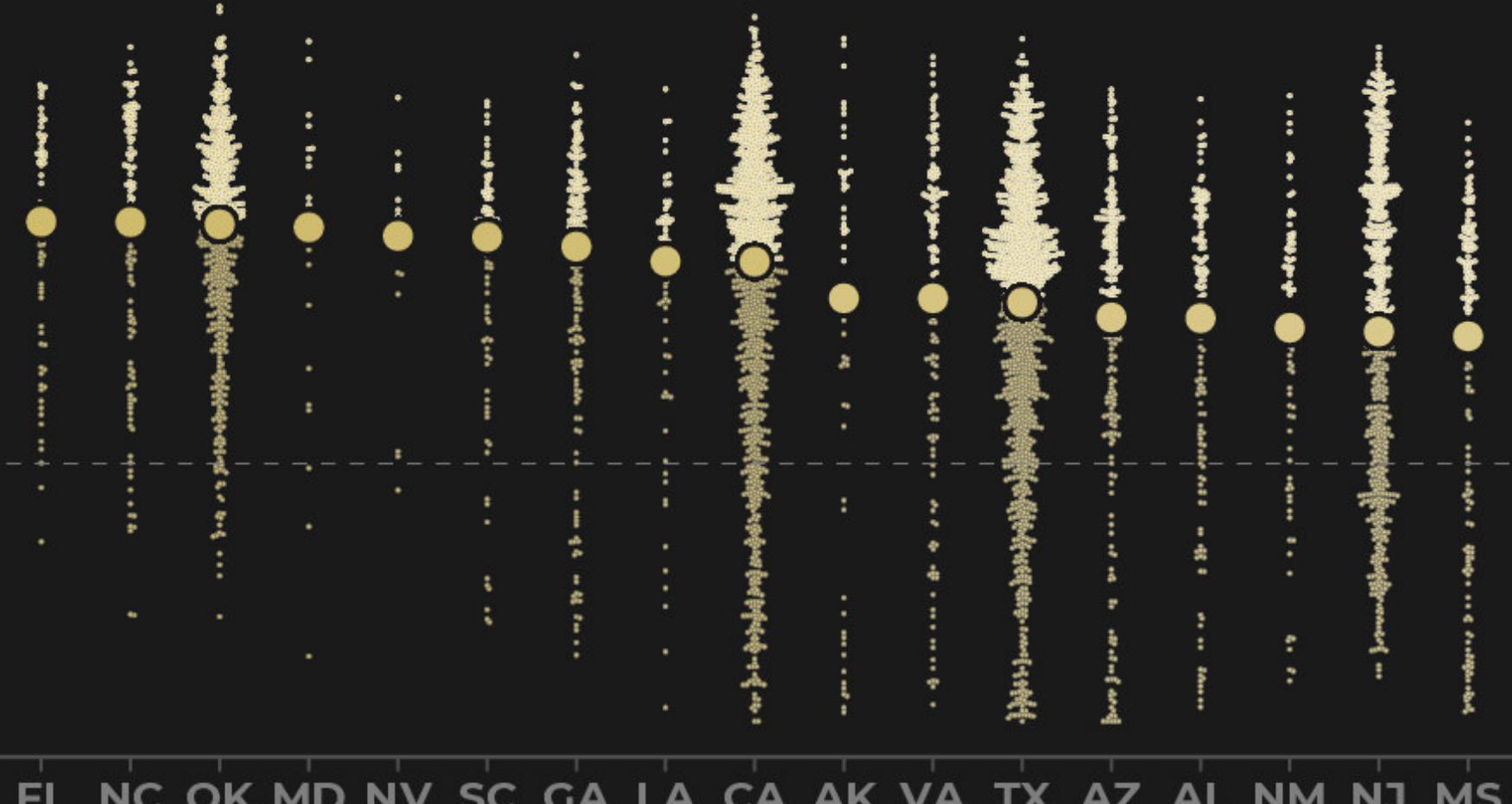
Larger dots represent each state's median diversity with darker colored points laying below and lighter colored points laying above this value.



Proportion of white students for all schools grouped per state, ranked from states with **few white students** to those with **many white students**.

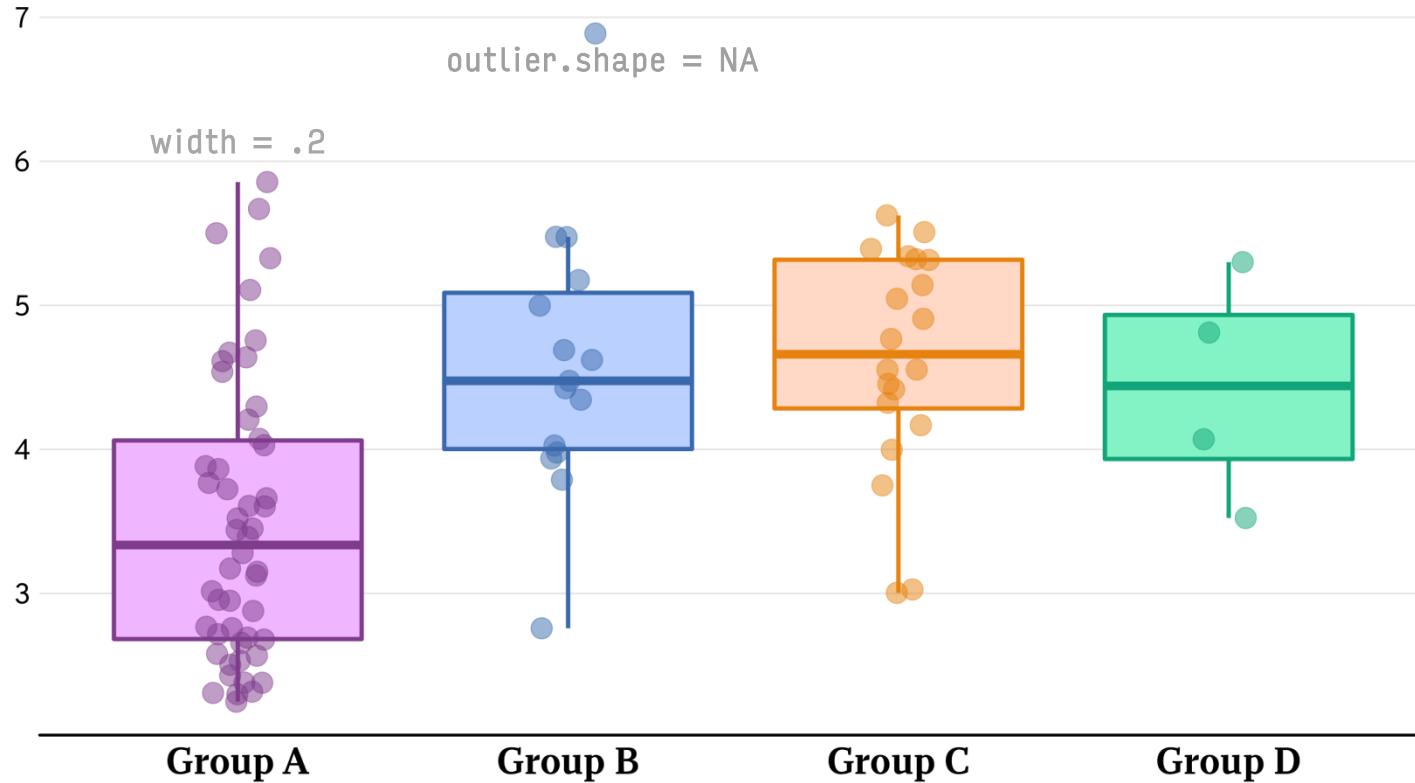
Larger dots represent each state's median proportion with lighter colored points laying below and darker colored points laying above this value.





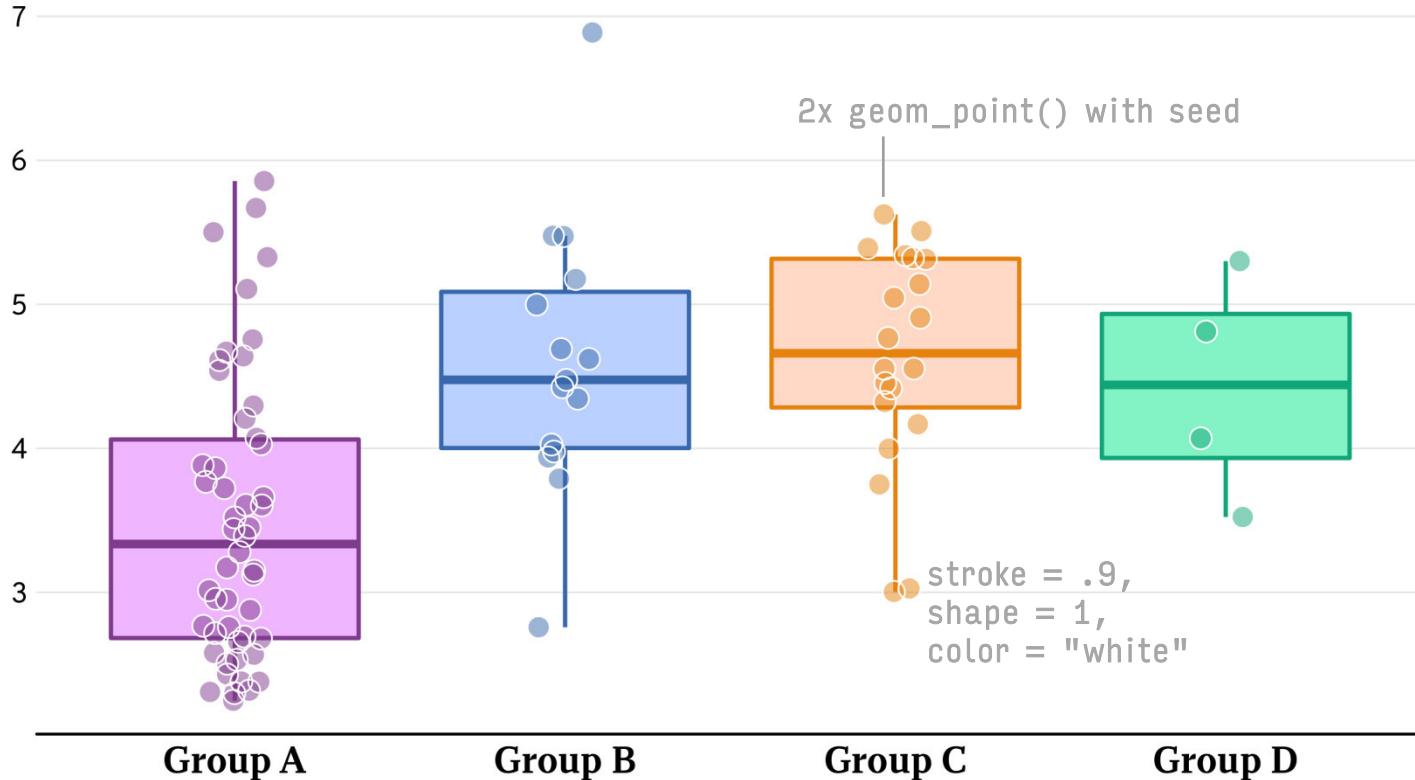
BOX PLOT X JITTER STRIPS

geom_boxplot() + geom_jitter()



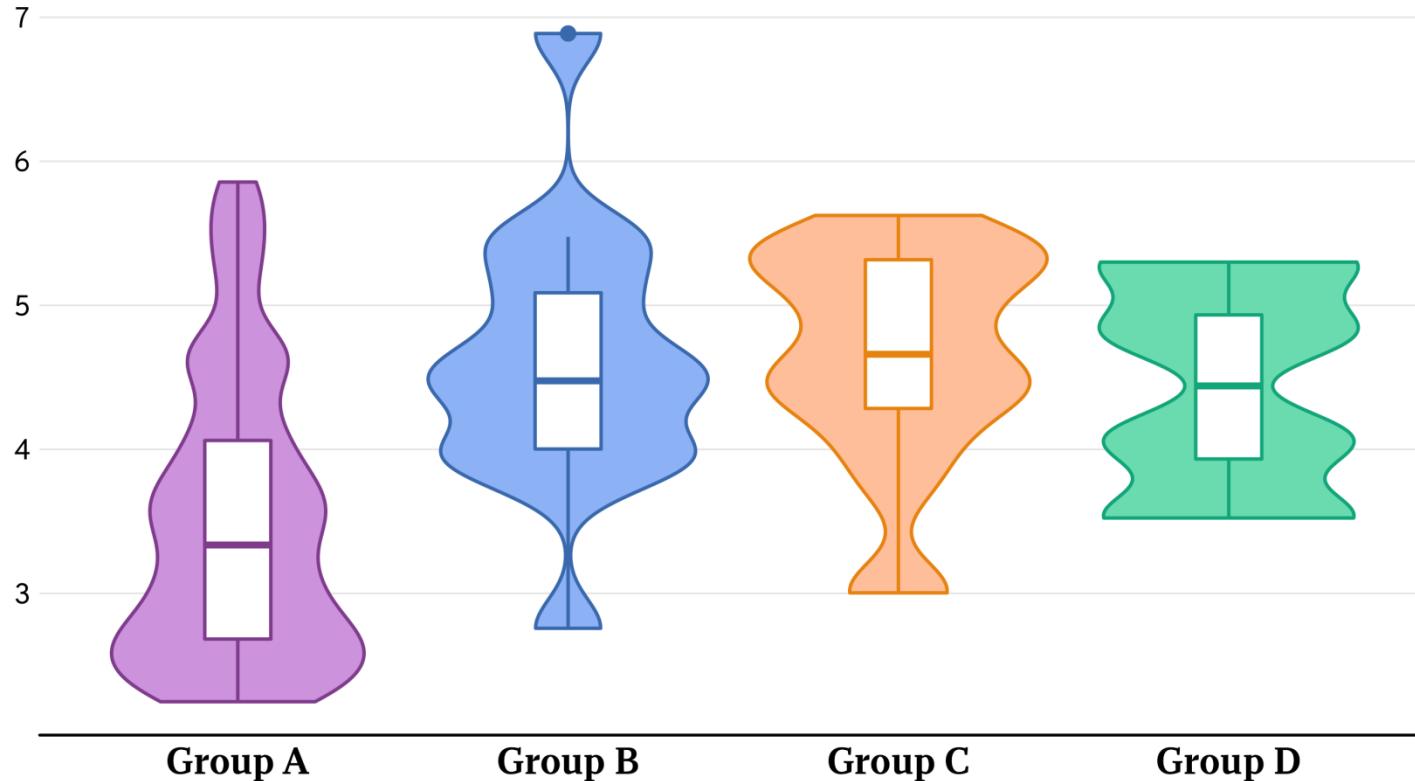
BOX PLOT X JITTER STRIPS

`geom_boxplot() + geom_point(position = position_jitter())`



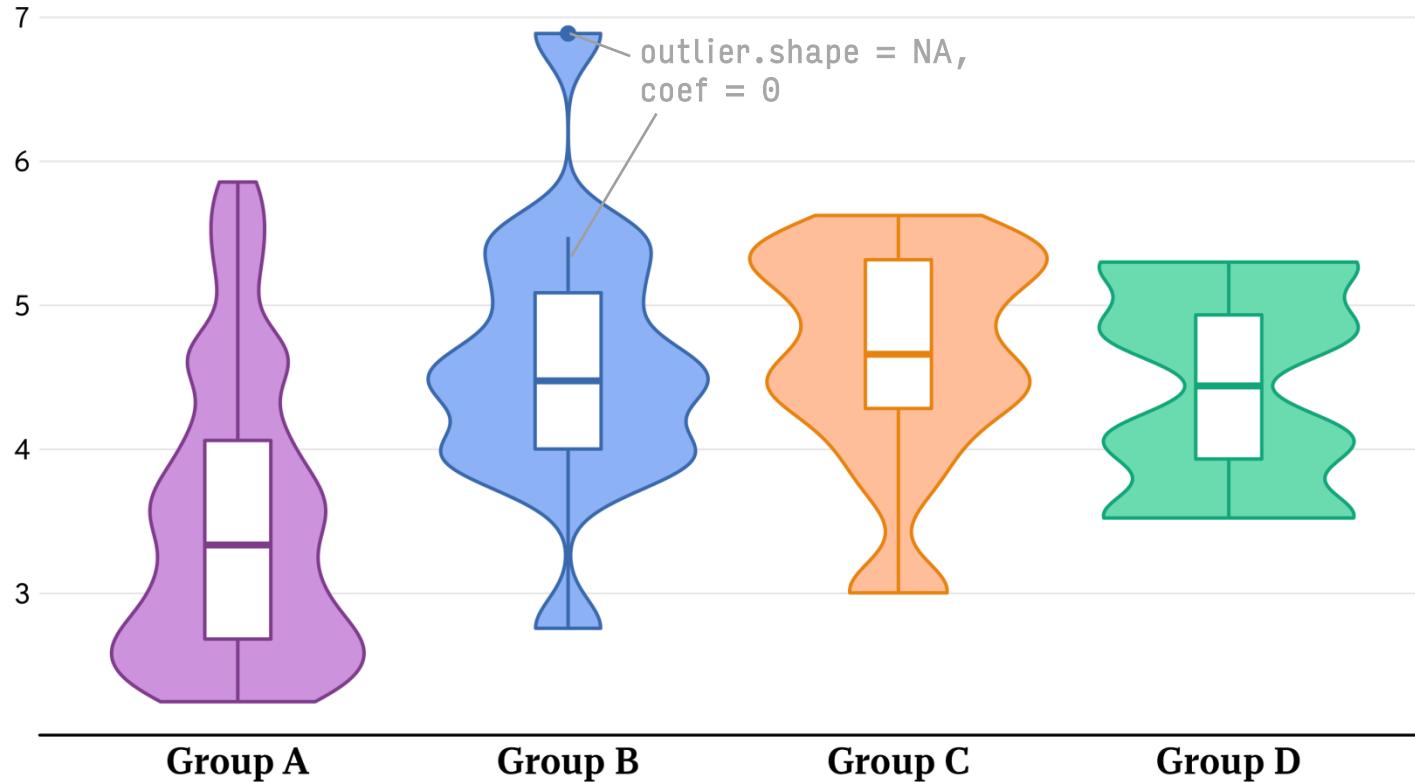
VIOLIN PLOT X BOX PLOT

geom_violin() + geom_boxplot()



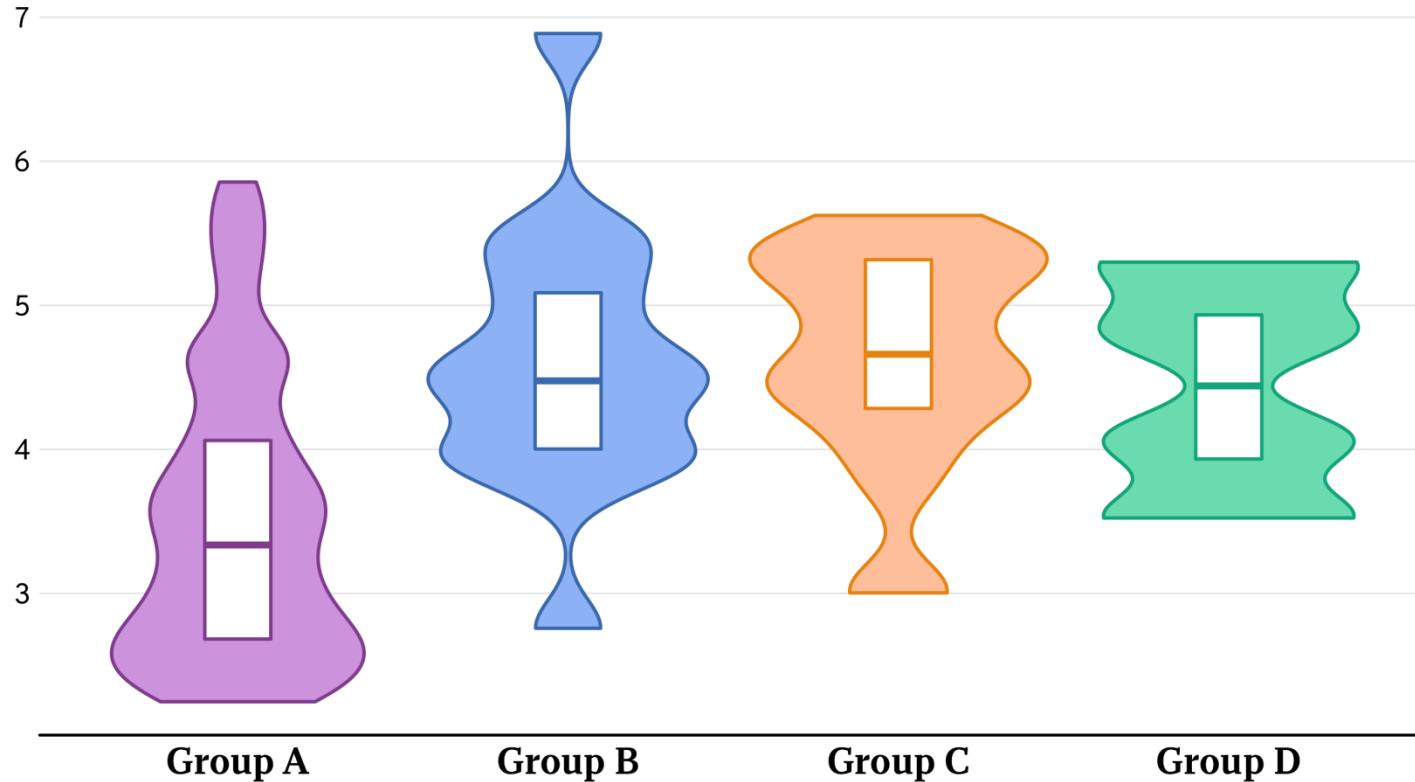
VIOLIN PLOT X BOX PLOT

geom_violin() + geom_boxplot()



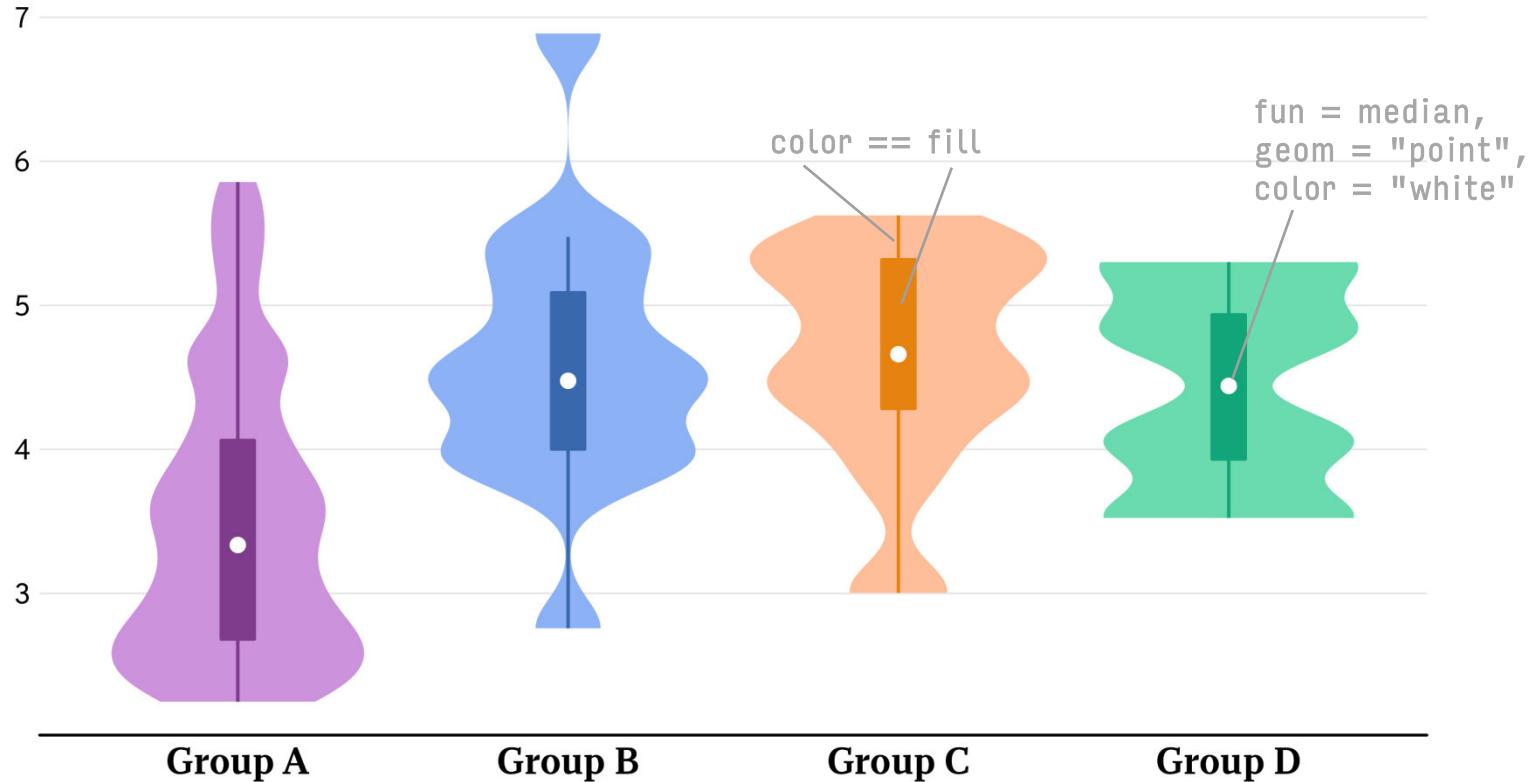
VIOLIN PLOT X BOX PLOT

geom_violin() + geom_boxplot()



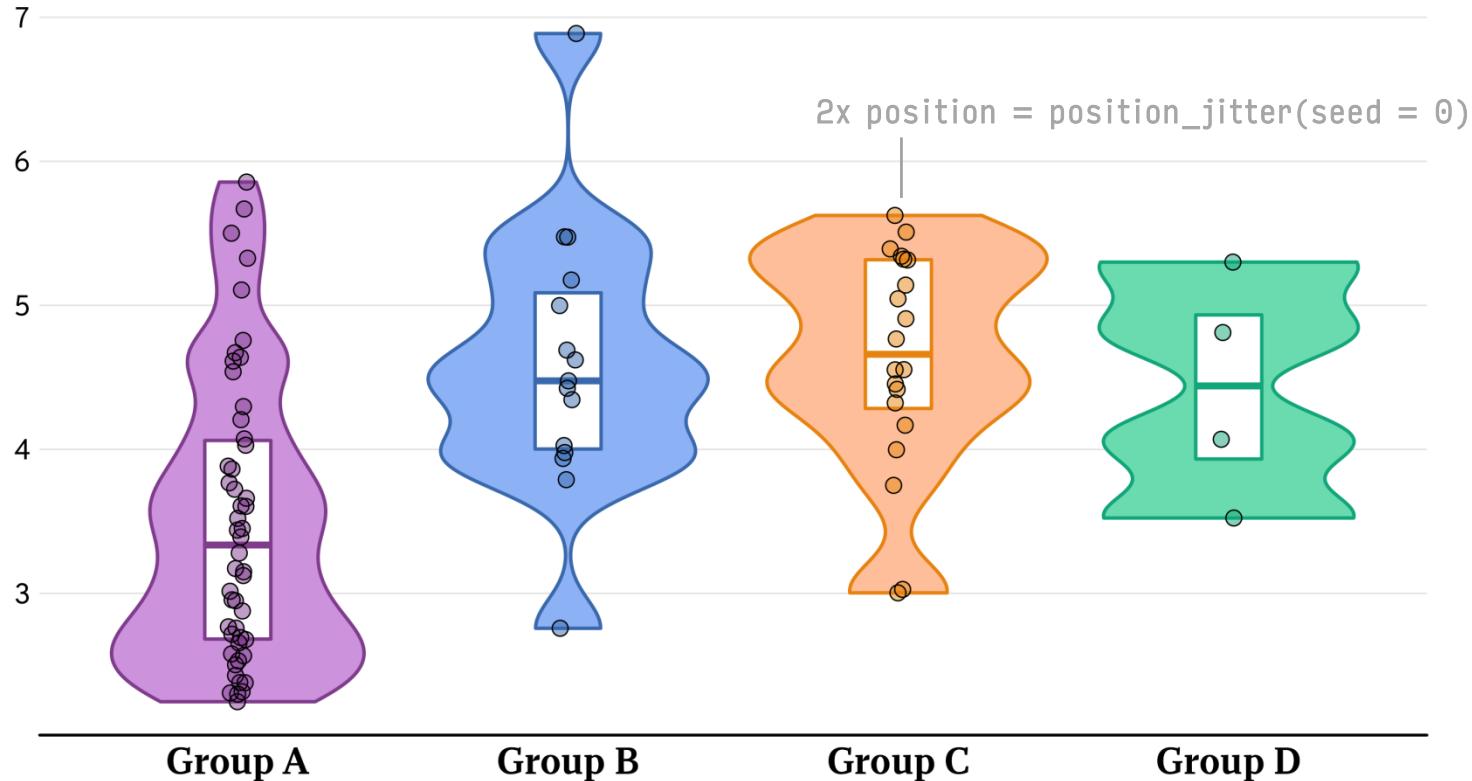
VIOLIN PLOT X BOX PLOT

geom_violin() + geom_boxplot() + stat_summary(geom = "point")



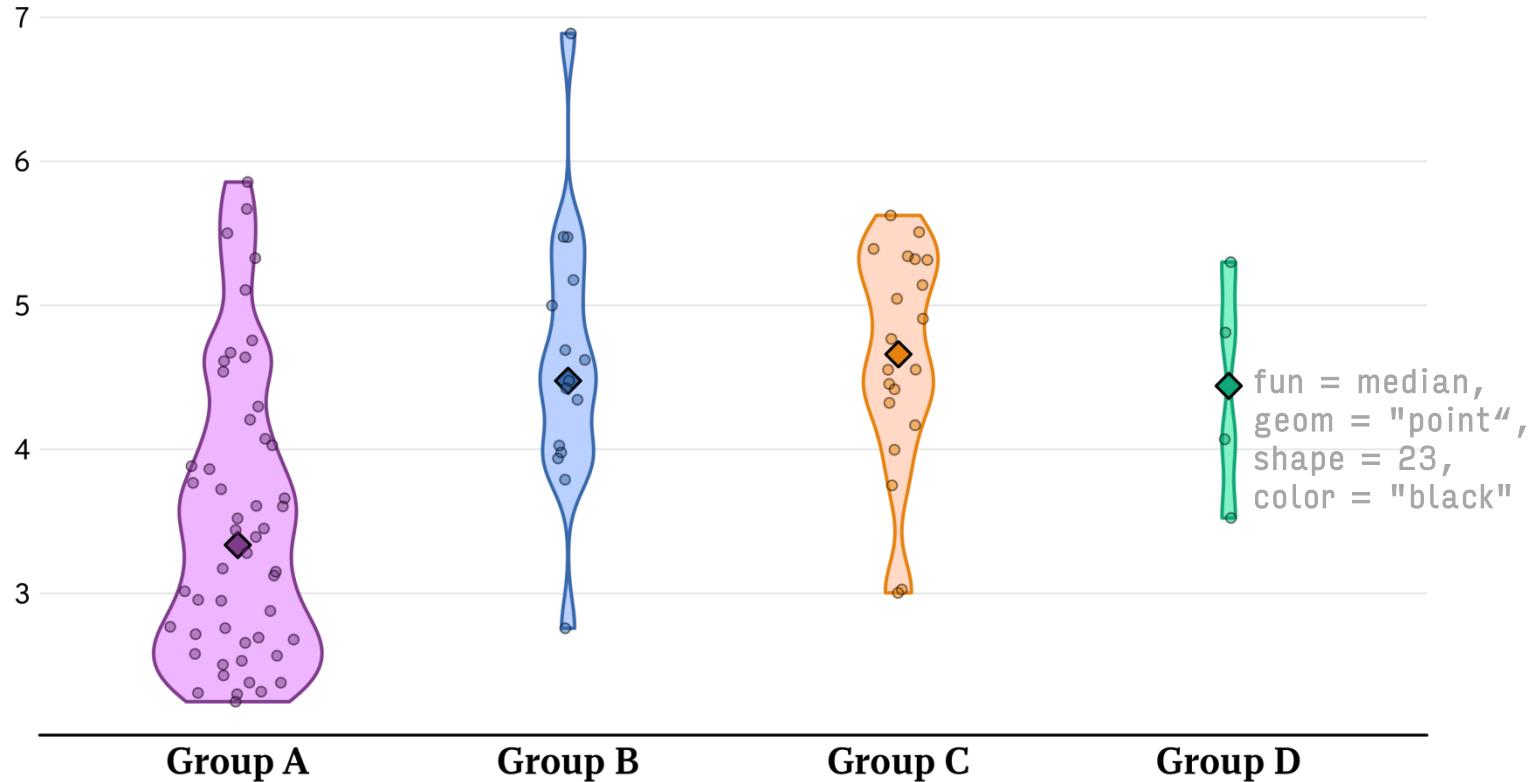
VIOLIN PLOT X BOX PLOT X JITTER STRIPS

geom_violin() + geom_boxplot() + geom_point()



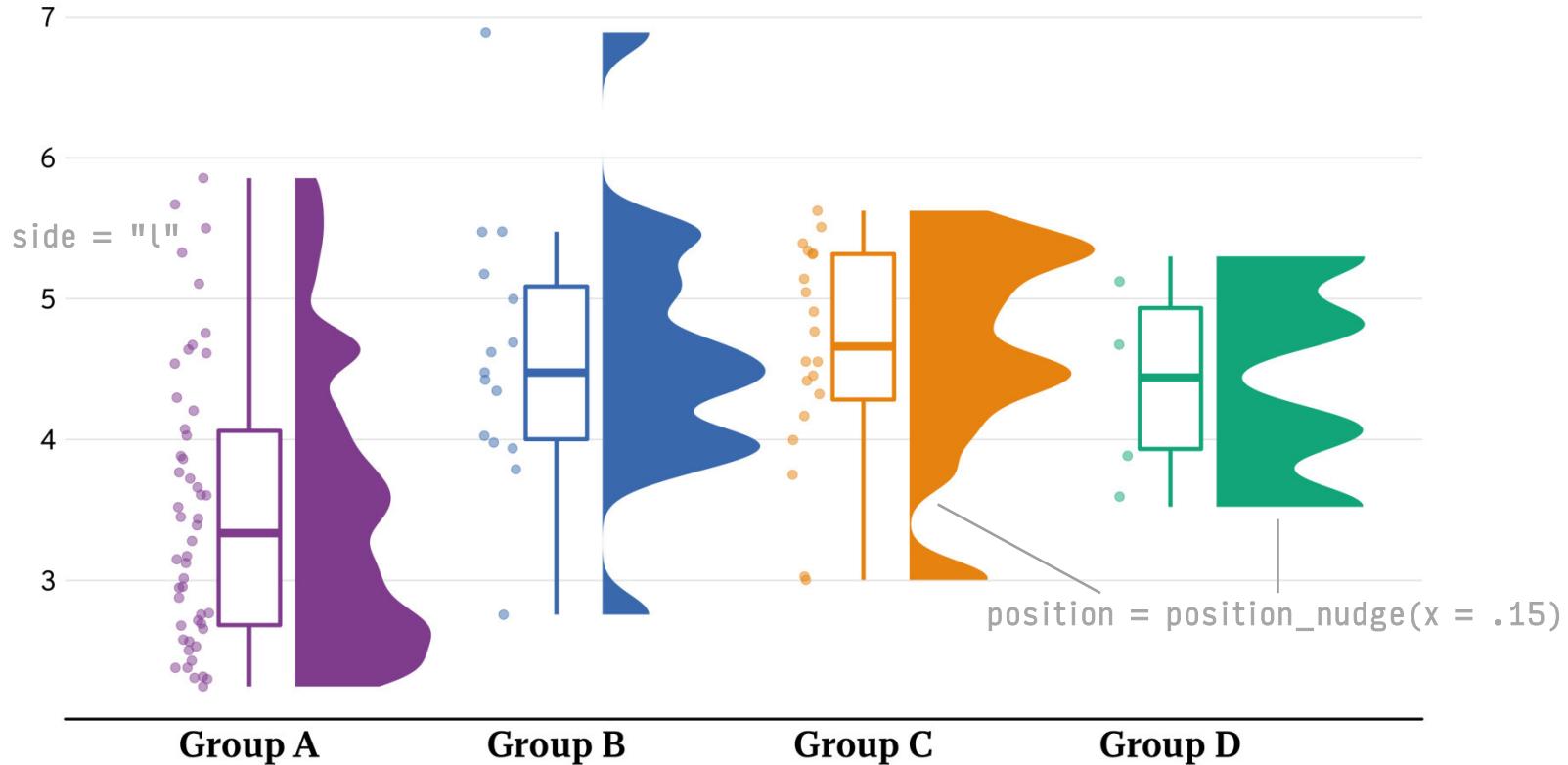
VIOLIN PLOT X SINA PLOT

geom_violin(scale = "count") + ggforce::geom_sina() + stat_summary()



RAINCLOUD PLOTS

gghalves::geom_half_point() + geom_boxplot() + ggdist::stat_halfeye()



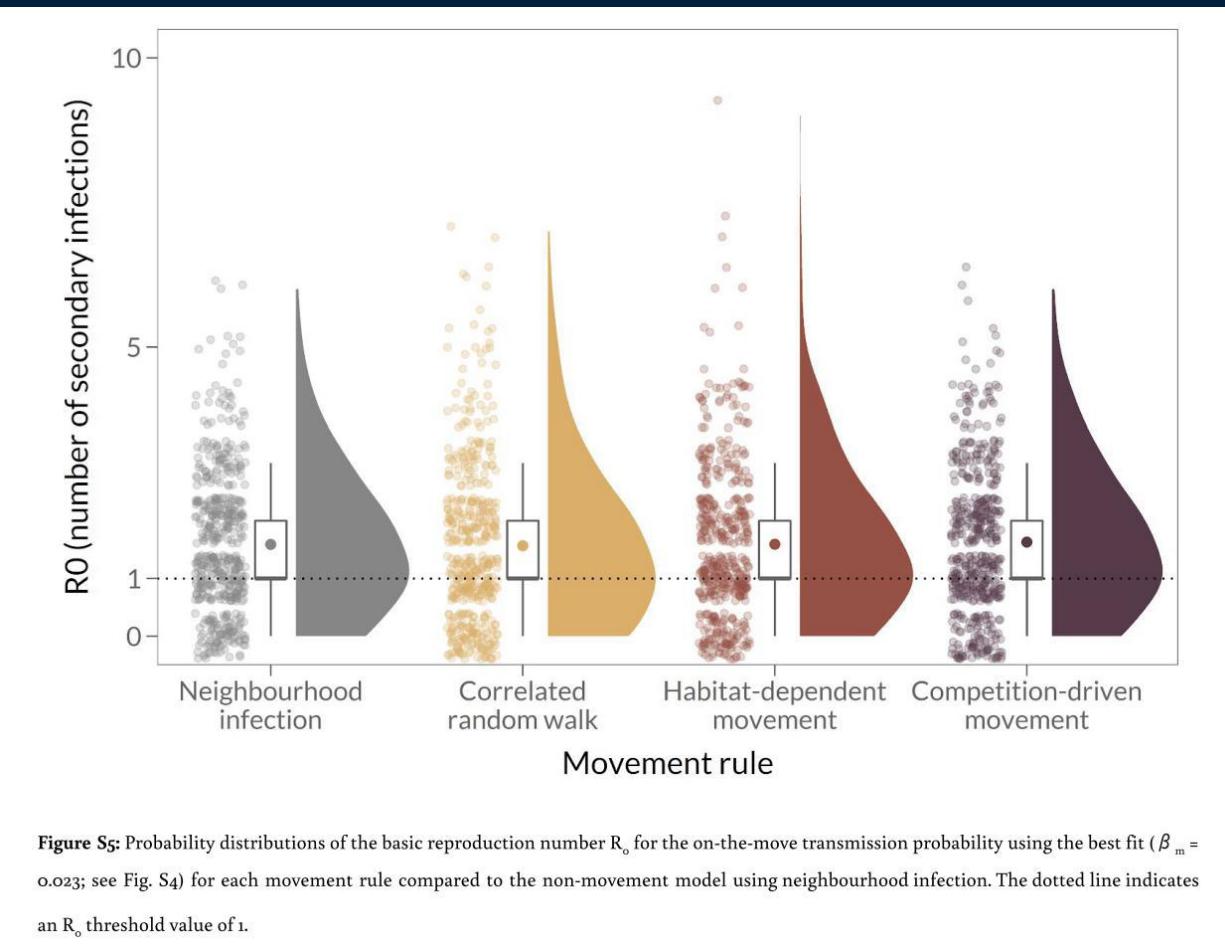
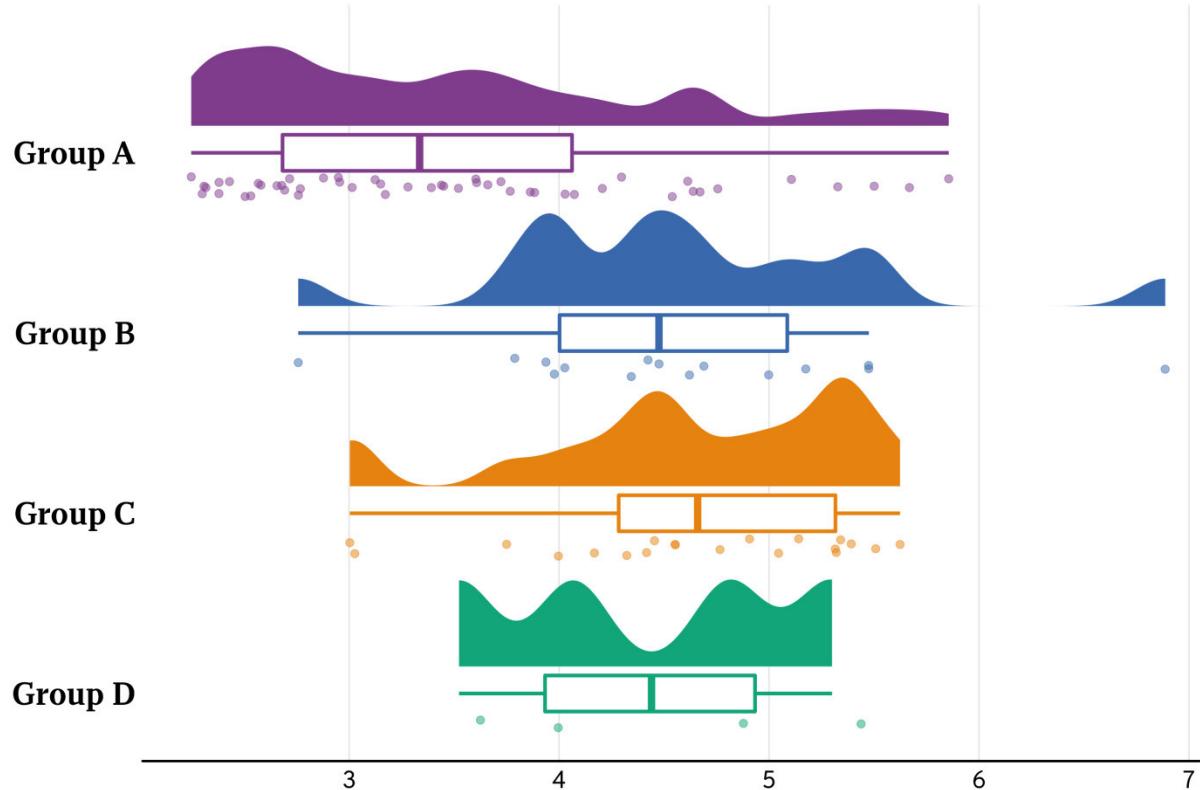


Figure S5: Probability distributions of the basic reproduction number R_0 for the on-the-move transmission probability using the best fit ($\beta_m = 0.023$; see Fig. S4) for each movement rule compared to the non-movement model using neighbourhood infection. The dotted line indicates an R_0 threshold value of 1.

RAINCLOUD PLOTS

gghalves::geom_half_point() + geom_boxplot() + ggdist::stat_halfeye()



Bill Ratios of Brush-Tailed Penguins (*Pygoscelis* spec.)

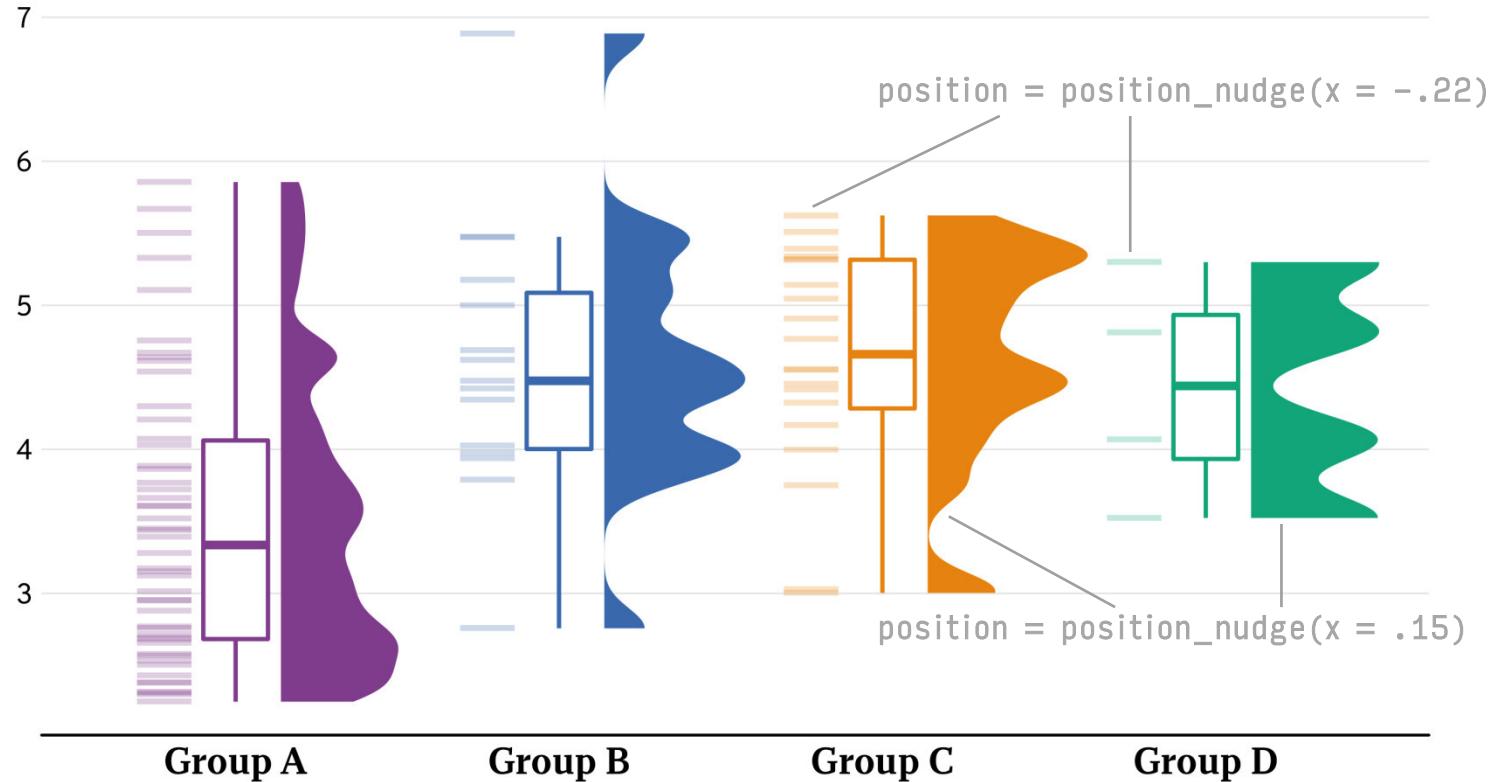
Distribution of bill ratios, estimated as bill length divided by bill depth.



Gorman, Williams & Fraser (2014) PLoS ONE DOI: [10.1371/journal.pone.0090081](https://doi.org/10.1371/journal.pone.0090081)
Visualization: Cédric Scherer • Illustration: Allison Horst

RAINCLOUD PLOTS WITH BARCODE STRIPS

`geom_point(shape = 95) + geom_boxplot() + ggdist::stat_halfeye()`

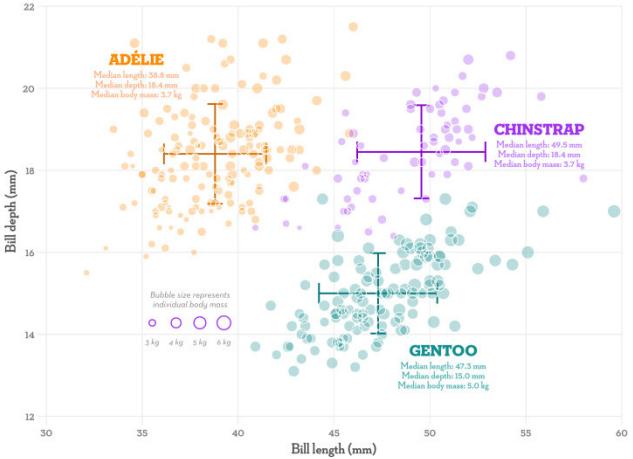


BILL DIMENSIONS OF BRUSH-TAILED PENGUINS

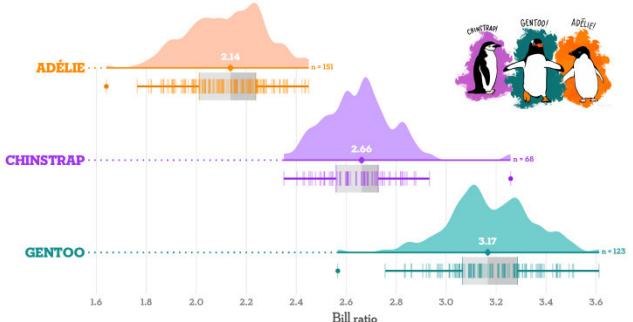
Pygoscelis adeliae (Adélie penguin) • *P. antarctica* (Chinstrap penguin) • *P. papua* (Gentoo penguin)



A. Scatterplot of bill length versus bill depth (median +/- sd)

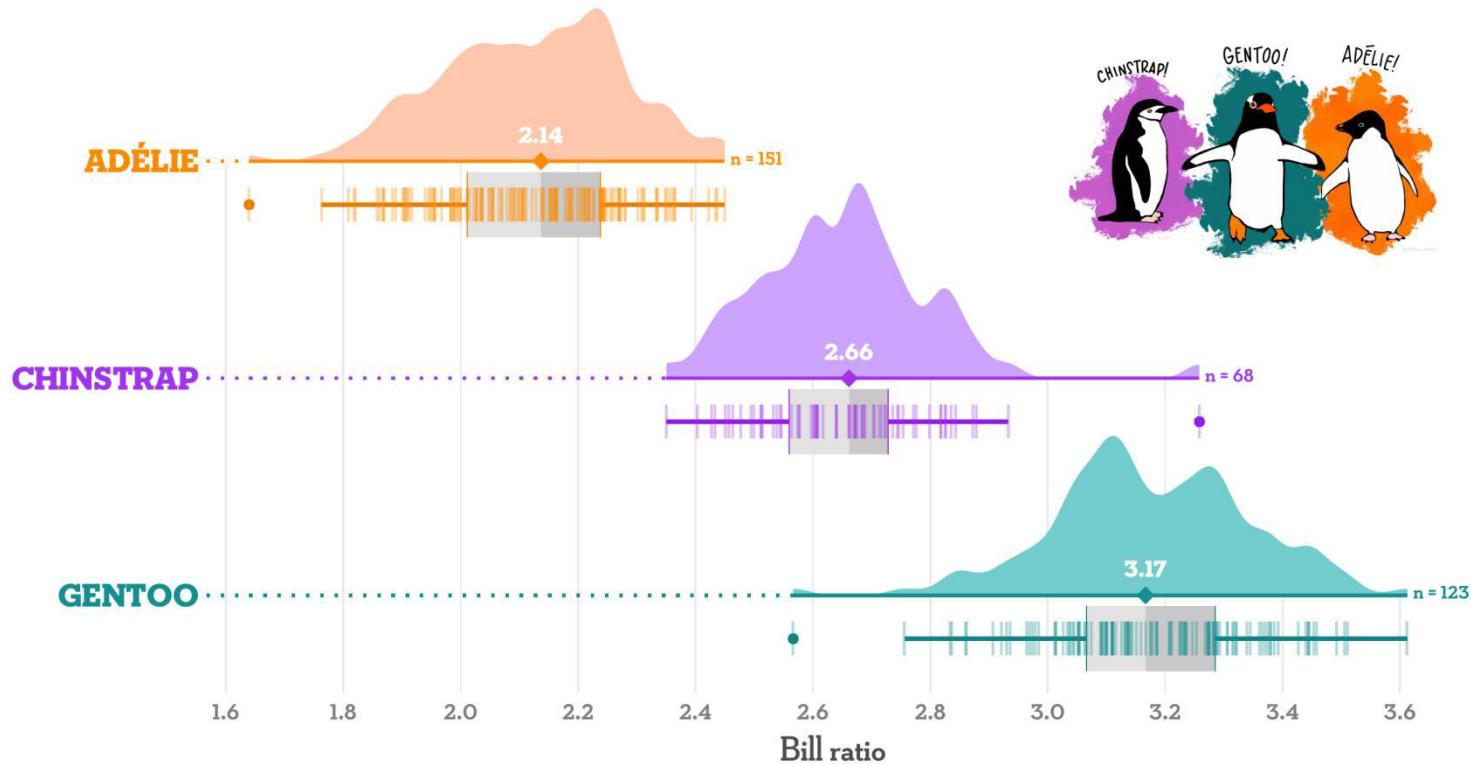


B. Distribution of the bill ratio, estimated as bill length divided by bill depth



Note: In the original data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal (upper) ridge of a bird's bill.
Visualisation Cédric Scherer + Data: Gorman, Williams & Fraser (2014) DOI: 10.1371/journal.pone.0090081 + Illustrations: Allison Horst

B. Distribution of the bill ratio, estimated as bill length divided by bill depth



Note: In the original data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal (upper) ridge of a bird's bill.

Visualization: Cédric Scherer • Data: Gorman, Williams & Fraser (2014) DOI: [10.1371/journal.pone.0090081](https://doi.org/10.1371/journal.pone.0090081) • Illustrations: Allison Horst

Not my cup of coffee...

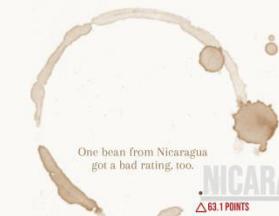
Each dot depicts one coffee bean rated by Coffee Quality Institute's trained reviewers. In addition, the multiple interval stripes show where 25%, 50%, 75%, and 100% of the beans fall along the rating gradient from 0 to 100 points. The rated coffee beans range from 59.8 points (Guatemala) to 89.9 (Ethiopia). Only countries of origin with 25 or more tested beans are shown. The red empty triangle marks the minimum rating, the black filled triangle indicates each country's median score.

Visualization by Cédric Scherer



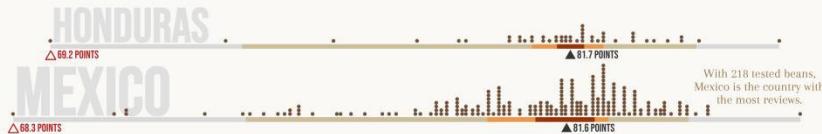
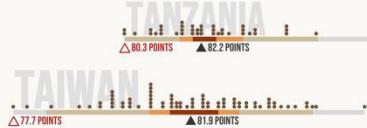
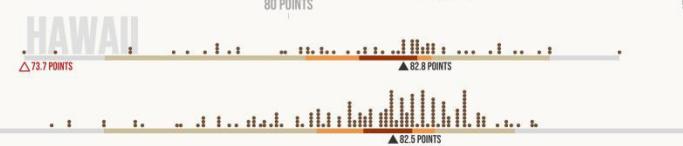
GUATEMALA

The coffee bean with the lowest rating has its origin in Guatemala.

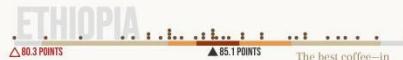


One bean from Nicaragua
got a bad rating, too.

NICARAGUA
▲ 63.1 POINTS



With 218 tested beans,
Mexico is the country with
the most reviews.



The best coffee—in terms of both median and maximum rating—is shipped to you from Ethiopia!

BRAZIL

△ 73.2 POINTS

Each dot depicts one coffee bean rated by Coffee Quality Institute's trained reviewers. In addition, the multiple interval stripes show where 25%, 50%, 95%, and 100% of the beans fall along the rating gradient from 0 to 100 points.

▲ 82.4 POINTS

HONDURAS

△ 69.2 POINTS

MEXICO

△ 68.3 POINTS

With 218 tested beans,
Mexico is the country with
the most reviews.

TAIWAN

△ 77.7 POINTS

▲ 81.9 POINTS

▲ 81.7 POINTS

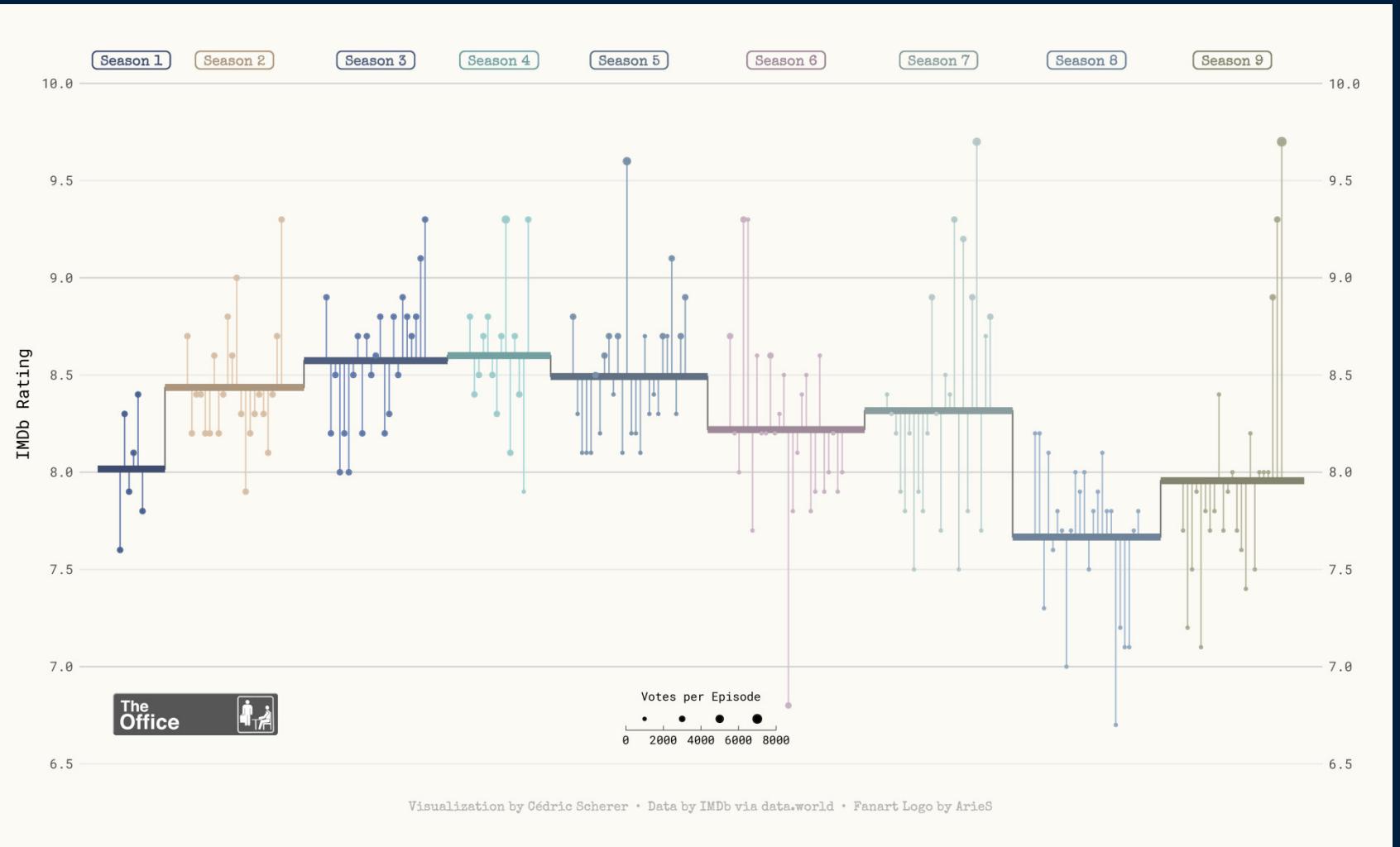
▲ 81.6 POINTS

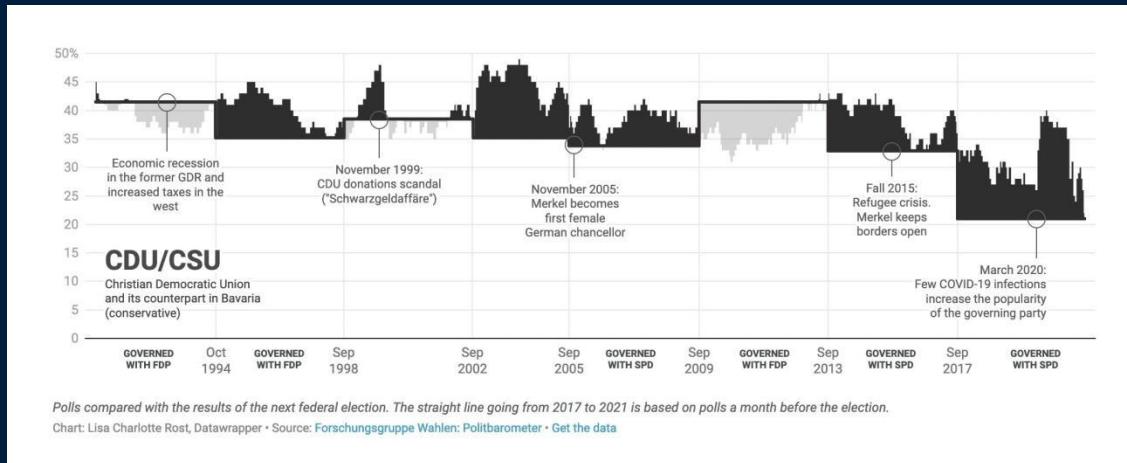
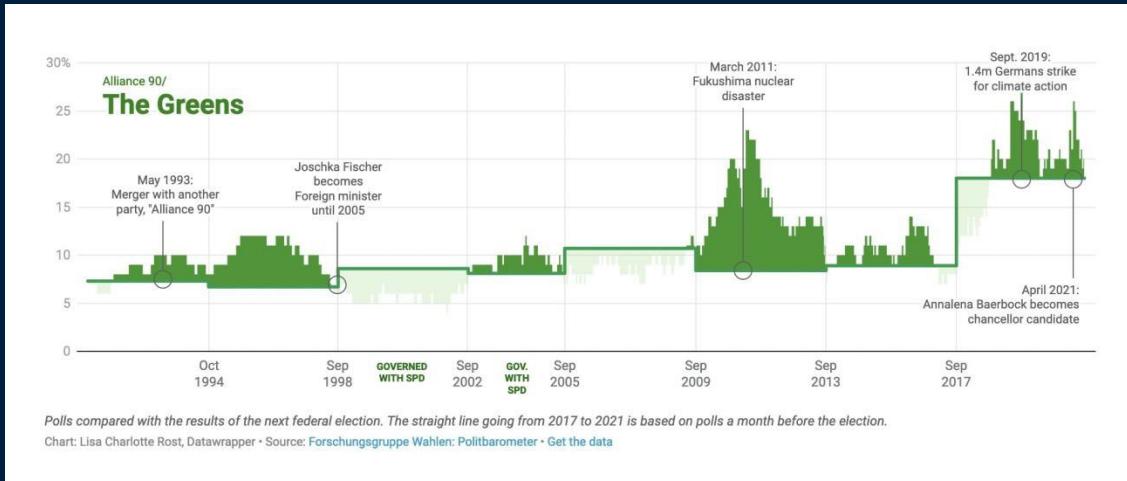
TANZANIA

△ 80.3 POINTS

▲ 82.2 POINTS

SOMETHING
Special





"Which German party is the most unlucky when it comes to election dates?" by Lisa C. Muth (DataWrapper)

"ANIMAL CROSSING" SEEMS TO BE FUN, NO MATTER IF YOU LIKE IT OR NOT

Or, more likely, users rating it low wrote things like "not much fun", "no fun at all" or "A fun game that I really wanted to play but what a **** multiplayer mode."



Low Grades (0-2)



Medium Grades (3-7)



High Grades (8-10)

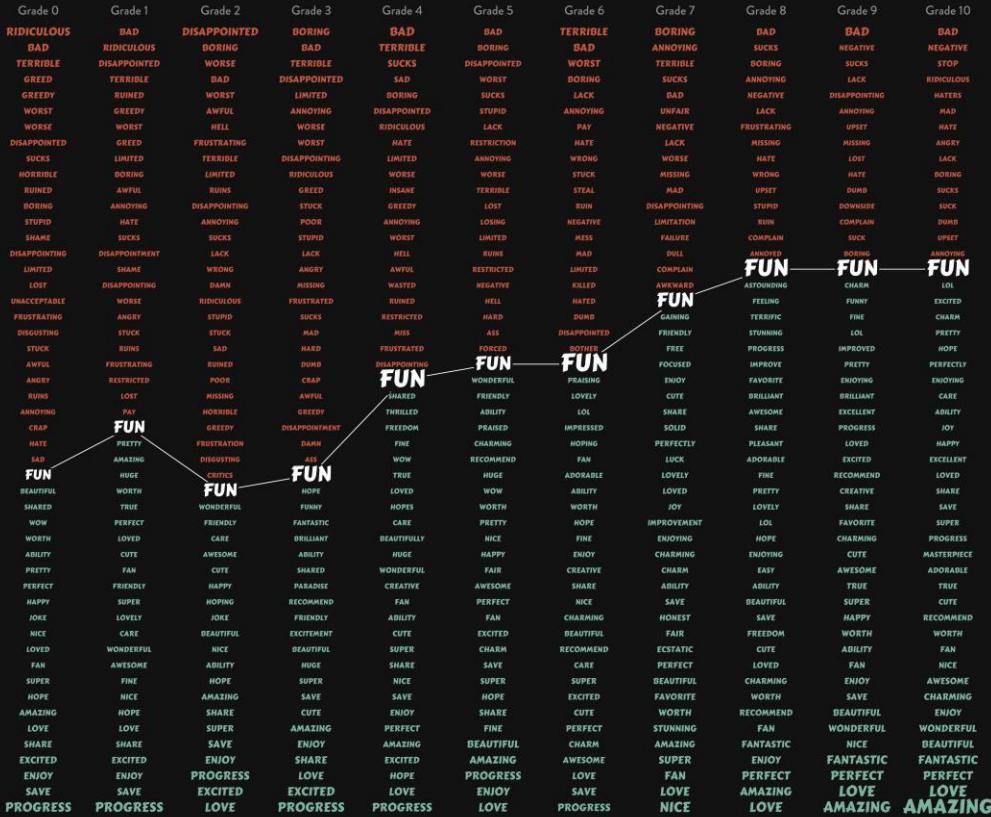
Based on a sentiment analysis of "Animal Crossing" user reviews, each wordcloud shows the 50 words that contributed the most to each grade category, either in a positive or in a **negative** way.

Visualization by Cédric Scherer • Data by Metacritic

"ANIMAL CROSSING" SEEMS TO BE FUN, NO MATTER IF YOU LIKE IT OR NOT

Or, more likely, users rating it low wrote things like "not much fun", "no fun at all" or "A fun game that I really wanted to play but what a **** multiplayer mode.

Based on a sentiment analysis of user reviews, each "word stripe" shows the 50 words that contributed the most to each grade category, either in a positive or in a negative way. The size of each word indicates its contribution per grade. The amount of words associated with positive sentiment increase, as we would expect, with higher grades. Fun, even though categorized as positive, is often used in both contexts.



CÉDRIC SCHERER

Data Visualization & Computational Ecology



A QUICK HOW-TO ON LABELLING BAR GRAPHS IN GGPLOT2

Bar charts are likely the most common chart type out there and come in several varieties. Most notably, Direct labels can increase accessibility of a bar graph. I got a request how one can add percentage labels inside the bars and how to highlight specific bars with `ggplot2`. This short tutorial shows you multiple ways how to do so.

POSTED BY CÉDRIC MONDAY, JULY 5, 2020

COLORS AND EMOTIONS IN DATA VISUALIZATION

As data visualization practitioners we are both engineers and designers. In our effort to create meaningful visualizations from our data, we transmit our message not only through the particular chart type and title we choose, but also through our choice of colors for the data itself and for any additional design elements.

POSTED BY CÉDRIC TUESDAY, JUNE 8, 2020

VISUALIZING DISTRIBUTIONS WITH RAINCLOUD PLOTS (AND HOW TO CREATE THEM WITH GGPLOT2)

Raincloud plots, that provide an overview of the raw data, its distribution, and important statistical properties, are a good alternative to classical boxplots. In this tutorial, I highlight the potential problem of boxplots, illustrate why raincloud plots are great, and show numerous ways how to create such hybrid charts in R with `ggplot2`.

POSTED BY CÉDRIC SUNDAY, JUNE 8, 2020

MY CONTRIBUTIONS TO THE FIRST #30DAYCHARTCHALLENGE

This April, Dominic Royé and I hosted the first `#30DayCharChallenge`, a data visualization challenge with the aim to create a chart every day of April with a given prompt. In total, we collected 1,960 contributions from around the world!

POSTED BY CÉDRIC SUNDAY, MAY 9, 2020

MY PERSONAL DATA VISUALIZATION YEAR 2020

Even though it was a crazy and exhausting year, there were also some good and exiting things happening. Therefore I've decided to take a short break on New Year's Day and look back at some of the positive moments of my personal data visualization journey during 2020.

POSTED BY CÉDRIC FRIDAY, JANUARY 1, 2021

THANK YOU!

cedricscherer.com
[@CedScherer](https://twitter.com/CedScherer)
github.com/z3tt

Buy me a coffee
buymeacoffee.com/z3tt