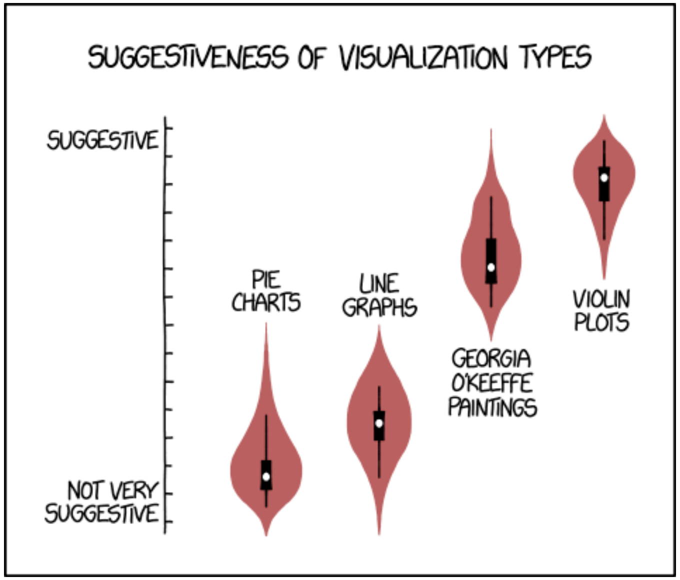
Introduction to Data Visualization

Cumulative Distribution Functions, Q-Q Plots
Many Distributions At Once, Proportions



Halil Bisgin, Ph.D.

Image: https://imgs.xkcd.com/comics/violin_plots.png

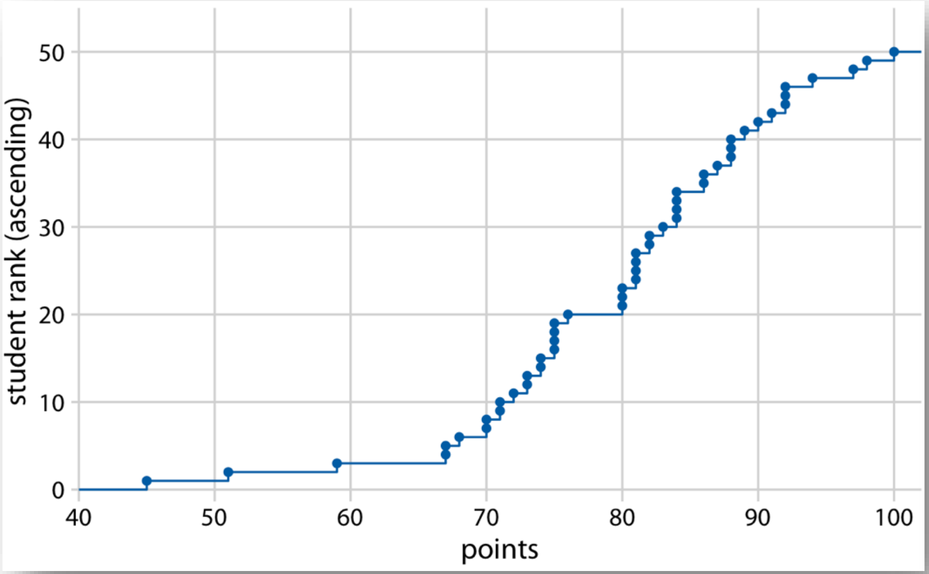
Empirical Cumulative Distribution Functions and Q-Q Plots

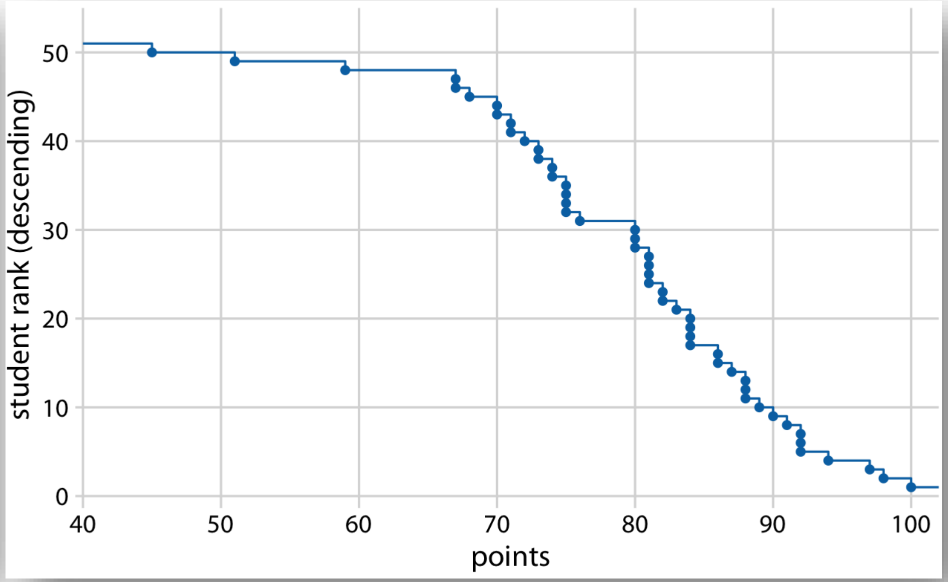
- Histograms and density plots are intuitive and appealing, but rely on parameters and not direct visualization of the data.
- We could plot all points, but unwieldy for large sets.
- More accurate and technical, but less intuitive way:
  - *empirical cumulative distribution functions (ECDFs)*
  - *quantile-quantile (q-q) plots*

Cumulative Distribution Functions

- Empirical cumulative distribution function of student grades for a hypothetical class of 50 students.
  - *Rank all students by the number of points they obtained, in*

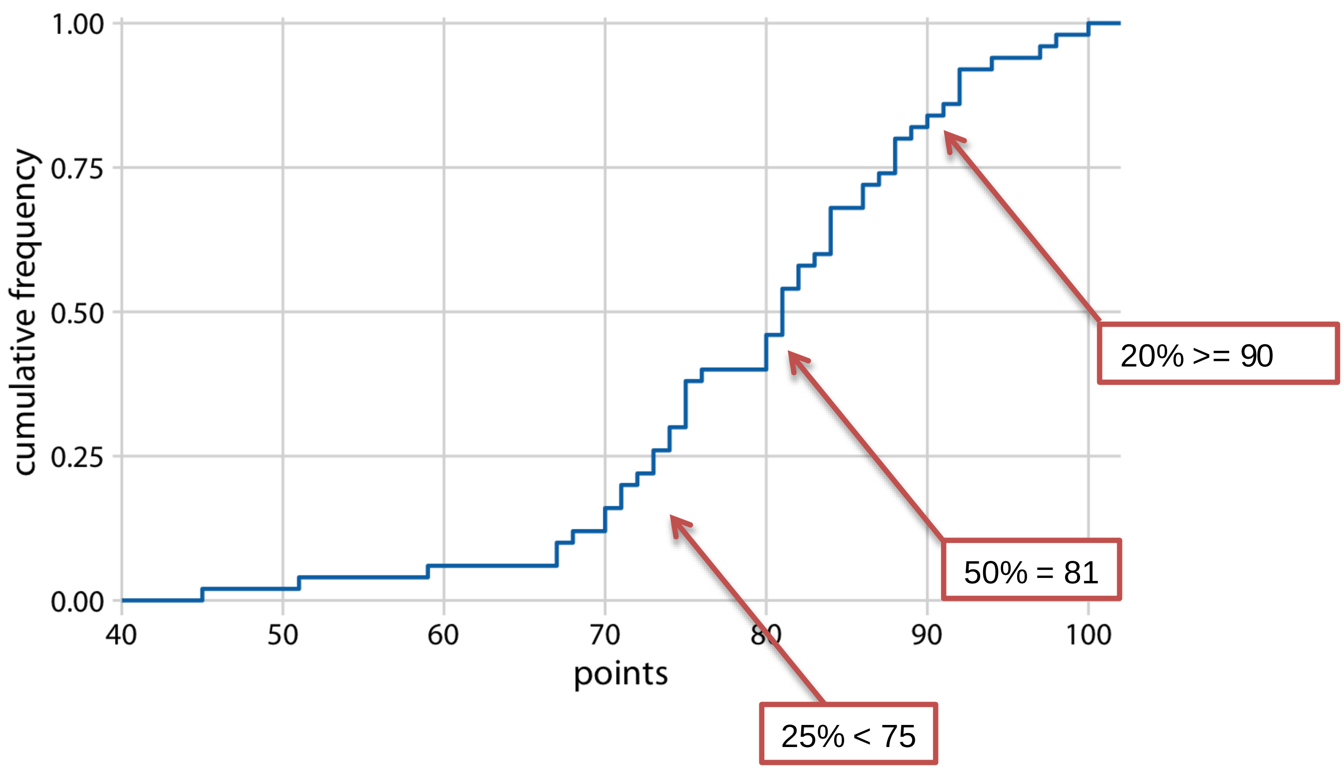
ascending order *or* descending order





Cumulative Distribution Functions

- Normalize the ranks by the maximum rank, so that the y axis represents the cumulative frequency.

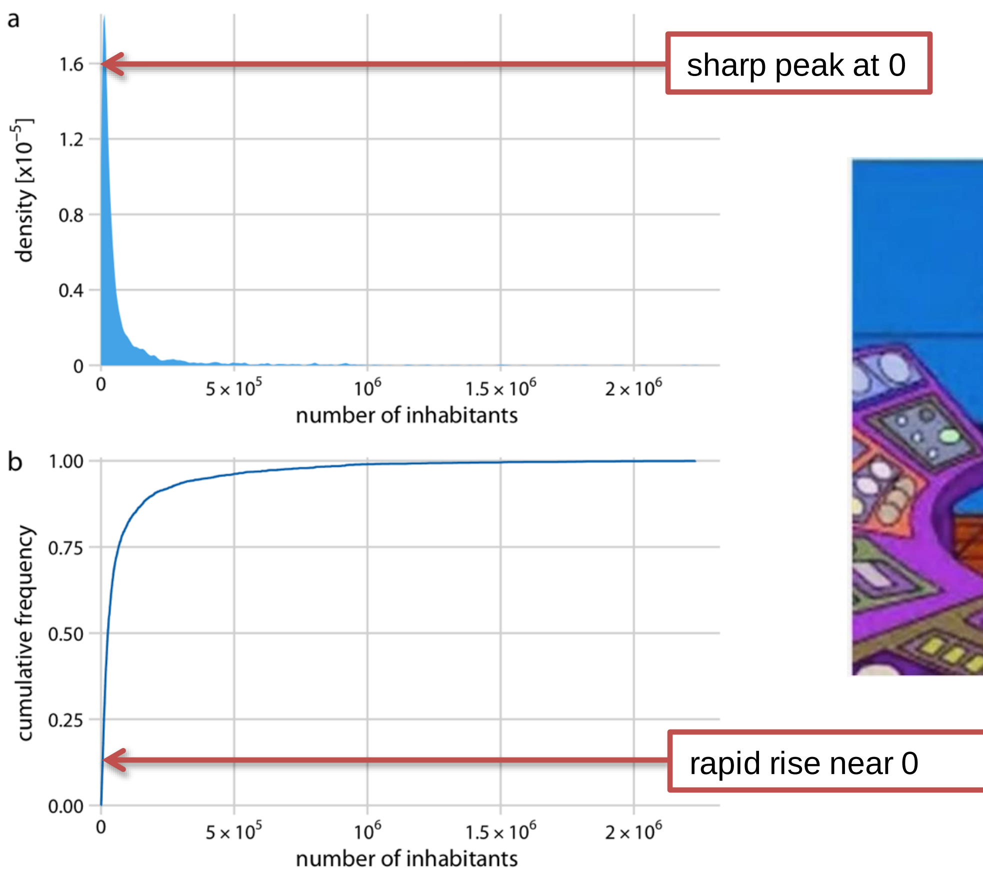


Highly Skewed Distributions

- Many empirical datasets display highly skewed distributions, in particular with heavy tails to the right, and these distributions can be challenging to visualize.
  - *the number of people living in different cities or counties,*
  - *the number of contacts in a social network,*
  - *the frequency with which individual words appear in a book*
- Right tail decays slower than an exponential function, meaning that very large values are not that rare
- Power Law

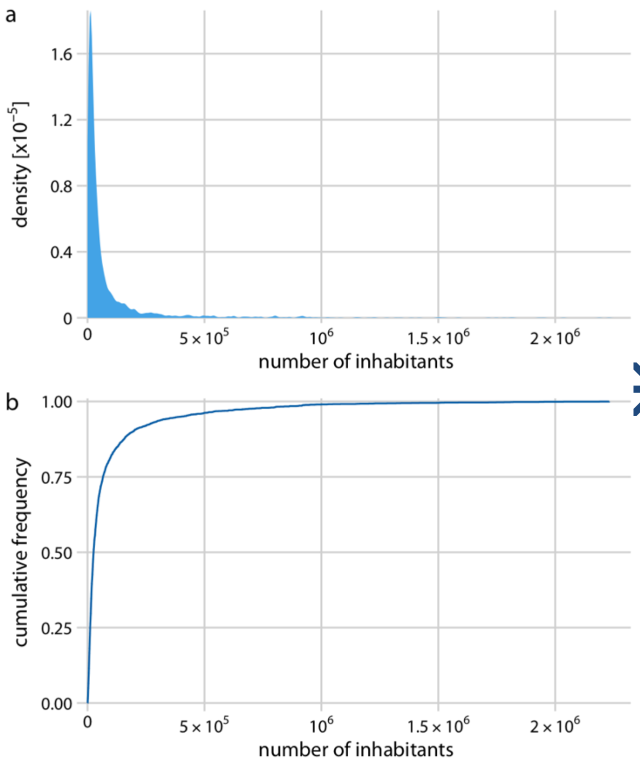
Highly Skewed Distributions

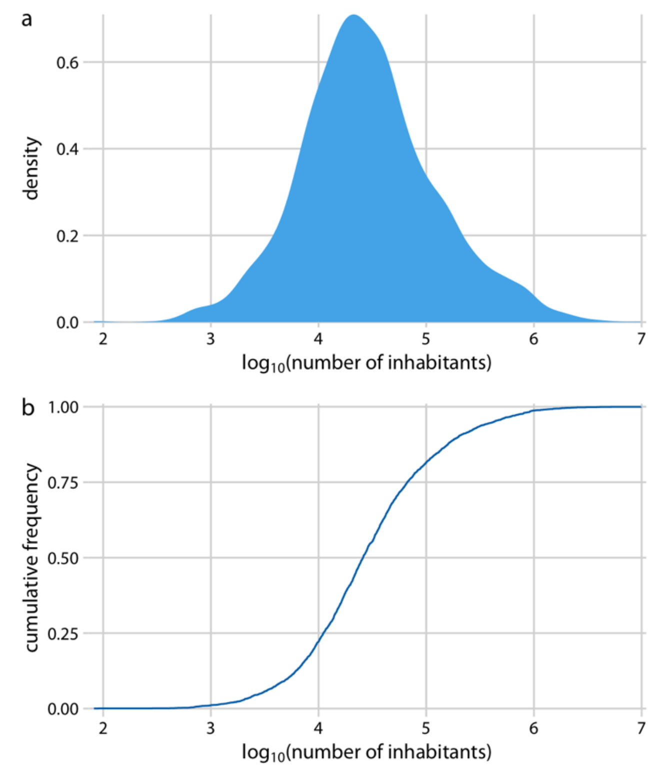
- The number of people living in different US counties according to the 2010 US Census.



Highly Skewed Distributions

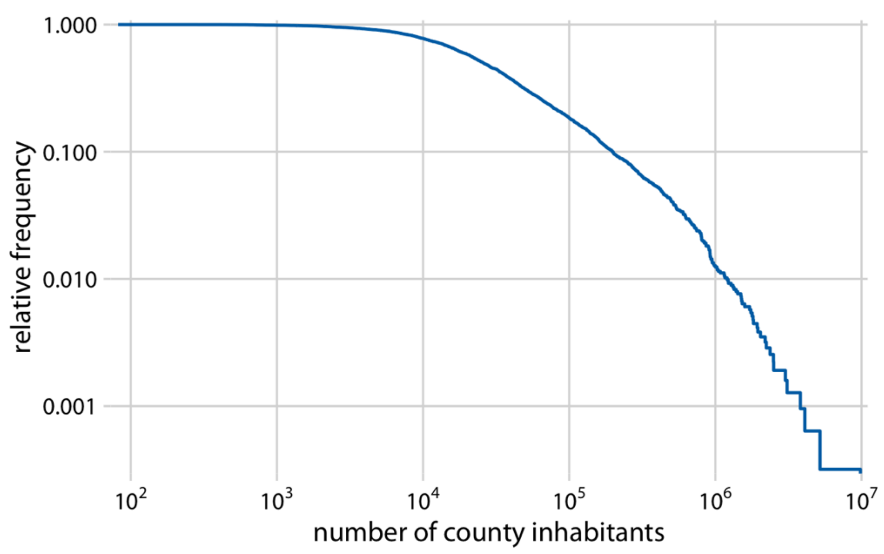
- Log transformation offers more interpretable charts.

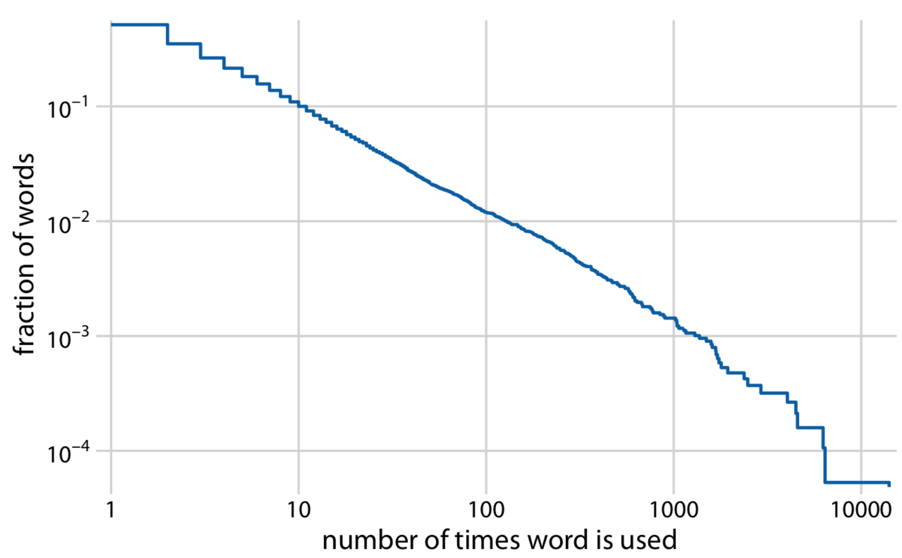




Highly Skewed Distributions

- To comply with power law, log-log of the descending ECD should be a straight line.



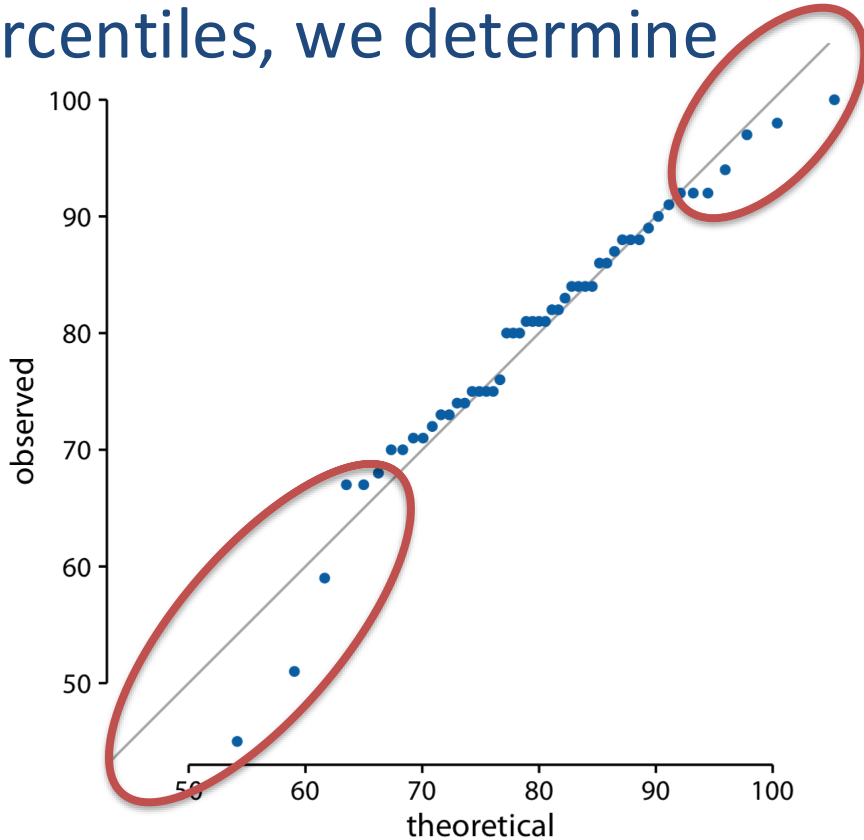


Quantile-Quantile Plots

- Quantile-quantile (q-q) plots are useful when we want to determine to what extent the observed data points follow a given distribution.
- Visualizes the relationship between ranks and actual values.
- We don't plot the ranks directly; rather, we use them to predict where a given data point would fall if the data were distributed according to a specified reference distribution.

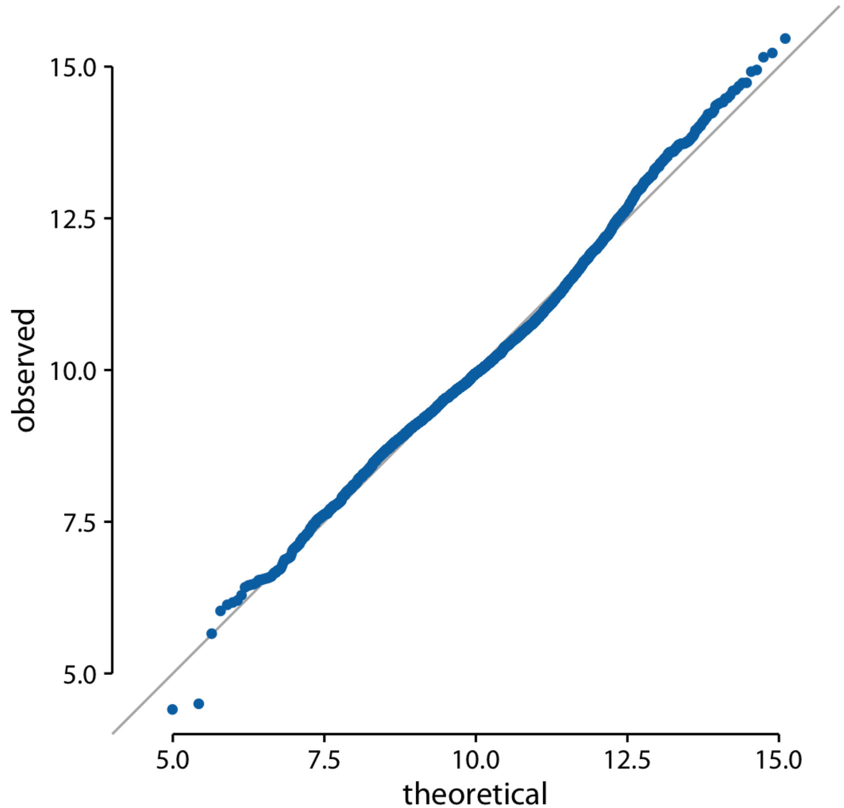
Quantile-Quantile Plots

- Suppose there are values with (μ=10, σ=3) and compare with normal distribution.
- Depending on theoretical percentiles, we determine the positions for the values.
- Then we plot their observed and theoretical positions.
- The more the points on the diagonal line, the more the distributions agree.



Quantile-Quantile Plots

- Do the population counts in US counties follow a log-normal distribution.

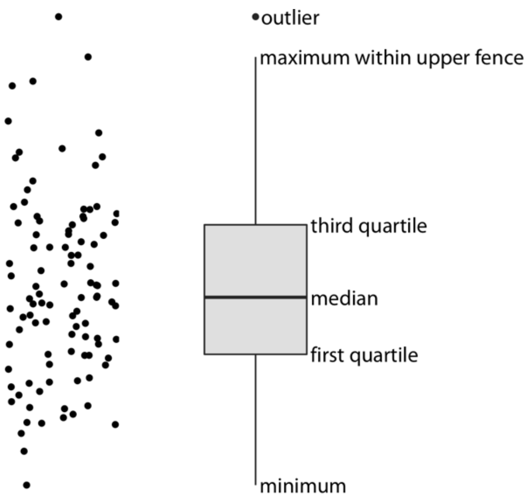


Visualizing Many Distributions at Once

- It is helpful to think in terms of the response variable and one or more grouping variables.
  - *Response variable is the one whose distributions we want to show.*
  - *The grouping variables define subsets of the data with distinct distributions of the response variable.*
    - *For temperature distributions across months, the response variable is the temperature and the grouping variable is the month.*
- The response variable can be along the vertical axis, and the horizontal axis.

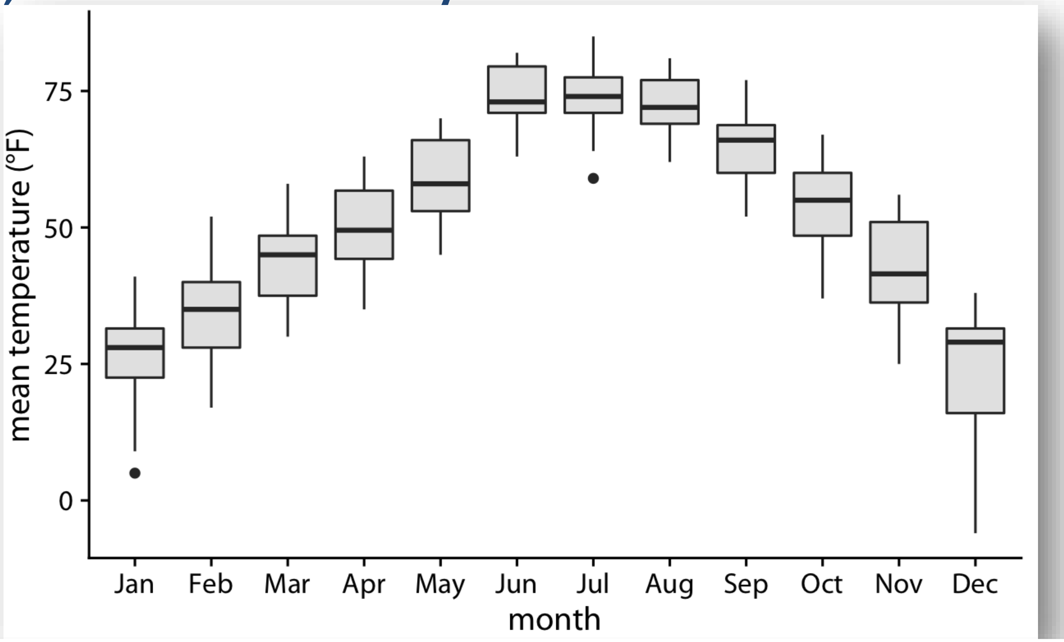
Along the Vertical Axis-Box Plot

- A boxplot divides the data into quartiles and visualizes them in a standardized manner.
- The vertical lines extending upwards and downwards from the box are called whiskers.
- The distances of 1.5 times the height of the box in either direction are called the upper and lower fences.



Along the Vertical Axis-Box Plot

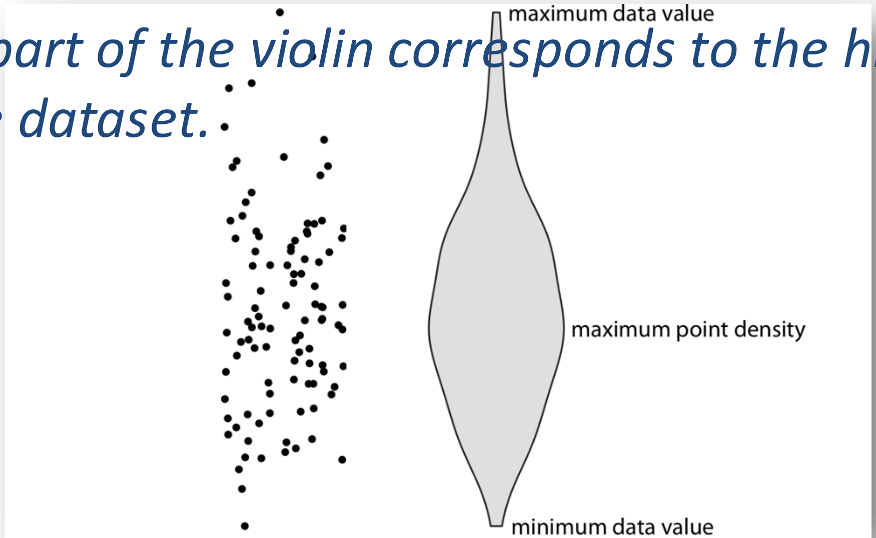
- Temperature is highly skewed in December (most days are moderately cold and a few are extremely cold) and not very skewed at all in some other months, such as in July.



The Lincoln (NB) temperature data, using boxplots.

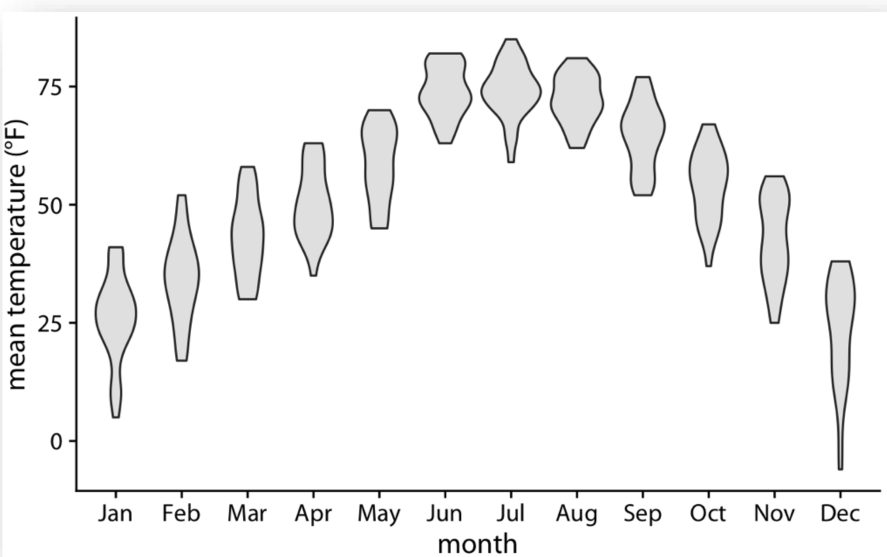
Along the Vertical Axis-Violin Plot

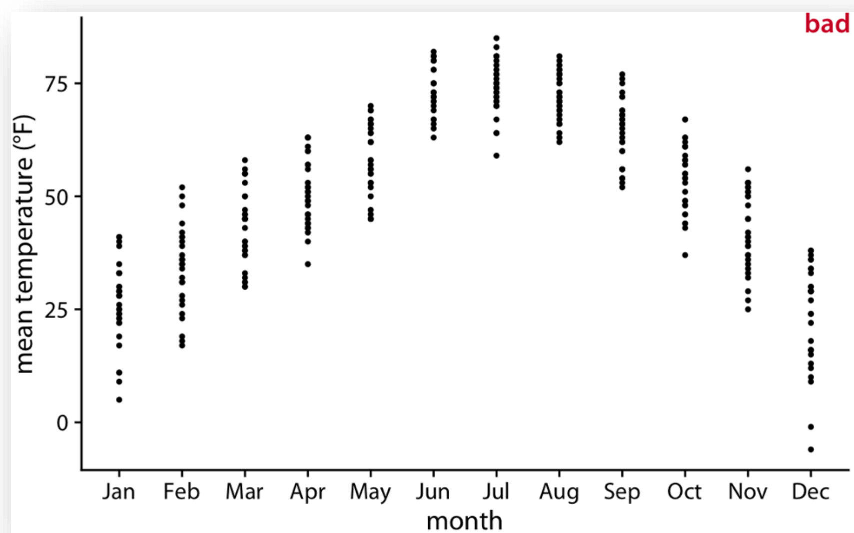
- Violin plots will accurately represent bimodal data.
  - *The width of the violin represents the point density at that y value.*
  - *It is a density estimate rotated by 90 degrees, mirrored, and therefore symmetric.*
  - *Violins begin and end at the min and max data values, respectively.*
  - *The thickest part of the violin corresponds to the highest point density in the dataset.*



Along the Vertical Axis-Strip Chart

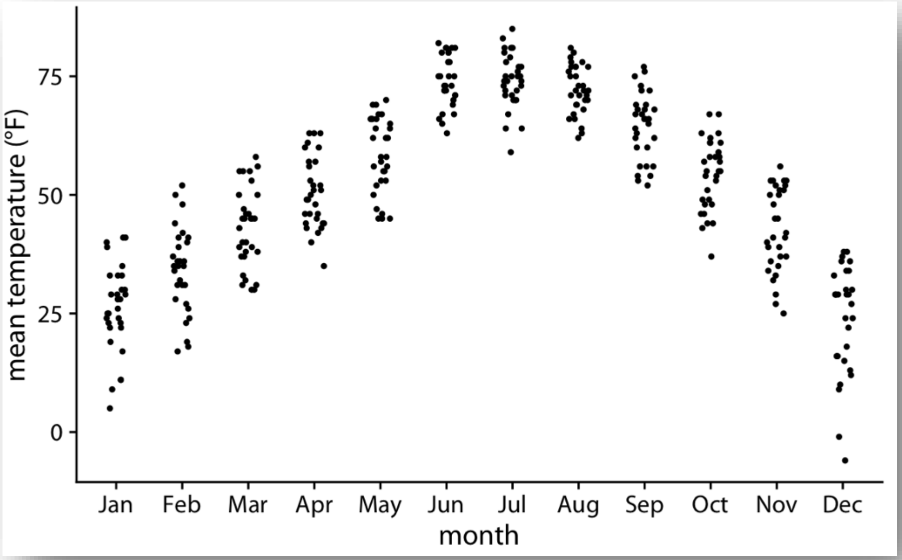
- Violin plots are derived from density estimates, they have similar shortcomings.
  - *They can generate the appearance that there is data where none exists, or that the dataset is very dense when actually it is quite sparse.*
- Strip chart overcomes this problem, but overplotting?

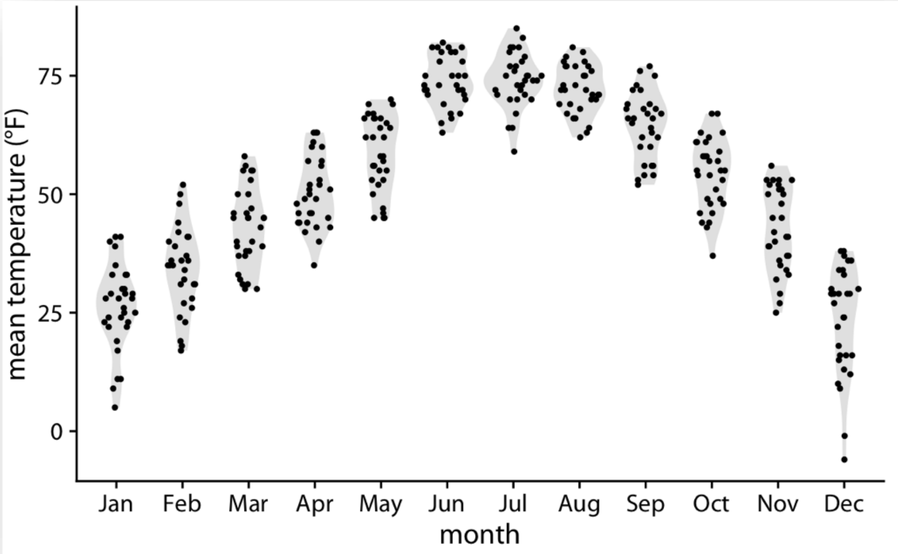




Along the Vertical Axis-Sina Plot

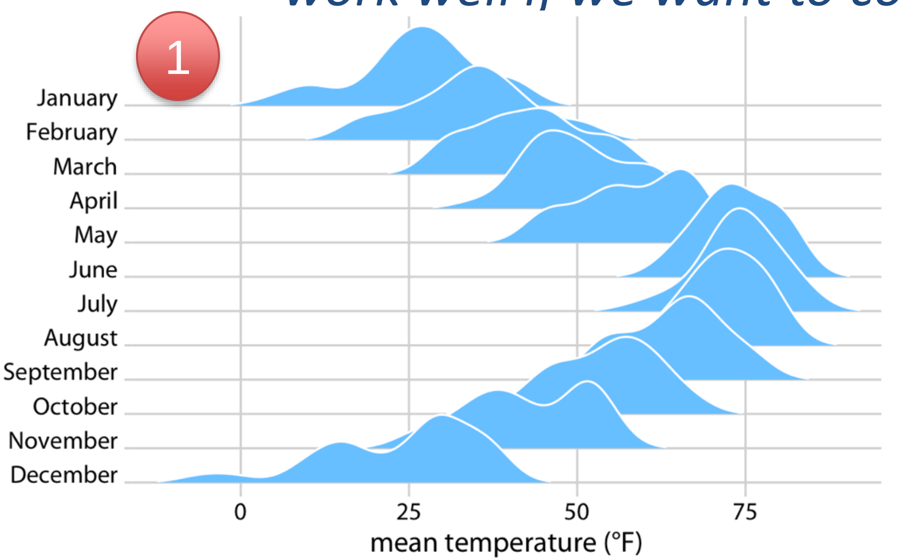
- We can spread out points by adding noise, *jitter.*
- we can combine the best of both worlds by spreading out the dots in proportion to the point density at a given y coordinate, *sina plot*

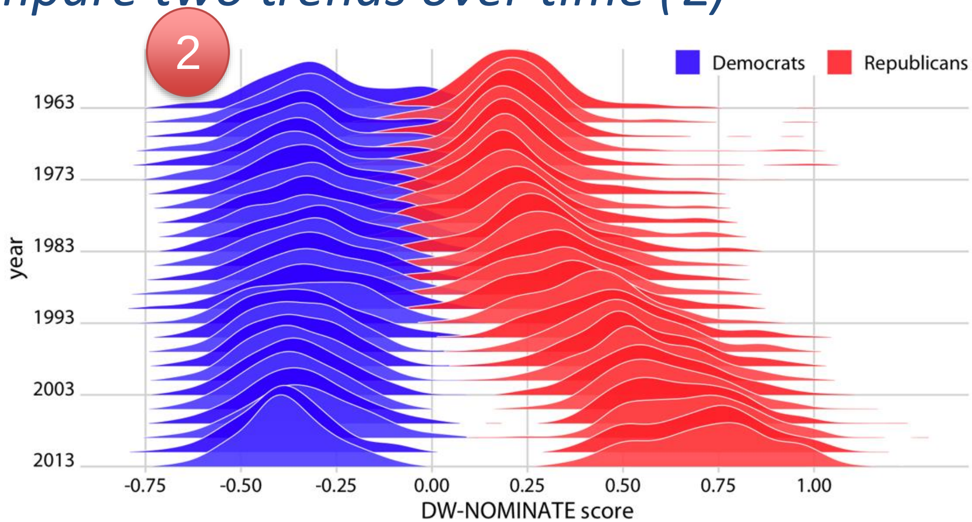




Along the Horizantal Axis-Ridgeline

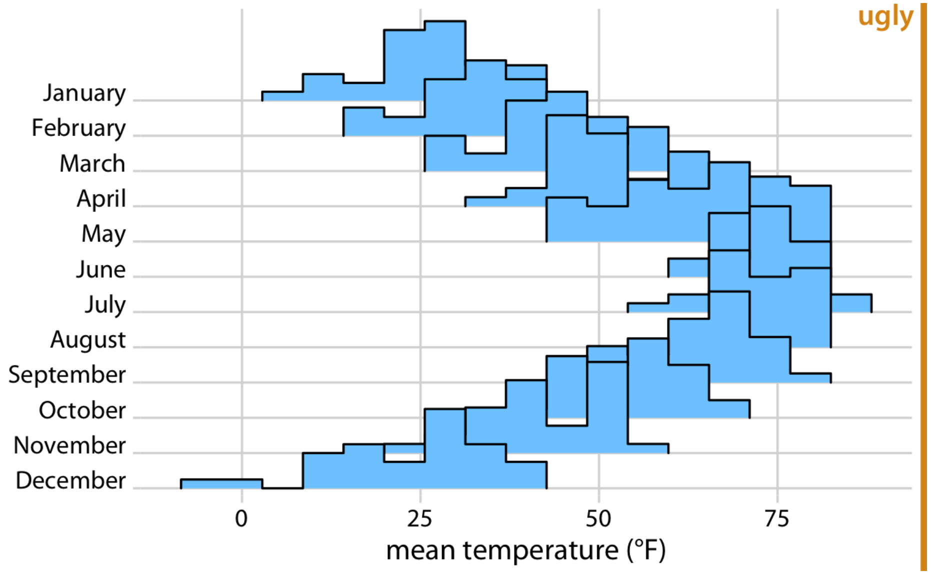
- The standard ridgeline plot uses density estimates.
  - *Quite closely related to the violin plot, but frequently evokes a more intuitive understanding of the data.*
  - *For example, the two clusters of temperatures around 35 degrees and 50 degrees Fahrenheit in November are much more obvious (1).*
  - *work well if we want to compare two trends over time (2)*





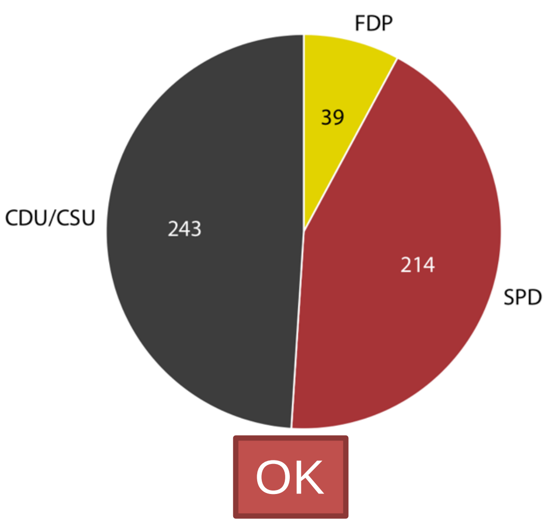
Along the Horizantal Axis-Histograms

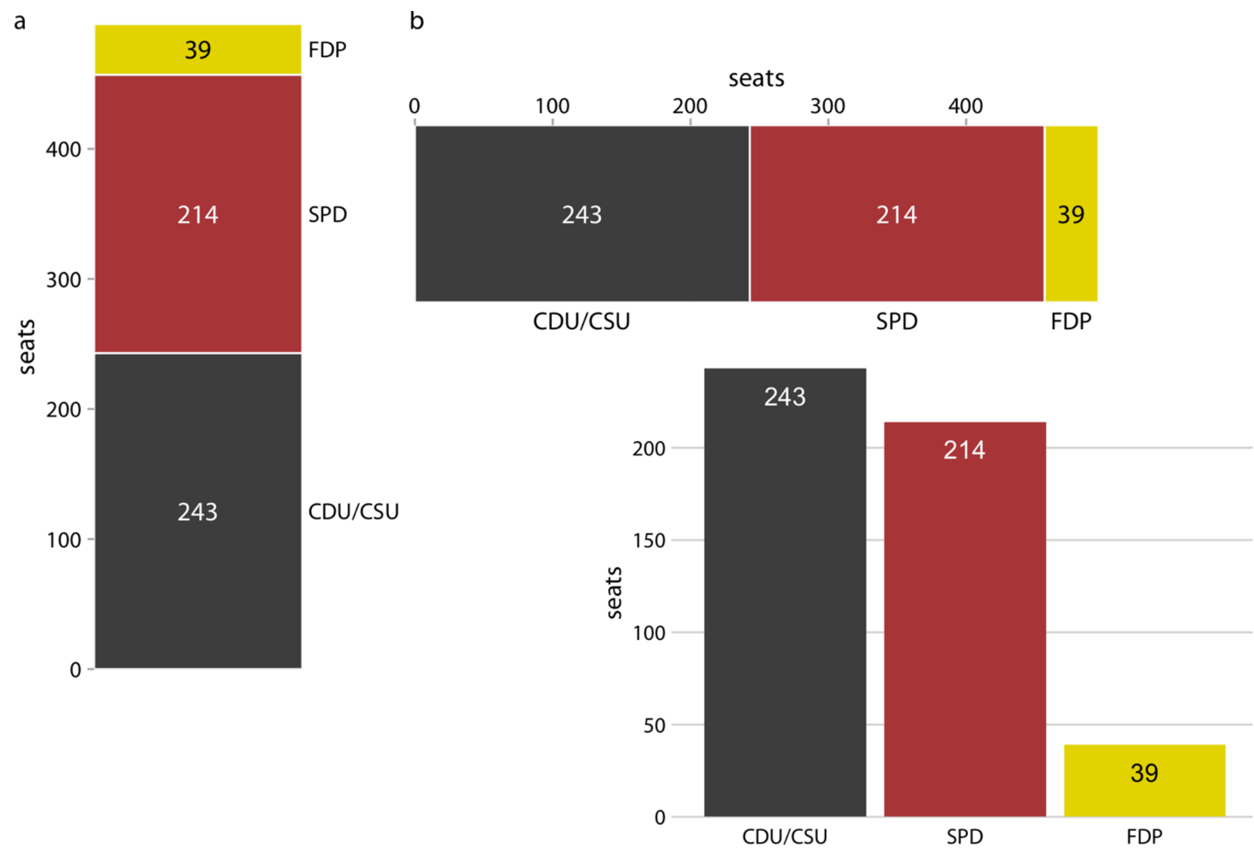
- In principle, we can use histograms instead of density plots in a ridgeline visualization. However, the resulting figures often don't look very good
- The bars from different histograms align with each other in confusing ways.



Visualizing Proportions

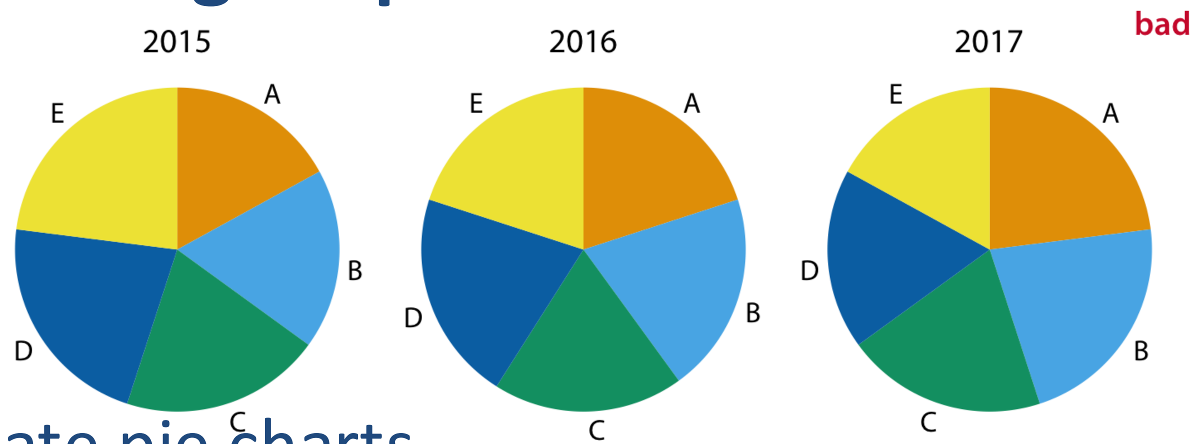
- Party composition of the eighth German Bundestag, 1976–1980



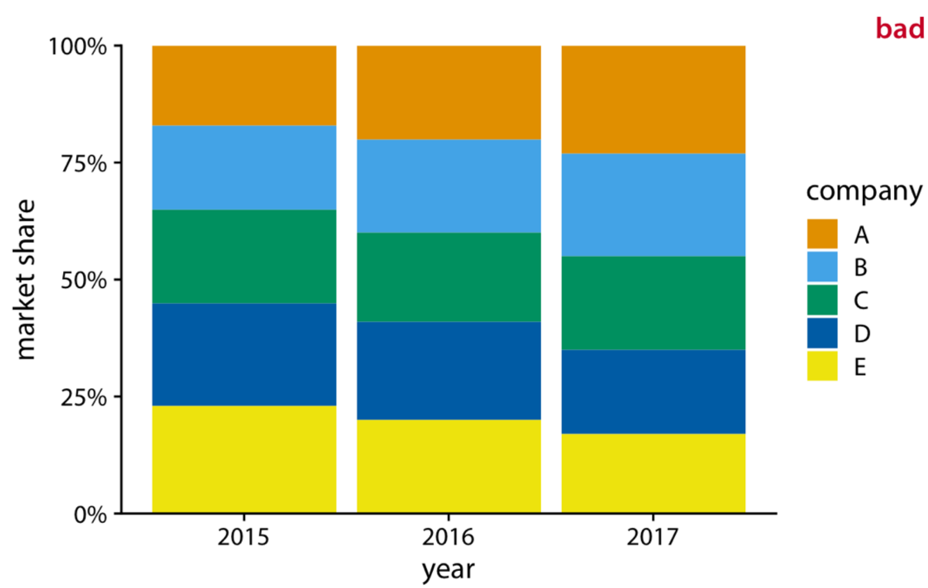


Visualizing Proportions

- Market shares

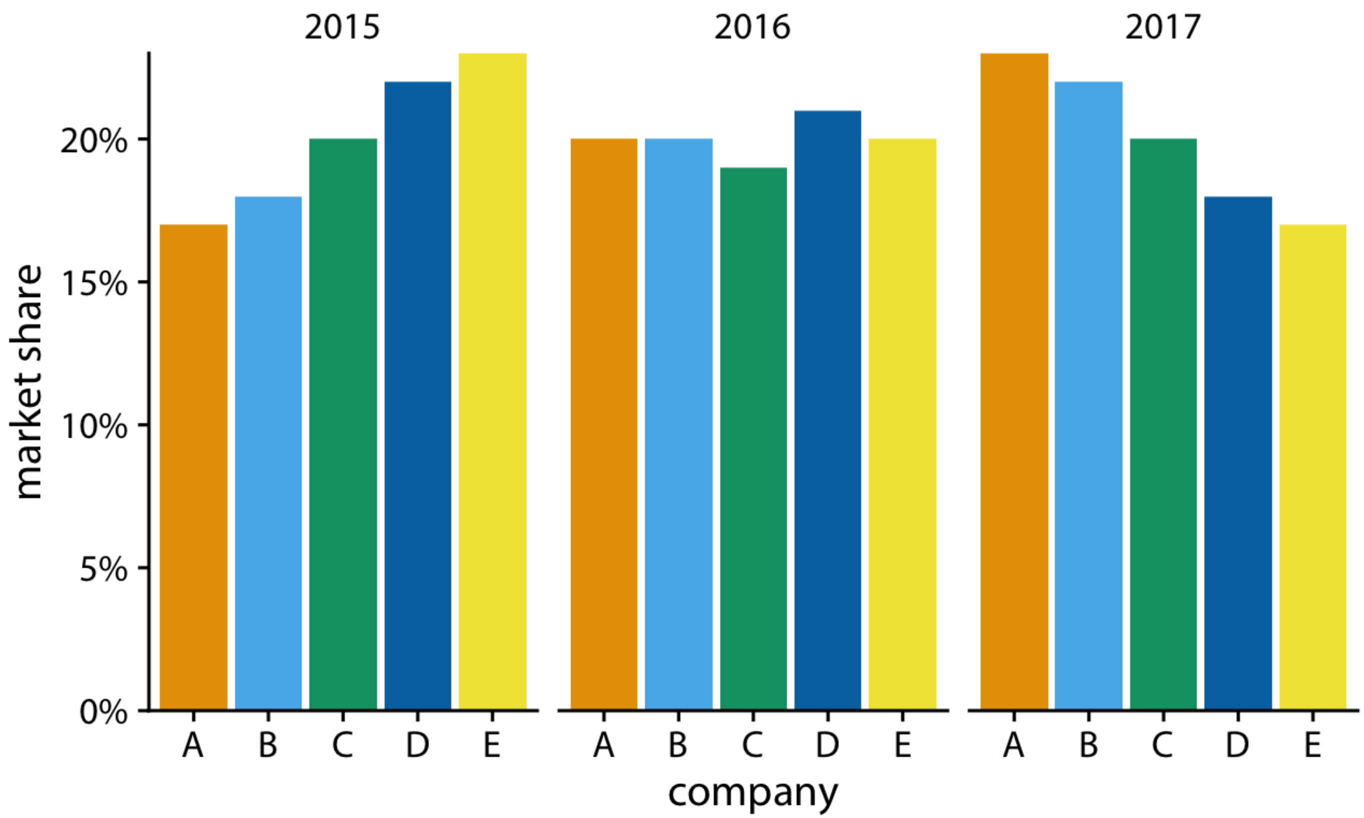


- Hard to differentiate pie charts



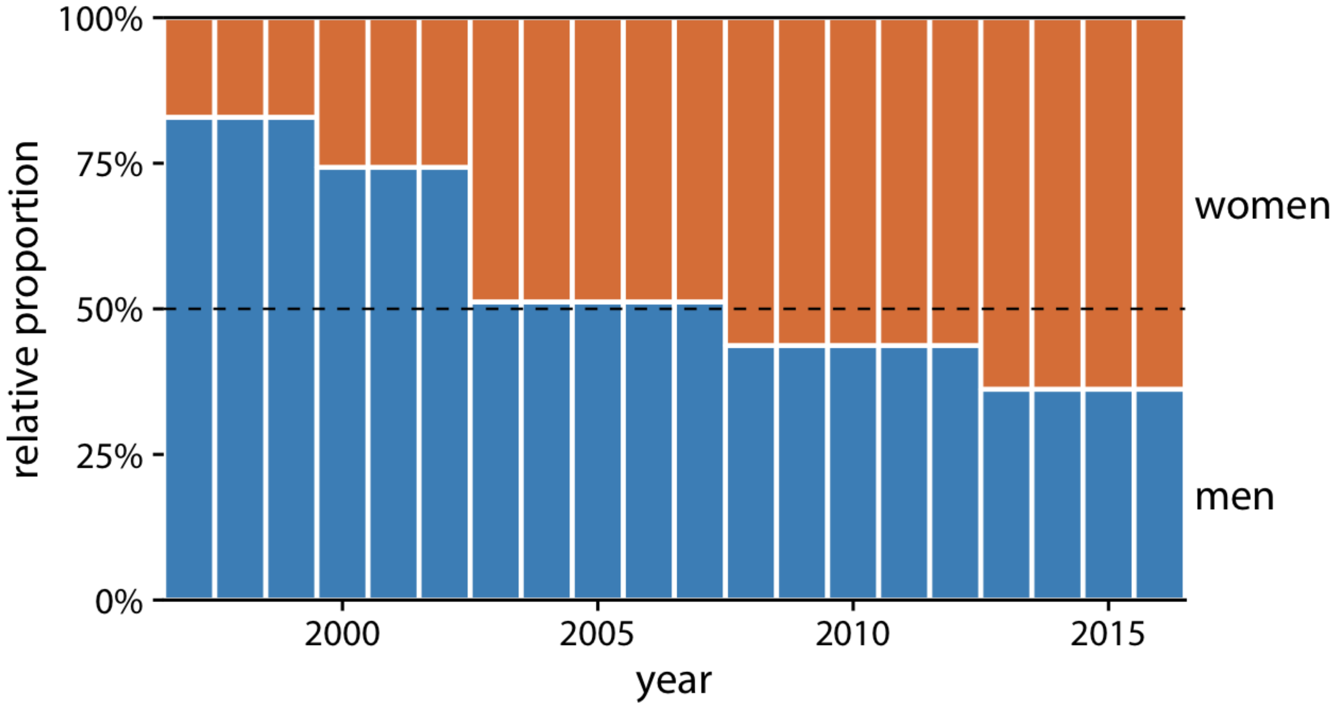
- Hard to follow B, C, and D

Visualizing Proportions



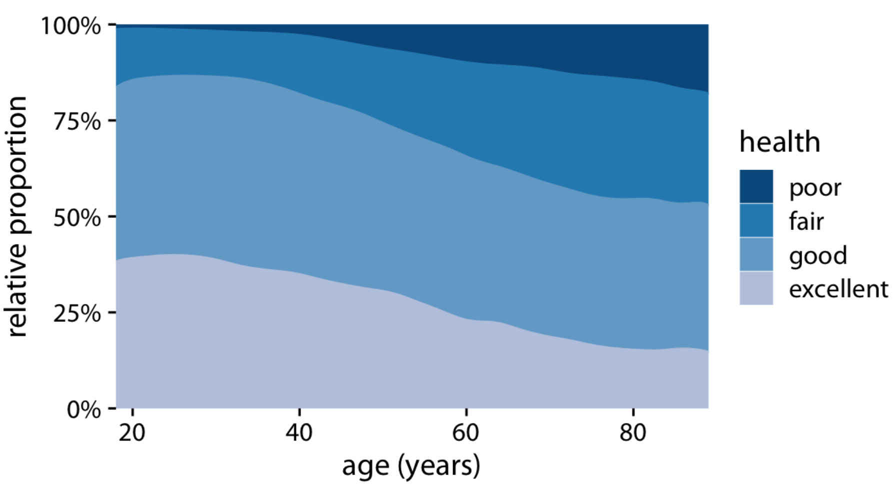
Visualizing Proportions

- The problem of shifting internal bars disappears if there are only two bars in each stack.
- Proportions are clear

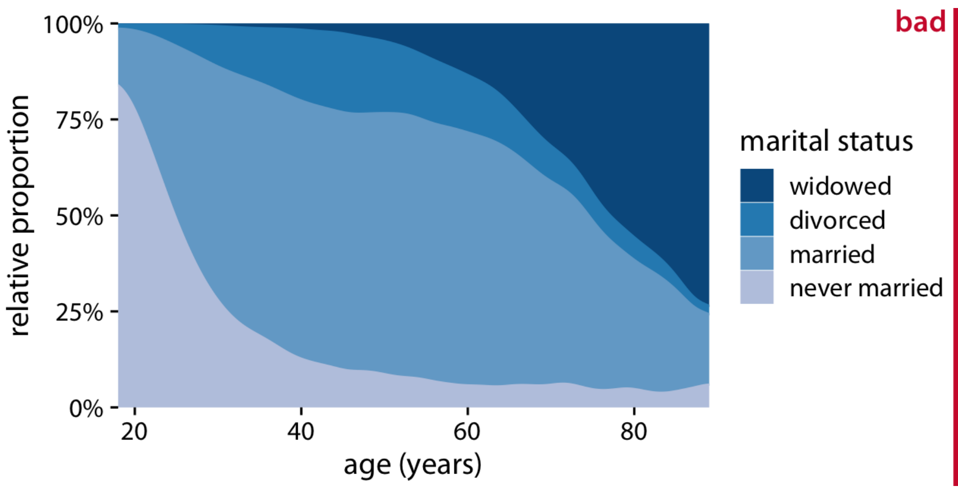


Visualizing Proportions

- Stack densities may (not) work



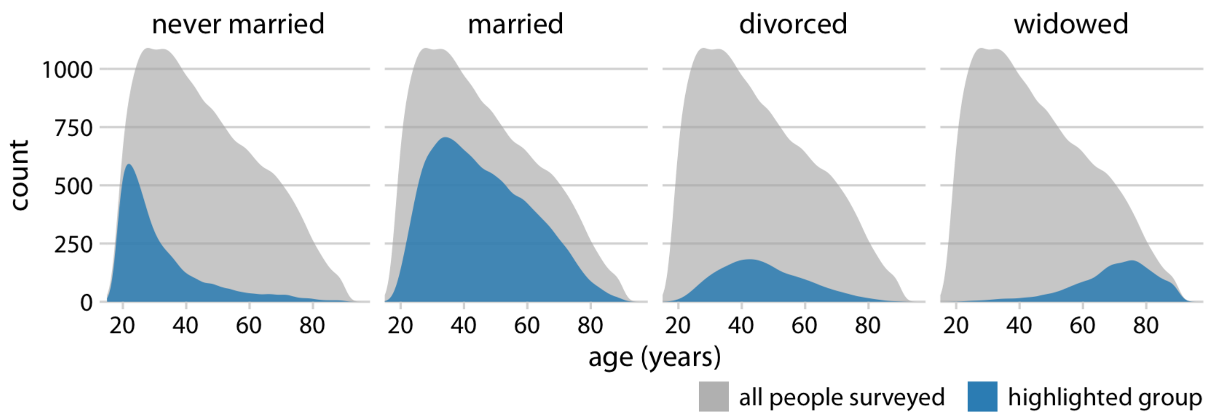
Health status by age



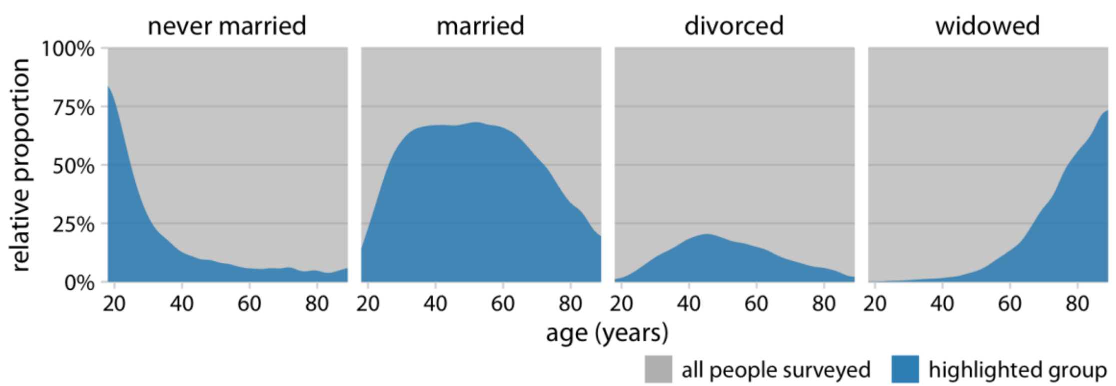
Marital status by age

Visualizing Proportions

- Better, but still not easy to determine relative proportions



- Much better

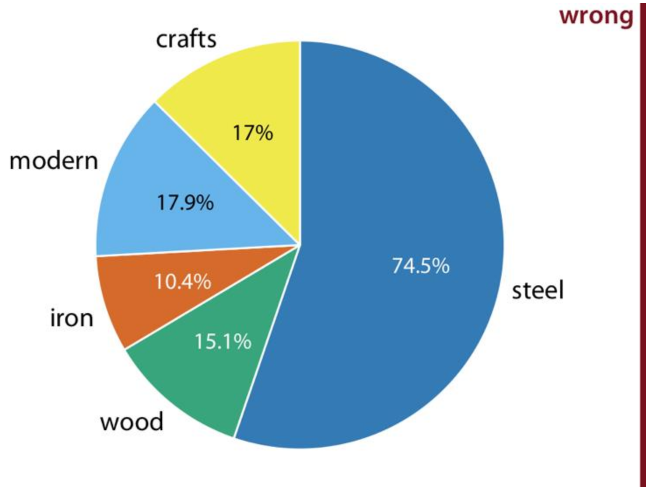


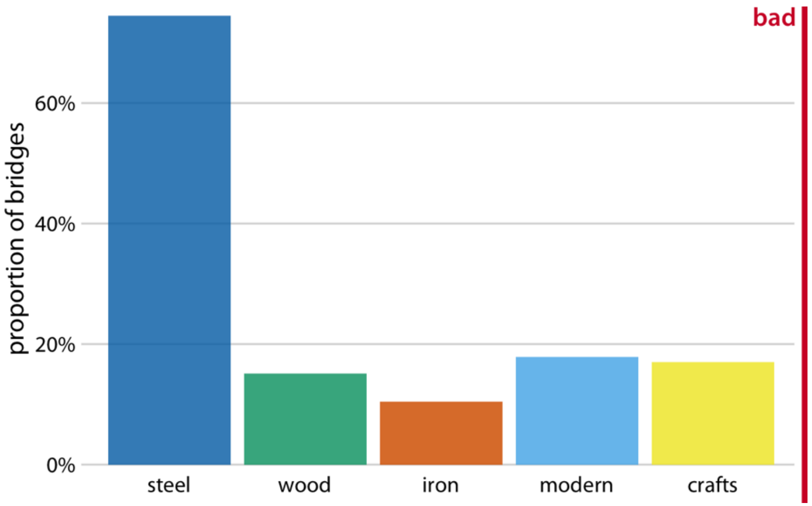
Nested Proportions

- We may want to drill down further and break down a dataset by multiple categorical variables at once.
  - *We could be interested in the proportions of seats by party and by the gender of the representatives.*

Nested Proportions Gone Wrong

- A dataset of 106 bridges in Pittsburgh.
  - *material from which they are constructed (steel, iron, or wood),*
  - *based on the year of erection, bridges are grouped into distinct categories, such as crafts and modern.*
  - *On which river?*



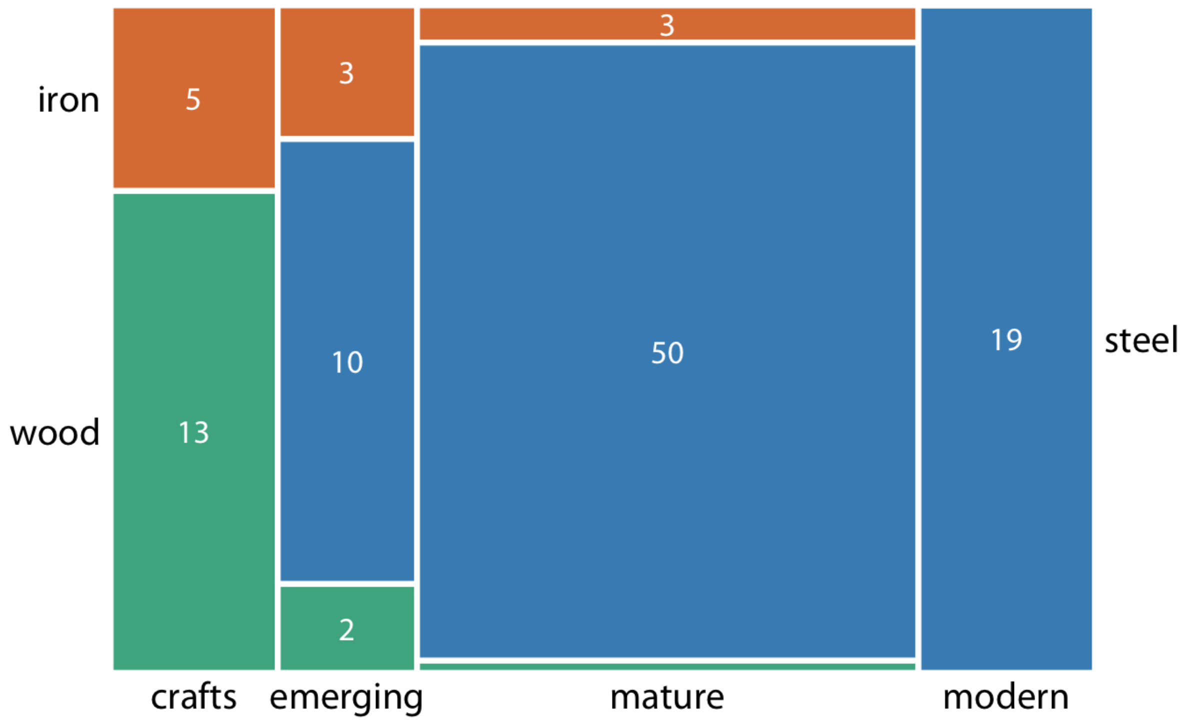


Sum exceeds 100% due to double count

It does not clearly indicate the overlap

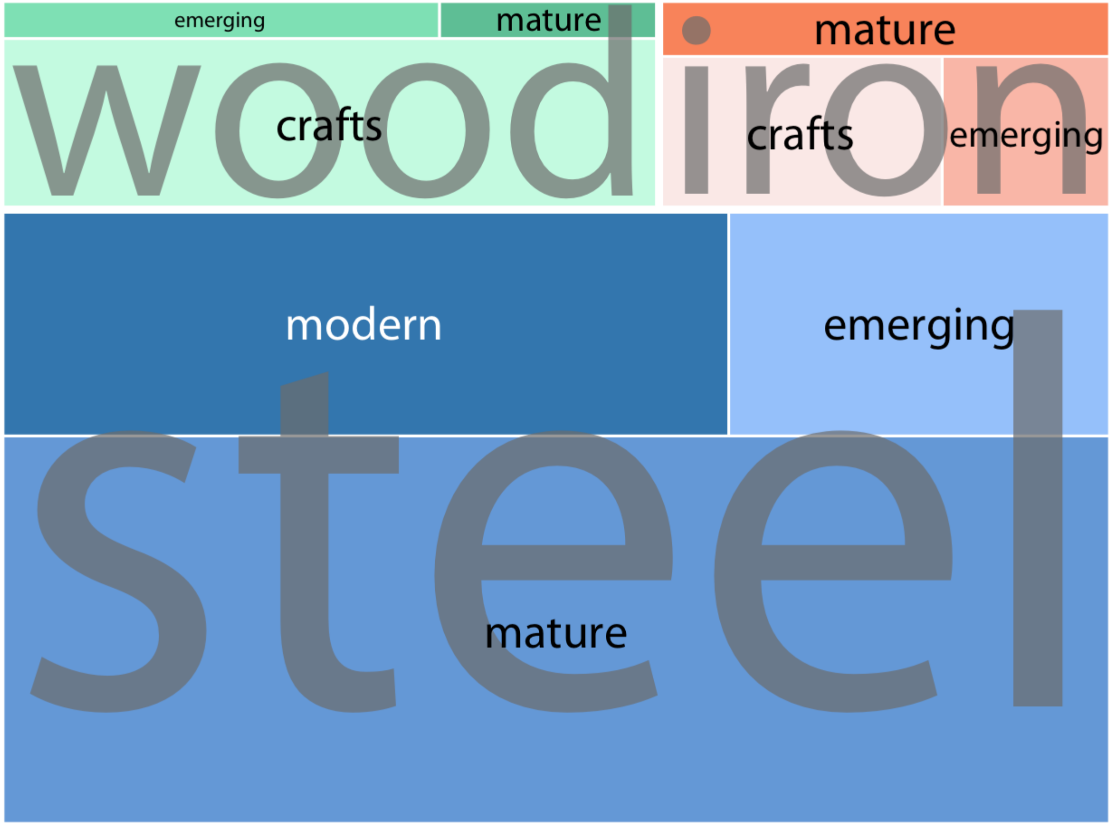
Nested Proportions-Mosaic Plot

- When there are overlapping categories, it is best to show explicitly how they relate to each other.

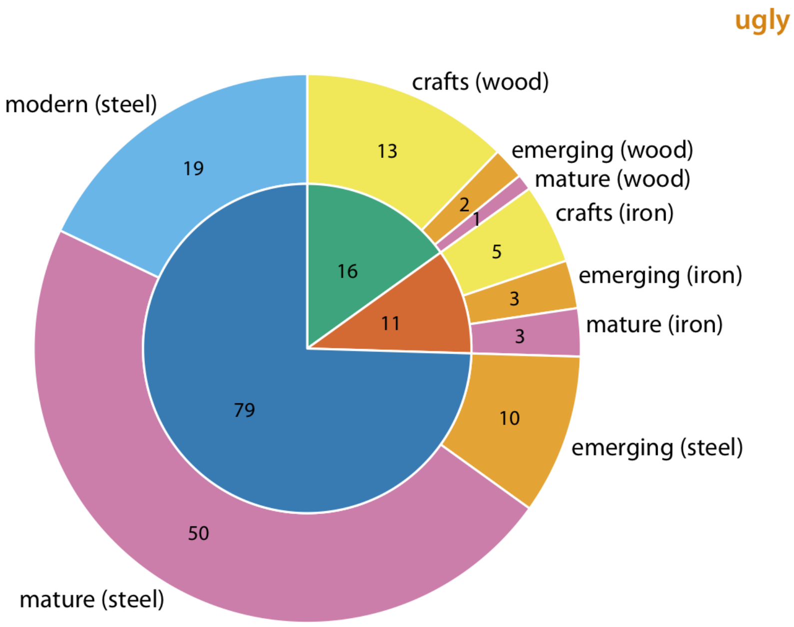


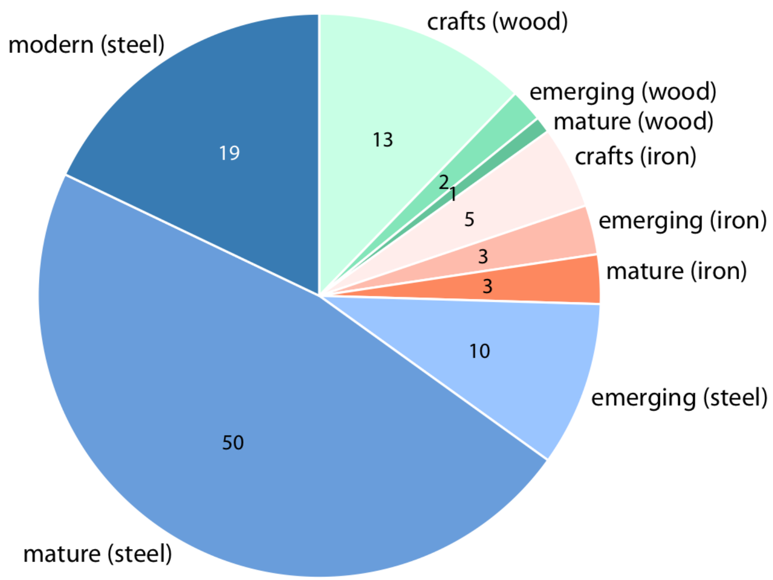
Nested Proportions-Treemap Plot

- Shows the counts for every possible combination



Nested Proportions-Nested Pies





Nested Proportions-Parallel Sets

- When more than two categorical variables, parallel sets plot can offer a less crowded view.
- It shows how the total dataset breaks down by each individual categorical variable by using shaded bands.