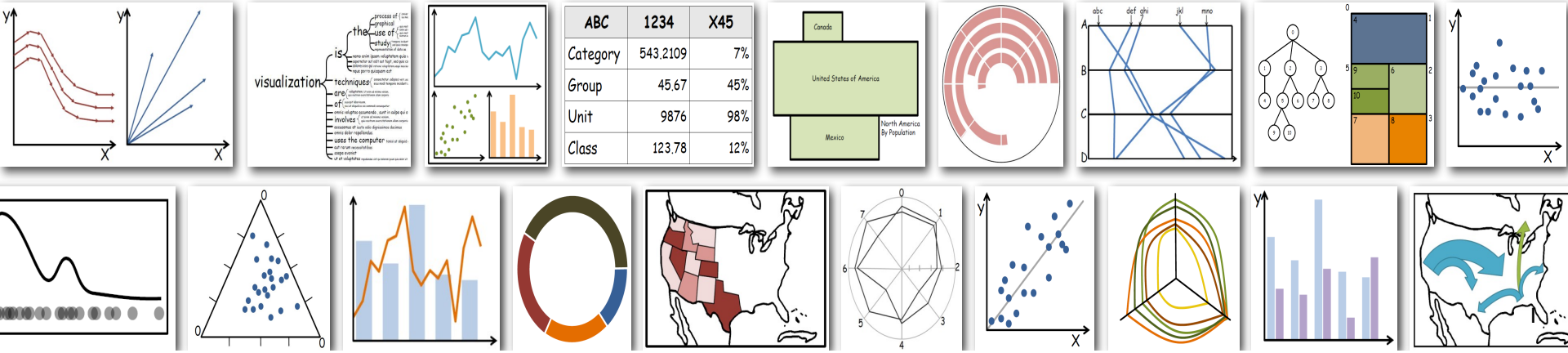


Visualizing **Distributions, Proportions & Nested Proportions**

DATA ANALYSIS & VISUALIZATION

FALL 2021

Dr. Muhammad Faisal Cheema

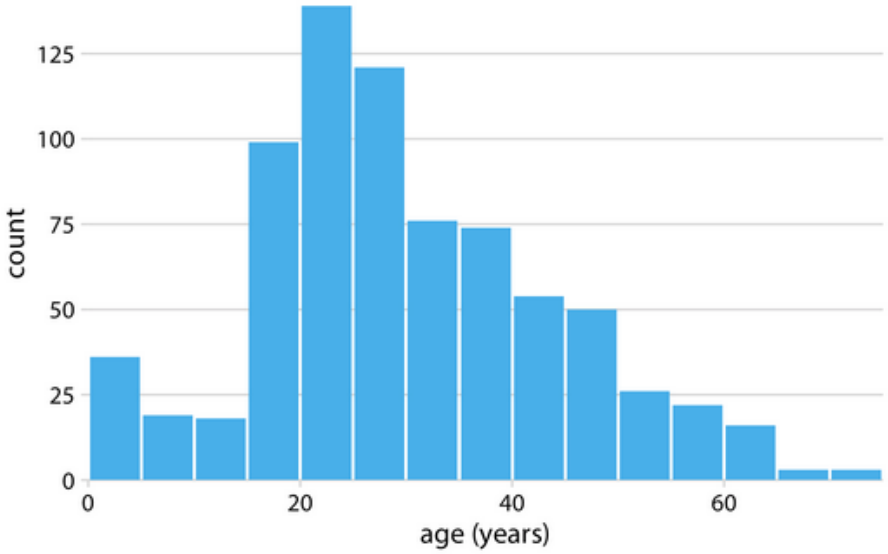
FAST NU

Visualizing (multiple) distributions

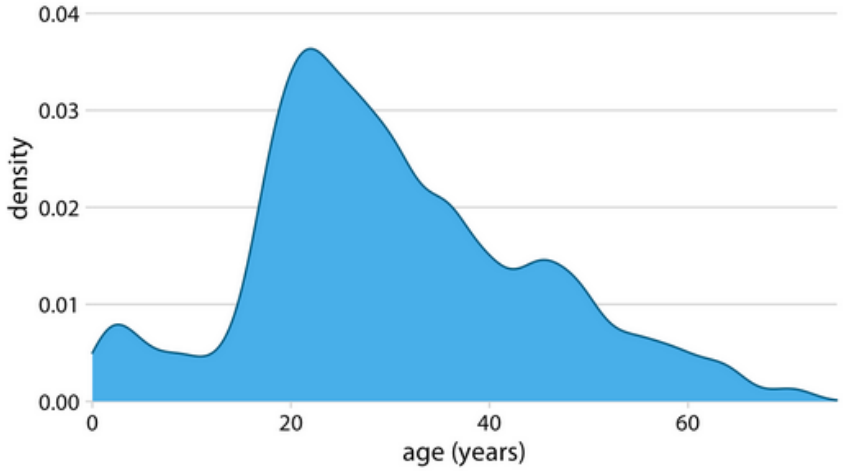
Visualizing distributions: Histograms and density plots

It is quite common, where one would like to understand how a particular variable is distributed in a given dataset.

Visualizing a single distribution



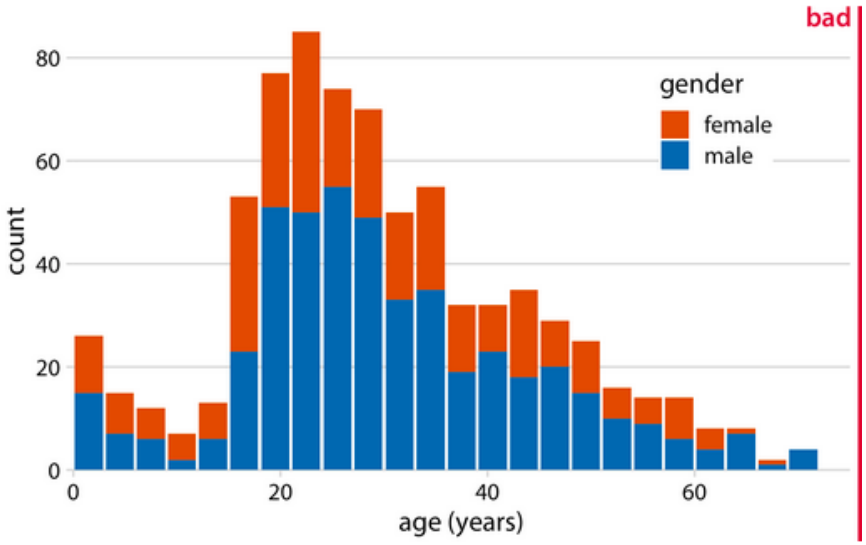
Histogram



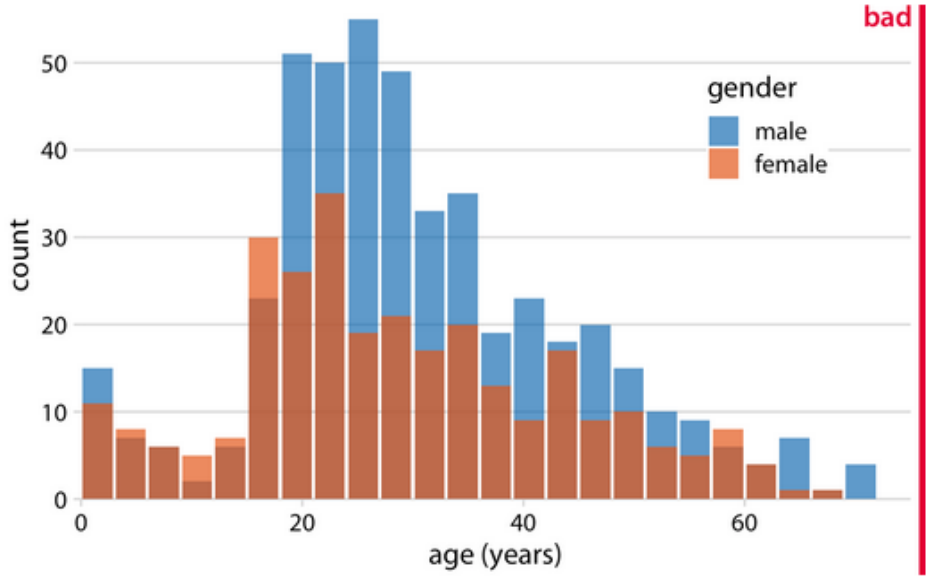
Density Plot

How do we visualize multiple distributions at the same time?

Visualizing multiple distributions at the same time

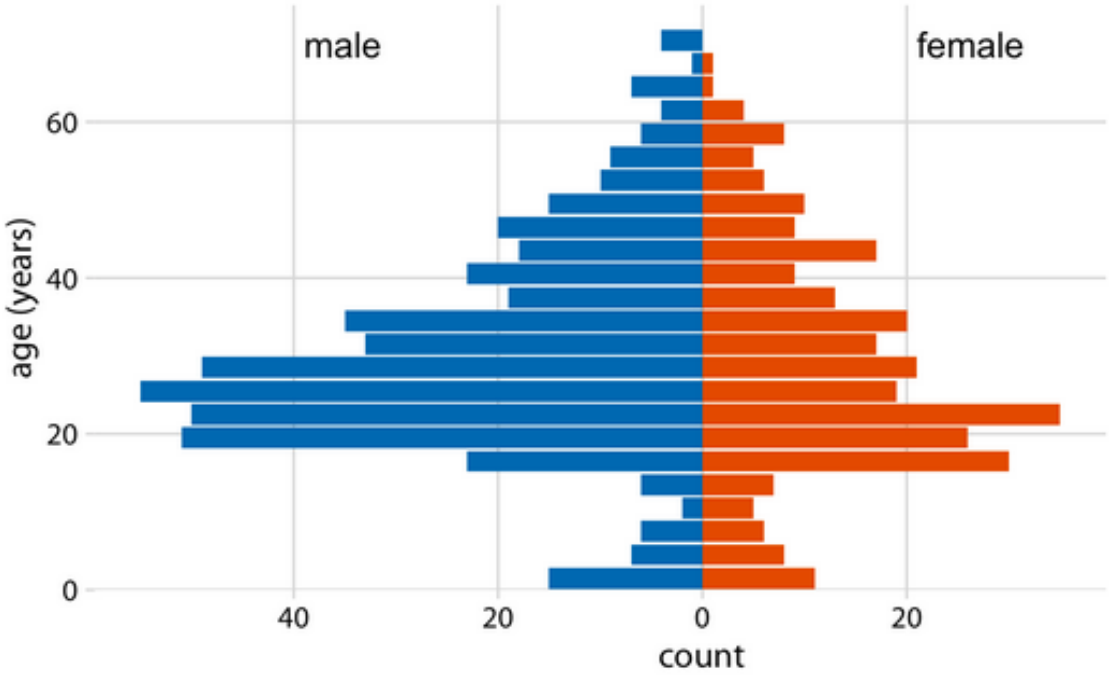


Stacked Histograms



Overlapping Histograms

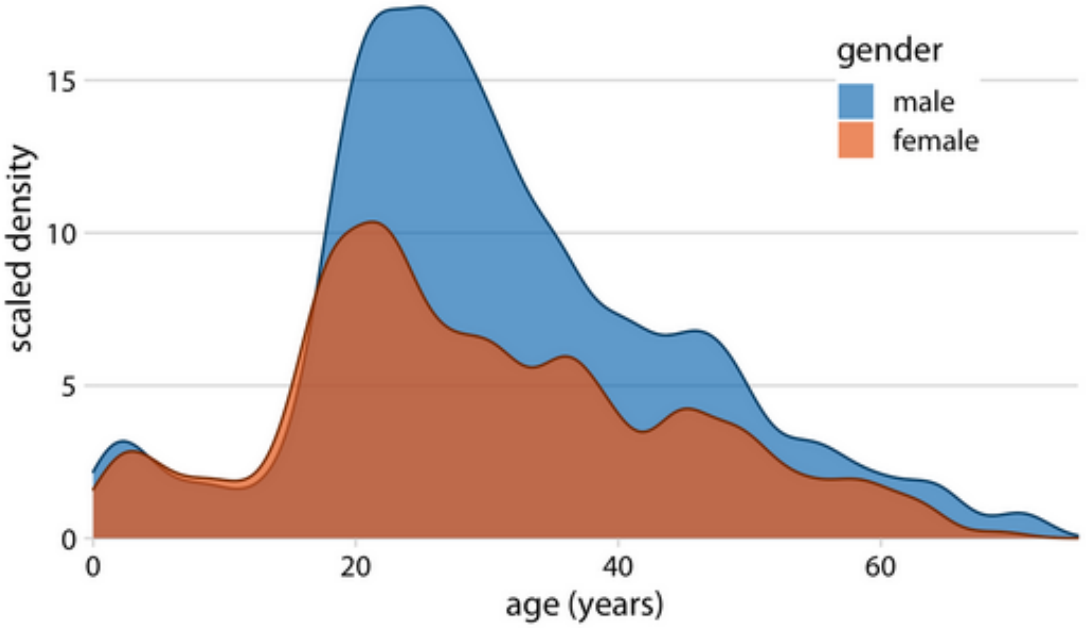
Visualizing multiple distributions at the same time



Pyramid Plot

How to construct?

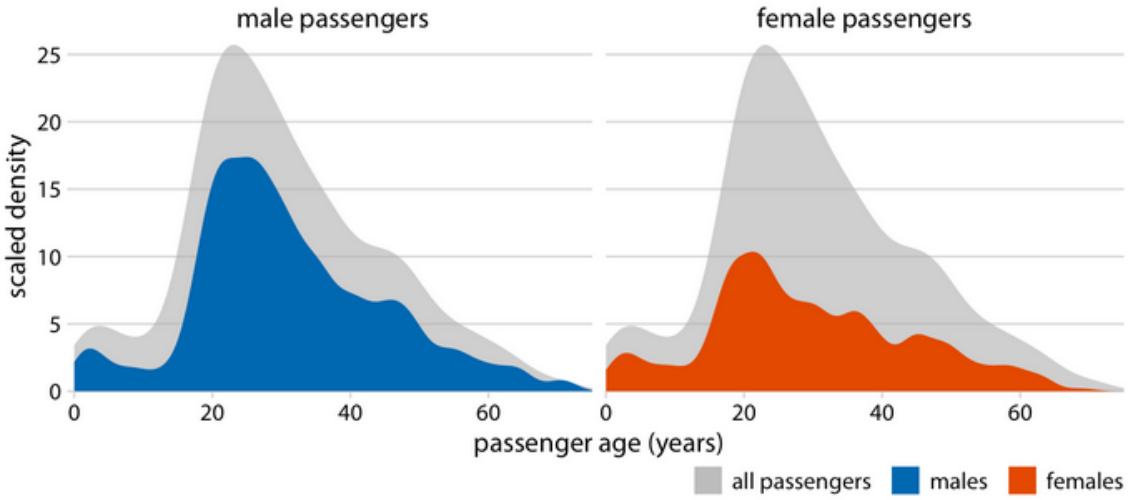
Visualizing multiple distributions at the same time



Overlapping Density plots

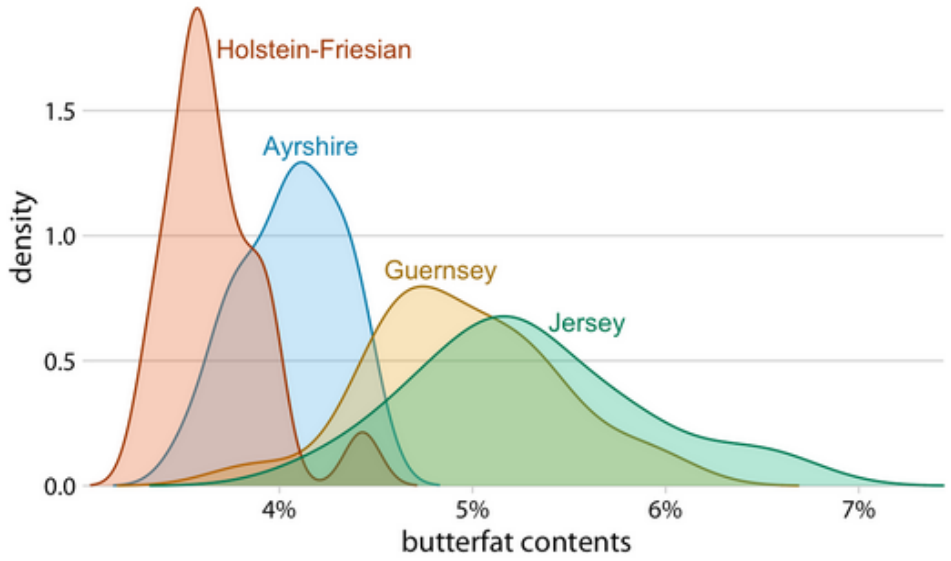
Visualizing multiple distributions at the same time

A solution that works well for this case is to show the distributions of two categories separately, each as a proportion of the overall distribution.



Visualizing multiple distributions at the same time

For multiple distributions, histograms tend to become highly confusing, whereas density plots work well as long as the distributions are somewhat distinct and contiguous.



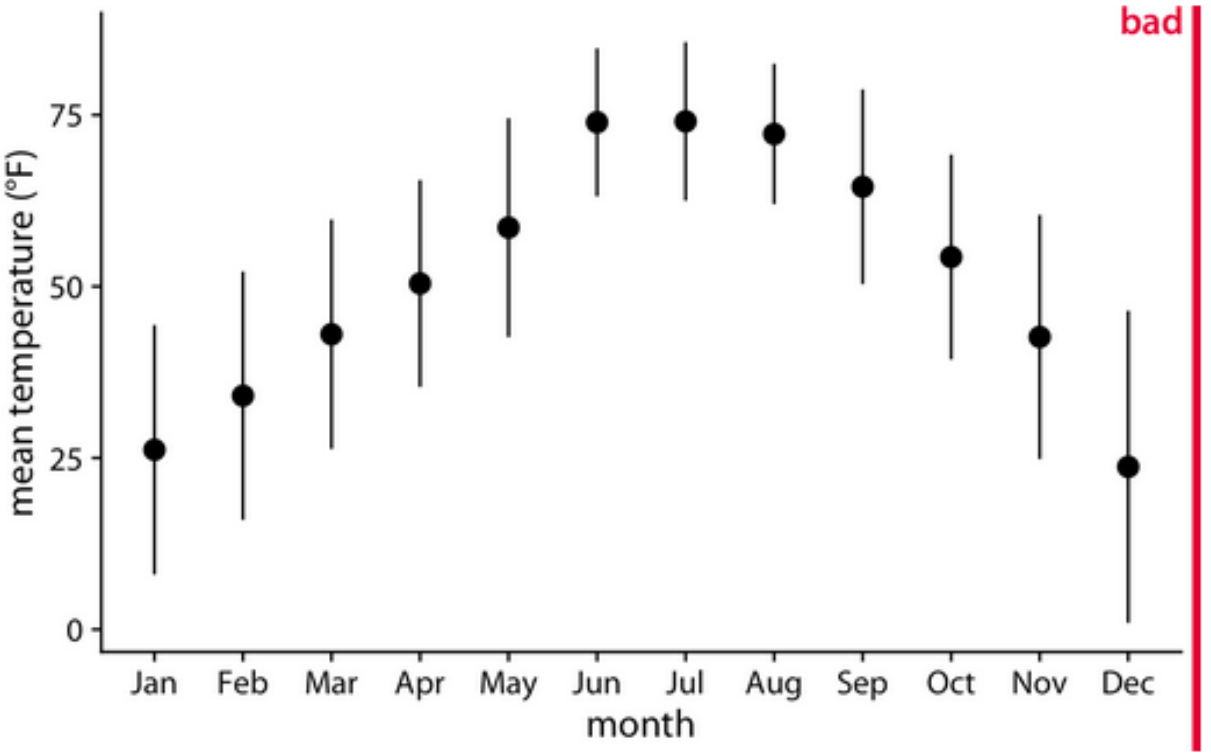
Visualizing many distributions at once

There are many scenarios in which we want to visualize multiple distributions at the same time.

Whenever we are dealing with many distributions, it is helpful to think in terms of the response variable and one or more grouping variables. The response variable is the variable whose distributions we want to show.

The grouping variables define subsets of the data with distinct distributions of the response variable.

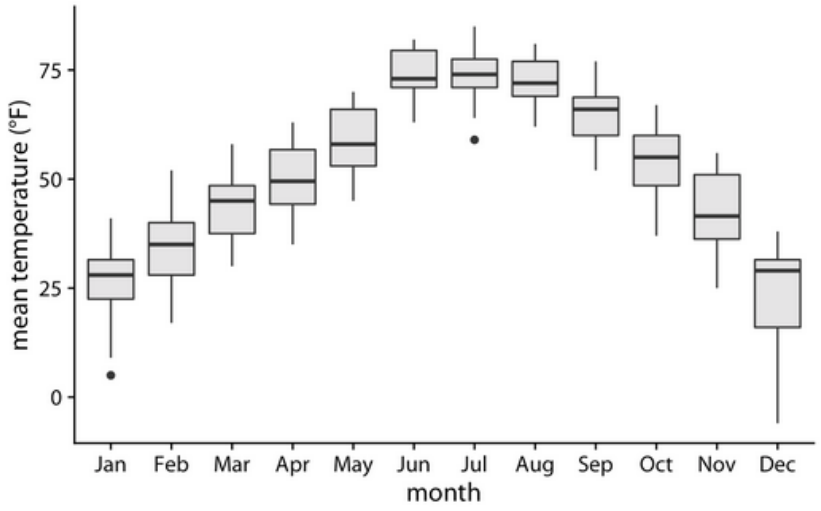
Visualizing distributions along the vertical axis

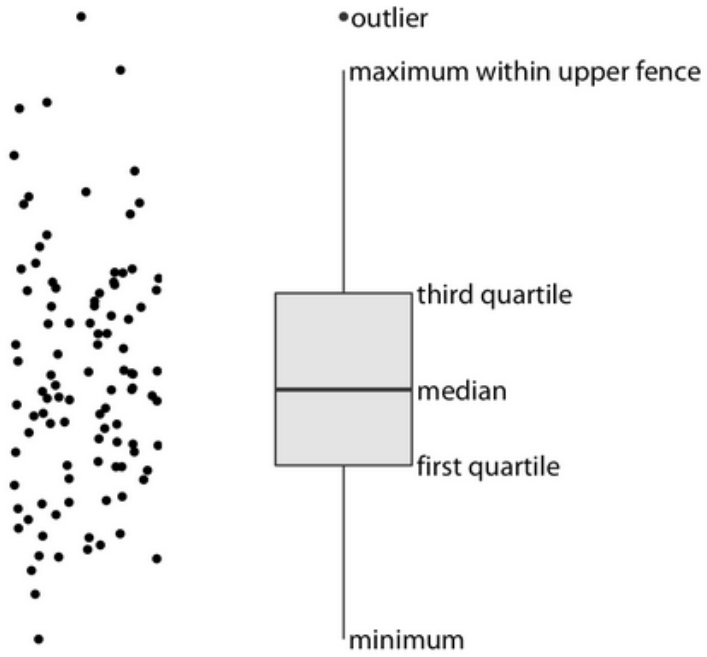


Mean/Median plots

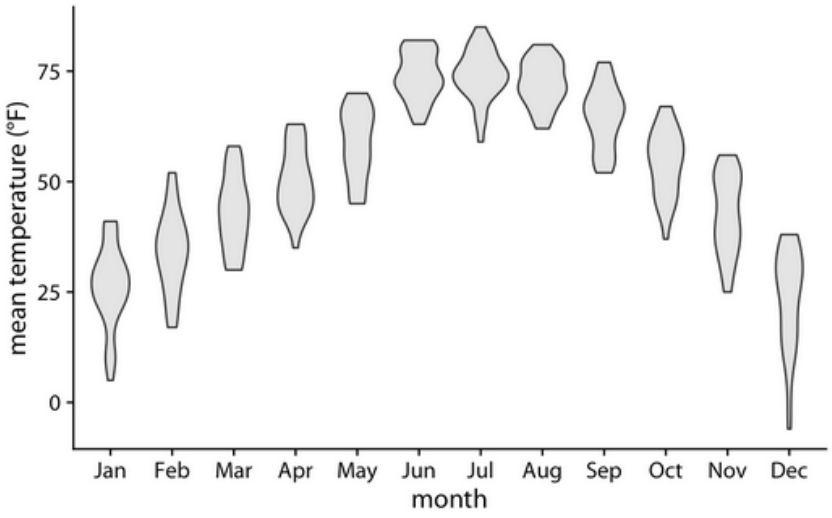
Visualizing distributions along the vertical axis

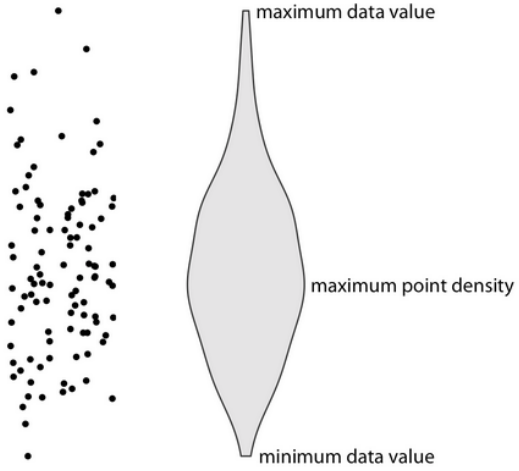
We can address all those shortcomings by using a traditional and commonly used method for visualizing distributions, the boxplot. A boxplot divides the data into quartiles and visualizes them in a standardized manner





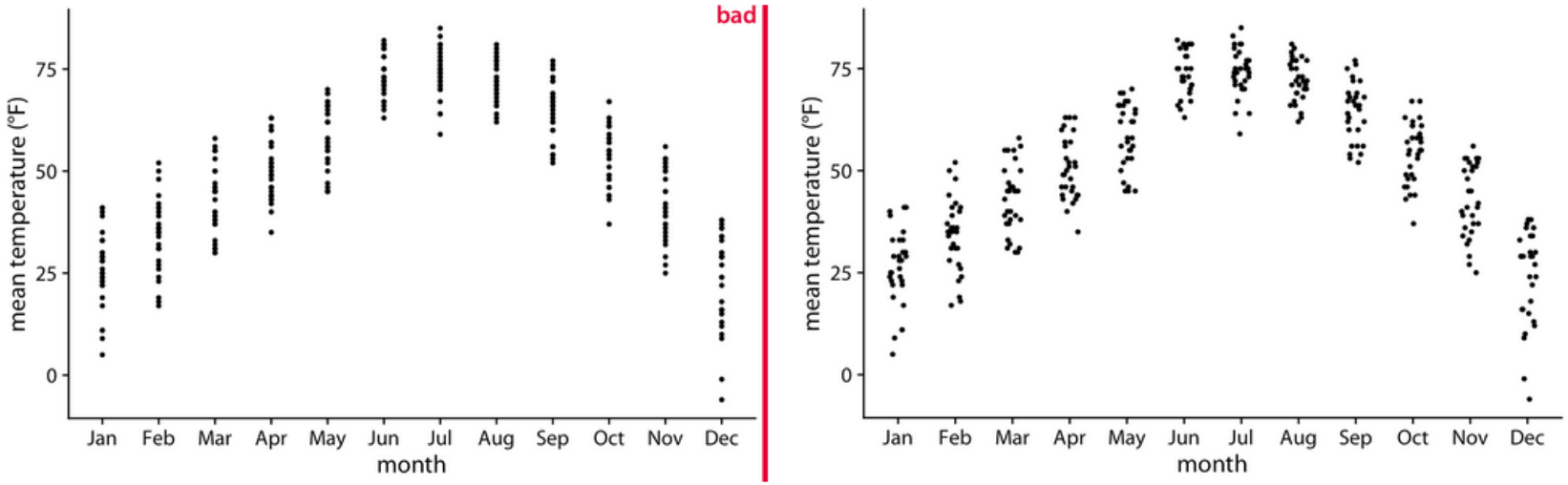
Visualizing distributions along the vertical axis





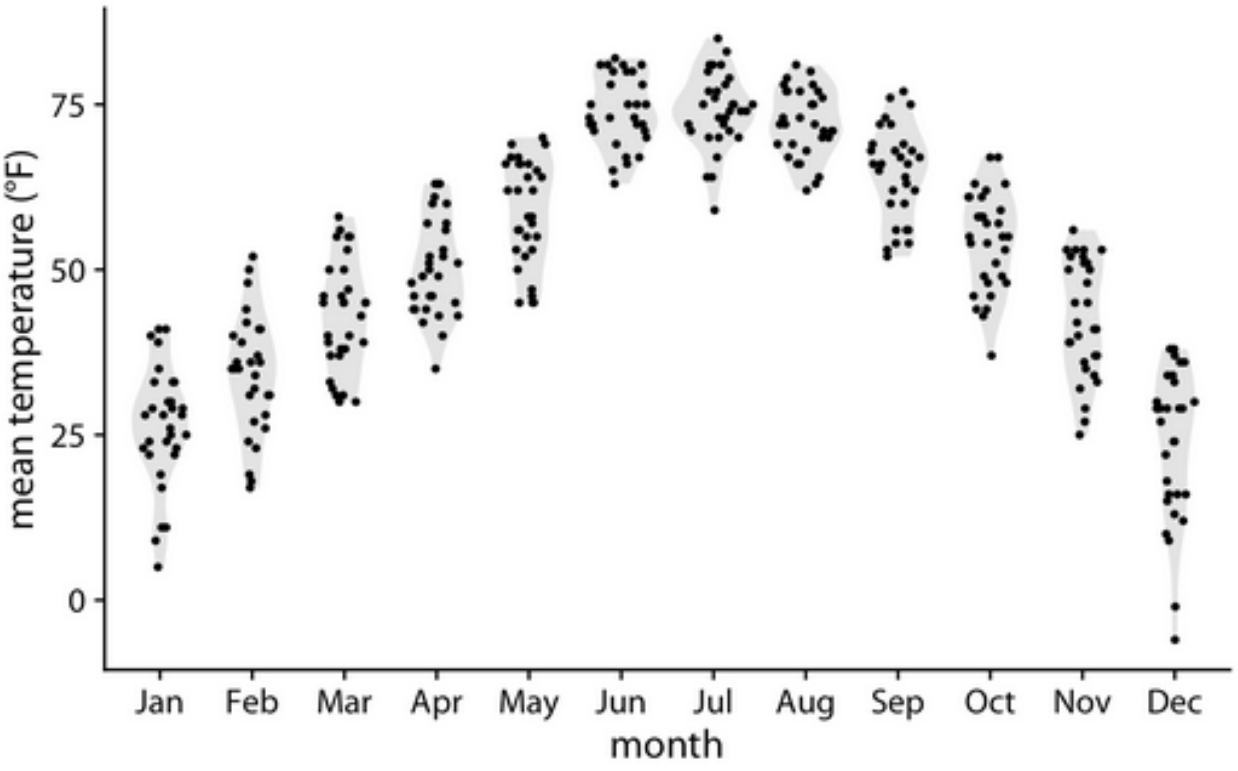
Violin plot

Visualizing distributions along the vertical axis



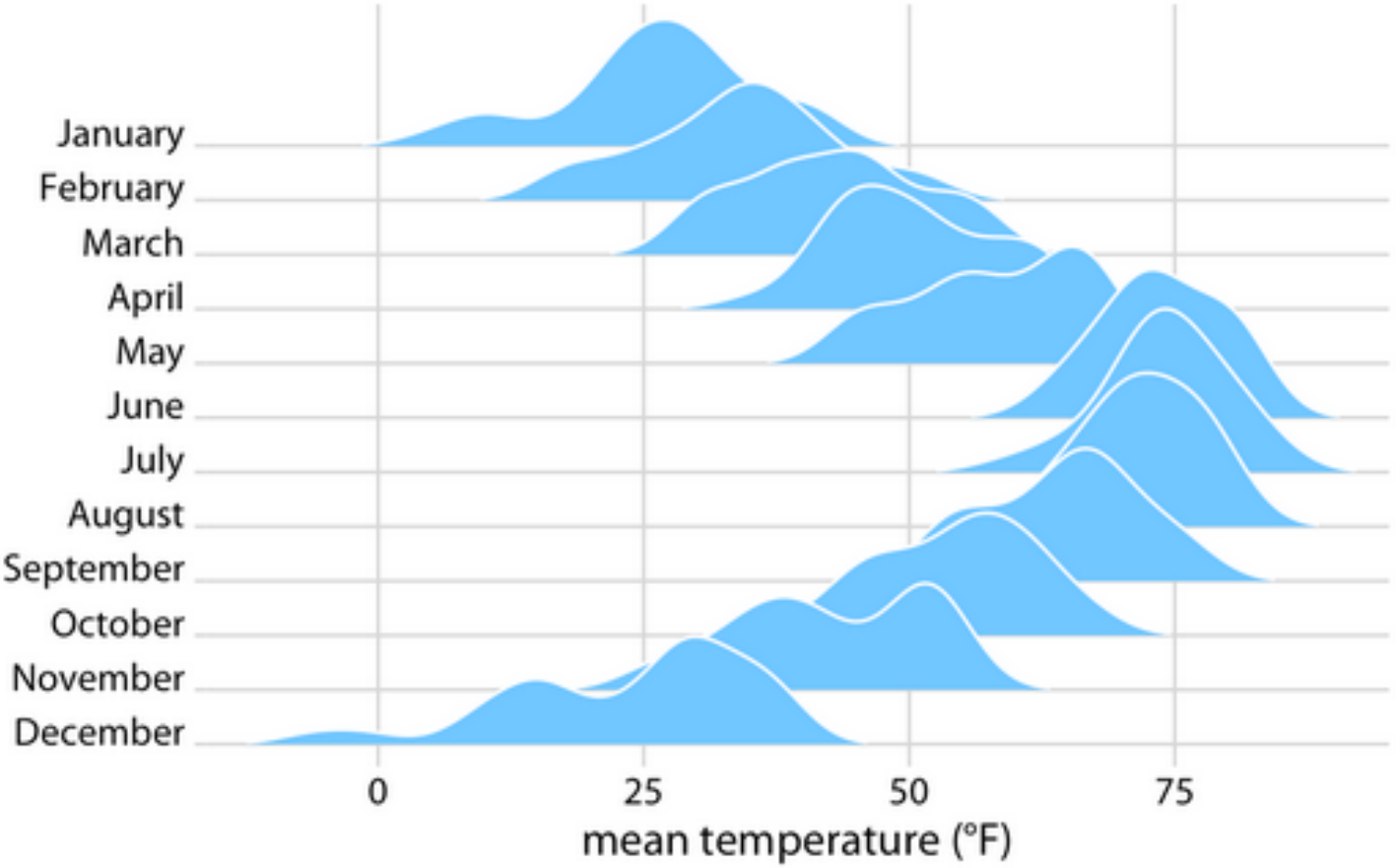
strip charts

Visualizing distributions along the vertical axis



sina plot

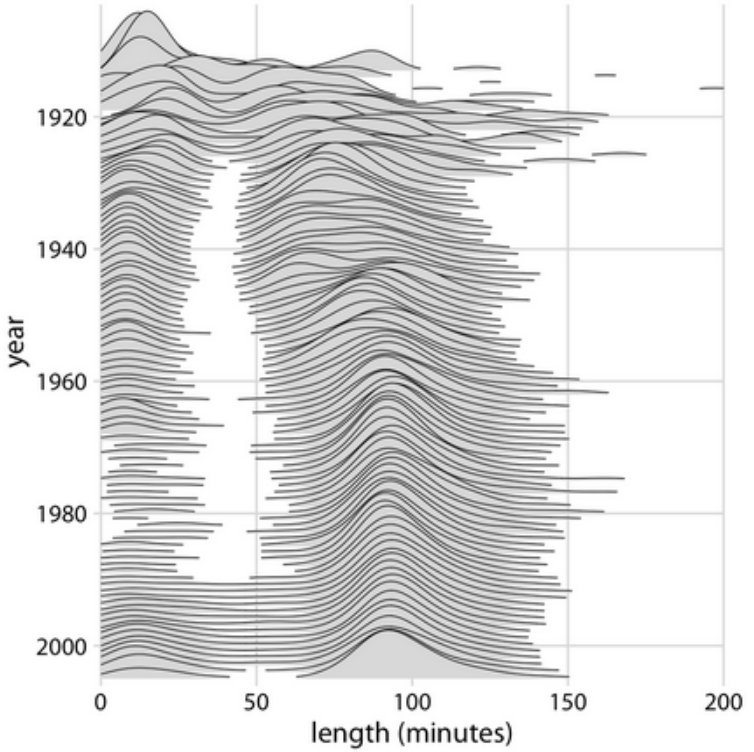
Visualizing distributions along the horizontal axis



ridgeline plot

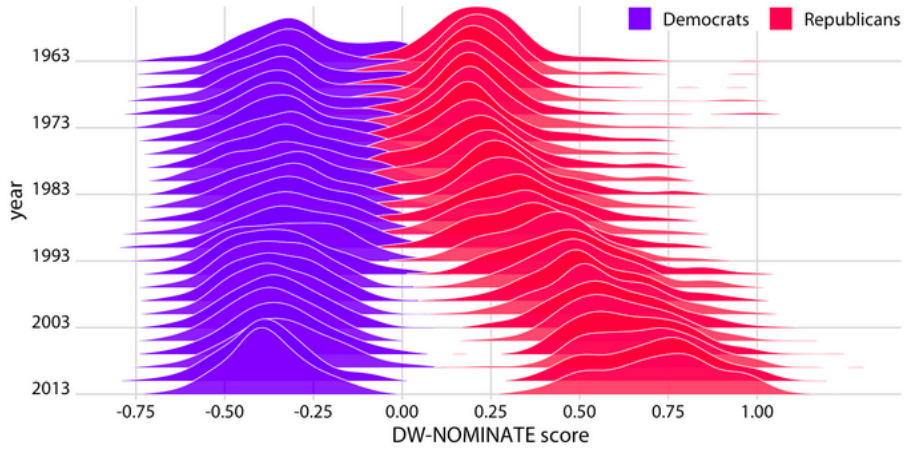
Visualizing distributions along the horizontal axis

Ridgeline plots scale to very large numbers of distributions.



Visualizing distributions along the horizontal axis

Ridgeline plots also work well if we want to compare two trends over time. We can make this comparison by staggering the distributions vertically by time and drawing two differently colored distributions at each time point.

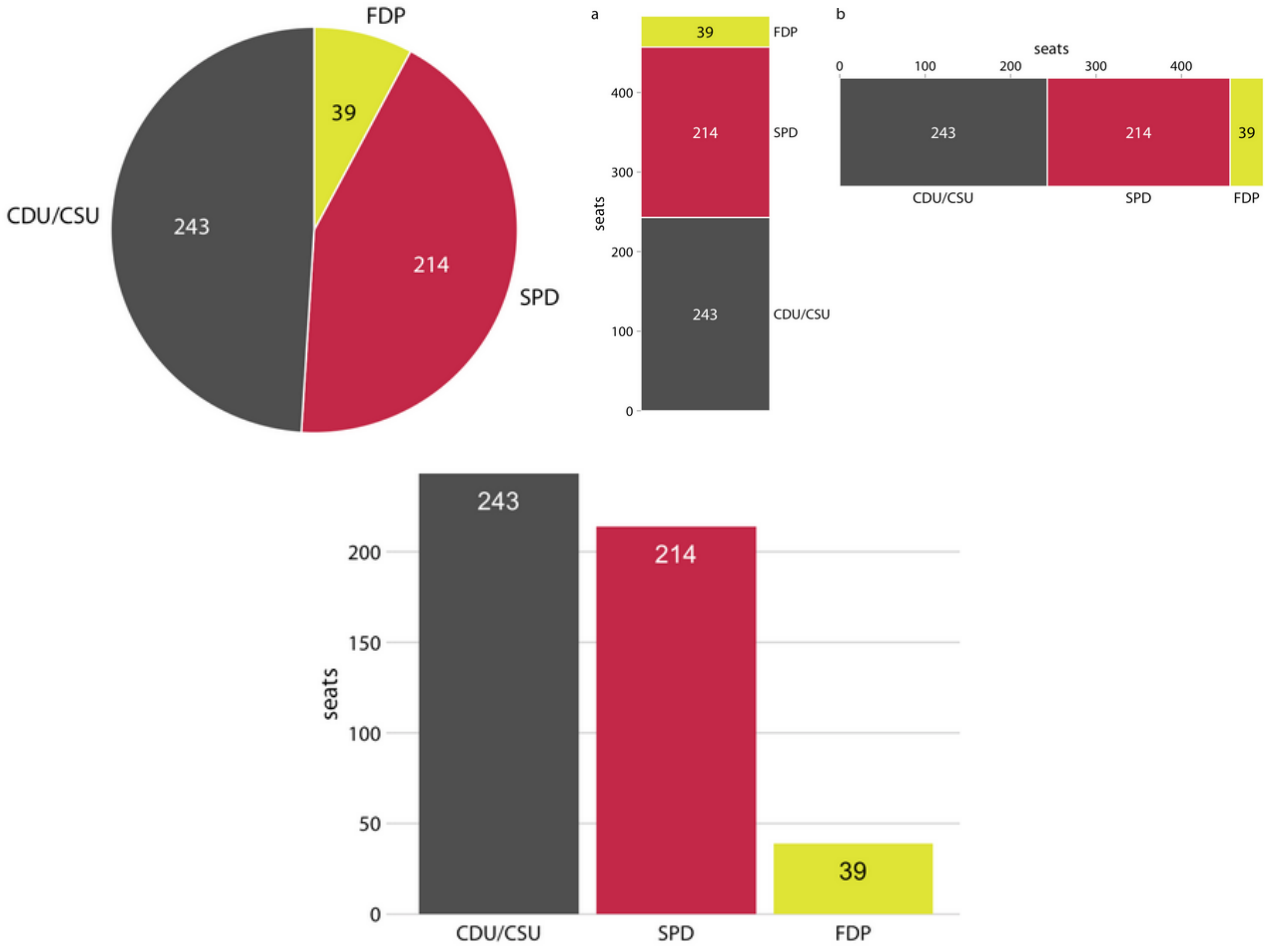


Visualizing proportions

Visualizing proportions

We often want to show how some group, entity, or amount breaks down into individual pieces that each represent a *proportion* of the whole. The archetypal such visualization is the pie chart, omnipresent in any business presentation and much maligned among data scientists. As we will see, visualizing proportions can be challenging, in particular when the whole is broken into many different pieces or when we want to see changes in proportions over time or across conditions

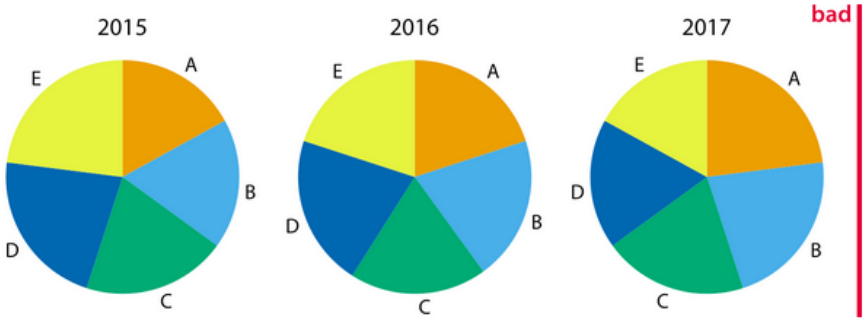
A case for pie charts

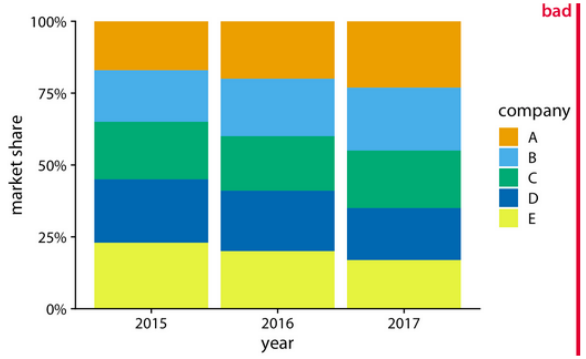


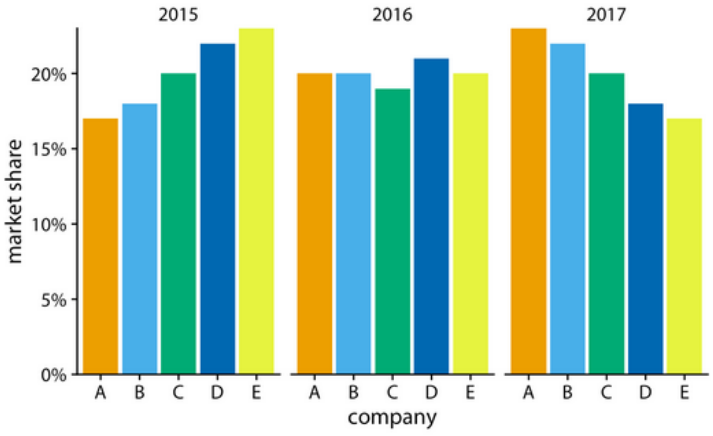
A case for pie charts

	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✔	✔	✖
Allows easy visual comparison of the relative proportions	✖	✖	✔
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✔	✖	✖
Looks visually appealing even for very small datasets	✔	✖	✔
Works well when the whole is broken into many pieces	✖	✖	✔
Works well for the visualization of many sets of proportions or time series of proportions	✖	✔	✖

A case for side-by-side bars



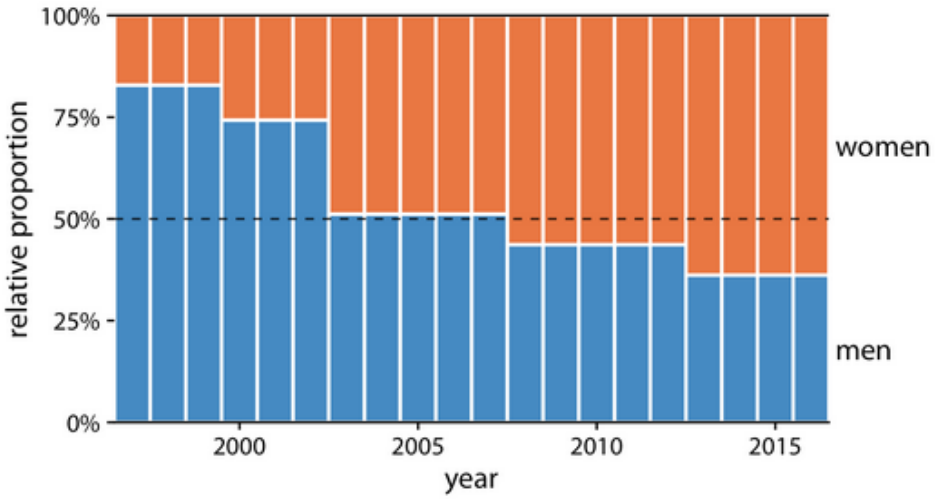




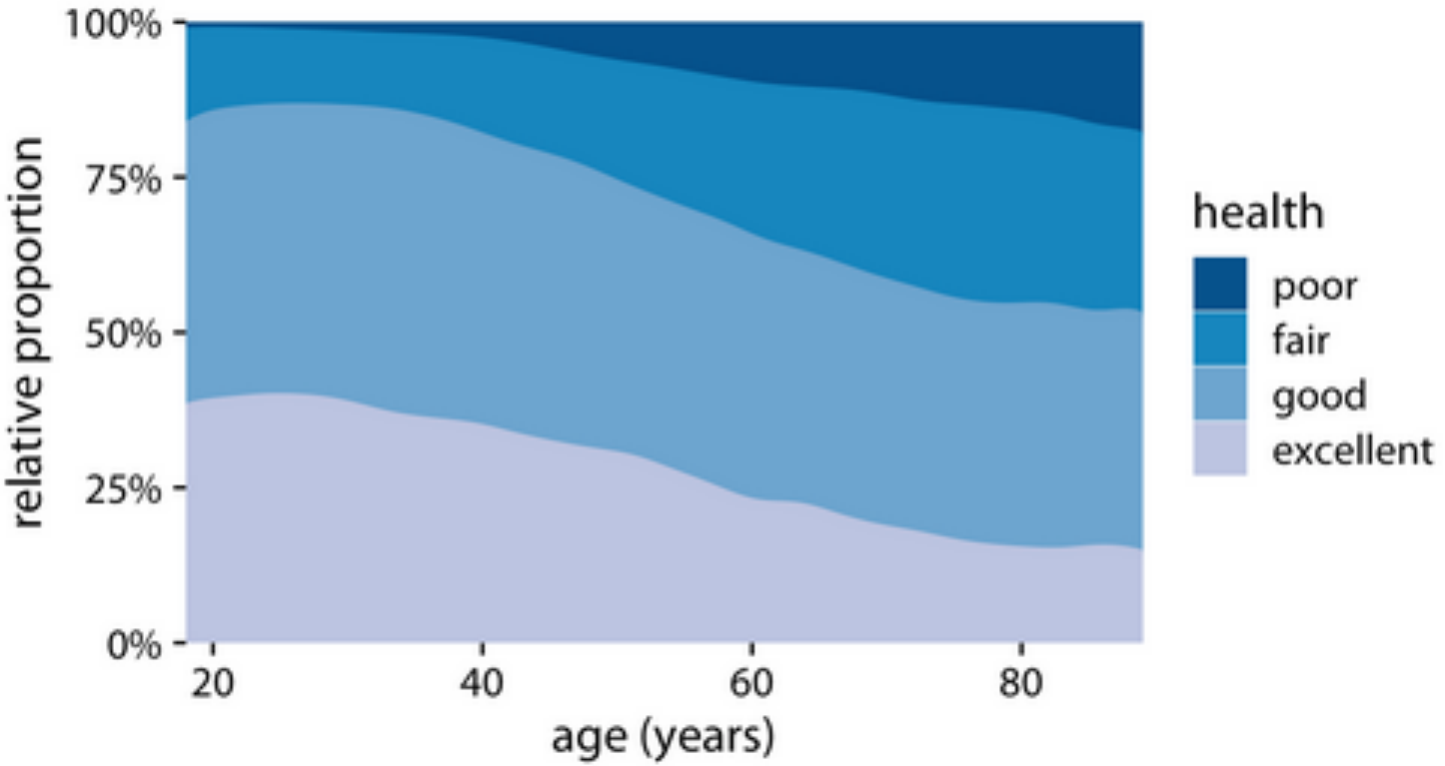
A case for stacked bars and stacked densities

Previously it was not recommend to use sequences of stacked bars, because the location of the internal bars shifts along the sequence.

However, the problem of shifting internal bars disappears if there are only two bars in each stack, and in those cases the resulting visualization can be quite clear.

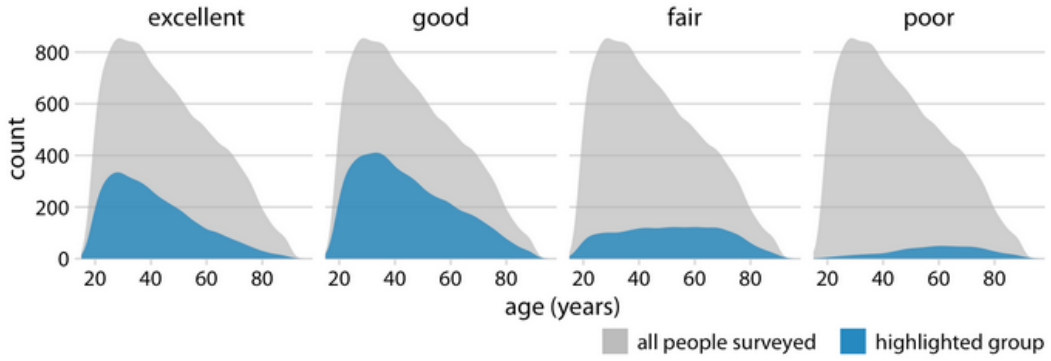


A case for stacked bars and stacked densities

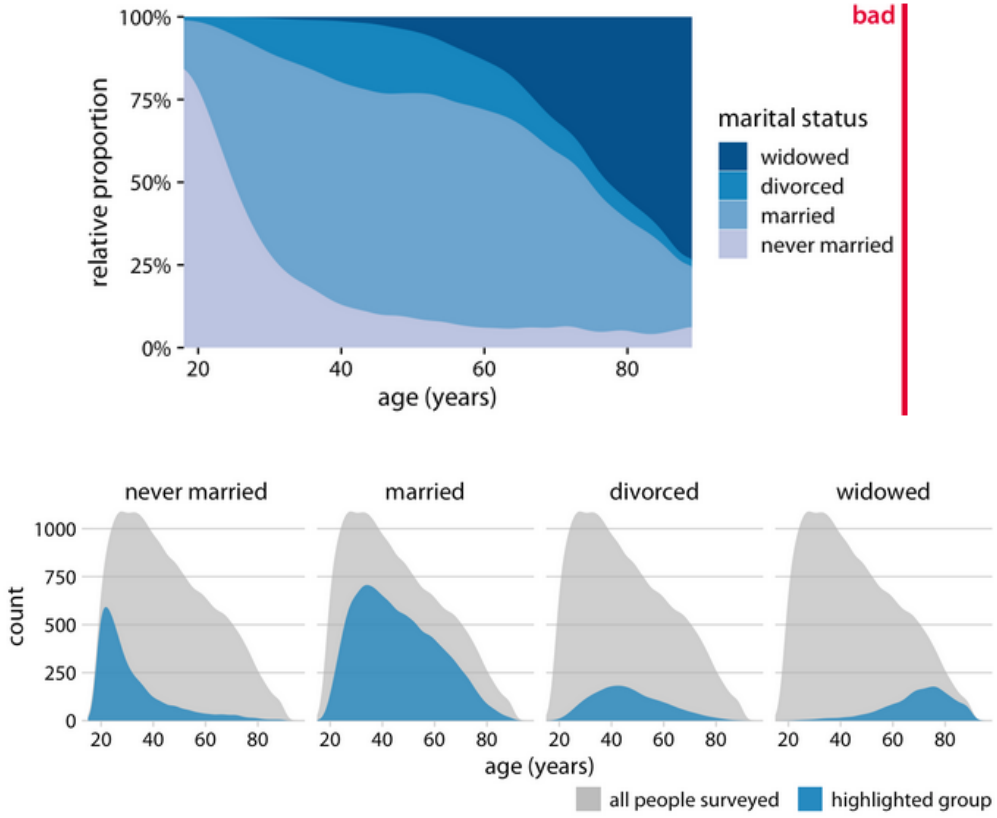


Visualizing proportions separately as parts of the total

Side-by-side bars have the problem that they don't clearly visualize the size of the individual parts relative to the whole and stacked bars have the problem that the different bars cannot be compared easily because they have different baselines. We can resolve these two issues by making a separate plot for each part and in each plot showing the respective part relative to the whole.

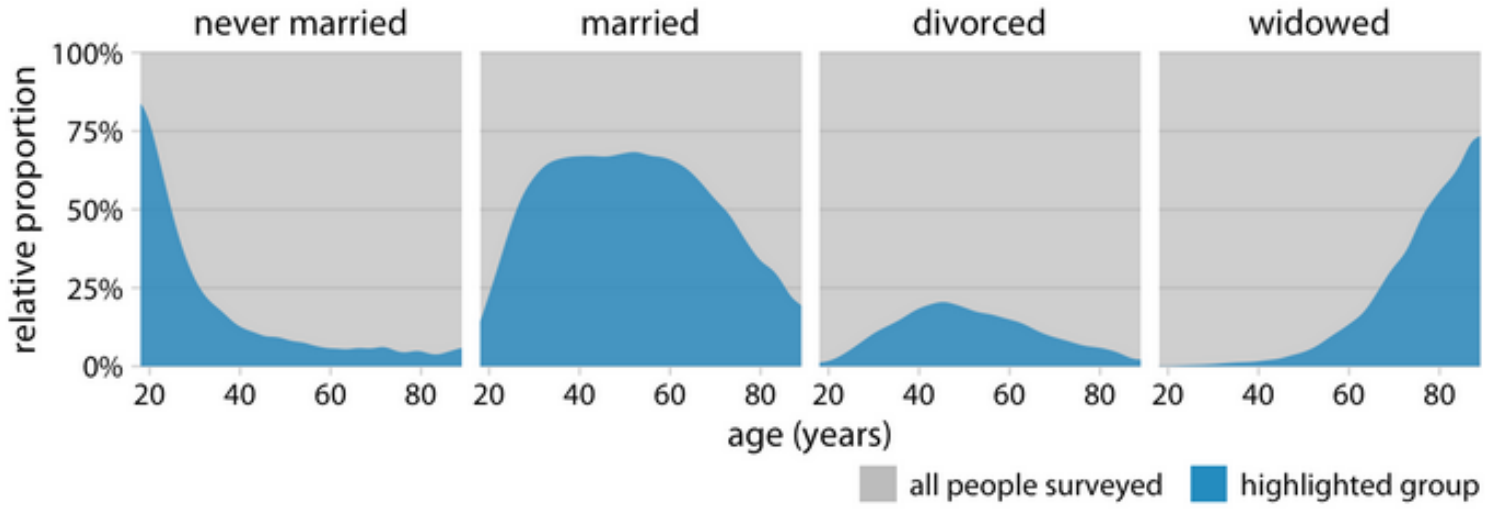


Visualizing proportions separately as parts of the total



Visualizing proportions separately as parts of the total

However, one downside of separate distribution plots is that this representation doesn't make it easy to determine relative proportions at any given point in time. For example, if we wanted to know at what age more than 50% of all people surveyed are married, we could not easily tell from the previous figure. To answer this question, we can instead use the same type of display but show relative proportions instead of absolute counts along the *y* axis



Visualizing nested proportions

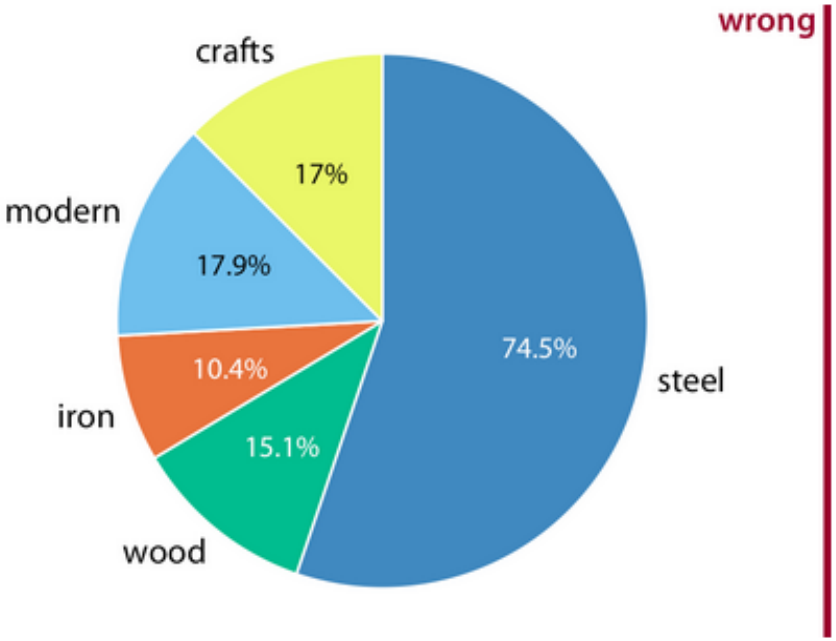
Visualizing nested proportions

It is not uncommon, however, that we want to drill down further and break down a dataset by multiple categorical variables at once. Such scenarios are refered as nested proportions, because each additional categorical variable that we add creates a finer subdivision of the data nested within the previous proportions.

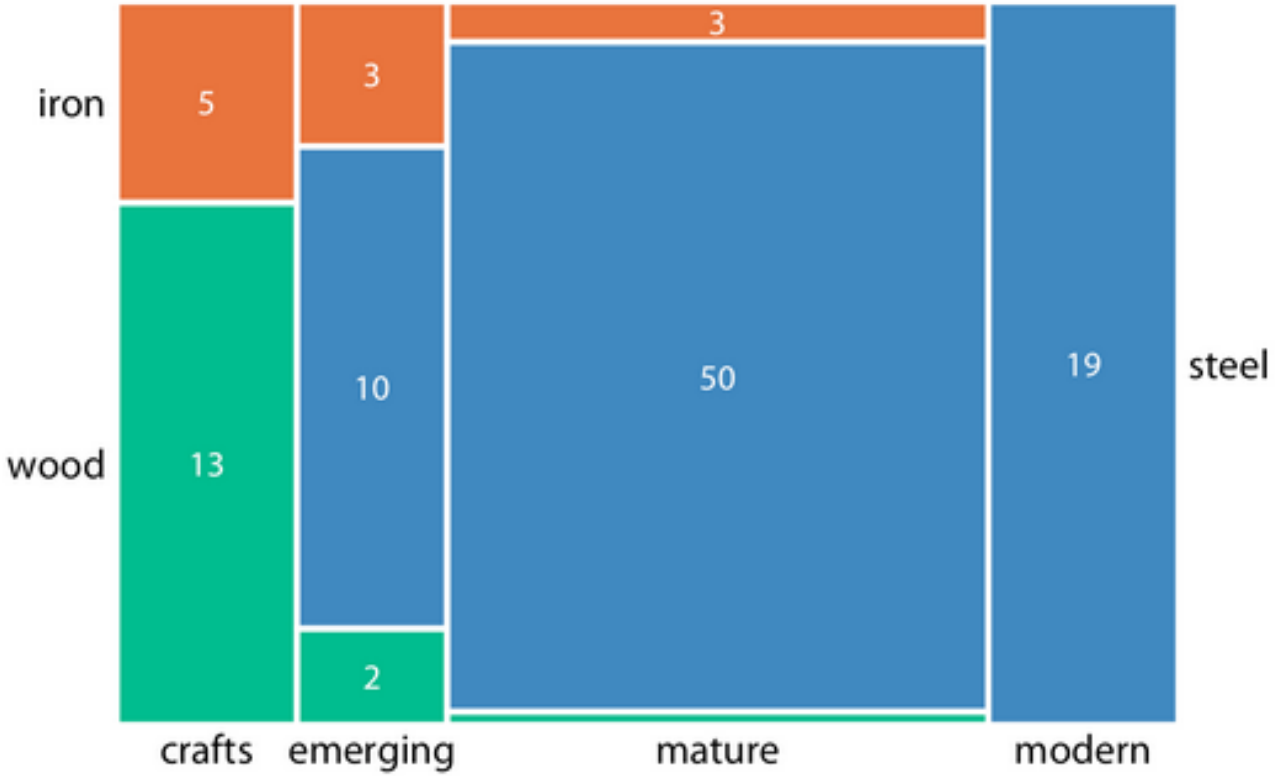
Nested proportions gone wrong

Let's assume we want to visualize both the two independent fractions of the given data. We might be tempted to do so by drawing a combined pie chart. However, this visualization is not valid. All the slices in a pie chart must add up to 100%, and here the slices add up to 135%.

We reach a total percentage in excess of 100% because we are double-counting bridges. Every bridge in the dataset is made of steel, iron, or wood, so these three slices of the pie already represent 100% of the bridges. Every crafts or modern bridge is also a steel, iron, or wood bridge, and hence is counted twice in the pie chart.

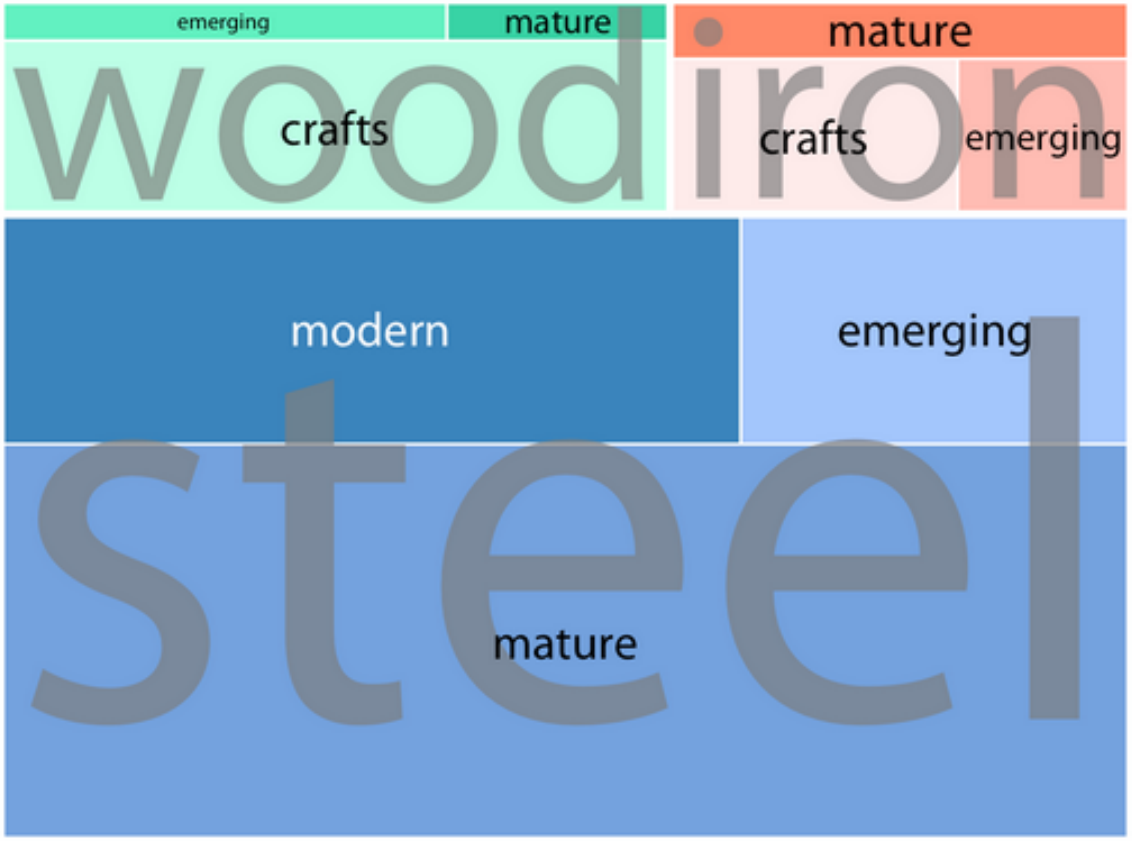


Mosaic plots and treemaps



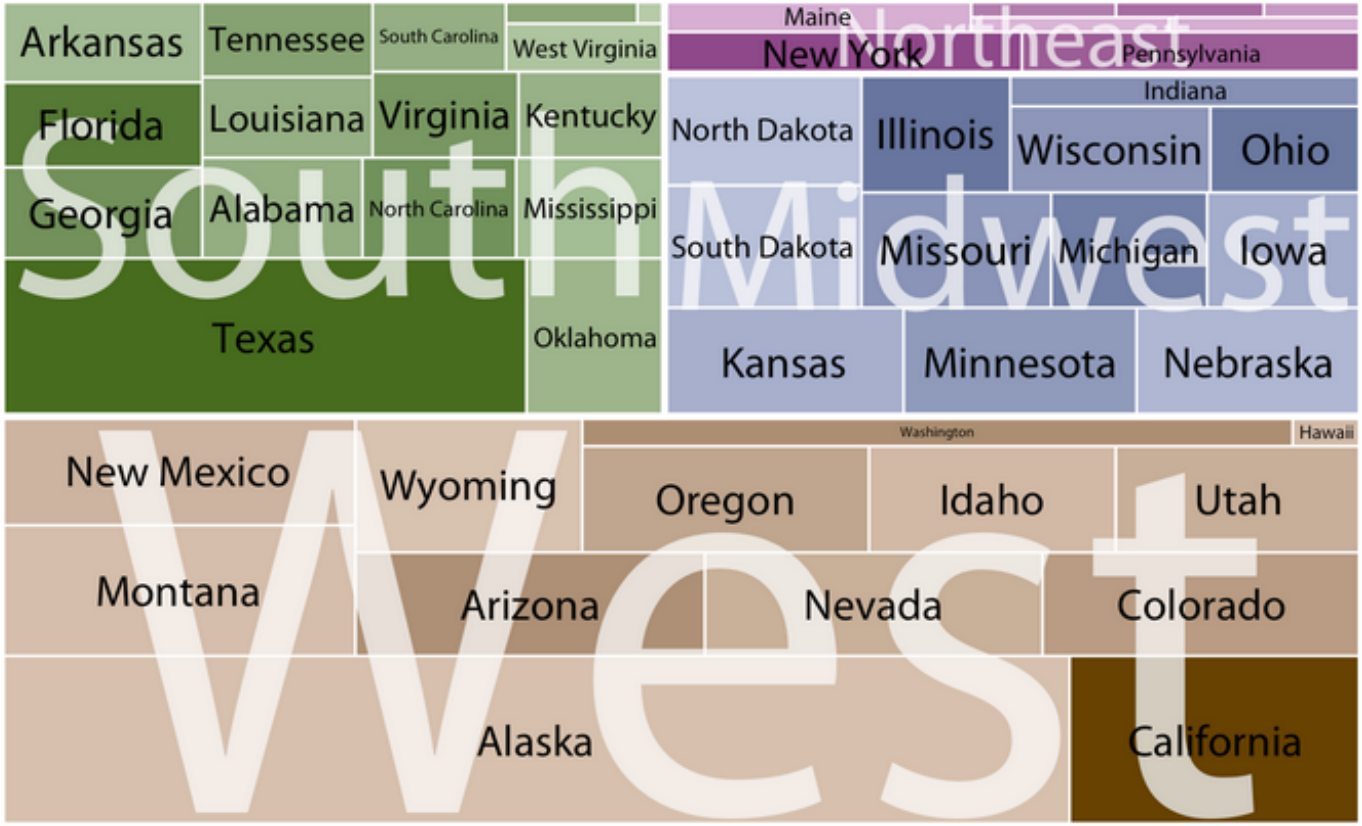
Mosaic Plots

Mosaic plots and treemaps

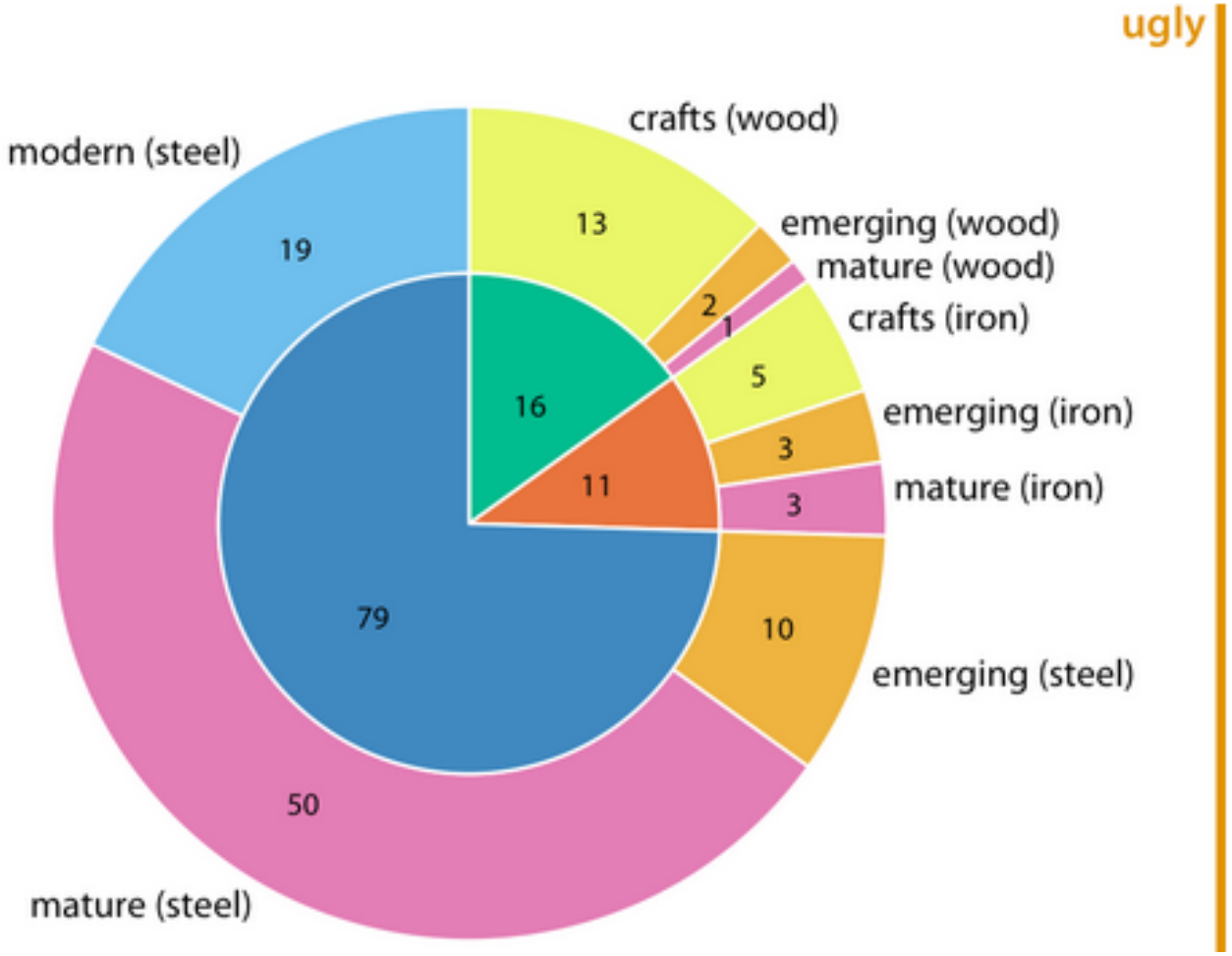


Treemaps

Mosaic plots and treemaps

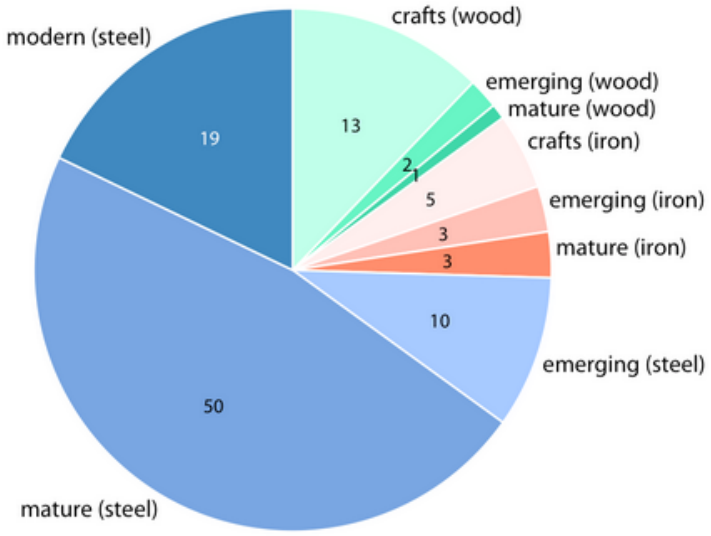


Nested pies


ugly
crafts (wood)
modern (steel)
13
19
emerging (wood)
mature (wood)
2
1
crafts (iron)
5
16
emerging (iron)
3
11
mature (iron)
3
79
10
emerging (steel)
50
mature (steel)

Nested pies

Alternatively, we can first slice the pie into pieces representing the proportions according to one variable and then subdivide these slices further according to the other variable. In this way, in effect we are making a normal pie chart with a large number of small pie slices. However, we can then use coloring to indicate the nested nature of the pie.



Parallel sets

When we want to visualize proportions described by more than two categorical variables, mosaic plots, treemaps, and pie charts all can quickly become unwieldy. A viable alternative in this case can be a *parallel sets plot*. In a parallel sets plot, we show how the total dataset breaks down by each individual categorical variable, and then we draw shaded bands that show how the subgroups relate to each other.