Visualizing nested proportions -

We can break down a dataset in multiple categorical variable at once, each dataset have subdivision of nested data.
For example(refers from previous pdf),

1. In the case of parliamentary seat, we could be interested in the proportion of seats by party and by the gender.
2. In the case of people's health status, we could breaks down health status further by marital status.

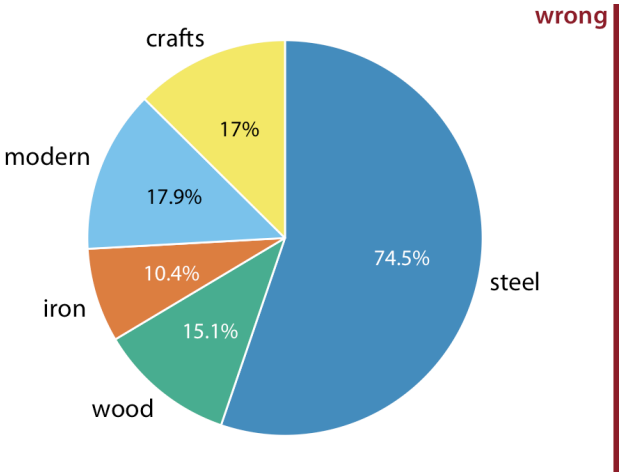
This scenarios can be referred as nested proportion because each additional variable has it's finer subdivision.

Approches - mosiac plots, trempas, ans parallel seats.

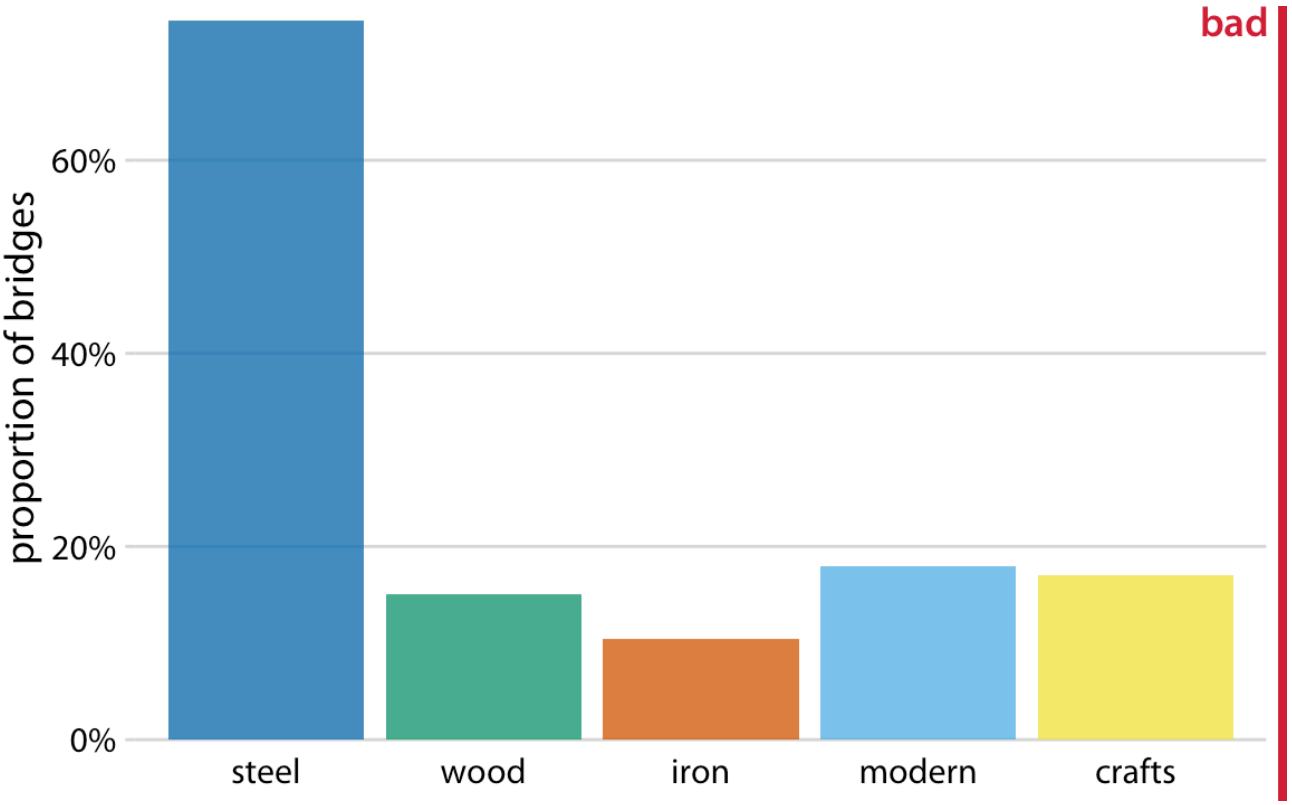
Nested proportions gone wrong -

In this portion, I will work with a dataset of 106 bridges. This dataset contains various pieces of information about bridges, such as the material from which they are constructed(steel, iron or wood) and the year when they were erected. Based on year of reaction, bridges are grouped into distinct categories, such as crafts bridges that were erected before 1870 and modern that were erected after 1940.

Let's assume we want to visualize both the traction of bridges made from steel, iron, or wood and fraction that are crafts or modern. We might be tend to draw a combined pie chart. However this is not a valid plot Because the total percentage of a pie chart must be 100% but it is 135% the reason behind is that we are double-counting the bridges.
Every bridge in the dates is made of still, iron, or wood, so these slices of the pie already represent 100% of bridges. Every crafts or modern is also a steel, iron, or wood bridge, and hence is counted twice in the pie chart.



Double counting is not the necessarily problem if total of proportion is 100%. A side-by-side bar plot can be next thinking of our mid to draw the dataset but it will also be 'bad' plot, because it does not immediately show that there is overlap among some of the categories. In a first seen it can interpreted as five separate categories of bridges, and that, for example, modern bridges are neither made of steel nor of wood or iron.

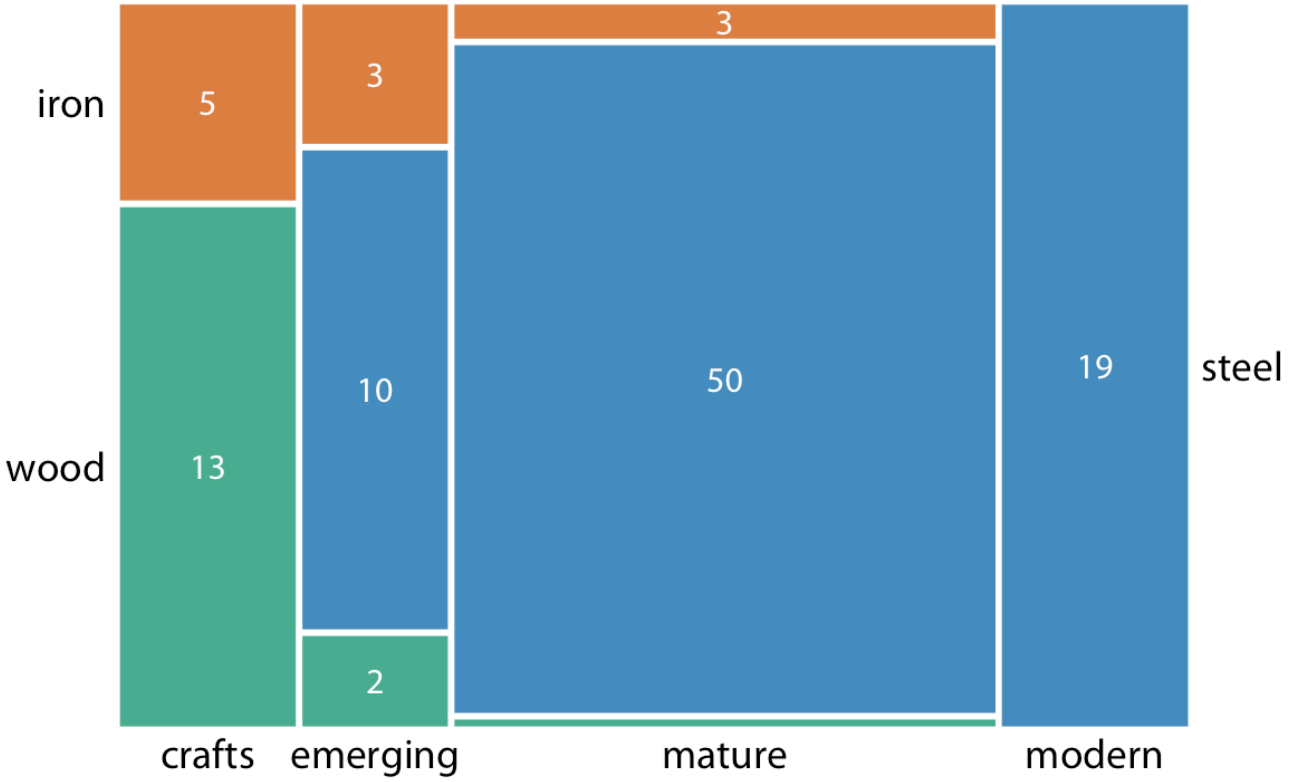


Mosaic plots treemaps -

Whenever we have categories that overlaps, mosaic plot are clear to show relation between categories. On first glance, a mosaic plot looks similar to a stacked bar plot. However, unlike in staked bar plot, in a mosaic plot both heights and widths of individual shaded areas vary.'

We can see two additional construction eras, emerging(from 1870 to 1889) and mature(1890 to 1939). In combination with crafts and modern, these construction eras cover all bridges in the dataset,

Every categorical variable shown must cover all the observation in the dataset.



Drawing mosaic plot-

To draw a mosaic plot,

- begin by placing one categorical variable along the x axis(here, era of bridge construction).
- Then place the other categorical variable along the y axis(here, building material) and within each category along the x axis subdivide the y axis by the relative proportion(in this, according to steel, iron, or wood).

The result is a set of rectangle whose areas are proportional to the number of cases representing each possible combination of categorical value.

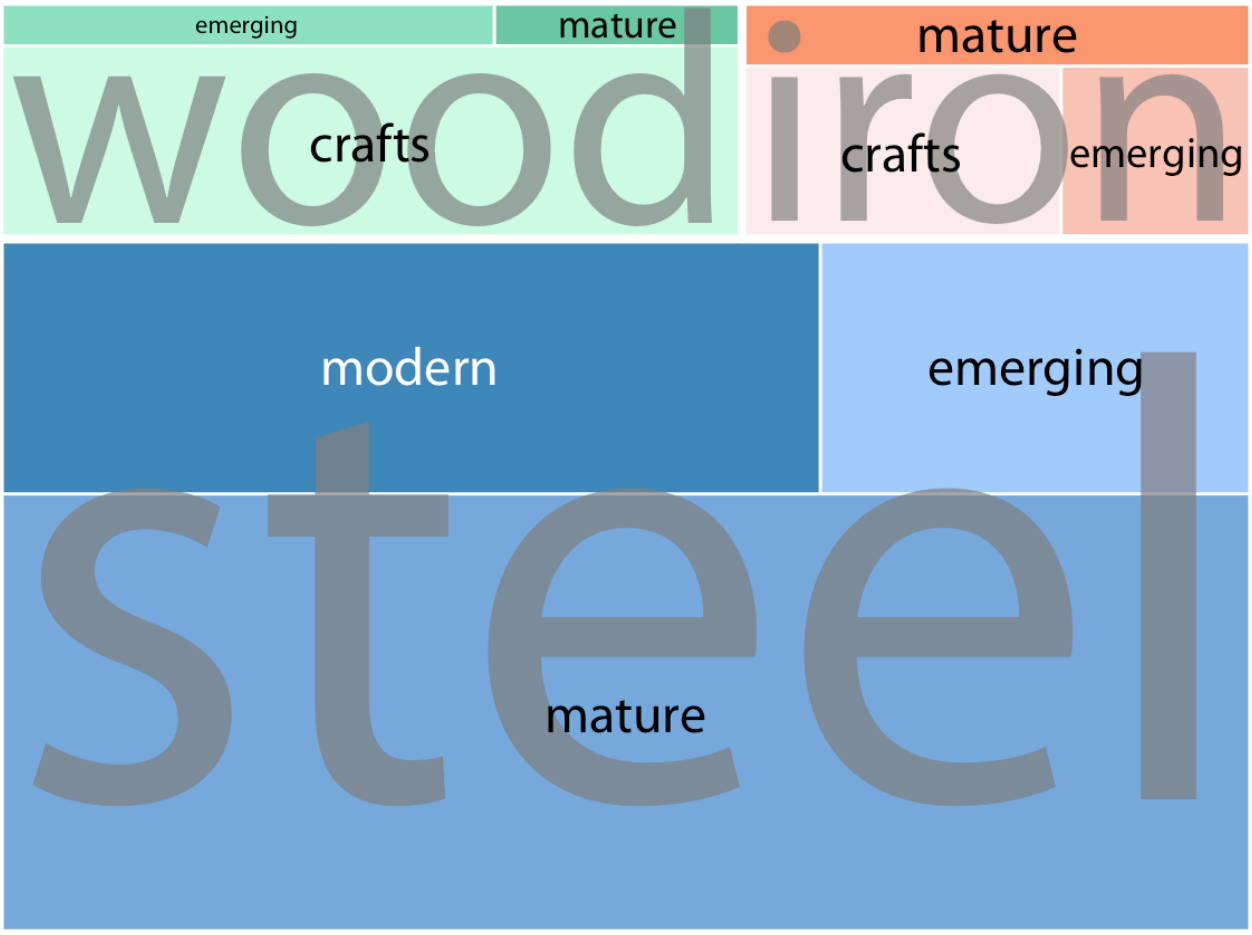
Treemaps-

In a treemap, we take an enclosing rectangle and subdivide into smaller areas represent the proportions. We recursively nest rectangles inside each other.

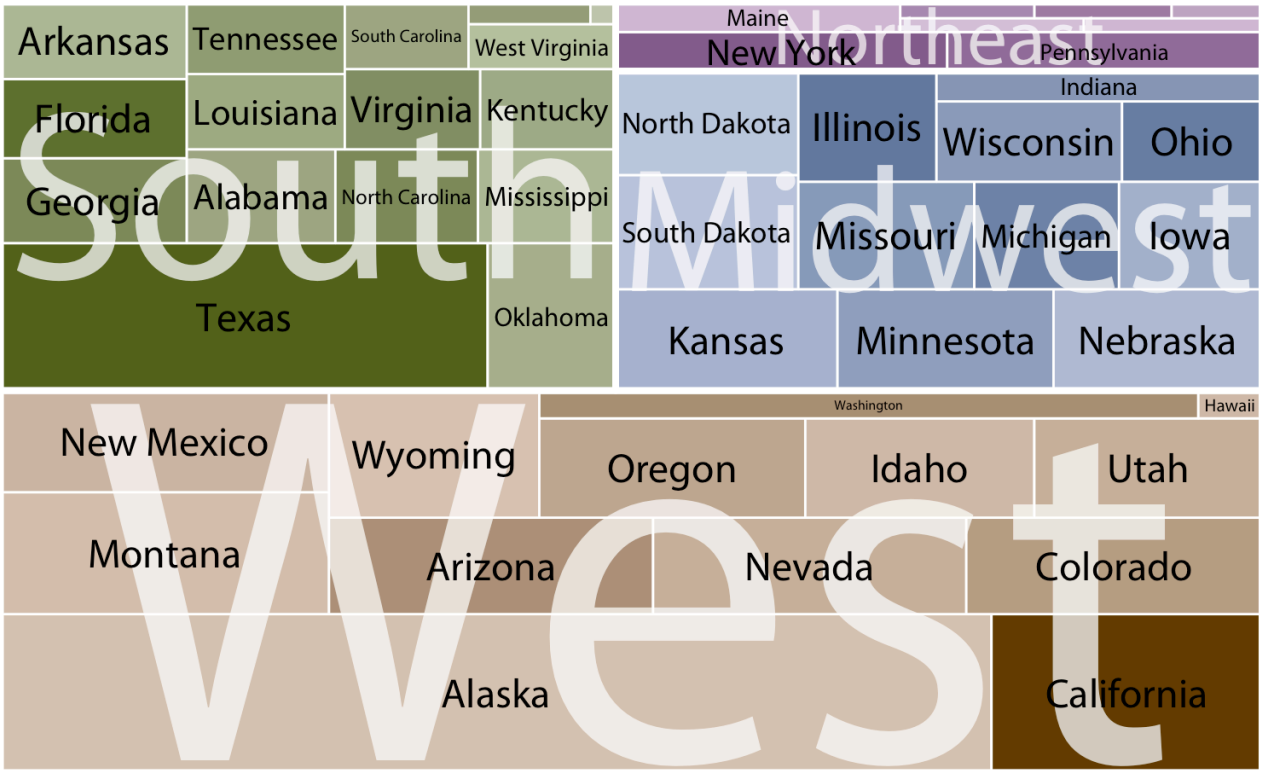
- Primary variable - For plotting above dataset we can subdivide the total area into three rectangles representing the three building materials wood, iron, and steel.

- Secondary variable - Then, we subdivide each of three rectangles further to represent the construction era(crafts, emerging, modern, mature).

Size of Nested rectangles shows the proportion of subdivided categorical values of a rectangle.
Here size of nested rectangles showing the number of brides made in that era with a particular material.



Treemaps tend to work well when the proportion is not described meaningful.
For example, we can separate the U.S into four regions(West, Northeast, Midwest, and South) and each region subdividing into distinct state. We can see that state in one region have no relationship in another region.



In the bridge dataset example we can see the relationship between material of bridge and era in which they constructed.
A={steel, iron, wood}, B={modern, emerging, mature, crafts}

A relation B =
{ {steel, modern}, {steel, emerging}, {steel, mature},
{iron, crafts}, {iron, mature}, {iron, emerging},
{wood, crafts}, {wood, emerging}, {wood, mature} }

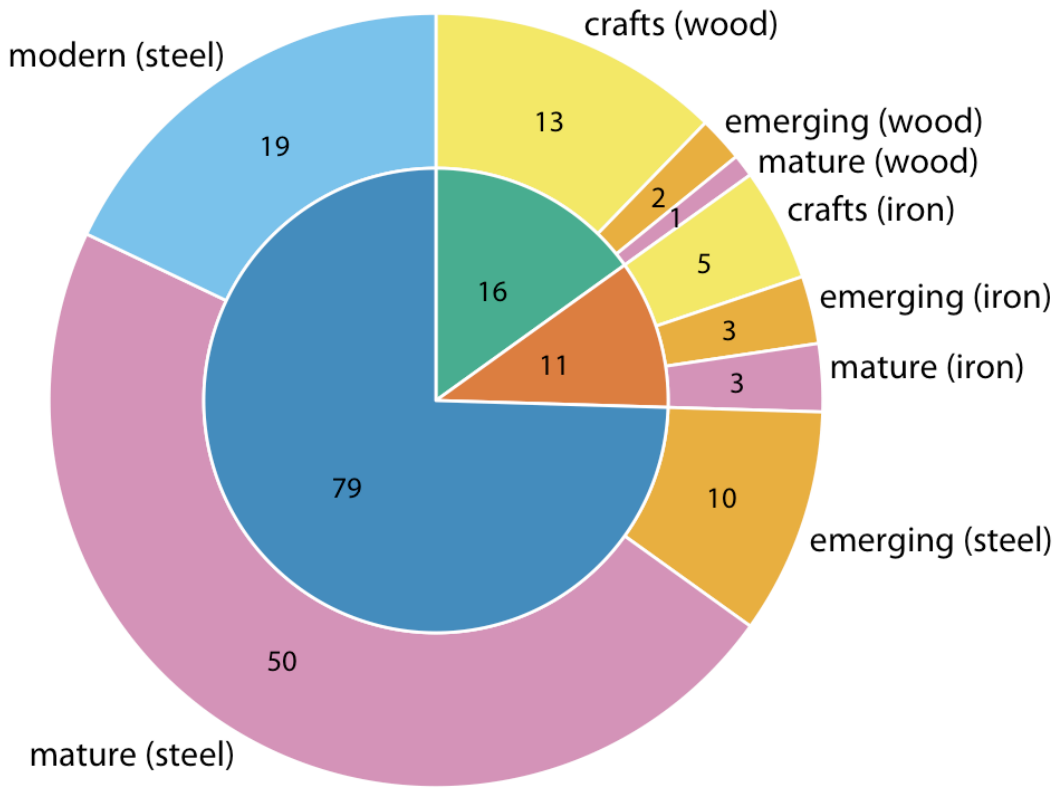
We can not see this type of relation in the above US dataset.

Nested pie charts -

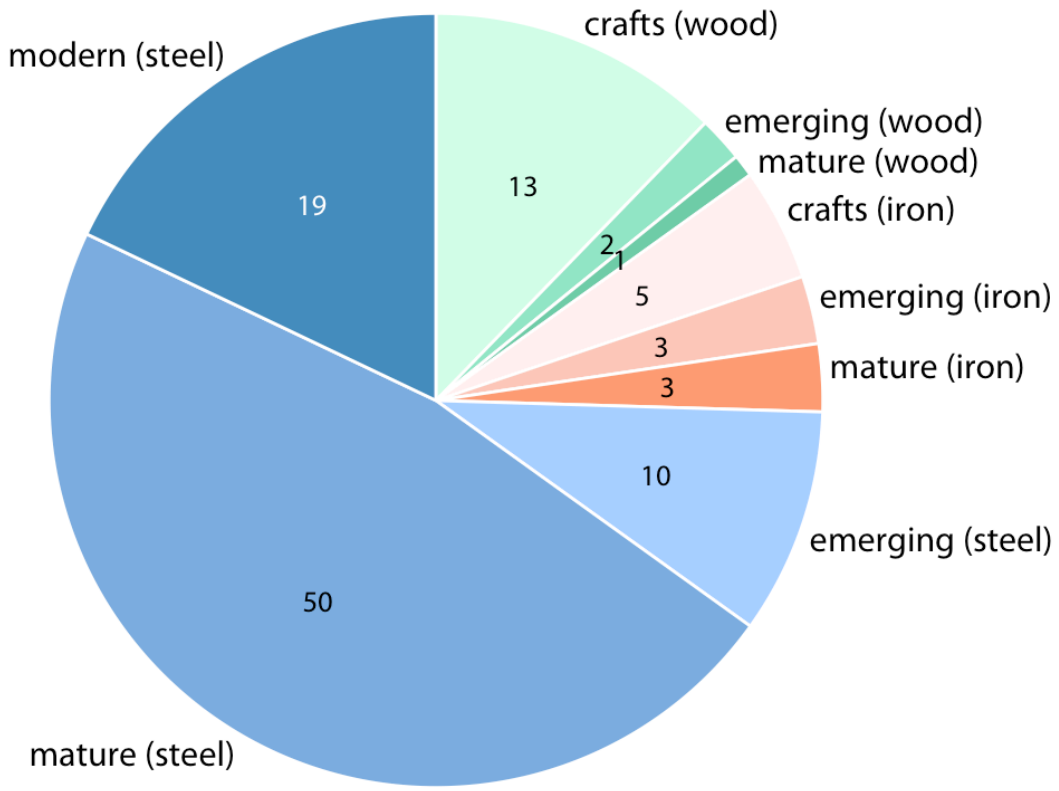
Both mosaic and treemaps are closely related to pie charts. Since they all use area to represent data values. The primary difference is the coordinates system, pie chart uses polar ordinates and mosaic or teemap uses Cartesian coordinate system.

We can plot a bridge dataset with nested pie chart but it falls in category of ‘Ugly’. The inner circle shows the building material and outer circle breakdown the slice of inner circle by the categorical value bridge construction era. The two separate circles obscure the fact that each bridge dataset in the dataset has both building material and era of bridge construction. In fact we are still doubling-counting each bridge.

ugly



Alternatively, we can first slice the pie into pieces representing the proportion according to one variable(here, material) and then subdivide these slices further to other variable(here, construction era). It is like we are making a simple pie chart with large number of small pie.
We are broke the primary variable(construction meterial) in colours and the secondary variable(construction era) in pies.

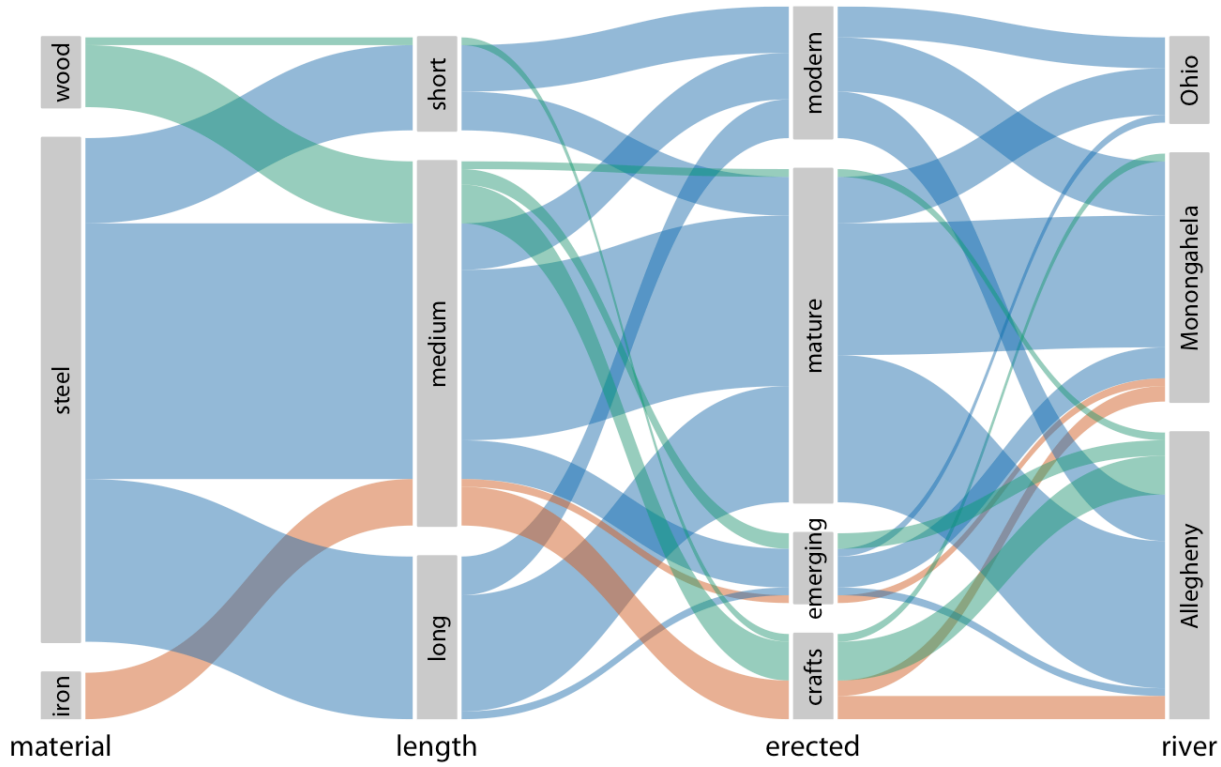


For two categorical variables-
I think mosaic is better then nested pie charts and treemap is better then mosaic plot.

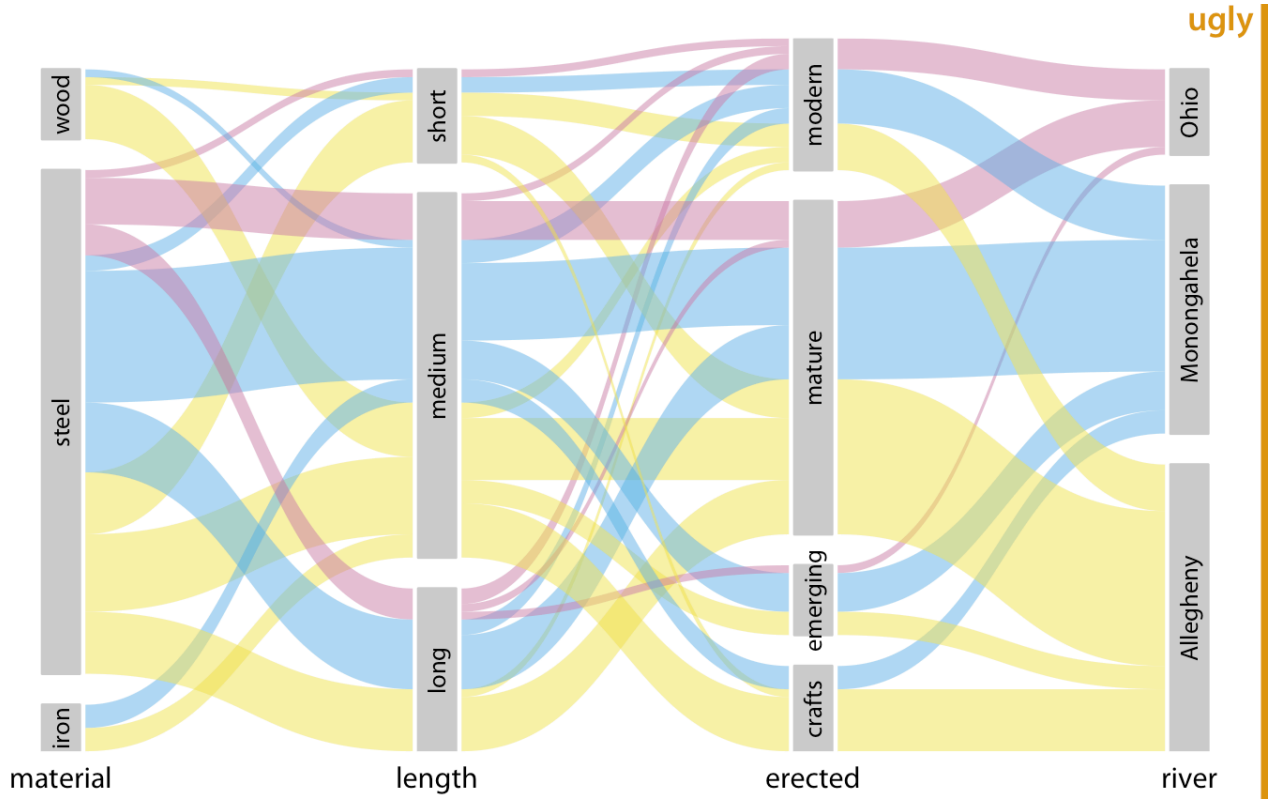
Parallel sets -

When we want to visualize proportion described by more then two categorical variables parallel sets are better the mosaic and treemap.

For example, I have broken down the bridges dataset by construction material(iron, steel, wood), length of each bridge(long, medium, short), era of construction(crafts, emerging, mature, modern), and the river each bridge spans(Allegheny, Moonogahela, Ohio).
The size of band(starting from primary categorical value) passing through the gate(secondary categorical value) shows the length of bridges, era, and river span for particular material.



The same visualisation looks quite different if we colour band by different criteria.



This is good but looking busy with many criss-crossing bands.

We can visualize the plot by changing the order of categorical values. It can minimise the amount of criss-crossing bands.