Oct 21, 2024

Data Visualization

Week 5. Visualizing proportions

Reminder

- visualizing distribution
  - histogram
  - kernel density
- visualizing several distributions
  - error bar
  - box plot
  - violin plot
  - ridgeline plot

Introduction

In practice, it may be necessary to show how a group or quantity is divided into individual parts, each representing a proportion of the whole. For example:

- the ratio of men and women in a community,
- the proportion of voters supporting different political parties in an election,
- the market shares of companies, etc.

The most commonly used chart types for visualizing such proportions are: ...

Pie-chart

A pie chart divides a circle into slices, where the area of each slice represents a part of the whole, proportional to the total.

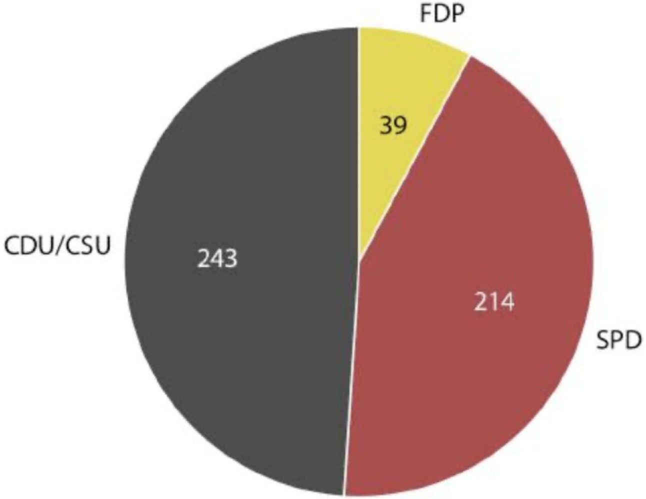


Figure 10-1. Party composition of the eighth German Bundestag, 1976–1980, visualized as a pie chart. This visualization highlights that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU. Data source: Wikipedia.

Pie-chart

Such data can also be visualized using stacked bar charts on either the horizontal or vertical axis.

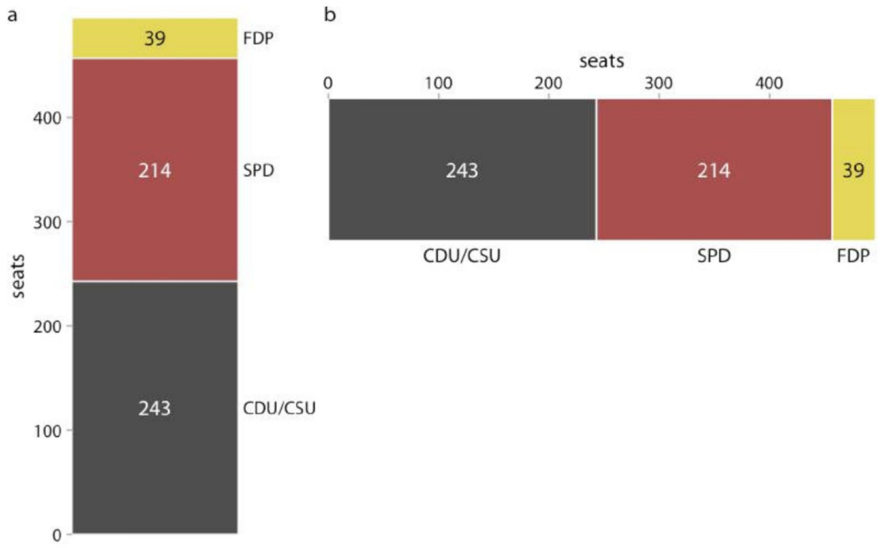
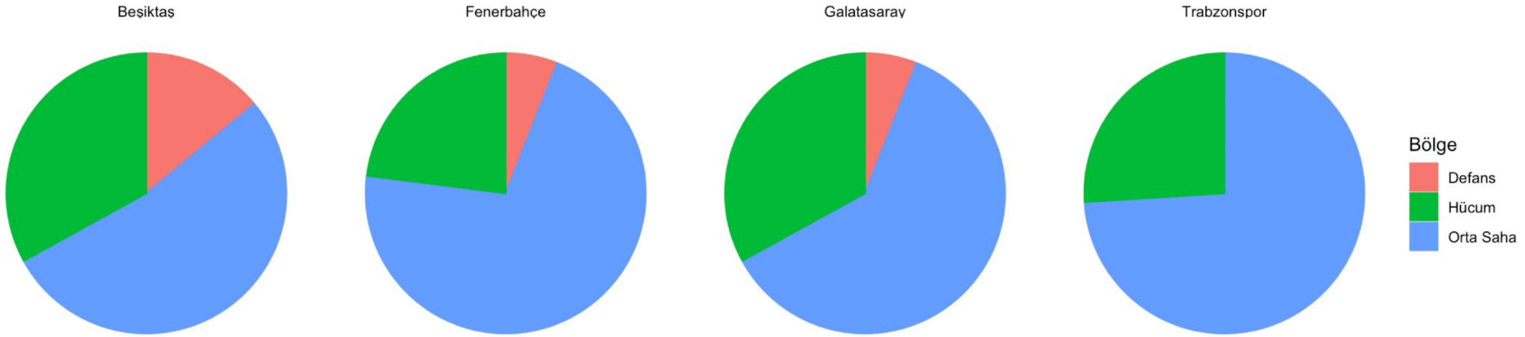


Figure 10-2. Party composition of the eighth German Bundestag, 1976–1980, visualized as stacked bars. (a) Bars stacked vertically. (b) Bars stacked horizontally. It is not immediately obvious that SPD and FDP jointly had more seats than CDU/CSU. Data source: Wikipedia.

Pie-chart

- Which is a better tool for visualizing proportions: pie charts or stacked bar charts is a frequently debated topic.
- Pie charts generally work better for visualizing simple ratios, such as half, one-third, or one-fourth.
- They do not perform well when there are many groups. In such cases, bar charts are a better alternative.

Problems in pie-charts



Barplot

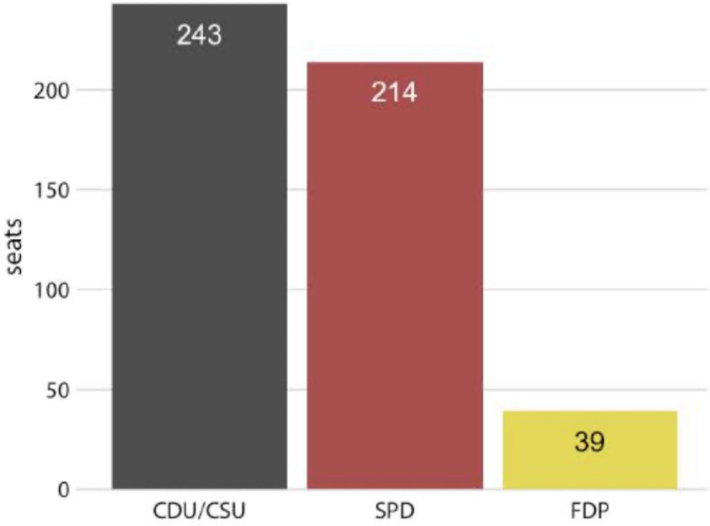
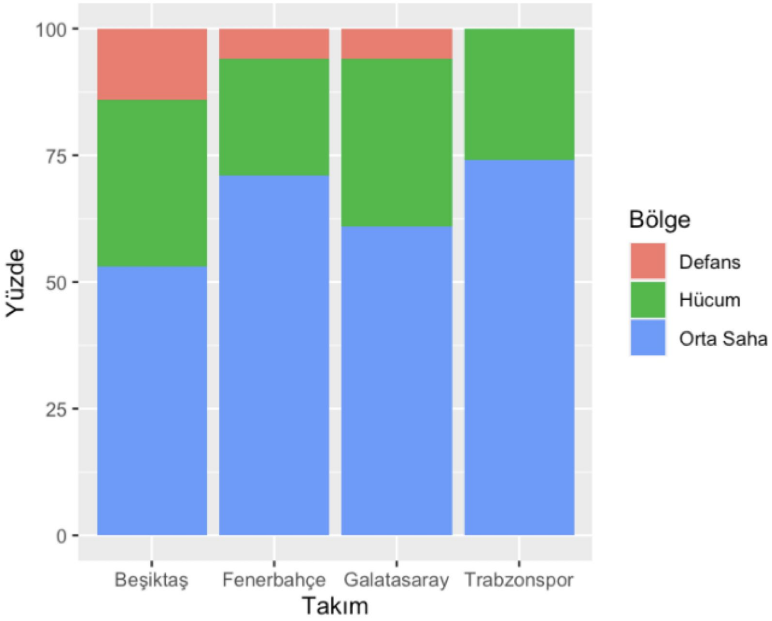
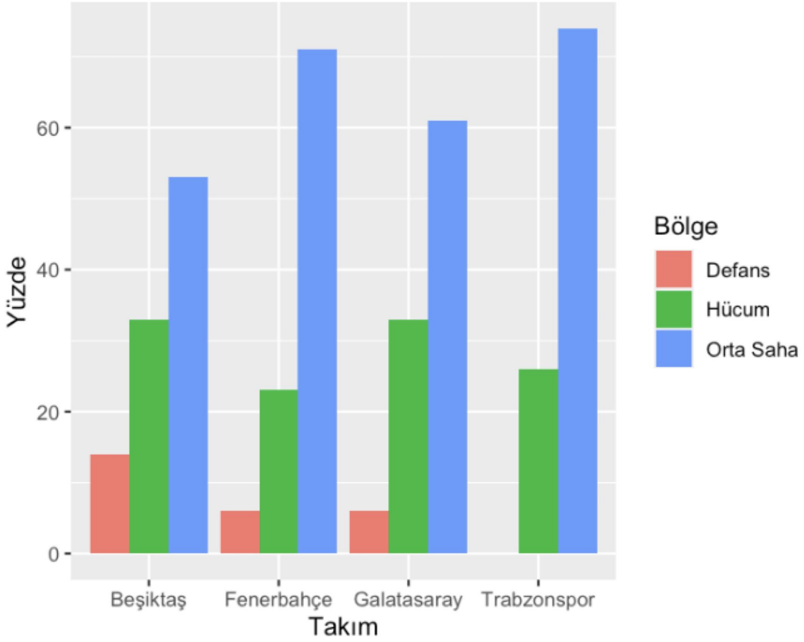


Figure 10-3. Party composition of the eighth German Bundestag, 1976–1980, visualized as side-by-side bars. As in Figure 10-2, it is not immediately obvious that SPD and FDP jointly had more seats than CDU/CSU. Data source: Wikipedia.

Problems in stacked barplots

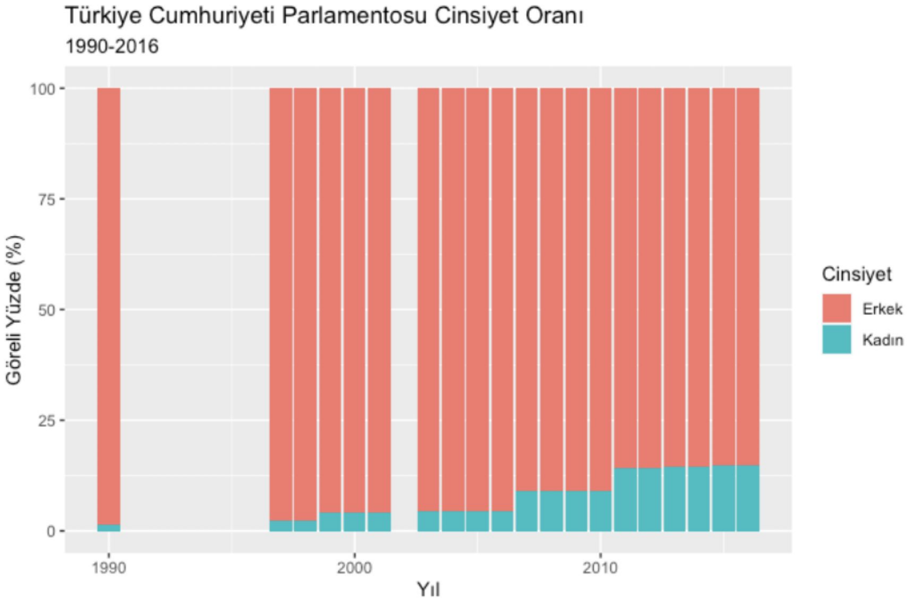


Problems in stacked barplots



Problems in stacked barplots

Stacked bar charts work well for visualizing proportions with only two groups.



Problems in stacked density plots

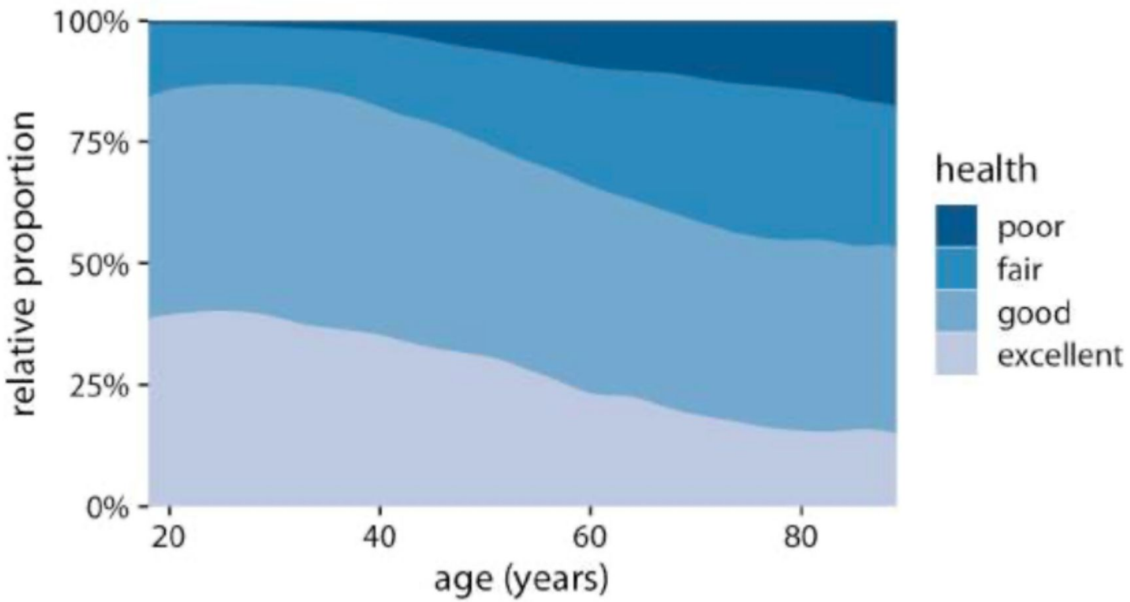


Figure 10-8. Health status by age. Data source: General Social Survey (GSS).

Visualizing proportions separately as parts of the total

Due to some issues with side-by-side charts, we can opt for visualizing the whole by breaking it down into parts.

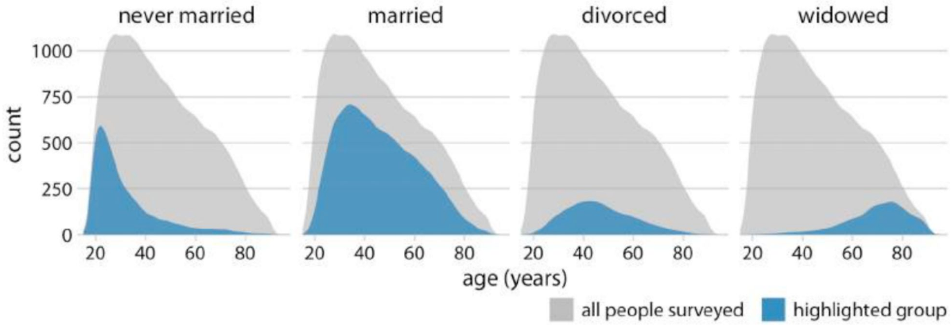


Figure 10-11. Marital status by age, shown as proportion of the total number of people in the survey. The colored areas show the density estimates of the ages of people with the respective marital status, and the gray areas show the overall age distribution. Data source: GSS.

Problems in visualizing proportions separately as parts of the total

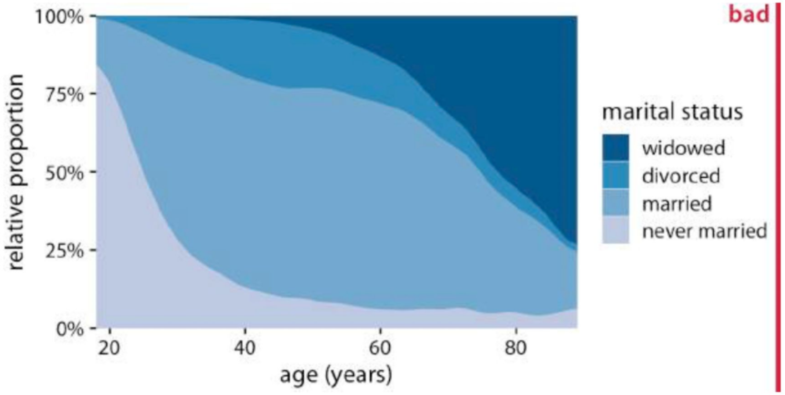


Figure 10-10. Marital status by age. To simplify the figure, I have removed a small number of cases that report as separated. I have labeled this figure as "bad" because the frequency of people who have never been married or are widowed changes so drastically with age that the age distributions of married and divorced people are highly distorted and difficult to interpret. Data source: GSS.

Problems in visualizing proportions separately as parts of the total

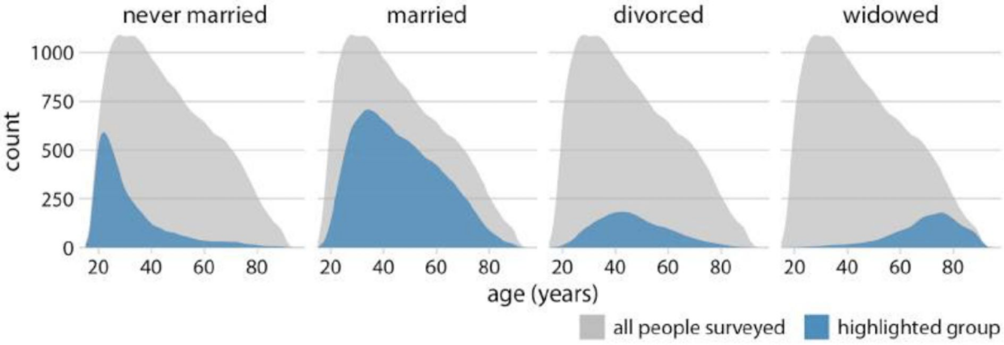


Figure 10-11. Marital status by age, shown as proportion of the total number of people in the survey. The colored areas show the density estimates of the ages of people with the respective marital status, and the gray areas show the overall age distribution. Data source: GSS.

Here, the issue is the difficulty in comparing the relative proportions of groups at any given point in time.

Problems in visualizing proportions separately as parts of the total

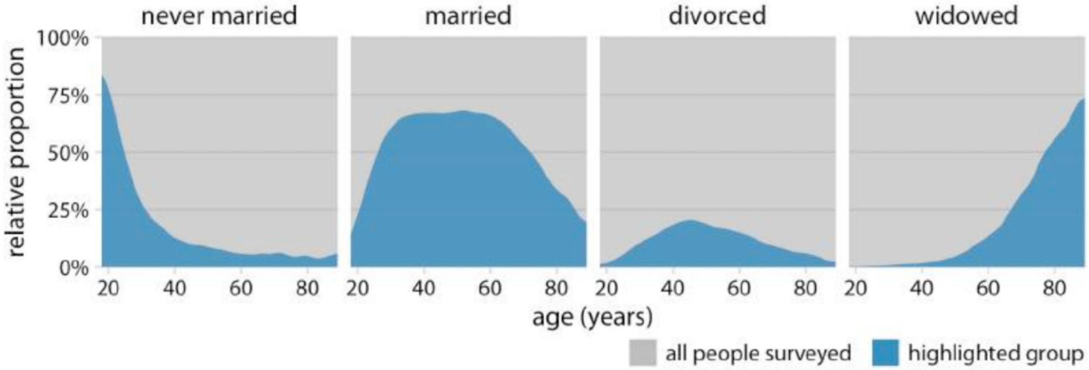


Figure 10-12. Marital status by age, shown as proportion of the total number of people in the survey. The areas colored in blue show the percent of people at the given age with the respective status, and the areas colored in gray show the percent of people with all other marital statuses. Data source: GSS.

Visualizing nested proportions

- We have discussed scenarios where a data set is divided based on a single categorical variable. It may be necessary to examine the data set with multiple categorical variables simultaneously and analyze the breakdowns.
- For example, the voting proportions of parties in a parliament may need to be examined both by parties and by the gender of the candidates. Such examples are referred to as nested proportions.
- Mosaic plots, treemaps, and parallel sets can be used to visualize nested proportions.

Mosaic plots

A mosaic plot consists of square areas representing two categorical variables and the proportions corresponding to the levels of the categorical variable.

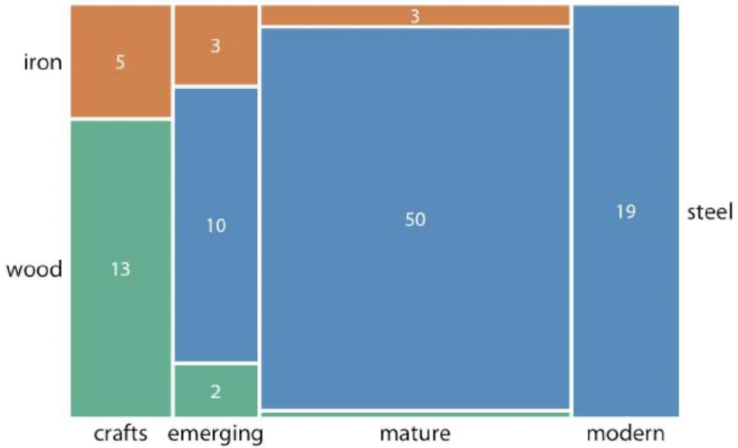


Figure 11-3. Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fenves.

Treemaps

Unlike mosaic plots, treemaps are created by placing smaller rectangles within larger ones.

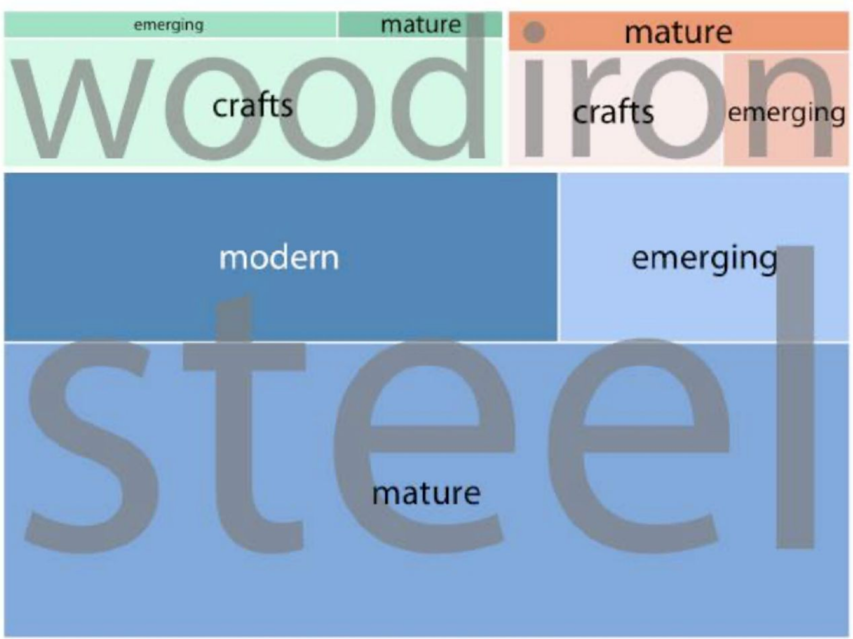
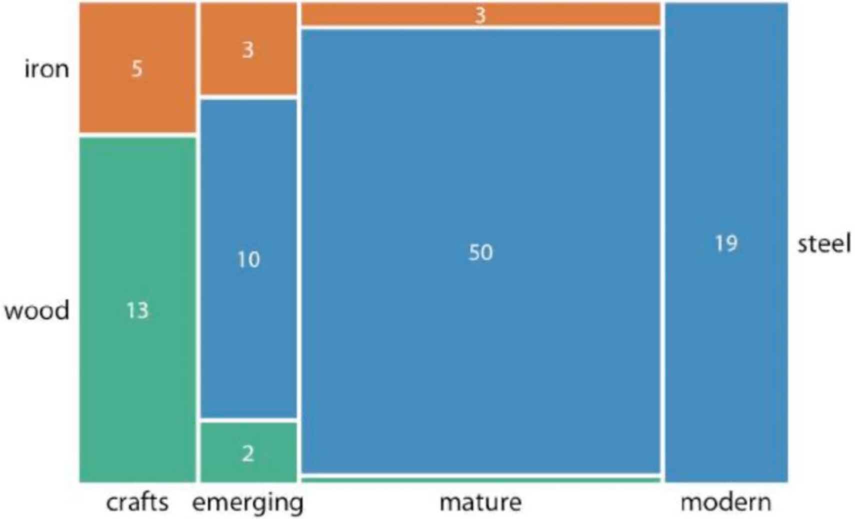
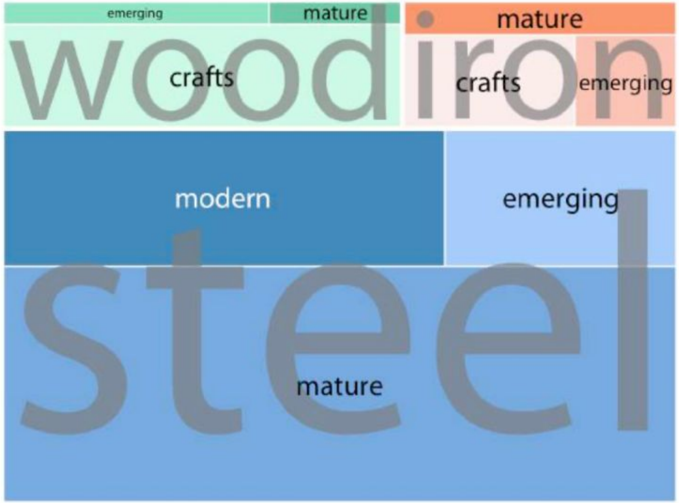


Figure 11-4. Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves.

Mosaic plots vs. Treemaps





Mosaic plots vs. Treemaps

The points of emphasis differ:

The mosaic plot highlights the temporal evolution of building material usage from the craft age to the modern era, while the treemap emphasizes the number of bridges made from steel, iron, and wood.

Sankey diagrams

Pie charts, mosaic plots, and treemaps can become harder to read as the number of levels in the categorical variable increases. In such cases, a Sankey diagram can be used.

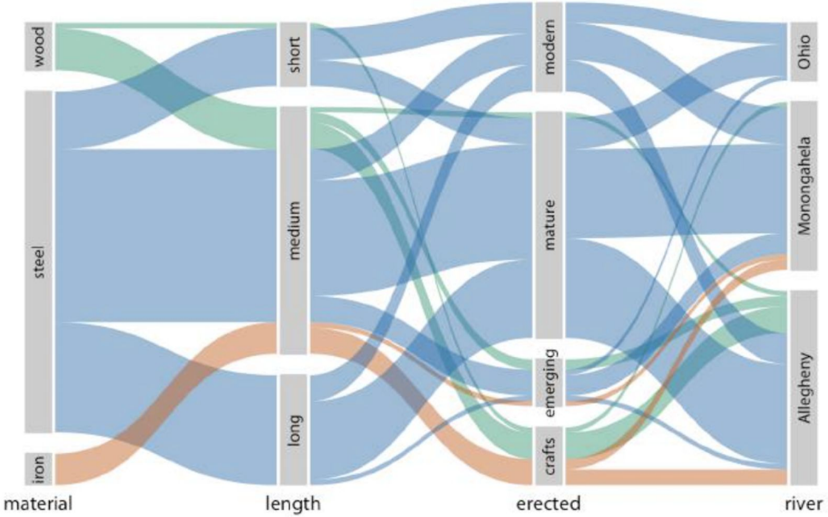
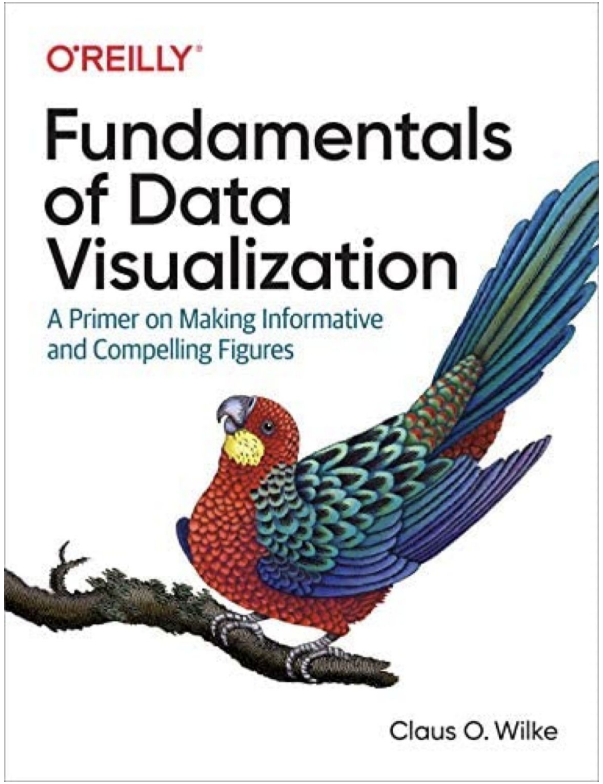


Figure 11-8. Breakdown of bridges in Pittsburgh by construction material, length, era of construction, and the river they span, shown as a parallel sets plot. The coloring of the bands highlights the construction material of the different bridges. Data source: Yoram Reich and Steven J. Fenves.

Reference

The notes and plots in the presentation are compiled from Claus O. Wilke's book, *Fundamentals of Data Visualization*.


O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke