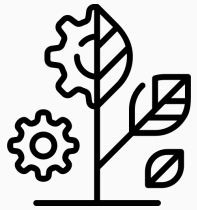
Lecture 1

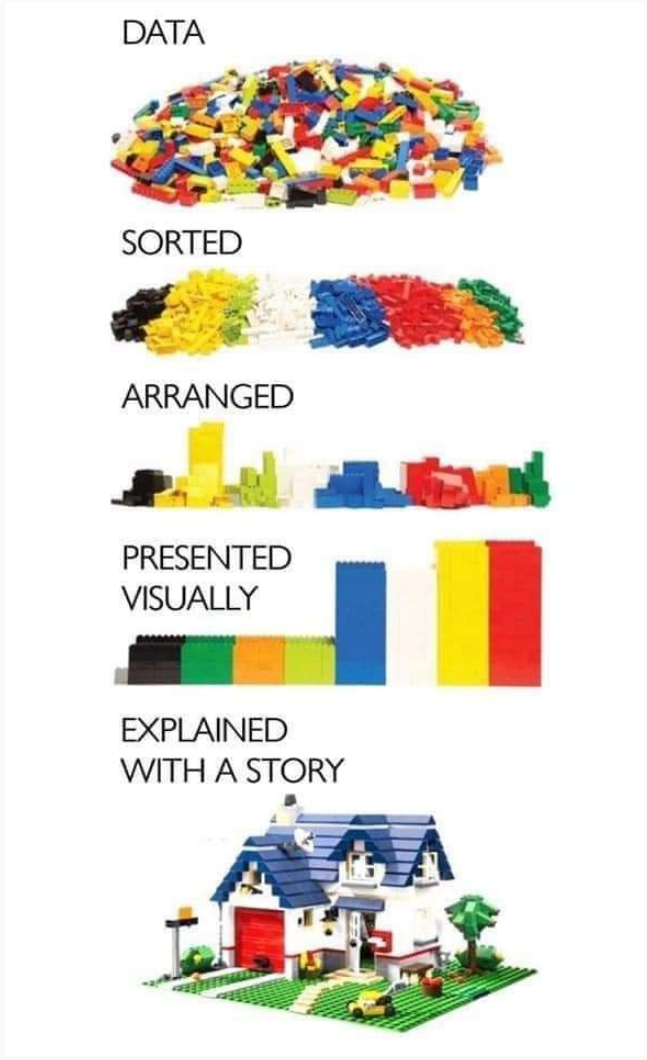
Data Vizualization in R

Principles of figure design

Victoria Mironova

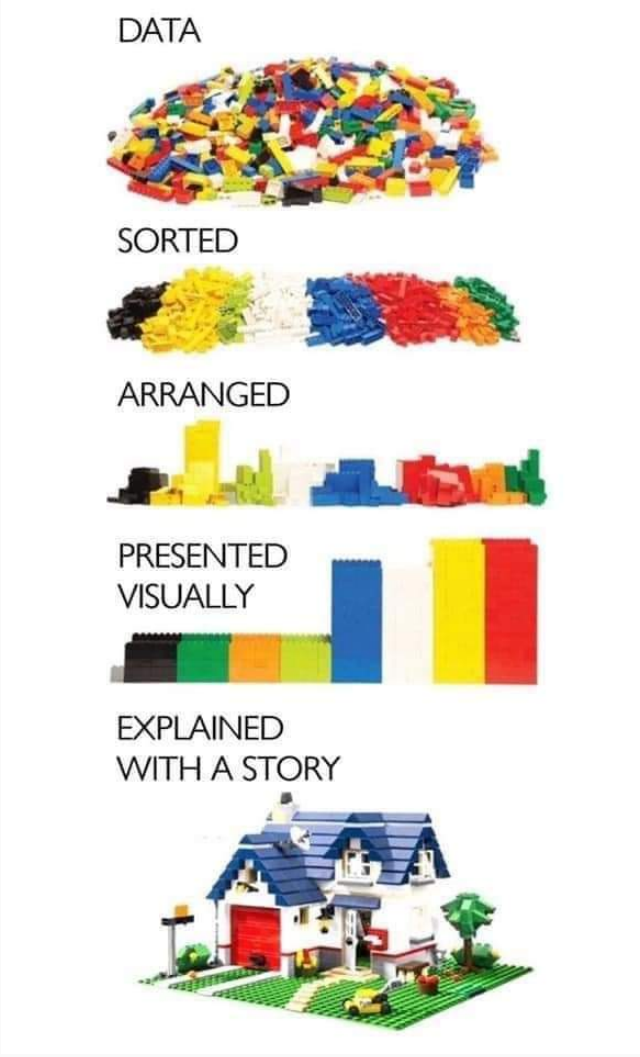
Associate Professor, Department of Plant Systems Physiology





source: unknown

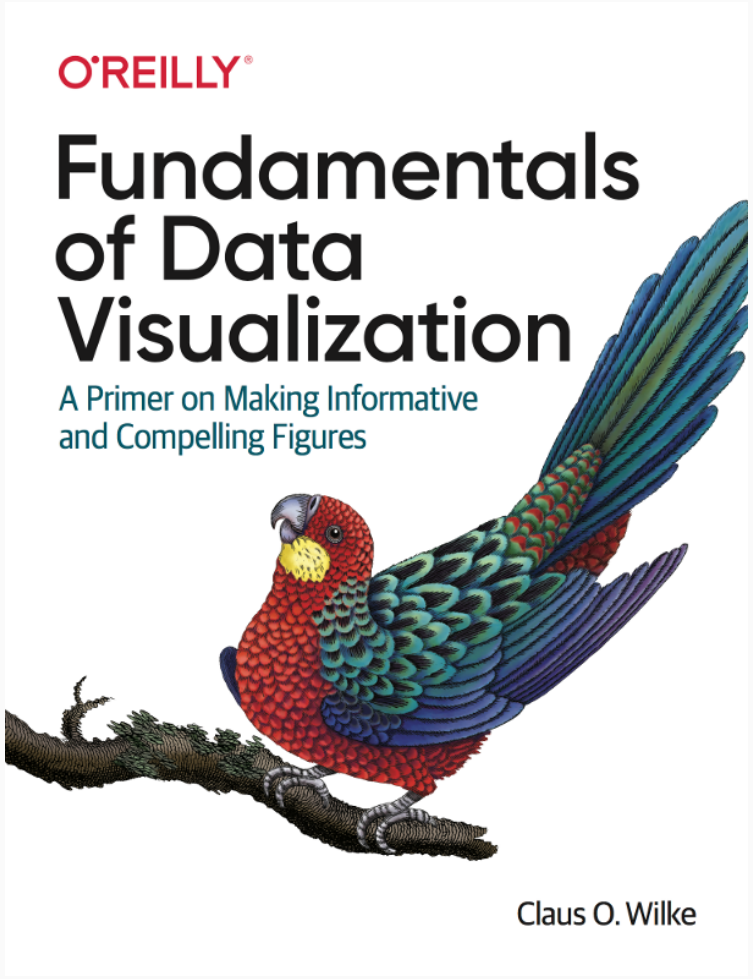
Learning goals



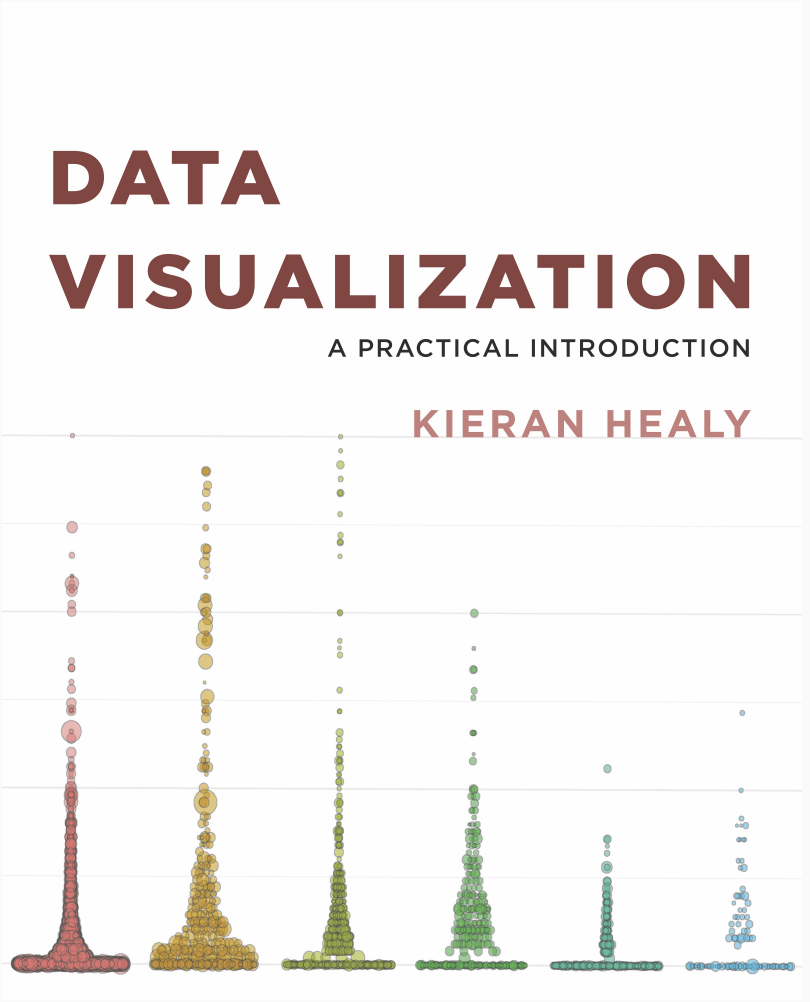
source: unknown

- Understand the basic principles behind effective data visualization.
- Create data visualizations in R using ggplot2
- Craft elegant visual presentations of data

Further reading



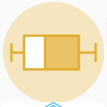
https://clauswilke.com/dataviz/



https://socviz.co/

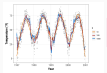
Further studying

Online tutorials:


https://r-charts.com/


https://r-graph-gallery.com/


https://www.data-to-viz.com/

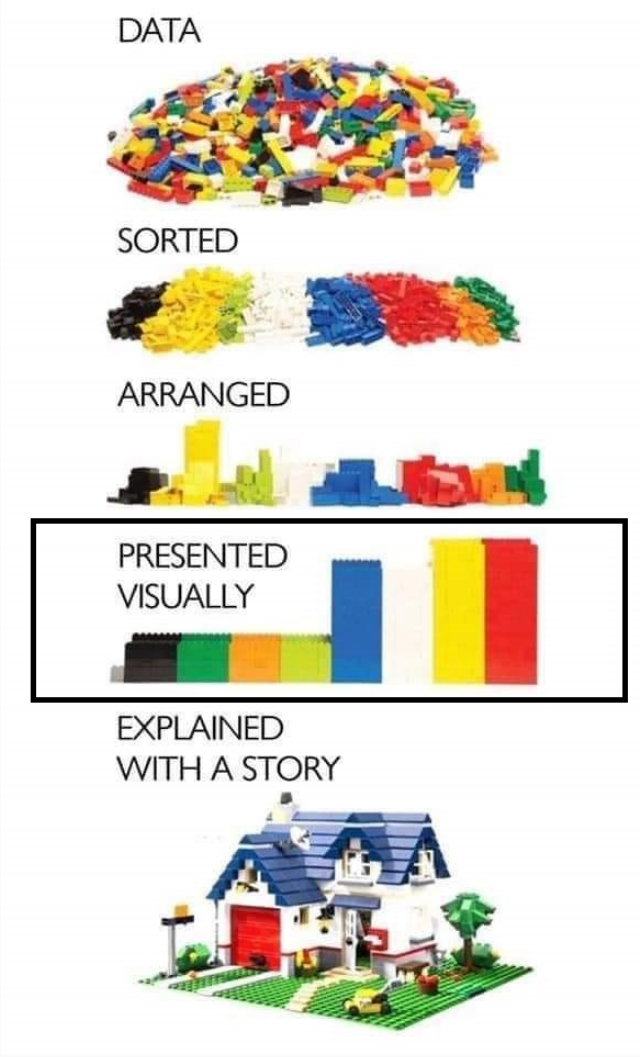

https://cedricscherer.netlify.app/2019/08/05/a-ggplot2-tutorial-for-beautiful-plotting-in-r/

...

3EC course in Radboud University:

NWI-BM083
Data Visualization for the Life Sciences
by K.W. Mulder

Course structure



source: unknown

Week 5:

Lecture 1. Principles of figure design.

Quiz 1.

Week 6:

Tutorial 1. ggplot2: plots and charts.

Quiz 2.

Week 7:

Tutorial 2. ggplot2: statistics, coordinate system, facets.

Tutorial 3. ggplot2: themes and styles.

Practice 1.

Quiz 3.

Week 8:

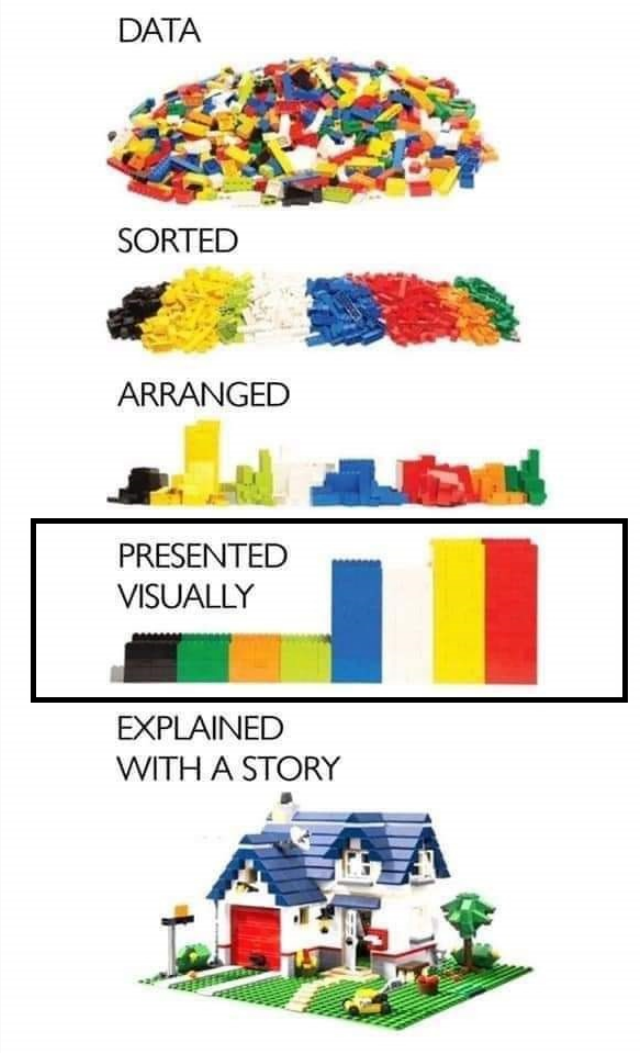
Practice 2. Project.

Practice 3. Project.

Practice 4. Project.

Assignment.

Evaluation criteria



source: unknown

Week 5:

Lecture 1. Principles of figure design.

Quiz 1. --> passed

Week 6:

Tutorial 1. ggplot2: plots and charts.

Quiz 2. --> passed

Week 7:

Tutorial 2. ggplot2: statistics, coordinate system, facets.

Tutorial 3. ggplot2: themes and styles.

Practice 1.

Quiz 3. --> passed

Week 8:

Practice 2. Project.

Practice 3. Project.

Practice 4. Project.

Assignment. --> passed

This is a hands-on course

You will learn how to visualize the data in R.

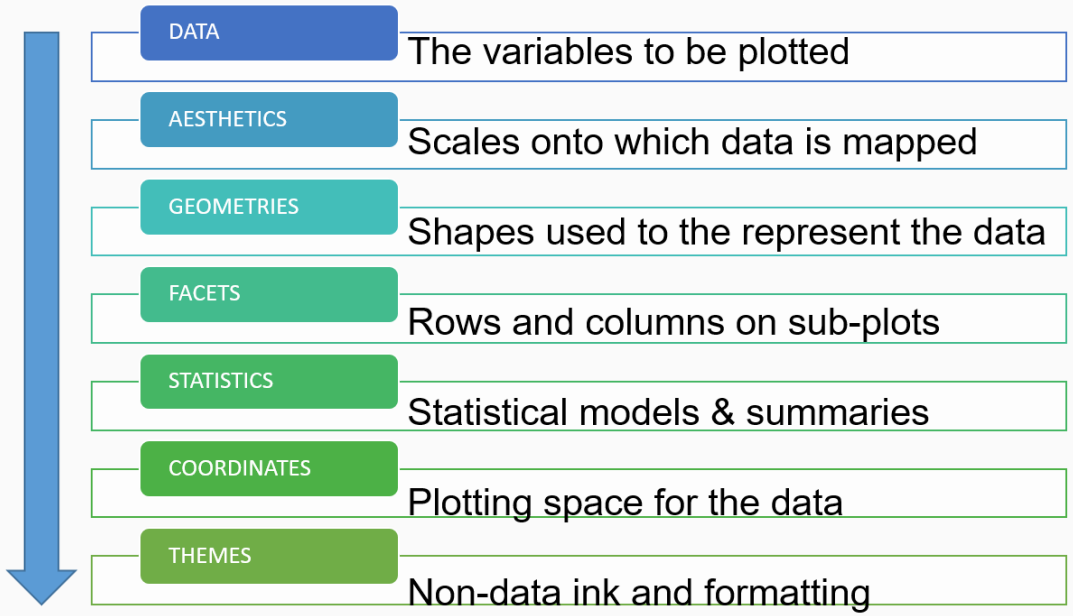
These slides were also designed in R and you can find the source code on GitHub:

```
library(tidyverse)
library(ggplot2)
library(cowplot)
library(viridis)
library(RColorBrewer)
```

Principles of figure design:

- Grammar of graphics
- Data
- Aesthetics
- Geometries
- Coordinates
- Color

Grammar of graphics



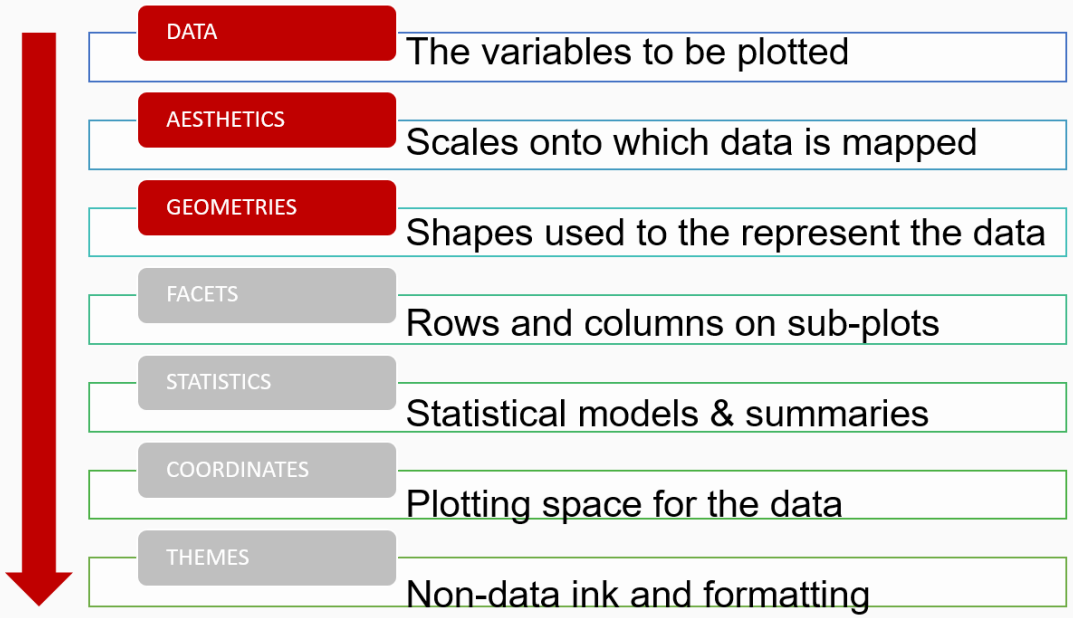
Layers in grammar of graphics

The *grammar of graphics* is a plotting framework developed by Leland Wilkinson (Grammar of Graphics, 1999) that dissects each component of a graph into individual layer.

There are two important principles:

- Graphics are made of distinct layers of grammatical elements
- Plots are built with appropriate aesthetic mappings to make these plots meaningful

Grammar of graphics



Layers in grammar of graphics

Three out of 7 layers are essential for any plot:

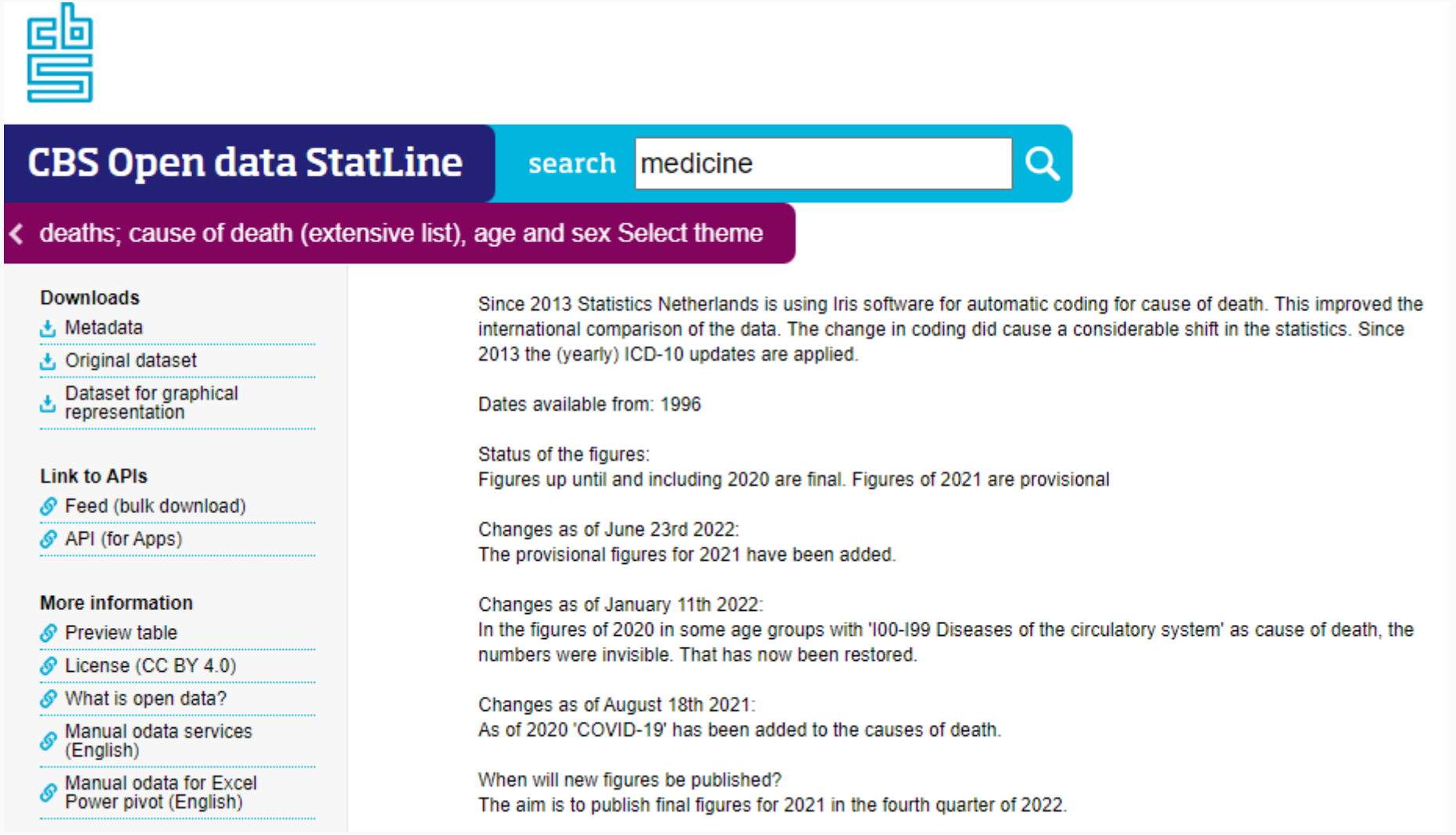
Data This is the dataset being plotted containing the variables to be plotted on the graph.

Aesthetics Aesthetics refers to the scales on which we map the data. Some common aesthetics to consider are axis, shape, size, and color.

Geometries Geom refers to the actual visual elements used for the data in the plot, such as points, lines, and bars.

Data

CBS open data: the source



https://opendata.cbs.nl/statline/

Data

Number of deaths in the population of the Netherlands by main underlying cause of death, by age and sex, 1996-2021

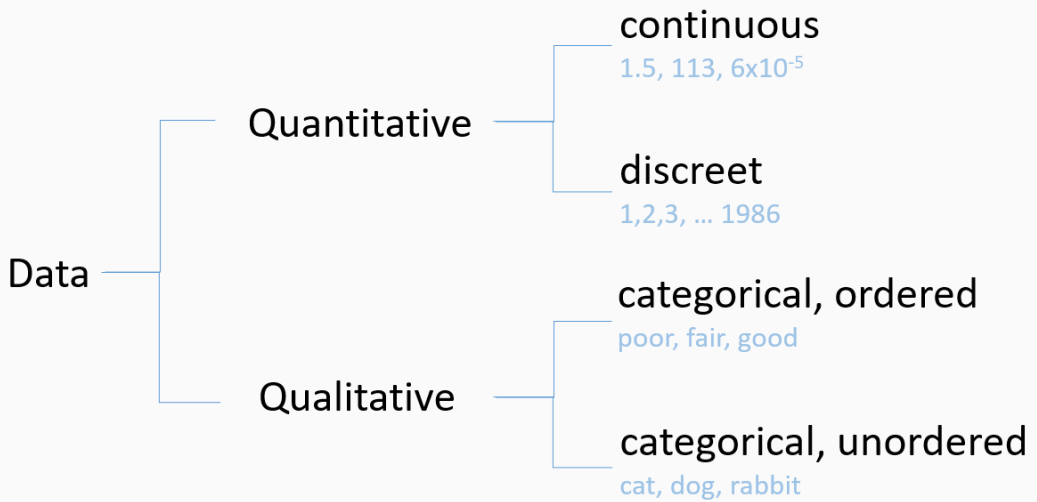
https://opendata.cbs.nl/statline/portal.html?_la=en&_catalog=CBS&tableId=7233ENG&_theme=1120

Identifier: 7233ENG

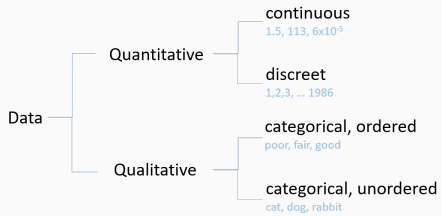
Short title: deaths; cause of death (extensive list)

Reference period: 1996-2021

Types of data



Types of data: example



Which type of data is present in different columns?

ID	Sex	Age	CausesOfDeath	Year	Deaths
2721092	Female	60-64	GenitourinarySystem	2006	28
2579640	Female	45-49	MusculSystem	2020	4
1115128	Male	1-9	Infections	2010	2
2624197	Female	50-54	SkinDeseases	2013	0
3005599	Female	90-94	CongenitalMalform	2021	2
2285664	Female	15-19	Endocr+Metabol	2000	3
2771707	Female	65-69	Perinatal	1999	0
2927153	Female	85-89	Infections	2017	420
1085319	Male	0	CirculatorySystem	1997	10
1931241	Male	85-89	GenitourinarySystem	2009	353

The data must be tidy to be plotted in gramar or

Following three rules makes a dataset tidy:

- variables are in columns,
- observations are in rows,
- and values are in cells.

source: R for Data Science 12.1

Is this dataset tidy?

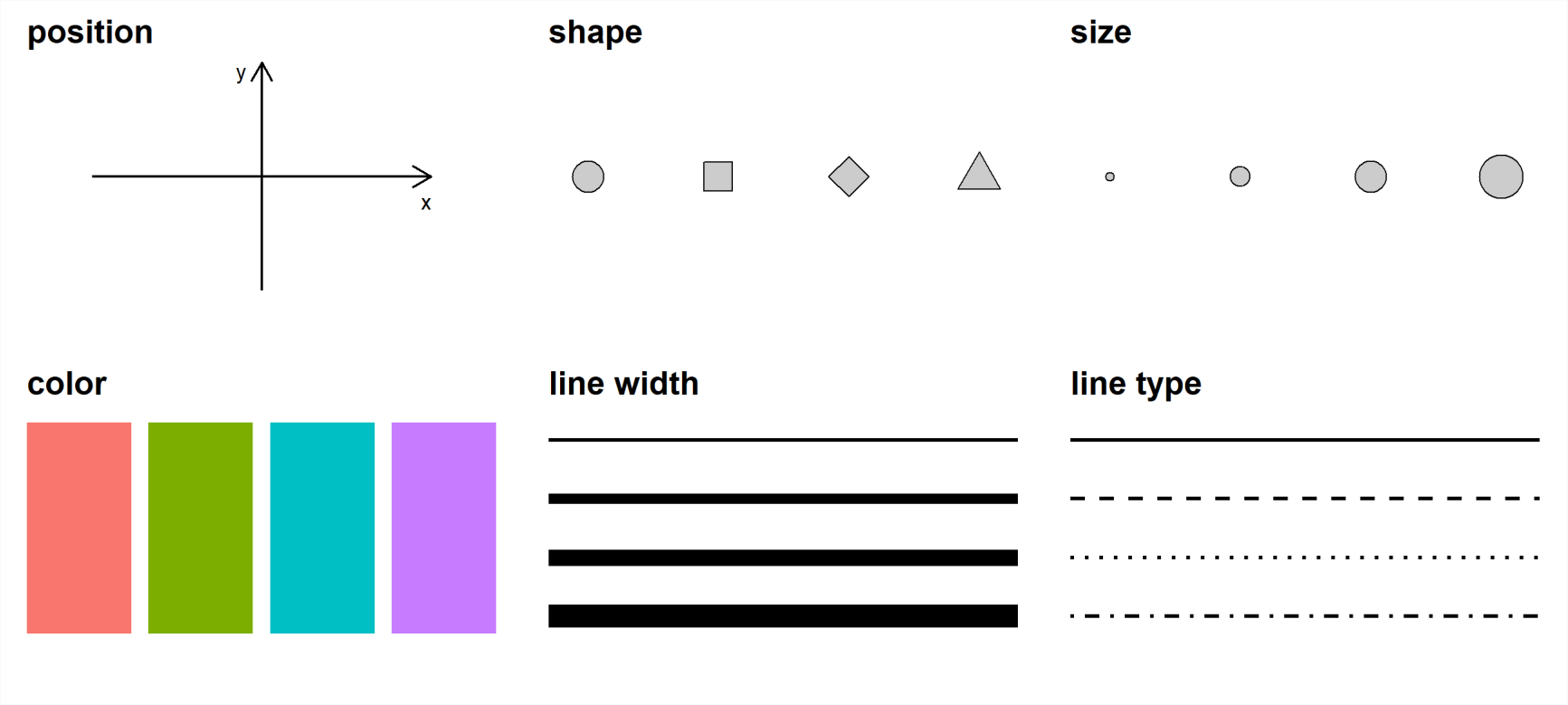
ID	Sex	Age	CausesOfDeath	Year	Deaths
1863018	Male	80-84	Neoplasms	2010	3601
2931645	Female	85-89	Neoplasms	2011	2477
1626221	Male	55-59	Infections	2021	37
1311905	Male	20-24	Psychological	2019	0
1327223	Male	20-24	GenitourinarySystem	1997	0
3034839	Female	95+	Eye+Adnexa	2011	1
2494512	Female	35-39	CongenitalMalform	2016	5
1325164	Male	20-24	MusculSystem	2018	0
2029877	Male	95+	CongenitalMalform	2001	0
1546251	Male	45-49	NervousSystem	2001	36

If you have doubts, check it in R for data science or in TowardsDataScience

Aesthetics

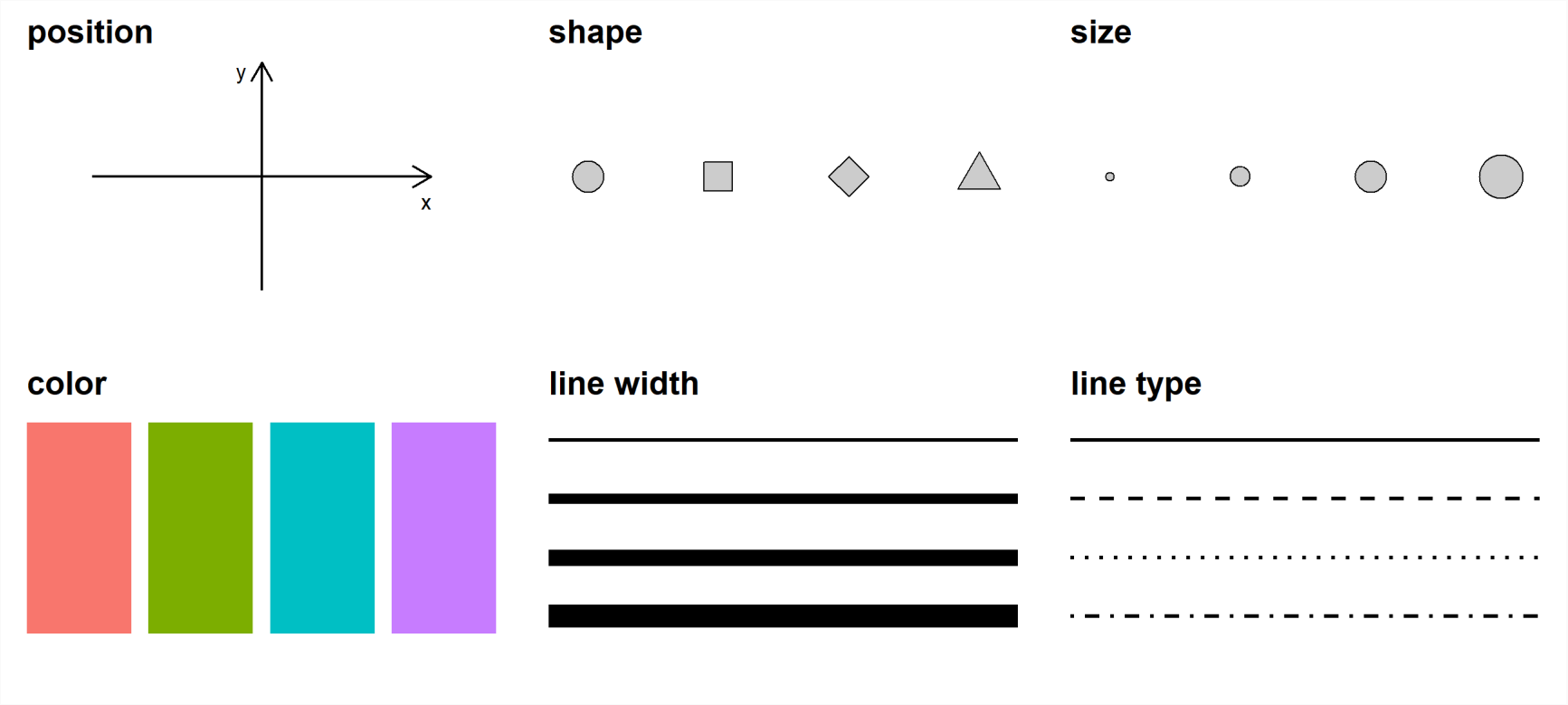
Channels for representing data

Aesthetics describe every aspect of a given graphical element.



Channels for representing data

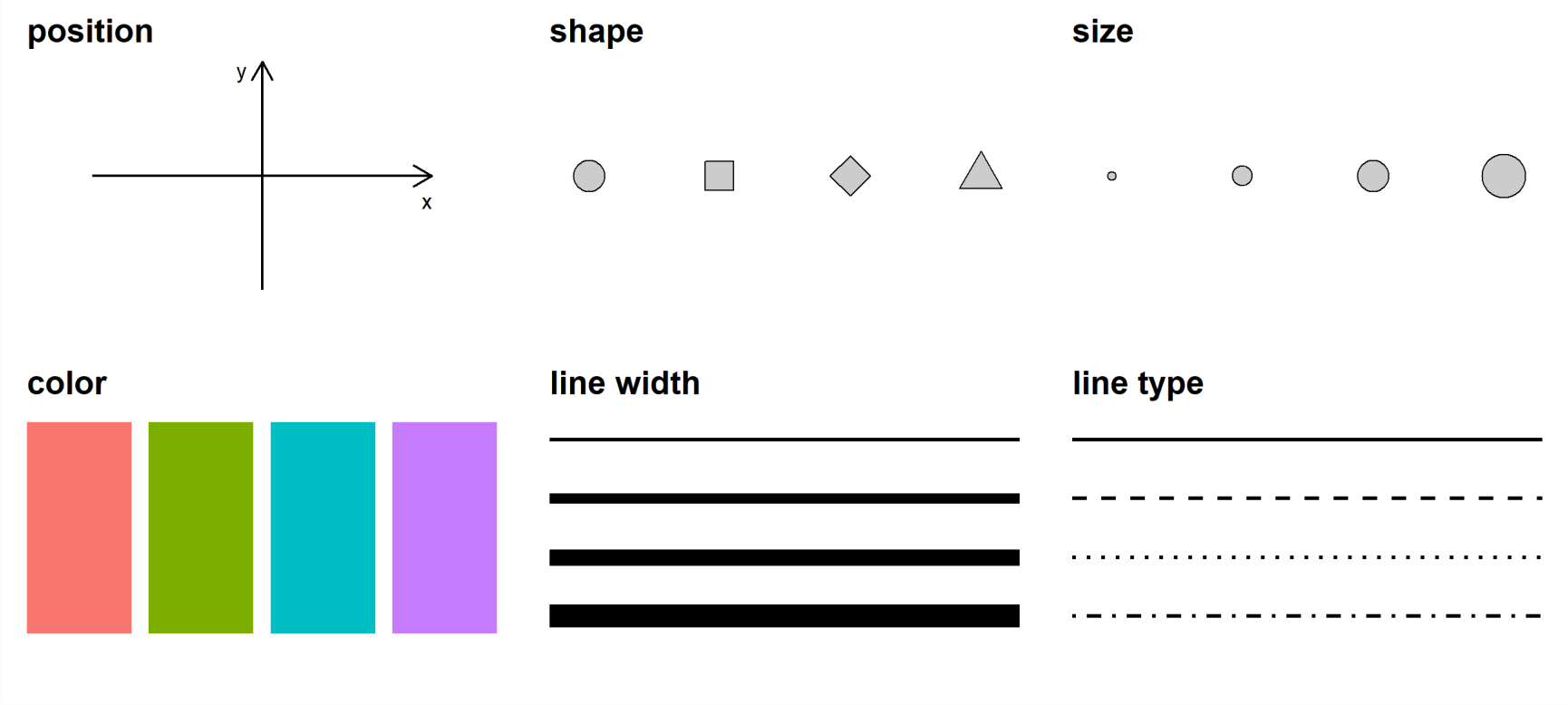
Aesthetics describe every aspect of a given graphical element.



This figure has been created using ggplot2, adjusted from: https://clauswilke.com/dataviz/

Channels for representing data

Aesthetics describe every aspect of a given graphical element.



This figure has been created using ggplot2, adjusted from: https://clauswilke.com/dataviz/

Scales map the data values onto aesthetics

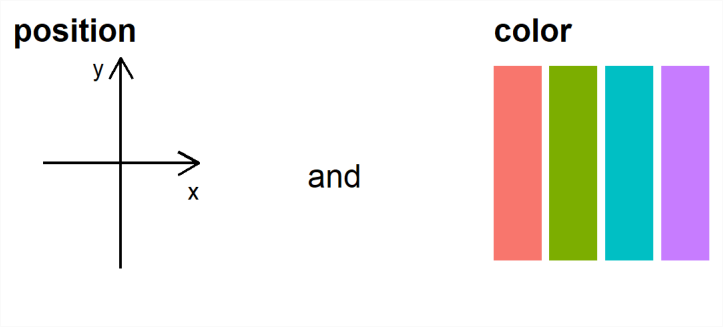
To map data values onto aesthetics, we need to specify which data values correspond to which specific aesthetics values.

Position: x = Year; y = Deaths; Color = CausesOfDeath.

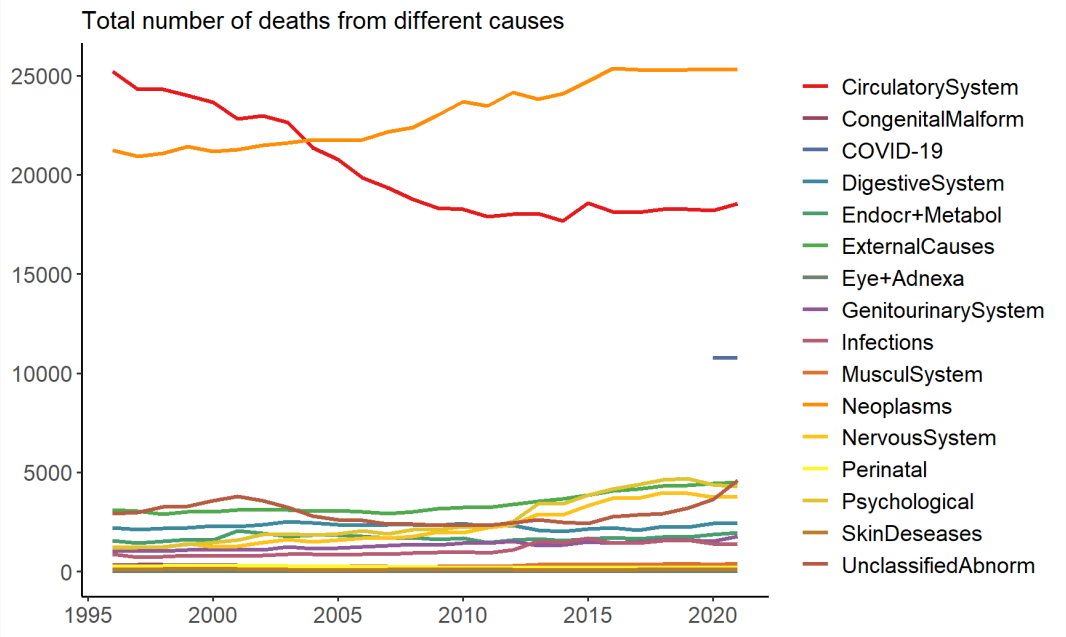
ID	Sex	Age	CausesOfDeath	Year	Deaths
1022190	Male	Total	Infections	1996	890
1022191	Male	Total	Infections	1997	736
1022192	Male	Total	Infections	1998	757
1022193	Male	Total	Infections	1999	814
1022194	Male	Total	Infections	2000	788
1022195	Male	Total	Infections	2001	773

Scaling data into 'position' and 'color' aestetics

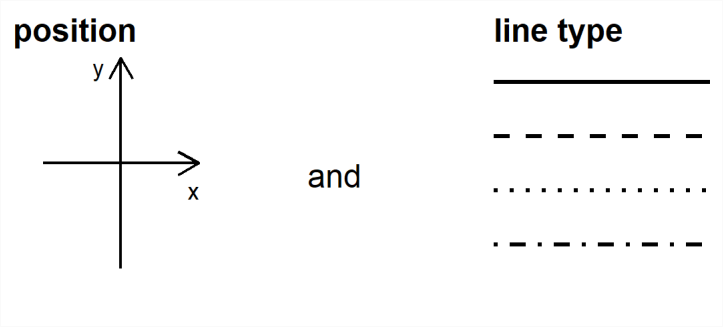
To map data values onto aesthetics, we need to specify which data values correspond to which specific aesthetics values.



Position: x = Year; y = Deaths;
Color = CausesOfDeath.

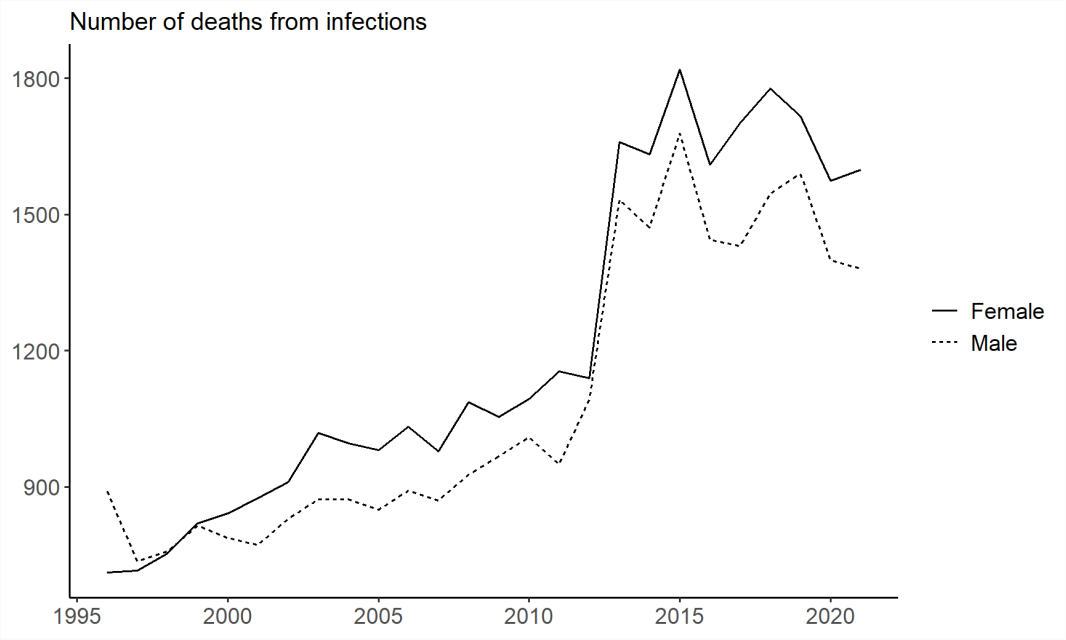


Scaling data into 'position' and 'line type' aestetics

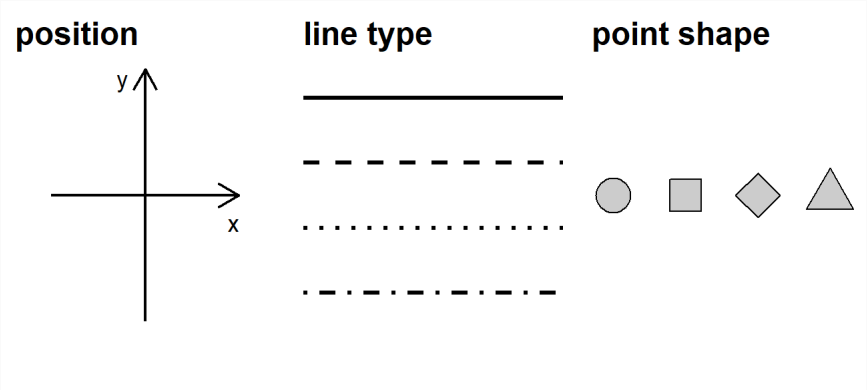


Position: x = Year; y = Deaths;

Line type = Sex.

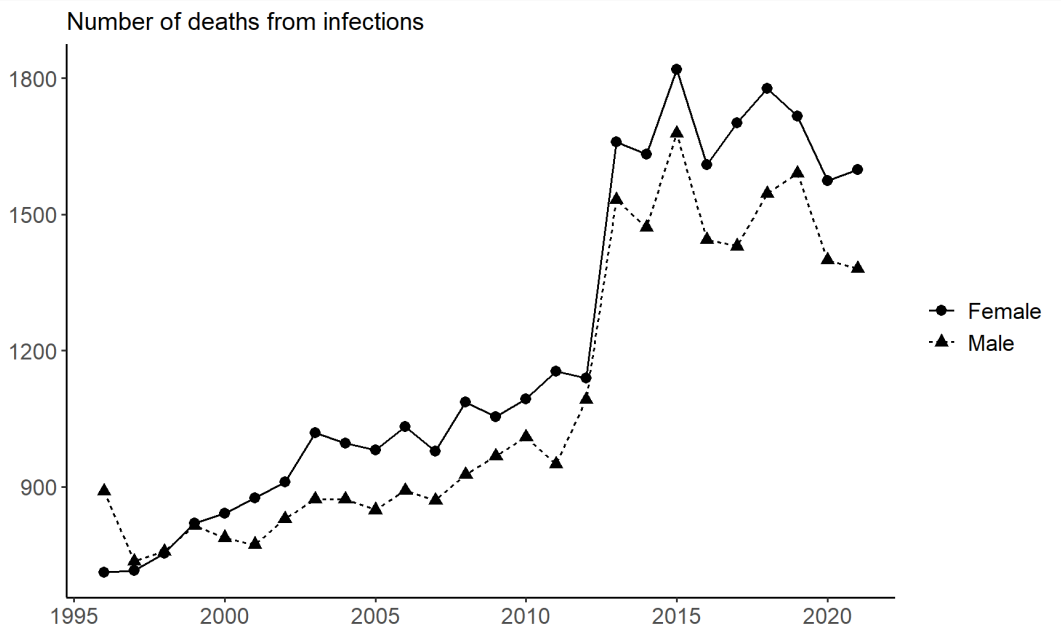


Scaling data into 'position', 'line type'&'shape'



Position: x = Year; y = Deaths;

Point Shape & Line Type = Sex.



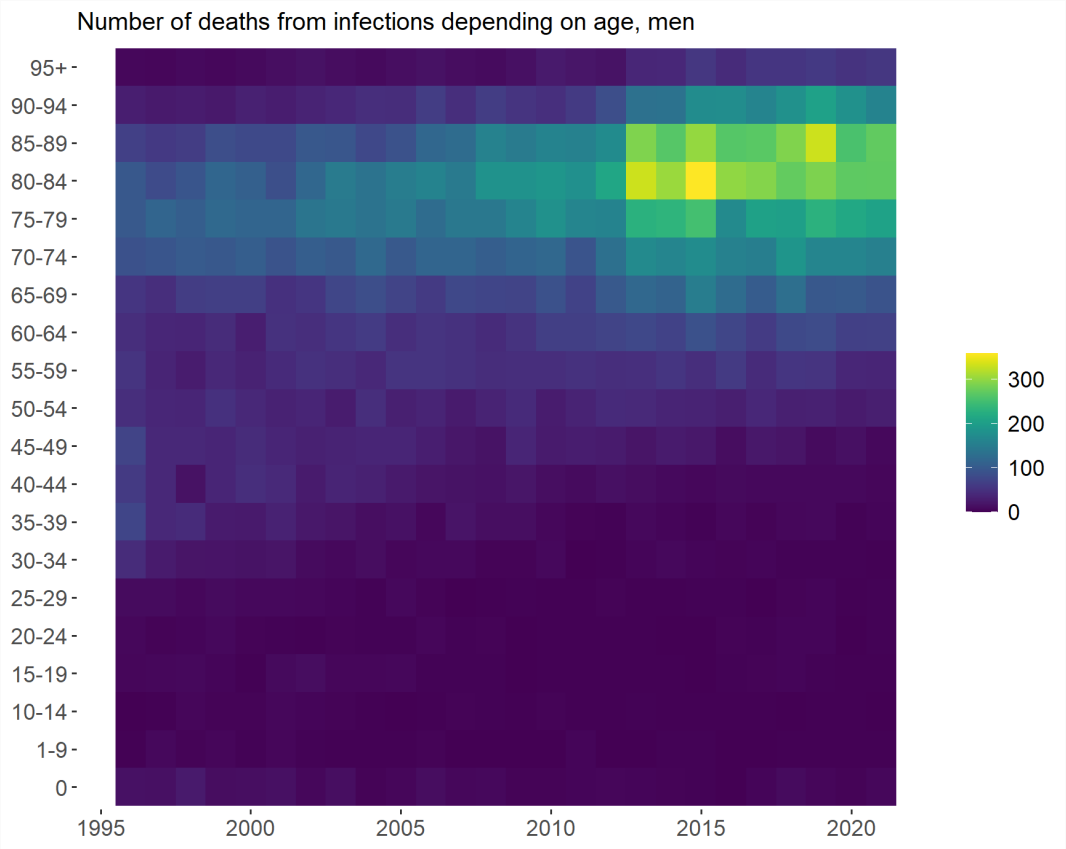
Scaling continuous quantitative data into color

ID	Sex	Age	CausesOfDeath	Year	Deaths
1022190	Male	Total	Infections	1996	890
1022191	Male	Total	Infections	1997	736
1022192	Male	Total	Infections	1998	757
1022193	Male	Total	Infections	1999	814
1022194	Male	Total	Infections	2000	788
1022195	Male	Total	Infections	2001	773

Color: Deaths;

Position x: Year

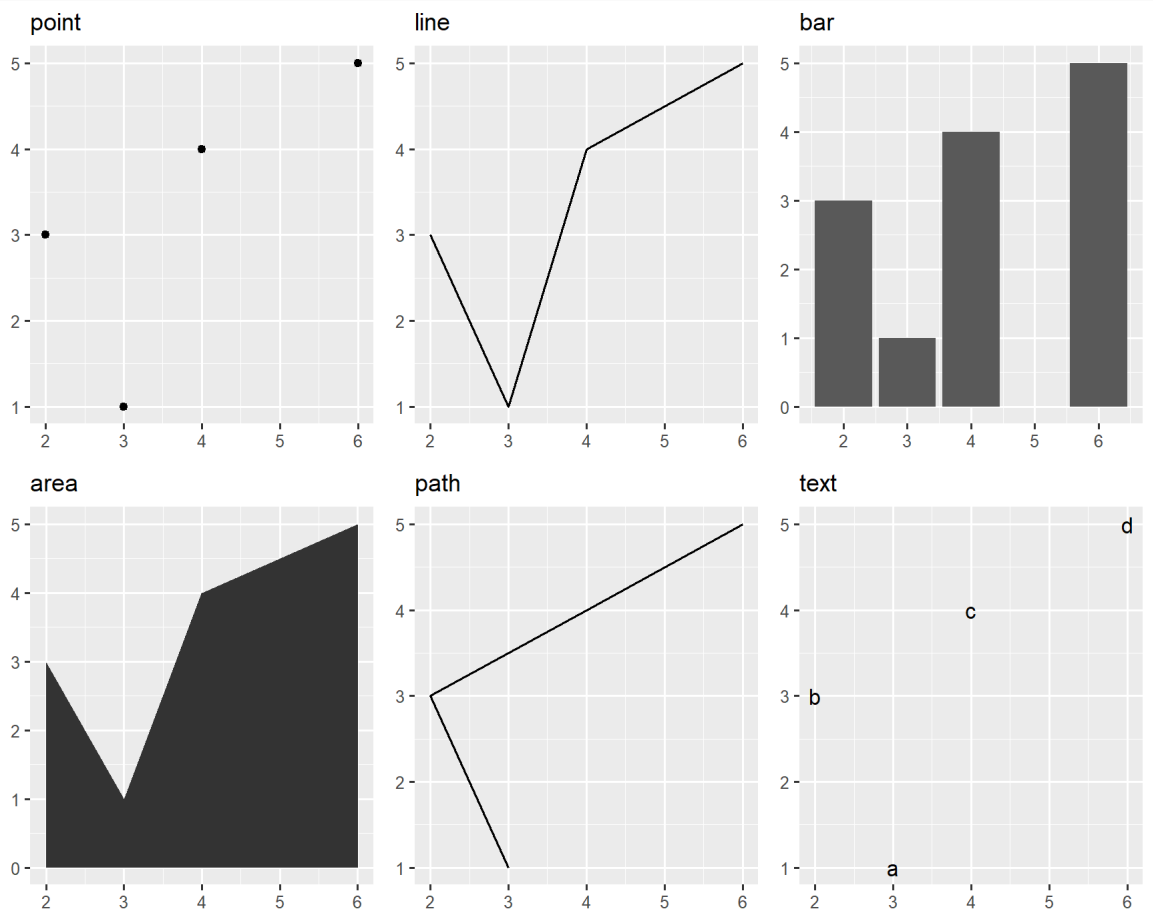
Position y: Age



Geometries

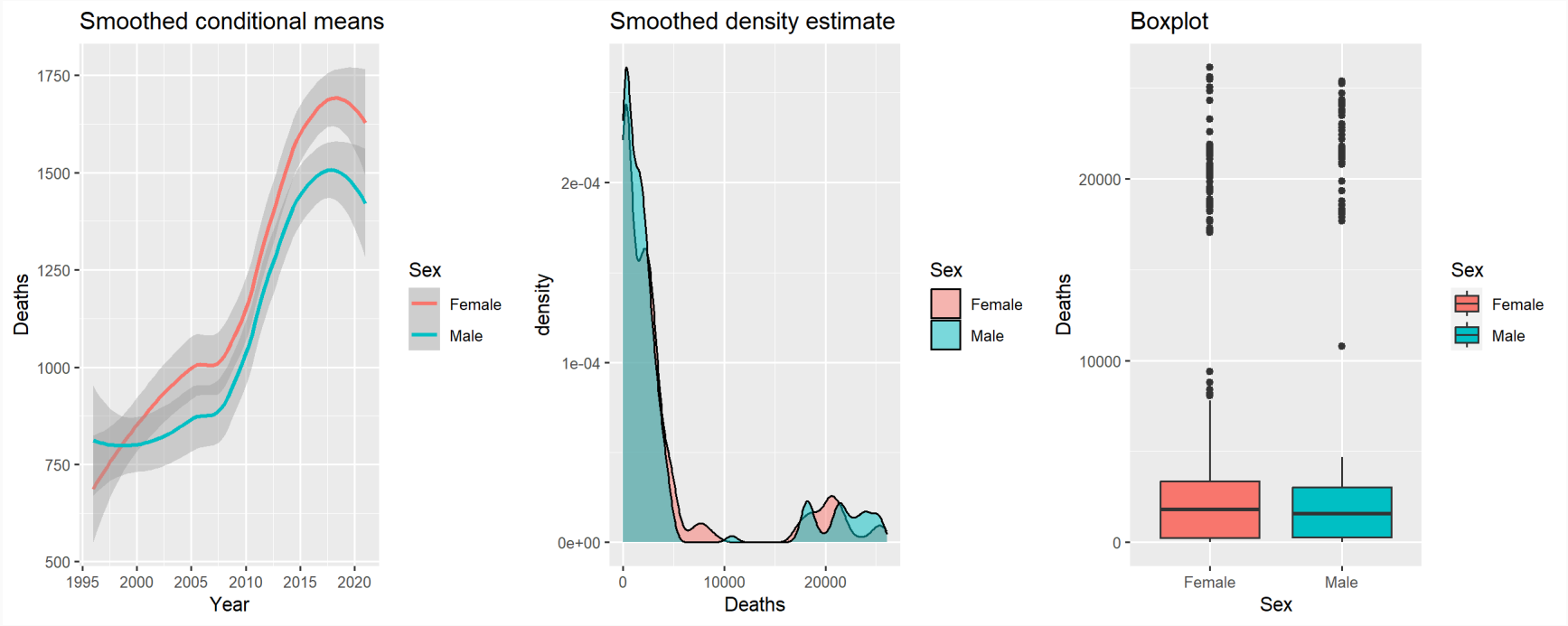
Geometries

Geoms can be roughly divided into individual and collective geoms. An individual geom draws a distinct graphical object for each observation (row). Here are the examples

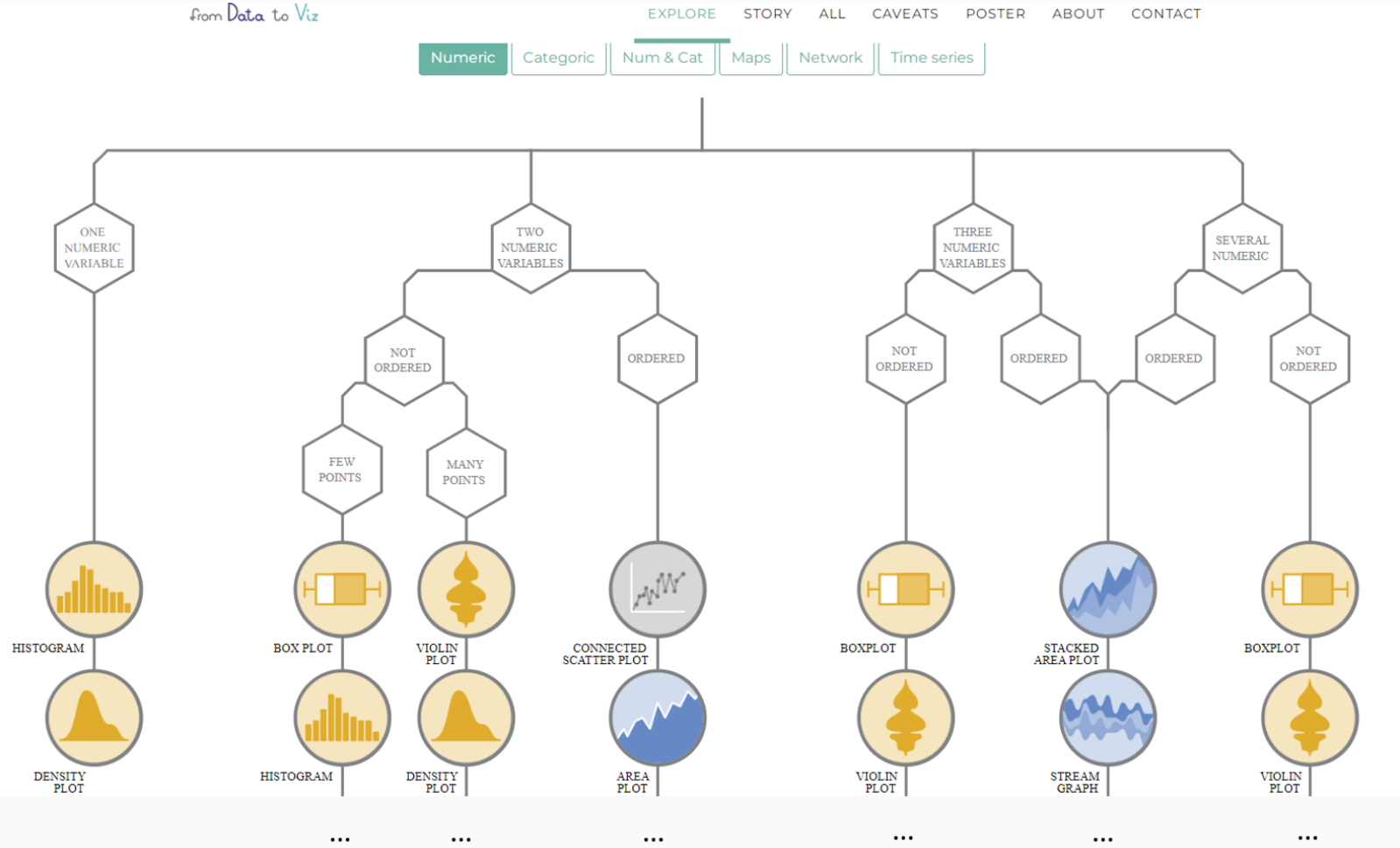


Geometries

A collective geom displays multiple observations with one geometric object. This may be a result of a statistical summary, like a boxplot, density plot, or histogram.



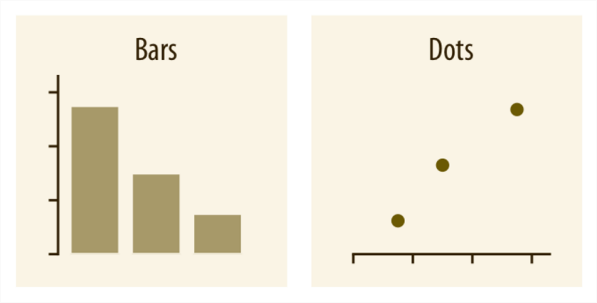
How to choose the geom type?



Source: from Data to Viz.

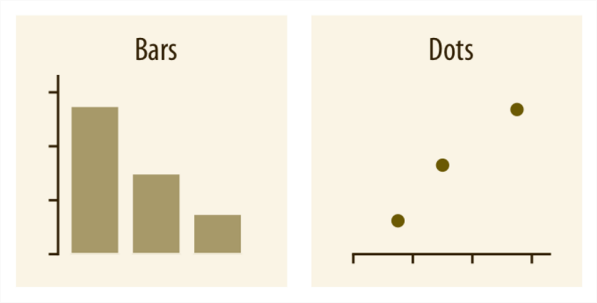
How to vizualize amounts?

The most common way to visualizing amounts is using bars, either vertically or horizontally arranged. Instead of using bars, one can also place dots at the location where the corresponding bar would end.

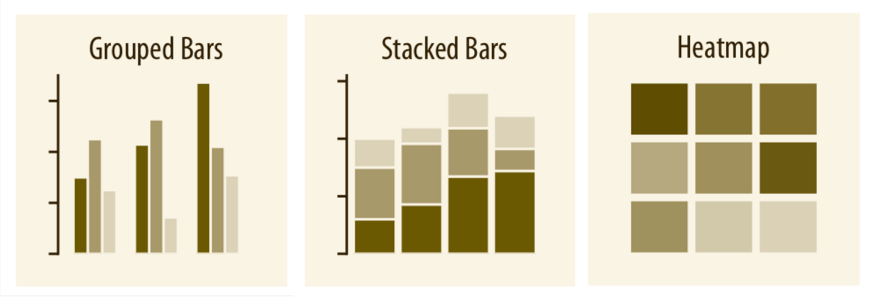


How to vizualize amounts?

The most common way to visualizing amounts is using bars, either vertically or horizontally arranged. Instead of using bars, one can also place dots at the location where the corresponding bar would end.



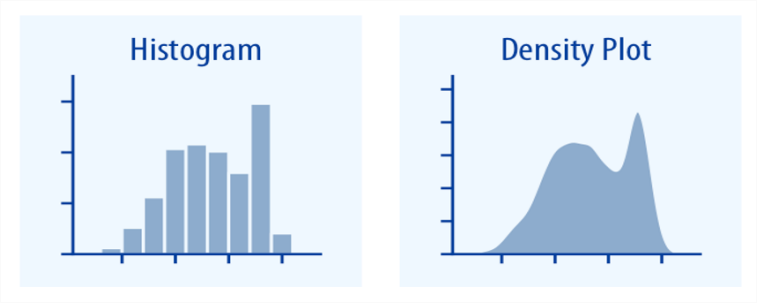
If there are two or more sets of categories for which one wants to show amounts, one can group or stack the bars. One can also map the categories onto the x and y axis and show amounts by color, via a heatmap.



Source: Fundamentals of Data Vizualization.

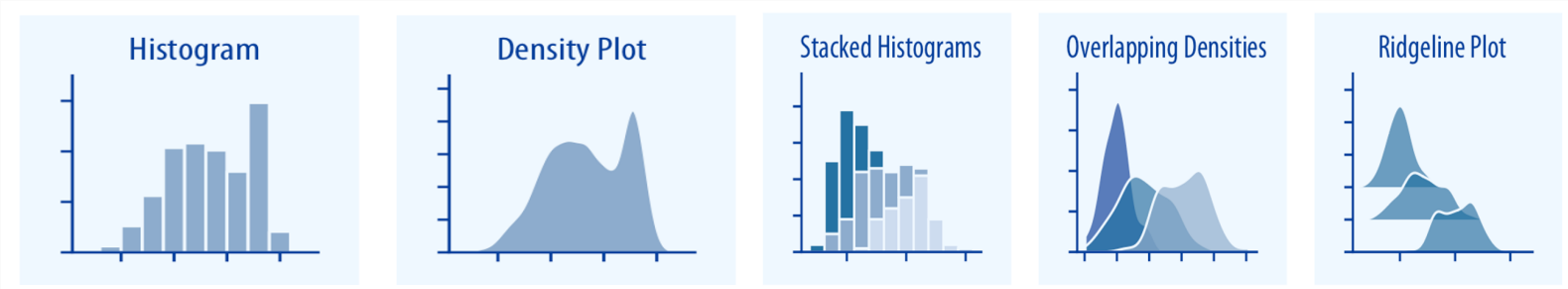
How to vizualize distributions?

Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading.



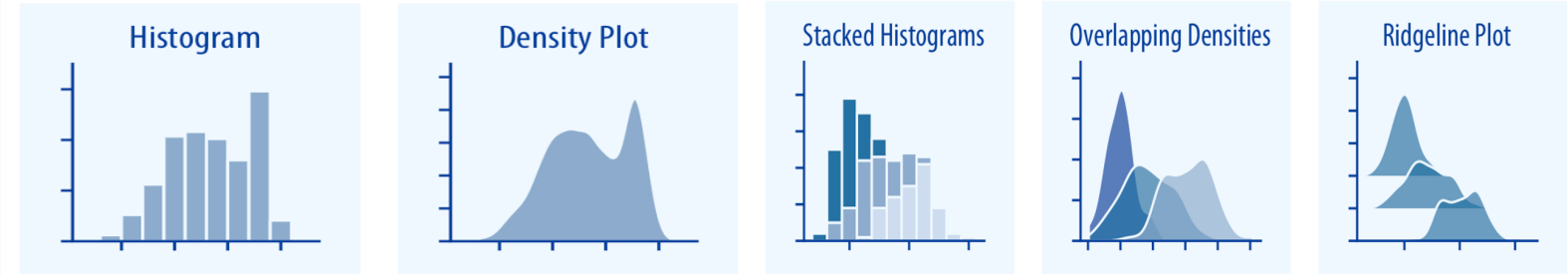
How to vizualize distributions?

Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading.

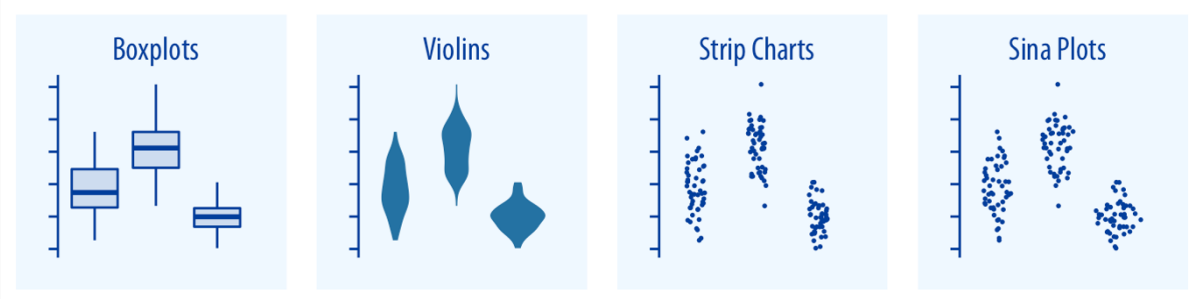


How to vizualize distributions?

Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading.



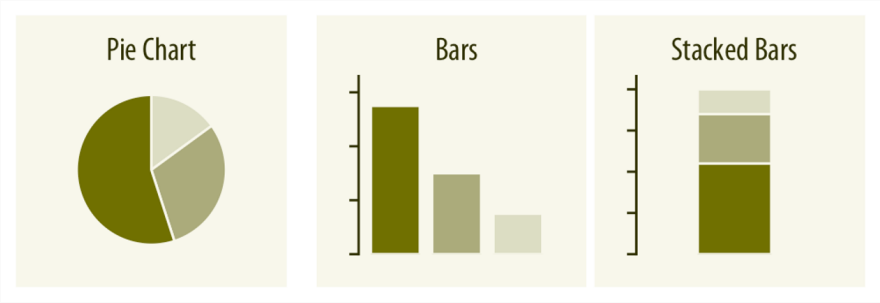
Boxplots, violins, strip charts, and sina plots are useful when one wants to visualize many distributions at once and/or if we are primarily interested in overall shifts among the distributions.



Source: Fundamentals of Data Vizualization.

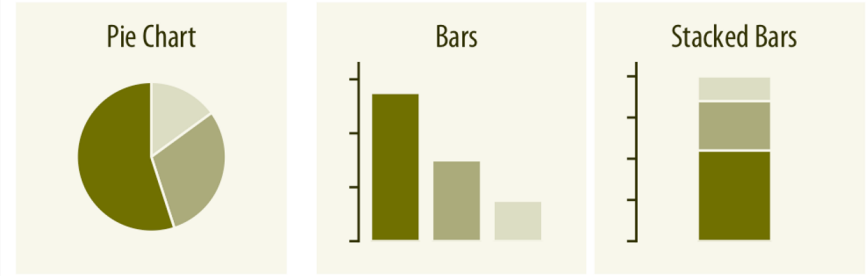
How to vizualize proportions?

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars.

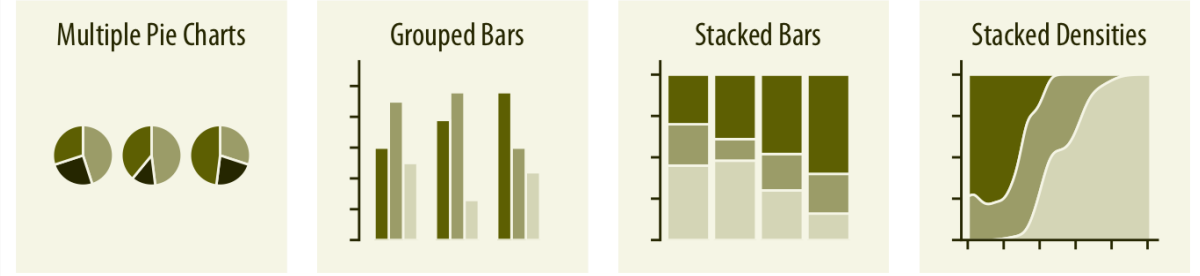


How to vizualize proportions?

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars.



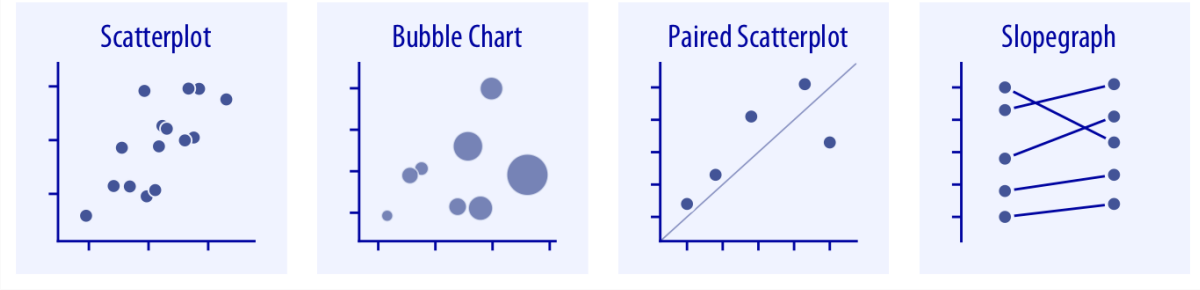
When visualizing multiple sets of proportions, pie charts tend to be space-inefficient and often obscure relationships. Grouped bars work well as long as the number of conditions compared is moderate, and stacked bars can work for large numbers of conditions. Stacked densities are appropriate when the proportions change along a continuous variable.



Source: Fundamentals of Data Vizualization.

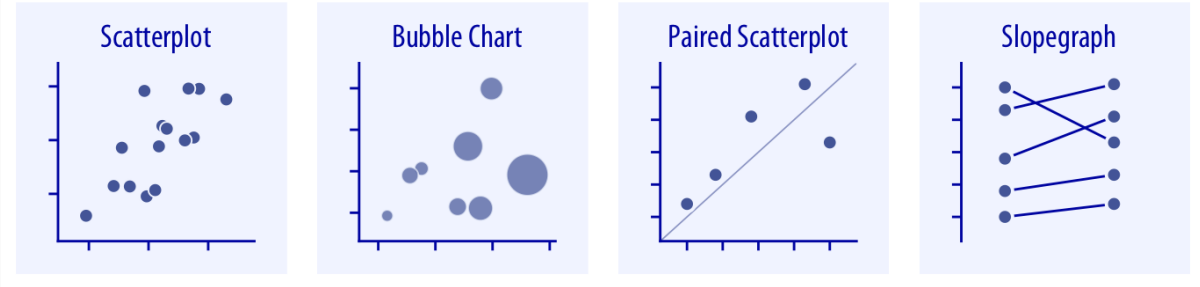
How to vizualize x-y relationships

Scatterplots are typical representation of x-y relationships. If we have three quantitative variables, we can map one onto the dot size, creating a bubble chart. If the variables along the x and the y axes are measured in the same units, it is helpful to add a line indicating x = y. Paired data can also be shown as a slope graph of paired points connected by straight lines.

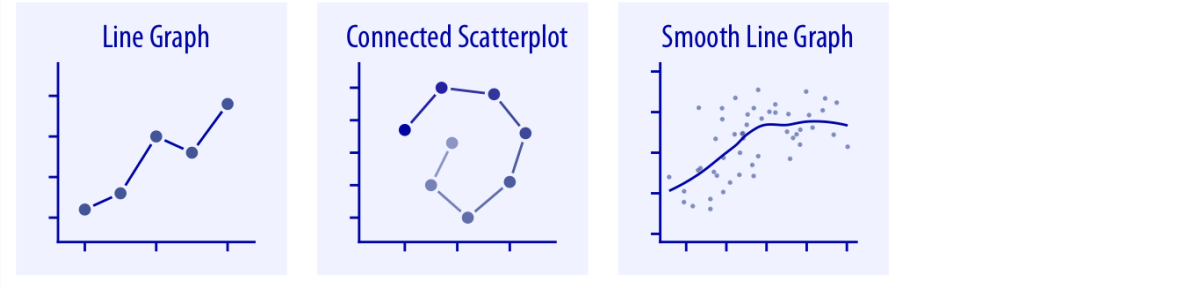


How to vizualize x-y relationships

Scatterplots are typical representation of x-y relationships. If we have three quantitative variables, we can map one onto the dot size, creating a bubble chart. If the variables along the x and the y axes are measured in the same units, it is helpful to add a line indicating x = y. Paired data can also be shown as a slope graph of paired points connected by straight lines.



When the x axis represents time or a gradually increasing quantity, we commonly draw line graphs. If we have a temporal sequence of two response variables, we can draw a connected scatterplot. We can use smooth lines to represent trends in a larger dataset.

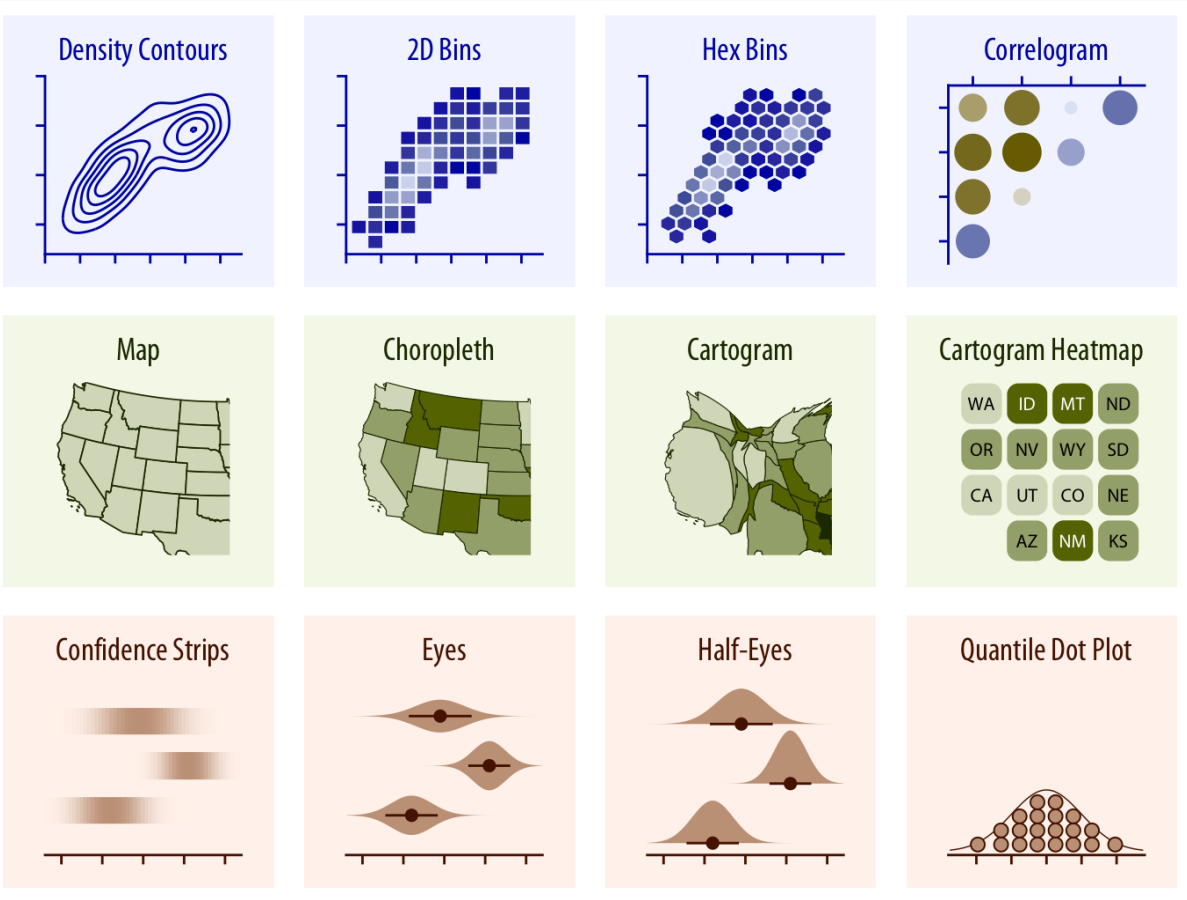


Source: Fundamentals of Data Vizualization.

There are more vizualization types...

...which we won't be able to overview in this small course.

But you can learn more on them in Claus Wilke's book on Fundamentals of Data Vizualization.



Color

Color

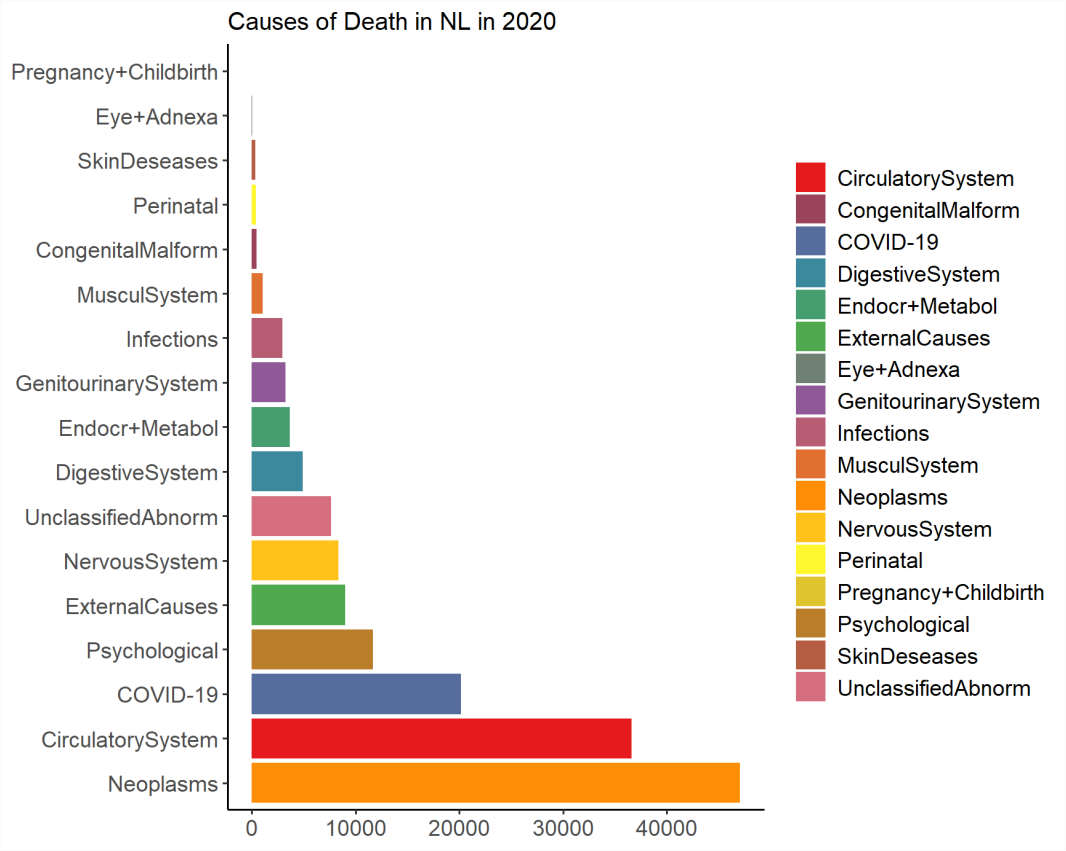
There are four fundamental use cases for color in data visualizations:

- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).

Color: case 1

There are four fundamental use cases for color in data visualizations:

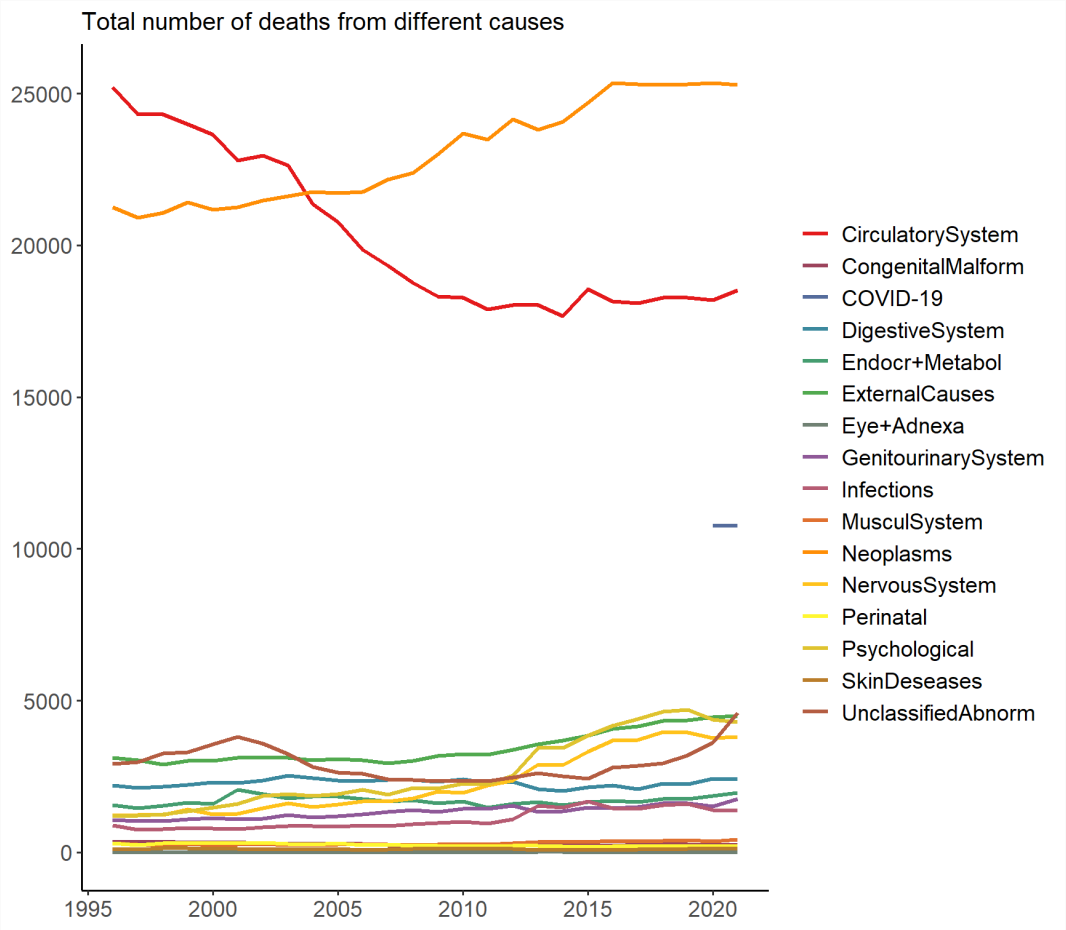
- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).



Color: case 1

There are four fundamental use cases for color in data visualizations:

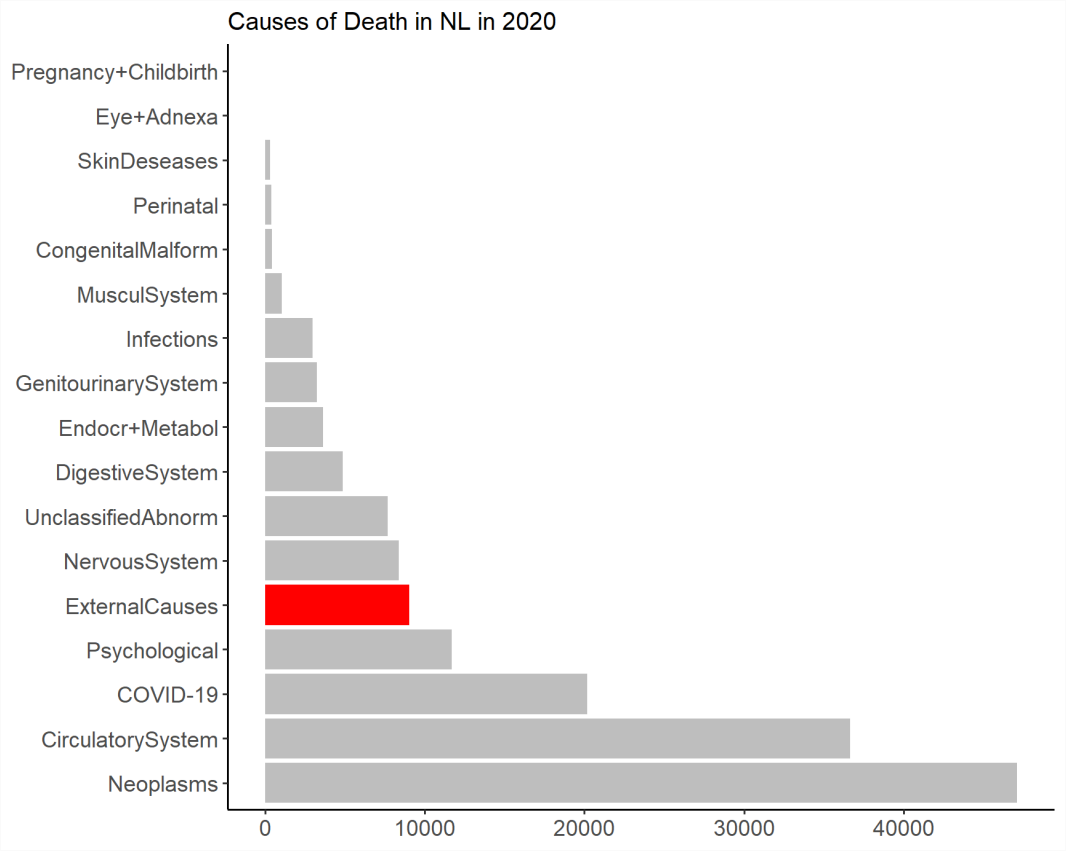
- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).



Color: case 2

There are four fundamental use cases for color in data visualizations:

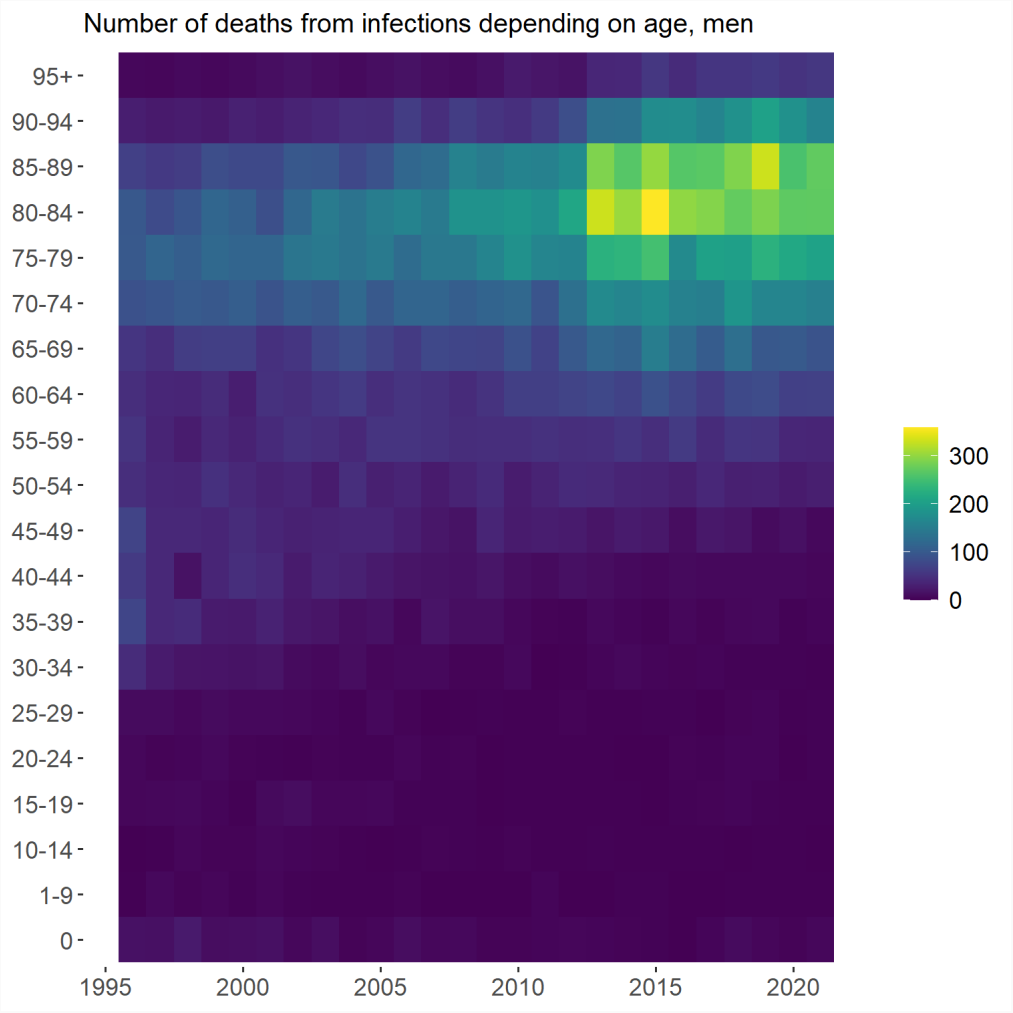
- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).



Color: case 3

There are four fundamental use cases for color in data visualizations:

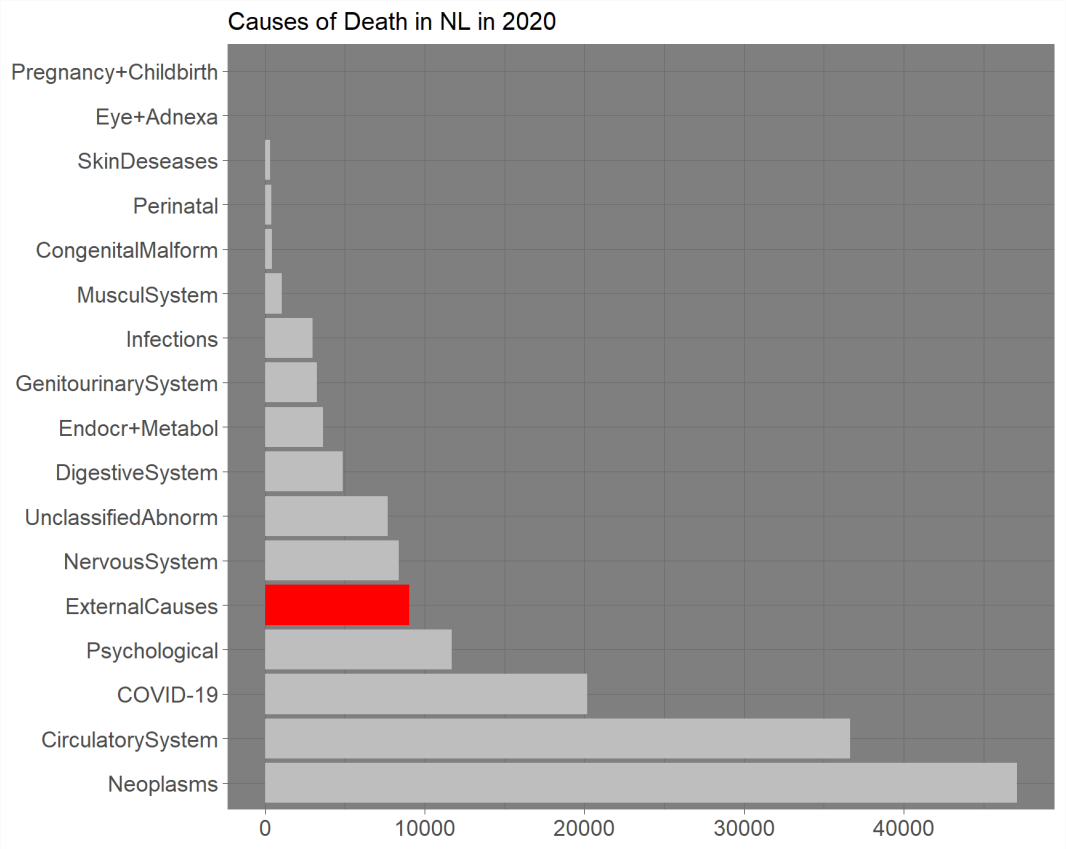
- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).



Color: case 4

There are four fundamental use cases for color in data visualizations:

- to distinguish groups of data from each other;
- to highlight relevant information;
- to represent data values;
- to adjust the layout (data-unrelated color).



To do before next week

- Pass the quiz on the Brightspace.
- Explore the source code of the lecture via GitHub
- Be sure that you have the most recent version of R installed.
- Download the dataset on causes of death in NL from CBS Open Portal